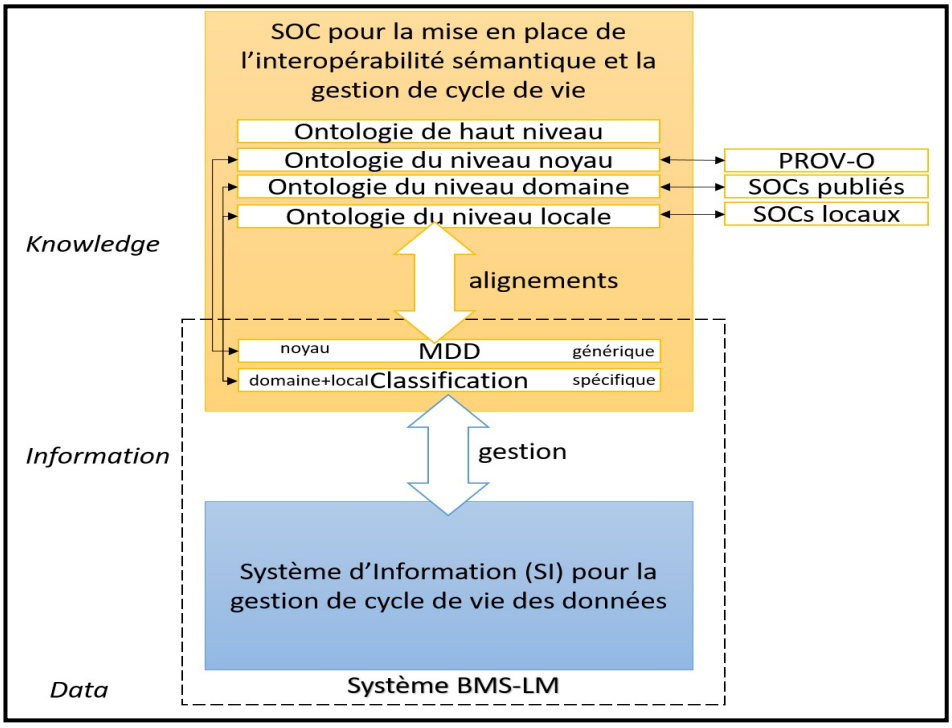
Gestion du cycle de vie des études multimodales de recherche biomédicale préclinique pour le partage et la réutilisation des données hétérogènes

Amel Raboudi-Souilem

Par **Amel RABOUDI-SOUILEM**

Gestion du cycle de vie des études multimodales de recherche biomédicale préclinique pour le partage et la réutilisation des données hétérogènes

Thèse présentée
pour l'obtention du grade
de Docteur de l'UTC



Soutenue le 31 mars 2021
Spécialité : Génie Industriel : Unité de recherche en Mécanique - Laboratoire Roberval (FRE UTC - CNRS 2012)
D2601

UNIVERSITE DE TECHNOLOGIE DE COMPIEGNE

ECOLE DOCTORALE N°71

Sciences pour l'ingénieur

Thèse de Doctorat

Amel Raboudi-Souilem

Gestion du cycle de vie des études multimodales de recherche biomédicale préclinique pour le partage et la réutilisation des données hétérogènes

Dirigée par Pr Benoît Eynard et Pr Bertrand Tavitian

Soumis pour l'obtention du grade de Docteur de l'UTC

Spécialité : Génie Industriel

Pour une soutenance le 31 mars 2021

Devant le jury composé de :

Jean Charlet, Chercheur HDR, Sorbonne Université ……..…(rapporteur)

Sebti Foufou, Professeur des Universités, Université de Bourgogne ……...…..(rapporteur)

Philippe Boutinaud, Directeur technique et innovation, Fealinx……...……...… (examinateur)

Anita Burgan, Professeur des Universités-Praticien Hospitalier, Université de Paris ……..……..(examinatrice)

Irène Buvat, Directrice de Recherche CNRS, Université Paris Saclay……..…….…(examinatrice)

Christophe Egles, Professeur des Universités, Université de Technologie de Compiègne …….(examinateur)

Benoît Eynard, Enseignant Chercheur HDR, Université de Technologie de Compiègne ….....(directeur de thèse)

Bertrand Tavitian, Professeur des Universités-Praticien Hospitalier, Université de Paris…….. (directeur de thèse)

Marianne Allanic, Consultante IA et santé PhD, Althenas ……(invité)

Daniel Balvay, Ingénieur de recherche, Université de Paris ……...…..(invité)

Alexandre Durupt, Maître de conférences HDR, Université de Technologie de Compiègne …….…...(invité)

Marc Joliot, Directeur de Recherche CEA, Université de Bordeaux ……....……....(invité)

Résumé

La recherche biomédicale se caractérise par des évolutions fréquentes de méthodes, de types de données et de personnel. Il en résulte une hétérogénéité importante des données de recherche : multisources, multimodales, pluridisciplinaires, multisites, etc. L'hétérogénéité freine le partage et la réutilisation de données scientifiques puisque la confiance dans celles-ci et leur compréhension sont en jeu. Pour améliorer la confiance dans les données, nous avons appliqué aux études de recherche biomédicale le paradigme de gestion du cycle de vie, basé sur une solution de « Product Lifecycle Management » (PLM) qui a son origine dans l'industrie manufacturière. Ainsi, nous proposons une démarche de gestion de données, avec un maximum de traçabilité de la provenance des données, tout au long du cycle de vie d'une étude de recherche biomédicale. Quant à l'amélioration de la compréhension des données, nous nous sommes focalisés sur la mise en place d'une interopérabilité sémantique entre les terminologies vernaculaires utilisées par les équipes de recherche d'un côté, et les standards, terminologies et ontologies (i.e. Systèmes d'Organisation de Connaissances (KOS)) publiées et reconnues par la communauté, de l'autre côté. Nous avons conduit nos recherches dans le Laboratoire de Recherche en Imagerie (LRI), spécialisé en recherche préclinique sur le petit animal, du centre de recherche PARCC. Le résultat est une ontologie multi-niveaux implémentée sur un système de BMS-LM (pour BioMedical Study – Lifecycle Management par analogie au PLM). Pour valider notre proposition, nous avons procédé à l'intégration des données et calculs scientifiques du laboratoire LRI dans le système BMS-LM. Nous avons appliqué nos méthodes à (1) des données issues de modalités différentes (TEP-TDM, Histologie, Protéomique) et (2) deux calculs scientifiques au laboratoire LRI : un premier pour la quantification de tissues histologiques et un deuxième pour l'analyse de la Réponse Impulsionnelle (RI) du cœur en imagerie TEP-TDM.

Mots-clés : recherche préclinique, recherche biomédicale, RDM, hétérogénéité des données, PLM, KOS, ontologie, annotation de données, interopérabilité sémantique, intégration de données, intégration de calcul scientifique, gestion du cycle de vie, imagerie multimodale, réutilisation de données scientifiques, partage

Abstract

Frequent changes in methods, data types and personnel are common is biomedical research. This results in significant heterogeneity of research data that become multi-source, multimodal, multidisciplinary, multisite, etc. Heterogeneity hinders the sharing and reuse of scientific data since trust in, and understanding of, the data is at stake. To improve data trust, we have applied the lifecycle management paradigm to biomedical research studies. It is based on a Product Lifecycle Management (PLM) solution from manufacturing industry. Thus, we propose a data management approach with maximum traceability of data provenance throughout the lifecycle of a biomedical research study. Regarding the understanding of data, we defined semantic interoperability between the vernacular terminologies used by research teams, on the one hand, and the standards, terminologies and ontologies (i.e., Knowledge Organization Systems (KOS)) published and recognized by the community, on the other hand. We conducted our research in the Imaging Research Laboratory (LRI) at PARCC, specialized in small animal preclinical research. The result is a multi-level ontology implemented on a BMS-LM system (for BioMedical Study - Lifecycle Management by analogy to PLM). To validate our proposal, we have integrated scientific data and computing workflows from the LRI laboratory into the BMS-LM system. We applied our methods to (1) data from different modalities (TEP-TDM, Histology, and Proteomics) and (2) two scientific computing workflows in the LRI laboratory: one for the quantification of histological tissues and a second for the analysis of the heart Impulse Response (IR) in PET-CT imaging.

Keywords: preclinical research, biomedical research, RDM, data heterogeneity, PLM, KOS, ontology, data annotation, semantic interoperability, data integration, scientific workflow integration, lifecycle management, multimodal imaging, scientific data reuse, sharing

À celui qui a choisi mon prénom,

À toutes les femmes privées d'éducation,

À tout enfant malade,

À la confiance de ma mère,

À la délicatesse de mon père,

À l'Amour,

Amel·

Remerciement

Je remercie l'ensemble des membres du jury, qui m'ont fait l'honneur de bien vouloir étudier mon travail et en particulier Jean Charlet et Sebti Foufou pour avoir accepté d'être rapporteurs de cette thèse, et Anita Burgan, Christophe Egles, Irène Buvat, et Marc Joliot d'avoir accepté de l'examiner.

Ainsi arrive la cérémonie solennelle de coutume pour remercier tous ceux qui ont permis à ce travail d'être ce qu'il est aujourd'hui en mars 2021. Durant tout le manuscrit de thèse, je fus coincée entre une rigueur scientifique d'une part et un style formel de l'autre. Dans ces quelques lignes, les plus personnelles de ce manuscrit, j'ai décidé de briser mes chaînes. Veuillez m'excuser, de cette spontanéité intentionnée.

Je tiens à remercier de prime abord mon ***Amel*** *intérieure qui ne lâche rien. Sans elle, je n'aurais jamais pu arriver jusque-là. Elle a su, malgré le long parcours parsemé de difficultés et d'épreuves, m'apporter un renfort, me guider vers la sortie, et me rassurer quant à l'utilité et l'apport, que mon travail a pour la Science. Ma chère Amel, je te remercie pour ton endurance, ta persévérance et ta patience ! Ce n'est guère une marque d'arrogance, mais au contraire, une humble reconnaissance !*

Au-delà de mon cercle existentiel, mes remerciements vont en premier lieu à ***Philippe****, qui a su se montrer chef, mais paternel, qui a pu m'épauler et me rassurer à chaque étape. Il a su devenir mon premier mentor professionnel, et m'encadrer avec tant de leadership. Merci de m'avoir aidée à tenir le cap, pour que je puisse gérer les périodes de doutes inévitables et fixer des objectifs atteignables. Je t'en remercie !*

Je remercie en second lieu ***Marianne****, qui a assuré un encadrement adéquat pendant ces années de doctorat. N'ayant pas la même vision des choses, nous avons pu avancer, avec nos différentes manières de penser. Vers la fin de la thèse, nous étions beaucoup plus en osmose. Je commençais, malgré mon entêtement, à lui donner raison, et elle commençait, à adapter son encadrement, à ma différente façon. Marianne, j'ai compliqué ta mission et je te demande pardon. Cependant, je pense que nous deux, nous avons réussi à nous adapter, et surtout mener à bien notre projet commun malgré l'enjeu. Je t'en suis reconnaissante.*

Mes remerciements à l'égard de ***Daniel*** *ne peuvent tarder encore. Il m'a soutenue dans toutes les phases de ma thèse. Son oreille attentive et son épaule consolatrice m'ont été d'un grand soutien. Il était mon ami professionnel pendant mes moments de stress émotionnel. Il a su me consoler et rendre mon parcours indolore. Il m'écoutait attentivement en me montrant mes raisons et mes torts. Daniel, tu es un homme en or, et pour l'équipe un trésor. Un peu comme le père spirituel de tous les doctorants, et ce n'est pas une métaphore. Merci d'être là et merci d'être toi.*

Mes remerciements vont après à ***Bertrand****, ma source d'inspiration. Tu es une source intarissable d'idées. C'est grâce à toi que j'ai pu apprendre à m'adapter, et être à l'écoute dans la conduite d'un projet. Tes conseils pratiques, tes propositions pragmatiques, ta façon d'analyser les situations et trouver la plus adaptée des solutions, ton investissement humain quand il s'agit d'aider un doctorant incertain, m'ont beaucoup appris. Tu es un homme de vérité, d'action, et de savoir. J'étais honorée d'être une des doctorants que tu as encadrés. Merci de m'avoir donné cette opportunité.*

Je m'envole à Compiègne où j'ai passé 10% de mon temps. Mes plus respectueux remerciements vont à ***Benoît****, mon directeur de thèse. Benoît m'a appris qu'une thèse est une loi, qu'il faut respecter à la lettre afin de la réussir dans la joie, sinon, c'est l'échec. Moi, avec mon esprit bordélique, je n'ai pas réussi à emprunter très tôt cette voie, et j'ai dû rattraper plusieurs règles à la foi. Un difficile apprentissage, mais au final un bel exploit, grâce à toi ! Tu es la sagesse incarnée, la perfection modérée, et le pragmatisme en soi ! Merci de m'avoir amenée au bon endroit, à chaque fois que mon*

itinéraire devenait étroit. Merci Benoît pour ta patience, ton suivi et tes conseils qui resteront gravés dans ma mémoire pour la vie. Je ne te remercierai jamais assez !

Toujours à Compiègne, j'adresse un remerciement très particulier à ***Alexandre****, mon encadrant qui me met à l'aise en me donnant des conseils pas des moindres. Avec Alexandre, la parole donne lieu à l'action et à l'ordre. J'ai toujours adoré discuter avec toi. J'étais certaine à chaque fois, que j'allais beaucoup apprendre, et c'était vrai. Je te remercie.*

Je remercie chaleureusement ***Pierre-Yves****, qui malgré nos vitesses de croisière très différentes, m'a encouragée et aidée avec des solutions pertinentes. Le lapin qui apprend comment sauter à la tortue plus lente. Même si je ne suis pas aujourd'hui au niveau que tu as espéré, ou que tu aurais pu réaliser, j'espère que j'ai pu être à la hauteur de tes attentes. Merci pour tout ce que tu as fait pour moi, je t'en suis reconnaissante, et bonne continuation dans ce monde de tortues toujours présentes !*

J'adresse un remerciement plein de notes d'amour, à mes compagnons de trajets : ***Thulaciga*** *et* ***Mailyn****. Thulaciga, ton énergie pleine de lueurs, ta gentillesse à plein cœur, et ton hospitalité, qui me rendent toujours confiante et sereine. Mailyn, ton rôle de grande sœur, toujours là pendant mes moments de soupir, pour me donner conseil et me rendre le sourire. Merci à vous mes chères perles rares.*

La liste de remerciement est longue certes, mais je ne peux pas la finir sans remercier mes collègues de Fealinx, de l'INSERM et de l'UTC. ****Fealinx*** *où j'ai appris à respecter et à comprendre les lois de l'entreprise explicites et implicites. Je remercie Thierry, Olivier, Nicolas B, Jérôme, Clément, Cyril, Fabien, pour les échanges enrichissants au sein de l'équipe R&D et pendant les hackathons internes. Je salue Arthur, Nicolas G, Thinh, Meven, Violaine, Ismael, Pierre, Katalin, Nam, Bertrand, Lionnel, qui m'ont accompagnée durant 50% de mon temps de parcours jusqu'aujourd'hui, et ont partagé avec moi les hauts et les bas.* ****L'INSERM*** *où j'ai appris que les relations humaines passent avant tout. Je remercie Thomas, Anikitos, Joevin, Gwennhael, Anais, Caterina, Thulaciga, Roberto, Sofia, Afef, Umit, Imad, Habib, Boris, Mitradeep, Nisreen, Omar, Imen, Alice, Julie, Louise, Fatou, Clément, Gabriel, et Laure. Et j'adresse un petit clin d'œil à Nabila qui a cru en moi et avec laquelle j'ai beaucoup appris.* ****L'UTC*** *où le mot clé est le sourire. Merci à Anne-Sophie, Émeric, Rachel, Julia, Stefan, Mohamed, Philippe, Matthieu, Julien, Joanna, Amélie, Magali, Nassim et Zohra.*

Arrivent ainsi les remerciements les plus chers à l'égard de ***ma famille*** *et de mes amis que je vais écourter, sinon je noircirai des pages faisant l'équivalent de ce document de thèse. La patience de ma mère la combattante, je la remercie. La bonté de mon père l'affectif, je le remercie. La bonne humeur de mes frères et sœurs qui remplissent ma vie de bonheur. Je leur en suis reconnaissante. La relation non possessive que nous avons battue, mon mari et moi, et enfin le sourire de mon fils qui j'espère comprendra un jour pourquoi il était contraint de vivre loin de mes yeux. Merci!*

Loin de mon pays natal, du soleil de la méditerranée, et de ma famille, ***mes amis*** *sont devenus ma deuxième famille. Je remercie Akram, Nour, Nadia, Salma, Abdelmouhaymen, Manel A, Mofida, Ibtihel, Fetia, Danielle, Asma, Sarah, Yasmin, Safa, Hiba, Souhir, Myriam, Abir, Souhaila, Dhouha, Wefa, Imen, Firas, Fatma Zahra, Manel G, Jihed, Chantal, Maram, Arij, Sarra, Sabrine, Yackolley. Je remercie tous ceux qui ont participé de près ou de loin à embellir, ensoleiller et colorier mon quotidien. Je vous aime tous!*

Table des matières

Liste des figures

(pour une impression noir et blanc, voir la version en ligne pour les couleurs et la netteté des images)

Liste des tableaux

Liste des abréviations

ACD	Acquisition Definition
ACQ	Acquisition Result
ADQIV	Annotation property, Data property, Quality, data Item, and Value specification
AGD	Agent Definition
AGR	Agent Result
AIF	Arterial Input Function
AQD	Device
API	Application Programming Interface
AWC	Active Workspace Client
BBR	Bibliographical reference
BOL	Beginning-Of-Life
BOM	Bill Of Material
BPMN	Business Process Model and Notation
BMI-LM	BioMedical Imaging – Lifecycle Management
BMS-LM	BioMedical Study – Lifecycle Management
CAO	Conception Assistée par ordinateur
CDD	Classification Décimale de Dewey
CDW	Clinical Data Warehouse
CLÉ	Cahier de Laboratoire Électronique
COTS	Commercial Off-The-Shelf
CSCW	Computer Supported Collaborative Work
CT	Computed Tomography
DICOM	Digital Imaging and COMmunications in Medicine
DIKW	Data, Information, Knowledge, Wisdom
DMP	Data Management Plan
DRIVE-SPC	Déploiement du Réseau d'Images du Vivant pour les plateformes d'imagerie Expérimentale - de Sorbonne-Paris-Cité
DUD	Data Unit Definition
DUR	Data Unit Result
DVC	Device
EHR	Electronic Health Record
EIF	Entreprise Interoperability Framework
EMR	Electronic Medical Record
EMA	European Medical Agency
EOL	End-Of-Life
ETL	Extract - Transform – Load
EXA	Exam Result
EXD	Exam Definition
^{18}F-FDG	2-désoxy-2-(^{18}F)fluoro-D-glucose
FAIR	Findable, Accessible, Interoperable, Reusable
Five Ws	Who ? do What ? Where ? When ? and Why ?
FTPH	Free Time Point Hermit
GDR	Gestion de Données de Recherche
GED	Gestion Électronique des documents
HIS	Hospital Information System

HL7	Health Level 7
i-c-e	information content entity
IRM	Imagerie par Résonnance Magnétique
ITD	Intervention Definition
ITR	Intervention Result
KBS	Knowledge Based System
KE	Knowledge Engineering
KM	Knowledge Management
KMS	Knowledge Management System
KO	Knowledge Organization
KOS	Knowledge Organization System
LIMS	Laboratory Information Management System
LOINC	Logical Observation Identifiers Names & Codes
LRI	Laboratoire de Recherche en Imagerie
MDD	Modèle De Données
MESH	Medical Subject Headings
MOL	Middle-Of-Life
MRI	Magnetic Resonance Imaging
MS	Mass Spectrometry
NCIT	National Cancer Institute Thesaurus
OBO Foundry	The Open Biological and Biomedical Ontology (OBO) Foundry
ODBC	Open DataBase Connectivity
OME	Open Microscopy Environment
OMERO	Open Microscopy Environment Remote Objects
Ontologie BFO	Basic Formal Ontology
Ontologie DCM	DICOM Controlled Terminology
Ontologie DOLCE	Descriptive Ontology for Linguistic and Cognitive Engineering
Ontologie FBbi	Biological Imaging Methods Ontology
Ontologie IAO	Information Artifact Ontology
Ontologie IOBC	Interlinking Ontology for Biological Concepts
Ontologie MS	Mass Spectrometry Ontology
Ontologie OBI	Ontology for Biomedical Investigation
Ontologie PATIT	Placental Investigative Technique
Ontologie PRO	Protein Ontology
Ontologie PROV-O	PROVenance Ontology
Ontologie QIBO	Quantitive Imaging Biomarker Ontology
Ontologie RO	Relation Ontology
OWL	Ontology Web Language
OWL-DL	Ontology Web Language - Description Logic
PACS	Picture Archiving and Communication System
PARCC	PAris Cadiovascular Research Center
PCD	Processing Definition
PCP	Processing Parameters
PCR :	Processing Result
PDM	Product Data Management
PET-CT	Positrons Emission Tomographie - Computed Tomography
PGD	Plan de Gestion de Données

PLM	Product Lifecycle Management
POC	Proof Of Concept
PUD	Processing Unit Definition
PUR	Processing Unit Result
QQOQCCP	Qui ? Quoi ? Où ? Quand ? Comment ? Combien ? Pourquoi ?
RDF	Ressource Description Framework
RDM	Research Data Management
RFD	Reference Data
RI	Réponse Impulsionnelle
RIS	Radiology Information System
SAD	Sample Definition
SADT	Structured Analysis and Design Technique
SAR	Sample Result
SCP	Secure CoPy
SE	Systèmes Experts
SGP	Subject Group
SI	Système d'Information
SKOS	Simple Knowledge Organization System
SOC	Système d'Organisation des Connaissances
SSH	Secure SHell
SSU	Study Subject
STL	Software Tool
STU	Study
SUB	Subject
SUV	Standard Uptake Value
SVN	SubVersioN
TAC	Tissue Activity Curve
TAL	Traitement Automatique de Langue
TEP	Tomographie par Émission de Positrons
TEP-TDM	Tomographie par Émission de Positrons-TomoDensitoMétrie
TIC	Technologies de l'Information et de Communication
UML	Unified Modeling Language
US	L'échographie ultrasonore
SADT	Structured Analysis and Design Technique
SGDT	Système de Gestion de Données Techniques
SNOMED-CT	Systematized Nomenclature Of MEDicine - Clinical Terms
SUMO	Suggested Upper Merged Ontology
VOI	Volume Of Interest
WFI	Workflow Input
XML	eXtensible Markup Language

Note : Les abréviations utilisées dans ce manuscrit ne sont pas toutes en français ou toutes en anglais. À chaque première occurrence d'un sigle, nous avons présenté les deux abréviations (française et anglaise) et nous avons choisi d'utiliser en fonction de la popularité l'un ou l'autre dans la suite du document.

Glossaire

Ce glossaire donne les définitions des termes « techniques » et précise leur sens tel qu'ils sont utilisés dans le manuscrit. Les {accolades} retrouvées dans les définitions renvoient à un autre terme du glossaire.

Attribut (informatique)	Variable informatique stockant les informations d'une {Classe (informatique)}.
Alignement	Ensemble des {correspondances <a, f, b>} utilisées pour convertir un jeu de données d'un premier {système} vers un deuxième {système} (Bonifati & Velegrakis, 2011).
Alignement automatique	Procédure informatique qui convertit un jeu de donnée d'un système vers un autre en utilisant l'ensemble des {correspondances <a, f, b>} entre les deux systèmes (Bonifati & Velegrakis, 2011).
Application cliente	Voir {Client (informatique)}
Application serveur	Voir {Serveur (informatique)}
Chaîne de traitement	Suite de nœuds de traitement représentant chacun une étape de traitement. Un nœud représente un script, un outil, une fonction, un programme logiciel. La succession de ces étapes permet d'effectuer une analyse sur les données (une quantification, un calcul statistique, etc.)
Changement	Le changement est défini dans cette thèse comme toute spécification (mineure ou majeure) qui apparaît après le démarrage d'une étude de recherche et qui n'a pas été définie dans son cahier des charges initial.
Classe (informatique)	Notion informatique qui consiste à encapsuler dans le même enregistrement, des informations appelées{Attributs}. Par exemple, la classe Personne a des attributs comme : genre, nom, date de naissance. Une Classe peut hériter les attributs d'une autre Classe dans une relation d'héritage entre Classe mère / Classe Fille.
Client (informatique)	Logiciel permettant d'accéder à des fonctionnalités fournies par un programme informatique distant, dit aussi « serveur ». Il est souvent associé à l'architecture réseau client-serveur.
Coffres forts électroniques	Emplacement électronique chiffré pour le stockage sécurisé des données et métadonnées.
Concept	Élément cognitif explicite utilisé par un ensemble de personnes (équipe de recherche, communauté scientifique, société, etc.)
Correspondance <a, f, b>	Triplet composé d'un terme « a », d'une fonction « f » de transformation de ce terme et d'un terme « b » : <a, f, b>. Le terme « a » appartient à un premier {système} et le terme « b » appartient à un deuxième {système} (Bonifati & Velegrakis, 2011).

Cycle de vie	Succession d'étapes, de phases, ou d'états que subit un objet d'étude (organisme vivant, produit commercialisé, produit en cours de fabrication, etc.) La notion de cycle de vie implique des changements, des événements, des procédures, et le suivi dans le temps depuis le début de la vie de l'objet jusqu'à sa fin d'exploitation. Sa ré-exploitation définit le début d'un nouveau cycle de vie.
Descripteur	Élément utilisé pour décrire un concept, le préciser, le définir, etc. Exemple : « format » décrit un « jeu de données », « masse » décrit une « souris », « bleu » et « rouge » spécifient le concept « couleur », etc.
Diagramme BPMN	Modélisation d'une succession d'étapes attribuées à des services et des personnes différentes dans un cadre collaboratif. Le diagramme sert à modéliser les processus de fonctionnement au sein du groupe (entreprise ou tout autre organisme) et aussi à simplifier les étapes d'utilisation ou de fonctionnement des applications logicielles. BPMN est un standard de modélisation reconnu.
Diagramme de classe	Diagramme informatique faisant partie des diagrammes de base du standard UML. Il permet de modéliser les {Classes (informatique)} et leurs {Attributs} ainsi que les liens entre eux : association, héritage, dépendance, composition et agrégation.
Diagramme de séquence	Diagramme informatique faisant partie des diagrammes de base du standard UML. Il permet de décrire un scénario d'utilisation d'une application logicielle ou un module logiciel. Il modélise des communications entre {Classes} et Acteurs (ceux qui interagissent avec le logiciel) via l'échange de messages : synchrone (avec blocage de l'émetteur jusqu'à réception de la réponse), asynchrone (sans blocage)
Diagramme SADT	Diagramme servant à modéliser une succession d'étapes en précisant pour chacune les entrées, les sorties, les variables de contrôles et les outils utilisés.
Essai clinique	{Recherche biomédicale} pratiquée sur des êtres humains volontaires et consentants. Elle comporte classiquement trois phases avant la mise sur le marché d'un nouveau médicament ou une nouvelle méthode de diagnostic ou thérapeutique.
Encodage	Représentation d'un élément du langage ou d'une situation réelle via un code informatique régi par un standard de modélisation.
Extract Transform Load (ETL)	Type de procédure informatique réalisant une transformation de données d'un premier format vers un deuxième format. Elle est paramétrable via des mises en correspondance.
Étude biomédicale	Situation expérimentale dans laquelle une hypothèse biomédicale est testée sur un animal (y compris Homo sapiens) ou sur un dérivé/prélèvement biologique de ce dernier. Elle se déroule en quatre étapes : (1) spécification de l'étude, (2)

	acquisition des données brutes, (3) analyse des données, et (4) publication des résultats. Dans ce manuscrit, une {recherche biomédicale} contient plusieurs études biomédicales. Par exemple : la recherche sur la toxicité d'un médicament peut être déclinée en plusieurs études : étude de l'effet de ce médicament sur le cœur de souris via des techniques in vivo, étude concernant un autre organe, étude utilisant des techniques d'intelligence artificielle, etc.
18F-FDG : 2-désoxy-2-(18F)fluoro-D-glucose	Glucose radioactif marqué au fluor 18. C'est un radiotraceur utilisé en imagerie TEP pour mesurer la consommation locale du glucose par les tissus.
Gestion de connaissances	Domaine qui étudie les connaissances et les expertises d'un groupe de personnes afin de les pérenniser sur le long terme. Les activités de gestion de connaissances sont organisées en un cycle qui commence par leur acquisition, puis leur stockage, puis l'application et l'utilisation de ces connaissances, leur partage, et enfin la création de nouvelles connaissances. Le cycle se referme avec la formalisation des connaissances nouvellement créées. Les connaissances gérées tout au long de ce cycle constituent la « mémoire » du groupe de personnes.
Gestion du cycle de vie	Démarche d'intégration de toutes les étapes, les changements, les événements, et spécificités du {cycle de vie} d'un objet d'étude.
Hétérogénéité des données	Caractère de données dites hétérogènes, c'est-à-dire composées d'éléments de natures différentes qui entraînent une incompatibilité freinant l'exploitation des données.
Hétérogénéité multiple	Lorsque plusieurs axes d'hétérogénéité entre les données sont présents, on parle d'hétérogénéité « multiple ».
Histologie	Domaine de la recherche biomédicale qui s'intéresse à l'étude des tissus biologiques. Exemple : visualisation au microscope de la structure fine d'un tissu prélevé d'un cœur ou une tumeur.
Imagerie TEP-TDM	Type d'imagerie qui repose sur la détection de la radioactivité, induite chez un organisme vivant, par des couronnes de détecteurs de rayons gamma. Elle permet le suivi et la quantification non invasive de la bio-distribution de substances radio-marquées. Exemple : le {FDG}. Elle est souvent couplée au TDM (TomoDensitoMétrie à rayon X) pour apporter une information morphologique.
Import de données	Dans cette thèse, l'import de données est défini comme l'action de transférer des données issues de sources différentes dans un système de gestion de données (e.g. le système BMS-LM). Ce transfert est accompagné d'un ensemble de procédures déclenchées automatiquement afin de permettre au système de gestion de données de reconnaître et gérer ces données.
In vivo	Terme latin qui signifie « au sein du vivant ». Il est utilisé pour désigner les expérimentations pratiquées sur des organismes vivants.

In vitro	Terme latin qui signifie « sous verre ». Il est utilisé pour désigner toute expérimentation réalisée en dehors d'un organisme vivant, à partir de matériel biologique.
Ingénierie de connaissances	Domaine dont l'objet est de modéliser les connaissances et le raisonnement humain afin de les exploiter dans des systèmes informatiques autonomes et intelligents. C'est l'une des disciplines de l'intelligence artificielle (IA), où elle représente un sous-domaine connu sous le nom de l'« IA symbolique ».
Instance (informatique)	Entité informatique créée en suivant le modèle donné par une Classe (informatique). Par exemple : L'Instance (Alexandra, Femme, 13/03/2012) est une instance de la Classe Personne (nom, genre, date de naissance).
Intégration de données	Ensemble des mécanismes et méthodes de lecture, traduction, transformation, et restructuration d'un jeu de données, ayant un format/structure en entrée, permettant de le gérer par un système de gestion de données (e.g. le système BMS-LM) et utilisant son format/structure de sortie.
Intégration de calculs scientifiques	Ensemble des mécanismes et méthodes de structuration en une chaîne de traitement, encapsulation dans un workflow, lancement, et exécution distante d'un calcul scientifique, ainsi que les méthodes d'échange de données et de traçabilité de ces échanges entre le programme (exécutant le calcul) et le système de gestion de données (fournissant les données d'entrée et recevant les données en sortie).
Intelligence Artificielle (IA)	Externalisation des facultés cognitives humaines dans un support indépendant de l'homme.
Interface logicielle (Nipype)	Type de fonction informatique qui permet de modéliser un workflow scientifique avec ses entrées, sorties et fonctions afin de permettre son identification et sa gestion par un gestionnaire de workflows (e.g. Nipype).
Interopérabilité sémantique	« Capacité pour plusieurs systèmes ou composants à échanger des informations et à utiliser les informations échangées » (IEEE, 1991). Dans le domaine de l'organisation des connaissances, elle est définie comme la « capacité de différents KOSs à communiquer de manière à préserver le sens ou la « signification voulue » », traduit de (Patel et al.,2015).
Knowledge Based System	Voir {Système à Base de Connaissances}
Knowledge Organization System	Voir {Système d'Organisation de Connaissances}
Mapping	Voir {Alignement automatique}
Matching	Procédure informatique qui permet de générer un ensemble de correspondances <a, f, b> entre deux systèmes (Bonifati & Velegrakis, 2011).
Mise en correspondance	Voir {Matching}

Modèle de donnée	Template (gabarit en français), pattern (patron en français), représentation, ou schéma respecté par l'organisation et l'instanciation des données dans une base de données. Exemple : la définition des colonnes dans une table.
Modèle de donnée orienté objet	Représentation informatique d'une base de données orientée objet, composée d'entités et leurs associations.
Namespace	Notion informatique qui consiste à étiqueter un ensemble d'entités informatiques avec le même nom – namespace - pour indiquer qu'elles appartiennent au même ensemble.
Objet (informatique)	Lorsqu'une Classe (informatique) est instanciée, l'entité informatique résultante s'appelle un objet.
Ontologie	Spécification formelle et explicite d'une conceptualisation partagée. Traduit de (Studer et al., 1998).
Organisation de connaissances	Étude de la manière dont les connaissances sont socialement organisées et de la manière dont la réalité est organisée. En d'autres termes, il s'agit de la classification et de la structuration des connaissances d'un domaine donné en prenant en considération ses liens avec le social et le réel. Traduit de (Hjørland, 2008).
Partage	Fait de rendre accessible des données à d'autres personnes. L'accès n'est pas limité à l'accès physique ou informatique et inclut aussi la signification et la compréhension de ces données.
Protéomique	Discipline étudiant l'ensemble des protéines (protéome) présentes dans un échantillon biologique à un instant t.
Protocole (biomédical)	Suite coordonnée et strictement définie d'étapes nécessaires pour réaliser une expérimentation, une acquisition ou une analyse de données.
Provenance	Information sur la nature d'une donnée : quand, où et comment a-t-elle été produite ; pourquoi et pour qui a-t-elle été exécutée, traduit de (Simmhan et al., 2005). Information sur la source d'une donnée définie en réponse aux {questions QQOQCCP}. Informations décrivant une donnée et son contexte d'origine et retraçant son historique depuis sa création jusqu'à son partage.
Pyramide de la connaissance	La donnée, l'information, la connaissance, et la compréhension ou « sagesse » forment la pyramide de la connaissance désignée par DIKW (Data-Information-Knowledge-Wisdom) (Rowley, 2007). Elle est utilisée dans les domaines de l'ingénierie des connaissances, des systèmes d'information et de l'intégration de données.
Questions QQOQCCP	Les questions Qui ? Quoi ? Où ? Quand ? Comment ? Combien ? Pourquoi ?
Recherche biomédicale	Ensemble d'{études biomédicales}

Recherche préclinique	La recherche préclinique est définie dans cette thèse comme toute recherche biomédicale pratiquée sur des animaux, avec ou sans débouchés thérapeutiques chez l'être humain.
Recherche translationnelle	Branche interdisciplinaire du domaine biomédical soutenue par trois piliers principaux : la recherche biomédicale, la pratique clinique et sa communauté. Son objectif est de combiner les disciplines, les ressources, l'expertise et les techniques afin de promouvoir l'amélioration de la prévention, du diagnostic et des thérapies. Traduit de (Cohrs et al., 2015). Elle désigne souvent le chemin pour le transfert des découvertes scientifiques du laboratoire au patient, appelée en anglais approche « bench-to-bedside »
Reporting	Documentation des expérimentations, des activités de recherche et des données scientifiques générant un ou des rapport(s).
Réutilisation ultérieure	Utilisation tardive et à long terme des données scientifiques : après la fin de l'étude de recherche au sein de laquelle elles ont été produites, ou après publication en « open access » par exemple.
Sémiotique	Étude des signes et du processus de la signification.
Serveur (informatique)	Programme informatique qui fournit à des applications logicielles – dites Clientes- des directives de fonctionnement, du contenu ou l'exécution d'une tâche demandée par un utilisateur. Il est souvent associé à l'architecture réseau client-serveur.
Système	« Ensemble d'éléments considérés dans leurs relations à l'intérieur d'un tout fonctionnant de manière unitaire. » Source : https://www.larousse.fr/dictionnaires/francais/syst%c3%a8me/76262
Système d'Organisation de Connaissances	Tout type de schéma d'organisation d'information et de gestion des connaissances, traduit de (Hodge, 2000). Structuration de termes et de significations permettant de modéliser la connaissance (pour un jeu de données par exemple, ou pour un domaine particulier, ou d'un point de vue général). Exemples : classifications, terminologies, vocabulaires structurés, glossaires, réseaux sémantiques, ontologies.
Système PLM	Système d'information et système de gestion de données avec traçabilité des processus exécutés sur les données et {workflows} collaboratifs. Au sein d'un groupe, il implémente la notion de gestion du cycle de vie des produits.
Système d'Information	Système d'échange d'informations entre personnes collaborant pour la réalisation d'une tâche précise et dans le cadre d'un projet commun. Système technico-social où la plateforme logicielle joue un rôle important dans la définition de la pratique commune de son utilisation.
Système à Base de Connaissances	Tout système composé d'une base de connaissances (régie par un SOC), d'un moteur de règles ou d'inférences modélisant le

	raisonnement humain et d'une interface homme-machine via laquelle le système interagit avec l'utilisateur.
Taxonomie	En biologie, classification des espèces dans une architecture structurée. Exemple : La classification des êtres vivants de Linné. La taxonomie informatique est une classification de l'information d'un domaine dans une architecture structurée.
Terme	Mot défini par son appartenance à un domaine ou à une discipline.
Thesaurus	Liste alphabétique de mots standards utilisés dans un domaine ou pour un sujet. Exemple : dictionnaire exhaustif comprenant le vocabulaire complet d'une langue.
Versionnage	Action de garder une trace des différentes versions d'une entité informatique (document, logiciel, etc.). Chaque trace – appelée version – est documentée en spécifiant les différents changements qu'elle apporte, et la ou les personnes responsables de ce changement.
Workflow	Suite d'activités aboutissant à la réalisation d'une tâche métier bien définie. Par exemple, la validation des feuilles de temps dans une entreprise, la réalisation d'une expérimentation en Histologie, …etc.

Introduction générale

La « Science Ouverte (*open science*) »

Le 3 janvier 2020, sept échantillons de lavages pulmonaires de patients atteints d'une pneumopathie mortelle arrivent dans le laboratoire du professeur Zhang à Pékin. Le 5 janvier, il partage la séquence d'un virus retrouvé dans ces échantillons avec ses collègues-virologues, dont le professeur Holmes en Australie, grand spécialiste de l'évolution des virus. Le 7, ils identifient un nouveau coronavirus et soumettent un article au journal Nature. Mais, en 4 jours, le nombre de cas reconnus a été multiplié par 200, il faut faire vite ! Holmes convint Zhang de mettre en ligne la séquence sur le site de « Science Ouverte (*open science*) » *virological.org* et l'annonce le même jour par un tweet historique. Immédiatement, des laboratoires européens et nord-américains se mettent au travail, un test diagnostique est mis au point en 10 jours, et… un vaccin en 10 mois ! Cette anecdote racontée par Holmes[1] illustre certes la puissance de la biotechnologie moderne, mais ce qui nous frappe aussi, c'est le rôle crucial qu'a joué le partage de données vérifiables et réutilisables pour combattre une pandémie qui va bientôt faire 10 000 morts par jour. Où en serions-nous aujourd'hui si la séquence de SARS-CoV-2 n'avait pas été rendue publiquement accessible à tous après avoir été analysée, contrôlée et vérifiée par des spécialistes ? S'il avait fallu attendre la publication par Nature et l'autorisation du gouvernement chinois ?

L'« open science » ou science ouverte est « un mouvement visant à rendre la recherche scientifique et les données qu'elle produit accessibles aux personnes de tous les niveaux de la société ». L'« open data » désigne également la mise à disposition des données de recherche en accès libre à grande échelle. Il s'agit d'un mouvement mondial, qui vise à augmenter la transparence et la reproductibilité des résultats scientifiques afin de garantir l'intégrité scientifique. Tout résultat scientifique doit être discuté et mis en question en permanence même pour les journaux prestigieux. Le taux élevé (56,1%) d'articles de la covid-19 publiés sans données patient (Raynaud et al., 2021), et la rétraction des articles de la covid-19, à cause de l'impossibilité d'audit des données brutes utilisées (jeu de données de l'entreprise américaine Surgisphere) (Ledford & Noorden, 2020) est alarmant. Ces événements récents révèlent l'importance de publier non seulement les résultats de recherche, mais aussi l'ensemble du processus, incluant les données brutes et les différents traitements opérés sur ces données, pour mieux comprendre la portée et la fiabilité des résultats. Pour cela, il ne suffit pas de donner accès à une liste de fichiers plus ou moins compréhensibles, il convient de fournir l'ensemble des informations décrivant l'étude de recherche. Elles doivent être lisibles et associées à des indications de provenance pour en fixer le contexte. Dans son livre « La recherche scientifique à l'ère des Big Data » (Leonelli, 2019), Leonelli Sabina parle de l'ère de la *post-vérité*, où à partir d'un même jeu de données, des études de recherche peuvent dresser des conclusions contradictoires, ce qui confirme l'importance du contexte. Le besoin d'améliorer à la fois la « confiance » et la « compréhension » des données partagées est donc particulièrement d'actualité en recherche scientifique, notamment en recherche biomédicale.

La donnée scientifique est centrale à la recherche biomédicale, sa bonne gestion est un facteur décisif de l'intégrité de cette dernière. Plusieurs initiatives en gestion de données scientifiques ou Research Data Management (RDM) ont vu le jour afin de conseiller le chercheur à propos du partage et de la réutilisation de ses données, en particulier, dans le cadre de la vision « open science » du programme Horizon 2020 de l'Union Européenne (DG RTD, 2017).

Dans le domaine de la RDM, les recommandations FAIR et le Data Management Plan (DMP) sont omniprésents. Paru dans (Wilkinson et al., 2016), les principes FAIR ont été vite adoptés et mis en pratique par les professionnels en gestion de données scientifiques. FAIR désigne quatre caractéristiques des données de recherche : F pour *Findable* ou « Trouvable », A pour « Accessible », I pour « Interopérable » et R pour « Réutilisable ». Elles doivent être satisfaites lors de la publication des

[1] https://www.medscape.com/viewarticle/943251

données scientifiques en « Open Data » ; une pratique recommandée par les journaux et les financeurs. Le DMP (Miksa et al., 2019), quant à lui, est un document qui est édité au début de l'étude de recherche et qui doit être mis à jour au fur et à mesure que l'étude avance. Il permet de sensibiliser les chercheurs à leurs stratégies de stockage et analyse de données ainsi qu'aux aspects de réglementation et de la préservation à long terme. Le Data Curation Center (DCC)[2], un centre au UK avec une portée international qui fournit des services pour la RDM, et l'INIST-CNRS[3], Institut de l'information scientifique et technique rattaché au CNRS et ayant pour mission de valoriser et partager l'information scientifique, ont mis en place respectivement le DMPTool[1] et le DMPOpidor[2] (version française se basant sur le DMPTool de DCC), pour permettre aux chercheurs de questionner d'une manière efficace l'environnement éthique, juridique, administratif, budgétaire, scientifique, organisationnel, et informatique des données de recherche qu'ils produisent (voir Figure 1).

En France, des initiatives à vocation pédagogique ont vu le jour, en l'occurrence, le projet DoRANum[3] qui représente un partenariat entre INIST et URFIST[4], un réseau national français de veille et formation à l'information scientifique et technique. Celui-ci met à la disposition des chercheurs, des institutions et des informaticiens, des ressources pédagogiques de formation sous la licence « Creative Commons (CC) » afin de clarifier la thématique de la gestion de données de recherche. Néanmoins, ces initiatives ne sont que des recommandations et des standards qui laissent aux chercheurs le libre choix des outils et actions au jour le jour. Ainsi, la mise en cohérence des pratiques quotidiennes avec les standards de gestion de données de recherche scientifique demeure encore assez floue pour les producteurs de données scientifiques.

Figure 1 Contenu d'un Data Management Plan (DMP)
(source https://doranum.fr/)

Multimodalités et Pluridisciplinarité en recherche biomédicale préclinique

Par ailleurs, plusieurs innovations et découvertes ont vu le jour ces dernières décennies et ont révolutionné la recherche biomédicale : le séquençage du génome humain, l'utilisation des cellules souches pour la synthèse des tissus et organes in vitro, l'impression 3D des prothèses personnalisées, l'implantation de cœur entièrement artificiel (qui est encore en essais cliniques), les avancées en

[2] https://www.dcc.ac.uk/about
[3] https://www.inist.fr/qui/institut/
[4] https://sygefor.reseau-urfist.fr/

techniques d'imagerie médicale, etc. En particulier, l'imagerie est une innovation majeure en recherche biomédicale et pour le diagnostic patient. En matière de diagnostic, elle est devenue un examen de routine hospitalière avec l'accroissement de son utilisation en clinique pour diagnostiquer un accident vasculaire cérébrale, une fracture d'os, un cancer, etc. En recherche, les techniques d'imagerie sont souvent les mêmes en préclinique et en clinique : à un micro-TEP correspond un TEP pour l'homme, à une IRM préclinique, une clinique, les scanners de lames sont les mêmes, etc.

Ainsi, plusieurs modalités d'acquisition de données sont à la disposition du chercheur dans le domaine biomédical, avec une forte présence de l'imagerie, et se déclinent en techniques *in vivo*, d'autres, *ex vivo* et d'autres, *in vitro*. Les techniques in vivo permettent de collecter l'information morphologique, fonctionnelle, métabolique et structurelle d'un tissu, un organe ou une zone d'intérêt d'une manière peu invasive, avec la possibilité d'un suivi longitudinal d'individus. Nous pouvons citer l'IRM, le TEP scan, l'échographie, le scanner X, la radio, etc. Concernant les techniques ex vivo, nous trouvons ceux analysant des échantillons biologiques intacts comme l'histologie, l'anatomopathologie, la visualisation aux microscopes. L'analyse d'échantillons biologiques broyés relève des techniques in vitro, telles que la protéomique, le métabolomique, le fluxomique et la génomique. Nous désignons dans la suite du document par *in vitro* les deux groupes de techniques *in vitro* et *ex vivo*.

Cette variété des techniques d'acquisition de données scientifiques impose une pluridisciplinarité et un multipartenariat inévitable au cours des études de recherche. Le radiologue, le physicien, le médecin, le biologiste, le statisticien et l'informaticien, travaillent à partir d'équipements de pointe et fournissent des expertises spécialisées tout en collaborant ensemble, afin de mener à bien leurs recherches. Les données issues de la recherche biomédicale deviennent de plus en plus complexes et hétérogènes. Elles sont donc multisources, multimodales, multiformats, multidisciplinaires, et nécessitent pour leur acquisition et leur analyse, une collaboration entre plusieurs profils de chercheurs, et un investissement financier non négligeable. D'où le besoin de partage et de réutilisation des données scientifiques afin de (1) accélérer la recherche en facilitant la collaboration entre chercheurs de différentes équipes et expertises, et de (2) renforcer l'intégrité des résultats scientifique.

L'exemple du vaccin contre la covid19 est stupéfiant quant à la rapidité avec laquelle le développement et le transfert de la technologie de vaccination par ARN ont été effectués du laboratoire au patient. Ce transfert de connaissances suit généralement le même schéma de recherche : recherche fondamentale, recherche préclinique, recherche clinique et transfert après autorisation aux pratiques de soins. En effet, la recherche clinique s'appuie très souvent sur les résultats de la recherche fondamentale et préclinique en amont. Notre thèse s'intéresse au domaine de la recherche biomédicale en général et à celui de la recherche préclinique sur le petit animal en particulier. La recherche préclinique est à la fois une recherche fondamentale dans plusieurs domaines biomédicaux, et une recherche avec des débouchés cliniques applicables à l'homme. Nous nous focalisons sur la gestion de données hétérogènes de la recherche biomédicale et préclinique afin de préparer leur partage et leur réutilisation dans le cadre de l' « open science ».

Gestion des données hétérogènes en recherche biomédicale

La « bonne » gestion des données scientifiques est un requis pour toute personne travaillant sur des projets de recherche. Elle est d'autant plus importante dans le contexte actuel où la gestion des données de recherche (RDM) est un sujet d'actualité. Elle inclut les activités historiques d'un chercheur comme : la détention d'un cahier de laboratoire papier ou électronique pour suivre les observations quotidiennes, la mise au point et la standardisation des protocoles utilisés dans un projet de recherche, la préparation des données scientifiques afin de les analyser ou traiter…etc. La liste s'alourdit de plus en plus avec les avancées technologiques et scientifiques à disposition des chercheurs en matière de calcul scientifique et de modalités d'acquisition des données. Ainsi, s'ajoutent à la liste des nouvelles tâches telles que le stockage, l'organisation, l'annotation, la classification, le partage, et l'utilisation des données scientifiques, de plus en plus complexes et hétérogènes.

Les chercheurs, même s'ils exécutent cette liste d'activités quotidiennes de gestion de données avec le maximum possible de rigueur scientifique et de contrôle qualité, semblent débordés par la complexification de cette gestion, qui n'est pas dans leur cœur de métier. Cette problématique engendre de faible annotation et documentation des données scientifiques qui augmente le risque de perte de l'information essentielle pour leur partage et réutilisation. Elle peut se décliner en sous-problèmes plus spécifiques comme suit :

- La prolifération des terminologies locales, ou « jargon », au sein d'un groupe de personnes et le manque de conventions rend difficile la « compréhension » des données par une personne externe.
- La perte d'informations essentielles à la réutilisation de données comme les paramètres d'acquisition ou d'expérimentation limite la « confiance » dans les données.

La « compréhension » des données partagées et la « confiance » dans les leurs sources sont primordiales pour leur partage et leur réutilisation. Nous traitons brièvement ces deux volets dans la suite de l'introduction.

Confiance dans les données scientifiques partagées

La perte d'informations essentielles à la réutilisation de données, comme les paramètres d'acquisition ou d'expérimentation, limite la confiance dans les données et freine leur réutilisation. Les informations décrivant une donnée et son contexte d'origine sont appelées les informations de « Provenance » d'une donnée, et retracent son historique depuis sa création jusqu'à son partage (Simmhan et al., 2005).

Des problèmes semblables ont émergé il y a plus de 20 ans en industrie lors de la mondialisation des processus de conception, fabrication et industrialisation dans l'industrie aéronautique et automobile. La collaboration multisite et les différentes langues et contextes réglementaires et juridiques entre les collaborateurs au sein d'une même entreprise ont confronté le domaine à la problématique de la « confiance » dans les données multisites, multimodales, multiformats, multidisciplinaires, et leurs sources. En conséquence, la notion de la « gestion de cycle de vie des produits » (Product Lifecycle Management ou PLM) est apparue dans le domaine de l'industrie manufacturière. Elle est définie comme une « démarche intégrée de gestion des informations des produits pendant toutes les étapes de sa vie : conception, fabrication, distribution, utilisation, maintenance, retrait ou recyclage » (Terzi et al., 2010) avec un maximum de traçabilité.

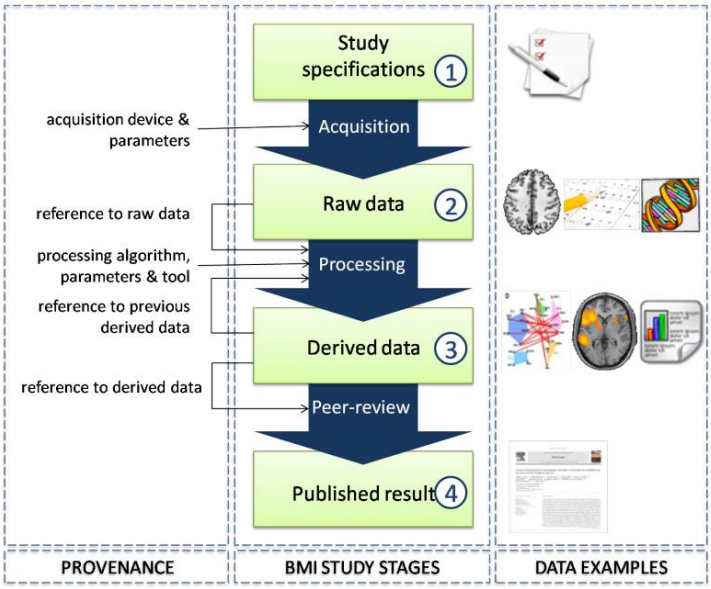


Figure 2 Étapes du cycle de vie d'une étude de recherche biomédicale avec traçabilité de la provenance (Allanic, 2015)

Plusieurs similarités entre industrie manufacturière et recherche biomédicale peuvent être relevées : hétérogénéité des données, pluridisciplinarité, collaboration complexe, etc. Lors de la thèse de la Dr Marianne Allanic (Allanic, 2015), le modèle de donnée BMI-LM (BioMedical Imaging – Lifecycle Management) a été proposé pour la gestion des données de recherche hétérogènes en neuroimagerie, en tenant compte de la traçabilité des données et de leur provenance. En plus de la provenance, ce modèle de donnée couvre toutes les étapes d'une étude – appelé « Cycle de vie » d'une étude de recherche biomédicale : (1) la spécification de l'étude, (2) l'acquisition des données brutes, (3) la production des données dérivées, et (4) la valorisation (voir Figure 2).

L'originalité des travaux de (Allanic, 2015) réside dans l'application de concepts développés dans un domaine (l'industrie) à un autre domaine (la neuroimagerie). Les résultats des travaux de (Allanic, 2015) sont encourageants quant à l'amélioration de la confiance dans les données hétérogènes qui ont été gérées via le modèle BMI-LM en neuroimagerie. Ainsi, nous adoptons la même hypothèse de recherche en l'élargissant à d'autres domaines de recherche biomédicale à savoir : la recherche préclinique.

Compréhension des données scientifiques pour réutilisation

Même avec un système de gestion de données, la compréhension des données partagées peut s'avérer un problème dans le cadre de données hétérogènes, provenant de plusieurs sources et domaines de recherche (imagerie, biologie, etc.). L'utilisation des termes locaux vernaculaires ou « jargon » pour l'annotation des données scientifiques rend difficile la compréhension de ces données par une personne externe (ou pas du domaine) et donc freine leur réutilisation. Pourtant, les termes vernaculaires sont fortement utilisés pour l'annotation des données en recherche scientifique pour deux raisons. La première est en lien avec le manque de temps et de sensibilisation pour l'utilisation des standards de domaine. La deuxième est en lien avec la nature même de la recherche scientifique qui évolue plus rapidement que les standards publiés dans le domaine en question, et qui impose aux chercheurs de définir eux-mêmes leurs terminologies locales pour annoter, gérer et partager leur recherche. Dans les deux cas, les terminologies locales freinent la réutilisation des données par un chercheur externe à l'étude, ou arrivé tardivement dans l'étude. Par conséquent, dans la pratique scientifique actuelle, la réutilisation des données s'effectue en étroite collaboration avec son producteur initial, ce qui est fortement recommandé mais pas toujours possible, en particulier à cause du fort turn-over rencontré dans la recherche publique. Par ailleurs, (Pasquetto, 2018) classifie ce type de réutilisation comme étant rare (p211).

Figure 3 La pyramide de la connaissance (Rowley, 2007)

L'étude des terminologies et des annotations des données relève des trois domaines de la gestion, de l'organisation, et de l'ingénierie des connaissances. Ces domaines traitent le lien entre les données, les informations qu'ils délivrent, les connaissances acquises par conséquent, ainsi que la compréhension

qui en découle. Nous en déduisons que la compréhension des données scientifiques après leur partage peut être améliorée en utilisant les avancées de ces trois domaines de la connaissance. La donnée, l'information, la connaissance, et la compréhension ou « sagesse » forment la pyramide de la connaissance (voir Figure 3). Elle est souvent désignée par DIKW (Data-Information-Knowledge-Wisdom) (Rowley, 2007) et est fortement utilisée en ingénierie des connaissances.

Nous nous focalisons sur la compréhension des données scientifiques sans avoir besoin d'interprètes, lors de leur partage et de leur réutilisation, en prenant en compte l'existence de terminologies locales qui, selon notre proposition, doivent être alignées avec les terminologies reconnues par la communauté scientifique. Ceci relève de l'« interopérabilité sémantique », identifiée par les principes FAIR cités auparavant, et élément phare des domaines de la connaissance.

Problématique de la thèse

Nous émettons une première hypothèse de recherche en énonçant que la gestion des données scientifiques et de leur provenance tout au long du cycle de vie d'une étude de recherche biomédicale prépare et facilite leur partage et leur réutilisation ultérieure en renfonçant la confiance dans les données. Cette hypothèse a été validée pour les données en recherche en neuroimagerie. Nous continuons de l'explorer et de la valider dans d'autres domaines en recherche biomédicale ; i.e. la recherche préclinique.

Notre deuxième hypothèse relève des domaines de la connaissance et énonce que la mise en place d'une interopérabilité sémantique entre les terminologies locales utilisées pour l'annotation des données et ceux reconnus par la communauté scientifique améliorera leur compréhension lors de leur partage.

Nos travaux de recherche ont été menés dans le cadre du projet de collaboration public-privé, DRIVE-SPC (Déploiement du Réseau d'Images du Vivant pour les plateformes d'imagerie Expérimentale de Sorbonne-Paris-Cité). Le partenaire industriel est l'entreprise Fealinx, anciennement Cadesis. Les partenaires académiques sont le Laboratoire de Recherche en Imagerie (LRI), désigné aussi comme l'équipe 2 du « Centre de Recherche Cardiovasculaire de Paris (PARCC) » sous tutelle INSERM, et rattaché à l'Université de Paris, et l'équipe SIPP (Système d'Information, Produit, Process) du laboratoire Roberval à l'Université de Technologie de Compiègne (UTC). Dans cette thèse, nous explorons nos deux hypothèses de recherche dans le cadre de la gestion de données hétérogènes issues de la recherche multimodale préclinique au laboratoire LRI.

Le manuscrit est organisé en 7 chapitres et trois grandes parties :

PARTIE A : État de l'art

Chapitre1 – présentation du cadre de nos travaux de recherche, du positionnement de nos travaux par rapport aux cycles de vie identifiés en recherche biomédicale, et présentation des modalités d'acquisitions, formats de données et outils d'analyses utilisés dans le laboratoire LRI et en recherche préclinique.

Chapitre2 – étude des systèmes et des besoins en gestion des données hétérogènes en recherche biomédicale (préclinique et clinique), en imagerie, en biologie, et en suivi journalier : les travaux identifiés dans la bibliographie sont présentés et analysés.

Chapitre3 – état de l'art en gestion des connaissances (KM), organisation des connaissances (KO) et ingénierie des connaissances (KE) : leurs applications pour l'interopérabilité sémantique en recherche biomédicale et en industrie sont présentées.

PARTIE B : Propositions et méthodes de mise en œuvre du paradigme BMS-LM (BioMedical Study - Lifecycle Management)

Chapitre4 – proposition et méthodes de mise en œuvre du système BMS-LM pour la gestion de cycle de vie des études de recherche biomédicales avec traçabilité des données hétérogènes et des calculs scientifiques ainsi que leur provenance pour renforcer leur confiance.

Chapitre5 – définition et méthodes de conception d'une représentation ontologique multi-niveaux (haut-*top*, noyau-*core*, domaine-*domain*, locale-*local*) afin de permettre une interopérabilité sémantique entre les différentes données manipulées tout au long d'une étude de recherche, et ainsi faciliter leur compréhension lors de leur partage et réutilisation.

PARTIE C : Application pour la recherche préclinique et discussion des résultats

Chapitre6 – applications et exemples du domaine préclinique pour la validation des différentes propositions de gestion de cycle de vie des études de recherche biomédicales (BMS-LM).

Chapitre7 – discussion des résultats, conclusion des travaux et présentation des perspectives de recherche

Chapitre I. Contexte et questions de recherche de la thèse

Cette thèse s'insère dans le cadre général de la gestion de données de recherche biomédicales (RDM) pour faciliter leur partage et leur réutilisation ultérieure. Elle a été menée dans un cadre de partenariat pluridisciplinaire qui fait intervenir plusieurs domaines de recherche : domaine biomédical, domaine informatique et domaine industriel. Ce chapitre présente les différents éléments de contexte qui ont participé à son élaboration. Les domaines d'application : la recherche biomédicale et la recherche préclinique sont présentées en premier. Les données hétérogènes au laboratoire LRI sont décrites ensuite. La gestion de cycle de vie en industrie et son application à la neuroimagerie sont détaillées en troisième. Le chapitre s'achève sur la description des différentes questions de recherche qui ont été posées dans le cadre de cette thèse.

I.1. RECHERCHE BIOMÉDICALE ET RECHERCHE PRÉCLINIQUE

Dans cette section, nous situons la recherche et l'imagerie préclinique, domaine d'expertise du laboratoire LRI, dans le contexte global de la recherche biomédicale.

L'article (L1121-1, 2008) du Code de la Santé Publique définit les recherches biomédicales comme étant : « les recherches organisées et pratiquées sur l'être humain en vue du développement des connaissances biologiques et médicales … ». La recherche pratiquée sur des humains est régie par plusieurs lois : (Loi Huriet-Sérusclat, 1988), (Loi Jardé, 2012). Ces lois visent à réglementer les pratiques en recherche biomédicale et permettent de garantir le respect des droits des êtres humains. D'ailleurs, avant même de commencer une étude de recherche, une autorisation doit être obtenue d'un « comité d'éthique de la recherche » rattaché au ministère et s'assurant que la recherche proposée respecte le cadre des lois. Tout au long de l'étude ensuite, des procédures en lien avec l'intégrité scientifique doivent être entreprises : la validation du cahier du laboratoire par un témoin, la documentation rigoureuse des expérimentations, etc.

Trois grandes catégories peuvent être identifiées en recherche biomédicale : la recherche clinique (souvent appelé essai clinique), la recherche préclinique (sur l'animal) et la recherche fondamentale. Souvent, l'appellation « Recherche Translationnelle » est utilisée pour désigner le chemin pour le transfert des découvertes scientifiques du laboratoire au patient, en anglais appelé approche « bench-to-bedside ». Cette appellation est une traduction de l'anglais « Translational Research (TR) ». Pour plus d'exactitude, l'appellation "Recherche Traductionnelle" aurait dû être utilisée. L'EUSTM (*European Society for Translational Medicine*) défini la « Translational Medicine (TM) », une autre appellation de la TR, comme (Cohrs et al., 2015):

« *An interdisciplinary branch of the biomedical field supported by three main pillars: bench side, bedside and the community. The goal of TM is to combine disciplines, resources, expertise, and techniques within these pillars to promote enhancements in prevention, diagnosis, and therapies*. » [5]

Dans le reste du document, nous allons utiliser l'appellation « recherche translationnelle ». Elle désigne alors tous les efforts de traduction en essais cliniques des recherches fondamentales et précliniques d'un nouveau procédé thérapeutique, ou une nouvelle technique de diagnostic ou de prévention. Elle peut ne jamais aboutir à une mise sur le marché. Ce n'est pas un processus défini dans les détails dès le départ

[5] En français : une branche interdisciplinaire du domaine biomédical soutenue par trois piliers principaux : la recherche biomédicale, la pratique clinique et la communauté. L'objectif de la MT est de combiner les disciplines, les ressources, l'expertise et les techniques au sein de ces piliers afin de promouvoir l'amélioration de la prévention, du diagnostic et des thérapies.

comme pour les produits industriels, mais un processus qui s'adapte et se modifie en fonction des nouvelles connaissances acquises au fur et à mesure que les études de recherche avancent. La Figure 4 montre une trame classique du long chemin nécessaire pour obtenir l'Autorisation de Mise sur le Marché (AMM) des innovations scientifiques en santé et explicite la différence entre les trois phases des essais cliniques communément référencées.

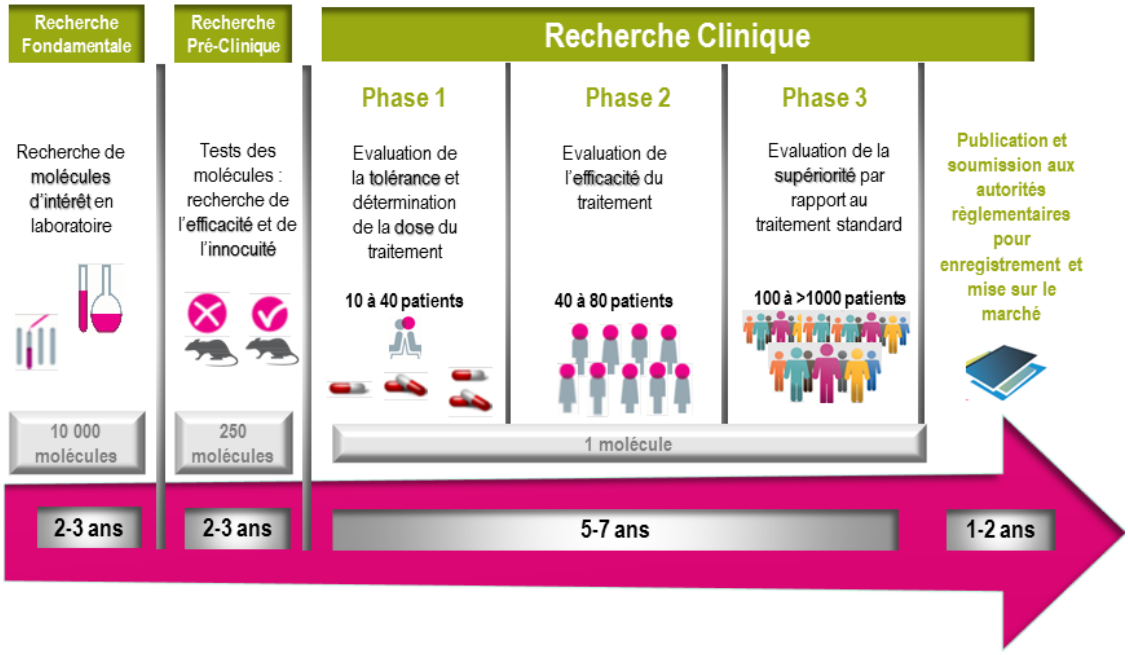


Figure 4 La recherche translationnelle et les trois phases des essais cliniques (source : https://www.icm.unicancer.fr/fr/recherche/la-recherche-clinique)

La recherche préclinique est définie ici comme toute recherche biomédicale pratiquée sur des animaux, avec ou sans débouchés thérapeutiques chez l'être humain. Elle est souvent sous-estimée et réduite à une étape de l'approche translationnelle en recherche biomédicale. Cependant, il s'agit d'un domaine de recherche à part entière avec ses spécificités, problématiques, protocoles, expertises, et méthodes. En effet, l'expérimentation sur des animaux implique : une gestion du séjour de l'animal du début jusqu'à son euthanasie, une gestion du riche panel d'expérimentations multimodales, des échantillons prélevés sur l'animal, ainsi que des analyses multiparamétriques et variées des données récoltées.

Dans une approche de recherche translationnelle, les résultats de la recherche préclinique sont décisifs pour la suite de l'étude biomédicale. Ainsi, l'analyse des données précliniques vise à être quantitative et accorde une importance capitale aux tests statistiques (Dillenseger, 2017).

La recherche préclinique peut être pratiquée in vivo (sur l'animal vivant) ou sur des cellules ou des échantillons biologiques. L'imagerie biomédicale préclinique in vivo est l'une des méthodes fortement utilisée et recommandée vu son caractère non ou peu invasif. L'imagerie biomédicale, en biologie et en médecine, est utilisée comme modalité de découverte d'informations in situ, non accessible par d'autres méthodes exploratoires (observation de surface ou prélèvements de fluides) dans les tissus biologiques d'un organisme vivant. Il s'agit d'un mode d'expérimentation et de recherche non invasive dans la majorité des cas et qui permet d'accéder à l'information morphologique, structurelle, fonctionnelle et moléculaire. Lorsque l'imagerie est effectuée sur des prélèvements biologiques elle est considérée une imagerie in vitro, une imagerie invasive. Nous distinguons, en effet, deux types d'imagerie :

- L'imagerie in vitro invasive : les images réalisées après extraction d'échantillons biologiques de l'organisme d'étude. Notamment, les images histologiques d'immunofluorescence ou de coloration tissulaire, les images de western-blot et les images par spectrométrie de masse (MSI).

- L'imagerie in vivo peu invasive : les images réalisées sur un organisme vivant sous anesthésie ou pas. Notamment, les images à rayon X, les échographies ultrasonores, les scans d'Imagerie par Résonnance Magnétique (IRM) et l'image de tomographie à émission de positon (TEP).

La Figure 5 (Khalil, 2017) cartographie le type d'informations morphologiques et biologiques que ces modalités d'imagerie véhiculent, depuis la morphologie des organes jusqu'aux informations génétiques et moléculaires en passant par l'information fonctionnelle, physiologique et métabolique.

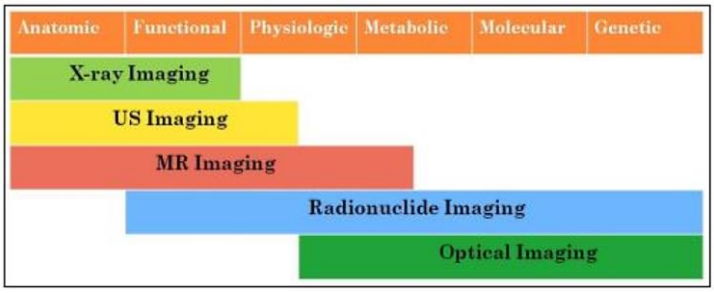


Figure 5 Cartographie des différentes modalités d'imagerie en fonction de l'information délivrée (Khalil, 2017)

Selon les statistiques françaises de 2019[6], la souris (Mus musculus) est l'animal le plus utilisé (61% des animaux utilisés en 2019), viennent ensuite les poissons (12% toutes espèces confondues), le rat (Rattus norvegicus 9%), et ensuite le lapin (Oryctolagus cuniculus 7%). La recherche préclinique se pratique majoritairement sur des animaux de petite taille, l'imagerie du petit animal constitue une branche importante de l'imagerie préclinique et de la recherche préclinique en général.

La Figure 6 positionne l'imagerie du petit animal dans le contexte général de la recherche biomédicale. Il est à noter que des alternatives à l'expérimentation animale sont en train de voir le jour comme l'utilisation des simulations numériques, d'organes artificiels (organoïdes) ou de montages micro-fluidiques plus ou moins sophistiqués. Dans le même ordre d'idée, les fantômes sont des objets géométriquement définis largement utilisés en imagerie pour calibrer l'acquisition et la reconstruction des images.

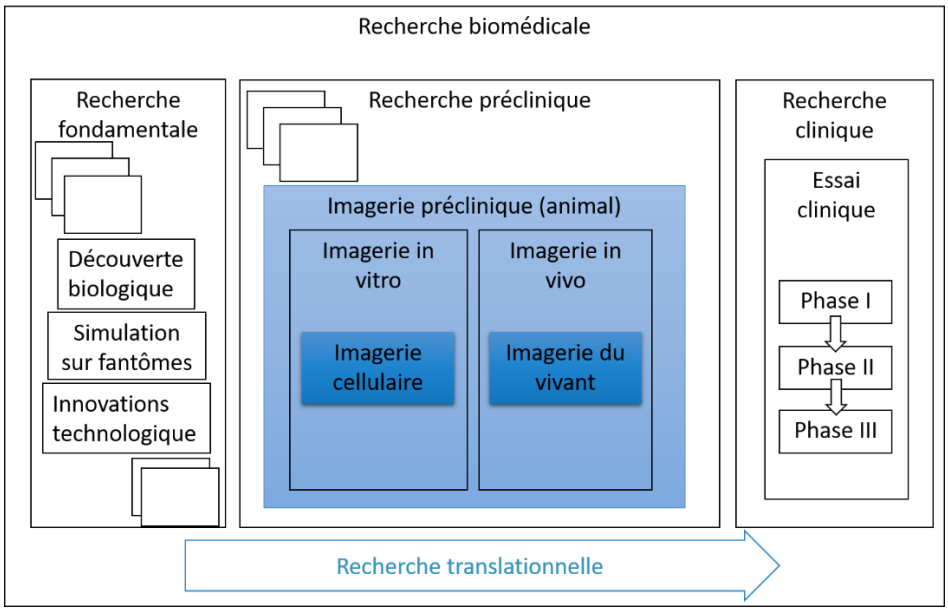


Figure 6 Relation entre imagerie préclinique et recherche biomédicale

[6] https://cache.media.enseignementsup-recherche.gouv.fr/file/utilisation_des_animaux_fins_scientifiques/85/8/STAT_2019_1371858.pdf

La recherche préclinique peut être vue comme étant une étape de la recherche translationnelle ou comme étant une discipline à part entière. Dans cette thèse, nous optons pour le deuxième point de vue et nous ne traitons pas les aspects translationnels. Notre domaine d'application est la recherche préclinique est et notre domaine de recherche est la gestion de données de recherche biomédicale.

I.2. DONNÉES HÉTÉROGÈNES DE LA RECHERCHE PRÉCLINIQUE AU LABORATOIRE LRI

La notion de données hétérogènes remonte aux années 70 avec les premiers essais d'harmonisation des différents systèmes de gestion de base de données existants (Adiba & Portal, 1978). Elle est aussi étroitement liée aux Systèmes d'Information (SI) au début de leur émergence. Ces derniers se sont confrontés à plusieurs sources et schémas de données qui ne sont pas forcément intégrables avec eux ni compatibles entre eux (Hidde, 1991).

La bibliographie abondante traitant le problème des données hétérogènes à des niveaux différents attribue, in fine, le terme « données hétérogènes » à toute situation où l'on observe une « incompatibilité ». Nous décrivons donc dans cette section les données et leurs formats, leurs modalités d'acquisition, et les outils d'analyse utilisés en recherche préclinique et dans le laboratoire LRI, afin de clarifier les facteurs d'hétérogénéités auxquels nous faisons face.

Les techniques d'imagerie in vivo telles que le TEP scan, l'IRM, et l'échographie ultrasonore sont décrites en premier. Après, des acquisitions de données in vitro, comme pour l'histologie, le western-blot et la protéomique, sont détaillées. Les explications fournies pour chaque modalité suivent le schéma suivant : brève présentation ou histoire, puis, explication du déroulé d'une expérimentation, avant, pendant et après, l'acquisition des données. Après la présentation de ces mécanismes d'acquisition (§I.2.1), nous nous focalisons, dans les sections suivantes (§I.2.2 et §I.2.3), sur les formats de données et les types d'outils utilisés en recherche préclinique. En l'occurrence, nous présentons une « photographie » à un instant t des outils utilisés au laboratoire LRI. Pour finir, nous dressons un bilan des données hétérogènes en recherche préclinique.

I.2.1. MÉCANISMES D'ACQUISITION DE DONNÉES EN RECHERCHE PRÉCLINIQUE

Dans cette section, nous présentons tout d'abord les acquisitions in vivo, et après ceux in vitro. Nous consacrons un dernier paragraphe pour introduire un autre type de données rencontré en recherche préclinique. Il s'agit de la documentation et le suivi journalier des résultats et expérimentations en lien avec l'activité de recherche quotidienne.

I.2.1.1. L'imagerie in vivo

Dans ce paragraphe, nous présentons trois modalités d'acquisition d'image in vivo. L'imagerie TEP, l'imagerie IRM et l'échographie ultrasonore. Pour chacune des modalités, un aperçu de son principe, de son histoire et de ses applications est effectué ainsi qu'une description du déroulement typique de l'examen d'imagerie suivit de l'analyse d'images.

I.2.1.1.1. L'imagerie TEP (Tomographie par émission de positons)

L'imagerie TEP est un socle de l'imagerie moléculaire in vivo. Elle repose sur la détection de la radioactivité induite dans un tissu biologique via l'utilisation de molécules radioactives ou radiomarquées appelées traceurs. Elle permet un suivi et une quantification non invasive de la biodistribution des substances radioactives in vivo.

Ter Pogossian a réalisé la première étude de TEP du cerveau humain en 1975 (Rich, 1997). L'intérêt croissant pour les études d'imagerie préclinique, la recherche fondamentale biomédicale et la recherche pharmaceutique a favorisé l'utilisation de la TEP chez les petits animaux. (Schnöckel et al., 2010)

L'imagerie TEP est utilisée en imagerie médicale clinique principalement pour la détection des tumeurs en oncologie, via le suivi du métabolisme pathologique du glucose radioactif dans le corps. La TEP est souvent couplée avec la TDM (TomoDensitoMétrie X) ou en anglais CT (Computed Tomography) afin de compléter l'information métabolique du TEP par l'information morphologique du TDM. La TEP-TDM (ou PET-CT) permet d'évaluer l'étendue du cancer et son stade (1, 2, 3...). Il permet aussi de différencier les lésions bénignes des lésions malignes. Il est aussi utilisé pour l'évaluation de la réponse de la maladie à la chimiothérapie ou à la radiothérapie et aussi pour la planification d'une biopsie, une chirurgie ou une radiothérapie (Almuhaideb et al., 2011).

^{18}F-FDG est le traceur utilisé en oncologie pour les utilisations citées auparavant. ^{18}F-FDG est une abréviation de 2-désoxy-2-(^{18}F)fluoro-D-glucose. Il s'agit du glucose marqué au fluor 18. Toutefois, une multitude de traceurs sont proposés pour le suivi de divers processus physiologiques (Graham, 2012). Ils sont pour l'instant beaucoup plus utilisés en recherche préclinique en attendant leur « translation » éventuelle du laboratoire au patient. En général, les traceurs ont une demi-vie (temps nécessaire pour que la radioactivité initiale diminue de moitié) limitée ce qui complique leur utilisation. Par exemple, la demi-vie du ^{18}F-FDG est de 110 min.

Un examen d'imagerie TEP suit un workflow (suite d'activités) bien défini qui passe par plusieurs étapes (Myers et al., 1999). En amont, la vérification de l'adéquation (résolution, sensitivité) du scanner TEP pour le besoin de l'étude. Au laboratoire LRI, un microTEP de résolution qui peut aller jusqu'à 0.7mm et de sensitivité maximale de 9% est utilisé. La configuration du microTEP via le logiciel du constructeur permet de sélectionner ou concevoir le protocole approprié pour la manip.

Ensuite l'examen lui-même, l'opérateur prépare le lit adéquat pour le petit animal et prépare l'animal selon un protocole défini et validé par le comité d'éthique animal (une chirurgie bien spécifique, une dose de traitement bien définie, etc.). L'animal doit être anesthésié et immobilisé pour limiter au maximum les artefacts dans l'image. Il y a l'injection du radiotraceur en mode intraveineux (ou rétro orbital, ou intrapéritonéal, ou par gavage (per os)) et le contrôle de la dose injectée en fonction du poids de l'animal et d'autres paramètres. En effet, comme la radioactivité est dangereuse à forte dose mais indispensable pour le déroulement de l'acquisition, une expertise pointue est demandée pour contrôler et suivre la dose injectée.

Cela étant, l'acquisition des données brutes commence quand la machine, l'animal et l'opérateur sont prêts. Et elle peut durer jusqu'à 1h ou plus. Les données brutes sont sous forme de coordonnées spatiales d'enregistrements d'événements qui correspondent à la détection de deux photons juxtaposés dans le détecteur du scanner appelé scintigraphie. Ce mécanisme est explicité dans la Figure 7.

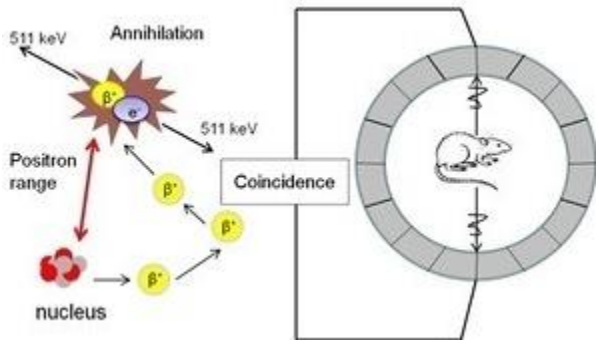


Figure 7 Principe de la scintigraphie en imagerie TEP
(Source : https://www.lbic.lu.se/platforms/preclinical-nuclear-medicine/basic-principles-petct-and-spectct)

Cette liste d'événements permettra plus tard de reconstruire l'image TEP en utilisant les données TDM et en suivant les étapes standards décrites par (Zaidi, 2007) dans la Figure 8.

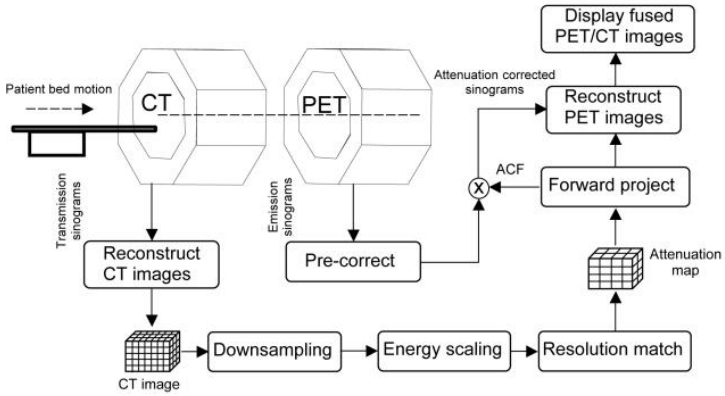


Figure 8 Workflow typique de reconstruction TEP (Zaidi, 2007)

Une fois l'acquisition achevée, le traitement et reconstruction des images suivent. Deux modes de reconstruction sont envisageables : le mode **dynamique** qui reconstruit la vidéo de l'acquisition et trace la migration de l'agent radioactive dans le corps et le mode **statique** qui se focalise sur la fin de la migration et les zones ou la radioactivité est la plus présente en fin d'acquisition, l'image sur la Figure 9 présente une radioactivité élevée en fin d'acquisition au niveau de la vessie et du cœur.

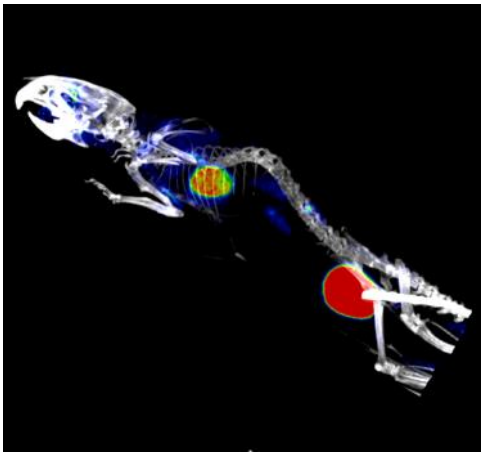


Figure 9 Reconstruction statique en 3D d'une image TEP d'une souris saine

Une fois les images reconstruites obtenues, l'opérateur analyse les régions d'intérêt (ROI) pour quantifier la radioactivité et ainsi avoir une information sur la consommation du glucose radioactive dans la région d'intérêt ROI (Region Of Interest) : tumeur, organe, vaisseaux…etc. L'unité de mesure est la valeur SUV (Standardized Uptake Value) ou le MBq (Mégabecquerel). PMOD[7] est le logiciel utilisé dans le laboratoire LRI. D'autres logiciels ouverts existent comme Amide[8] ou ImageJ[9].

Enfin, les résultats d'analyse et les images sont interprétés par des groupes d'experts ou individuellement à la vue de la question d'étude initiale afin de permettre la découverte de nouvelles connaissances.

[7] https://www.pmod.com/web/
[8] http://amide.sourceforge.net/
[9] https://imagej.nih.gov/ij/

I.2.1.1.2. L'Imagerie par Résonance Magnétique (IRM)

L'Imagerie par Résonance Magnétique (IRM) est une technique d'imagerie médicale découverte par Peter Mansfield et Paul Lauterbur, Prix Nobel 2003 de physiologie ou médecine, dans les années 70[10]. Elle repose sur le principe de la Résonance Magnétique Nucléaire (RMN), une propriété magnétique de certains atomes qui, une fois excités par un champ magnétique, émettent un signal magnétique qui les caractérise avec deux temps de relaxation (T1 et T2).

L'IRM est une technique d'imagerie complexe qui permet d'accéder à des informations morphologiques et fonctionnelles grâce à un paramétrage complexe des séquences d'acquisitions qui peuvent être pondérées en T1, T2, comme dans la Figure 10 ci-après, ou aussi densité de proton, etc. L'atome d'hydrogène présent notamment dans les molécules d'eau (H2O) est à l'origine de la résonance magnétique mesurée en IRM. Plus la présence de l'eau dans les tissus observés est forte, plus le signal détecté est fort, i.e. la zone qui lui correspond, en image IRM pondérée en T2, est plus blanche.

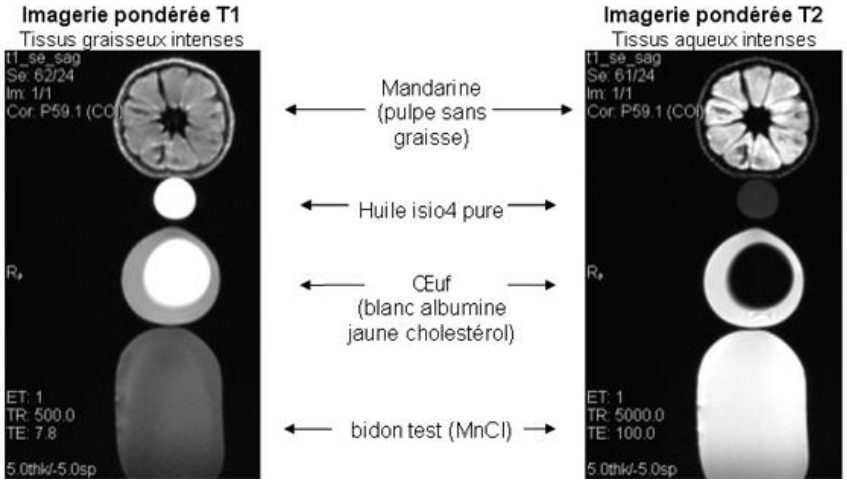


Figure 10 Différence entre acquisitions IRM pondérées en T1 ou en T2 (source : http://irmcardiaque.com/index.php?title=Temps_relaxation)

Vu son caractère peu invasif, l'IRM est beaucoup utilisée en clinique, elle représente un examen clinique de routine. Elle permet de visualiser le système nerveux, les muscles, le cœur, et les tumeurs en utilisant différentes séquences IRM adaptées à l'organe pour visualiser différentes caractéristiques dans les coupes imagées par l'opérateur. Typiquement, un examen IRM en recherche contient entre 5 et 10 séquences. La série d'images ci-après (Figure 11) présente chacune un type de séquence IRM avec des applications en gynécologie, oncologie, et neurologie.

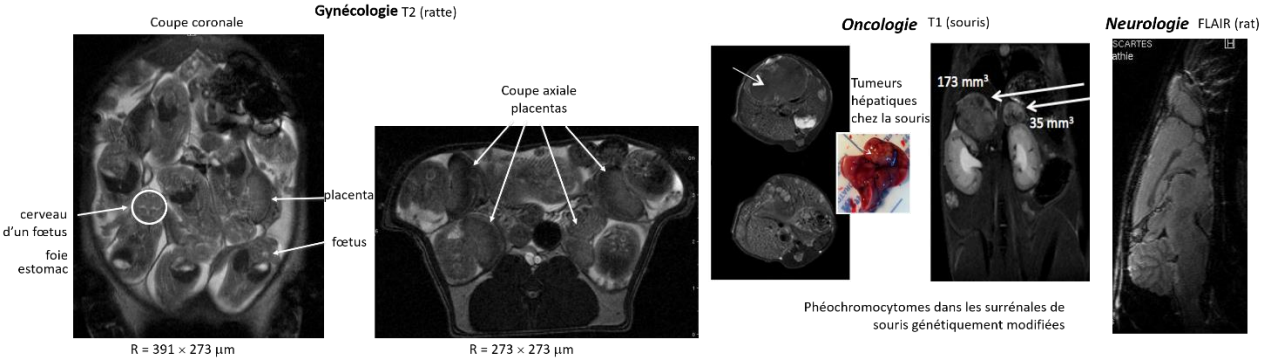


Figure 11 Série d'images précliniques en IRM dans différents domaines (source : plateforme IRM du PARCC)

[10] http://www.fonar.com/timelineofmri.htm

Nous pouvons aussi utiliser des agents de contraste qui modifient l'intensité IRM en fonction de leur concentration locale. Les agents de contraste les plus courants sont constitués d'un atome de Gadolinium intégré dans une plus grande molécule.

Le déroulement de l'acquisition des images IRM en préclinique s'effectue selon un workflow (suite d'activités) bien précis. En amont, l'opérateur prépare la machine d'IRM en choisissant l'Antenne et le Fourreau de gradients (bobines du champ magnétique) adéquats à l'animal, sujet de l'examen d'IRM. Ce dernier est anesthésié sous Isoflurane (ou autre produit anesthésiant). Tout au long de l'expérimentation, il faut veiller à régler le taux d'Isoflurane de l'animal pour le maintenir à un niveau bas, mais suffisant. L'animal est après logé sur le berceau et positionné au centre de l'aimant. Une séquence d'acquisitions d'images appelée « Tripilot » doit être effectuée comme étalonnage afin de vérifier que tout est bien installé et positionné. L'antenne ne doit être branchée qu'une fois l'animal et l'antenne en bonne position. Après cette préparation méticuleuse, l'opérateur est prêt à acquérir les séquences d'images prévues pour son expérimentation. Des exemples d'images issues de ces séquences sont présents Figure 11 : séquence T1, séquence T2, séquence FLAIR, etc. Elles varient selon l'organe étudié.

Une fois acquises, les images sont visualisées, recadrées et traitées par des outils propriétaires fournis avec la machine d'IRM ou par des outils « maison » développés par le laboratoire pour un besoin plus précis, qui a émergé de l'étude en cours. L'outil phare utilisé au laboratoire LRI est Matlab, comme décrit plus tard dans ce chapitre.

I.2.1.1.3. L'échographie ultrasonore (US)

L'échographie est la technique d'imagerie la plus accessible et la moins coûteuse pour les patients en imagerie clinique. L'ultrason permet de capturer des images morphologiques du corps humain. Elle a une longue histoire en pratique clinique surtout pour le suivi d'une grossesse et les investigations d'urgence. Elle a été introduite dans les années 50 et a été beaucoup étendue et adaptée à des applications plus pointues comme par exemple, l'imagerie Doppler pour l'observation de la circulation sanguine, l'élastographie afin de tester la rigidité des tissus et ainsi détecter les tumeurs (Newman & Rozycki, 1998) ou l'imagerie ultrasonore ultrarapide (IUU) développée il y a plus de 20 ans à l'Institut Langevin (Tanter & Fink, 2014).

Le principe physique d'une échographie US médicale est la propagation des ondes mécaniques avec des fréquences ultrasonores intermédiaires à hautes (2-20 MHz en imagerie clinique (Azhari, 2010)). Les ondes ultrasonores pénètrent les tissus mous avec une vitesse de 1540m/s (Briguet et al., 2014). Elles sont atténuées en fonction de la densité des tissus et de leur rigidité. Les changements dans la densité du tissu créent une réflexion des ondes ultrasonores, appelée « écho ». Le retour de cet écho à la sonde de l'échographe crée un signal électrique permettant de déterminer le temps de voyage du signal et ainsi la distance entre le transducteur et le tissu.

L'imagerie ultrasonore Doppler est un type d'échographie utilisée principalement pour explorer le réseau artériel et le réseau veineux. Elle a le nom du physicien autrichien Christian Doppler qui a découvert l'effet Doppler en 1842 (Newman & Rozycki, 1998). Elle mesure le changement de la fréquence ou longueur d'onde entre fréquence émise et fréquence reçue d'un objet en mouvement. La plupart du temps, ce sont les globules rouges.

Comme dans les autres techniques d'imagerie, il est possible d'utiliser des agents de contraste en échographie. Par exemple, les microbulles, des minuscules bulles d'air, proposées dans les années 60 sont d'une très grande utilité pour délimiter l'aorte (Tranquart et al., 2007).

L'examen d'échographie se déroule un peu différemment des autres examens. L'opérateur examine le patient et visualise les images en même temps. Pour cela, il a besoin d'une sonde ultrasonore, d'un gel,

d'un écran de visualisation des images collectées et des différents systèmes informatiques et électroniques qui permettent de stocker les images et contrôler l'échographe.

Une fois l'échographe en place, l'opérateur dépose le gel sur la zone d'intérêt pour l'examen et commence l'exploration des tissus avec la sonde. Pendant l'examen exploratoire, des images peuvent être capturées quand un élément d'intérêt est détecté. L'accessibilité des tissus et le caractère mobile de la sonde fait de l'échographie, l'outil d'exploration morphologique par excellence. Cependant, il dépend fortement de l'opérateur, ce qui diminue drastiquement la reproductibilité d'un examen d'échographie. Des propositions pour la stabilisation de la sonde, et la mise en place d'examens reproductibles sont aussi en train de voir le jour notamment via l'utilisation d'un robot pour le déplacement de la sonde.

L'échographie est utilisée aussi en recherche. Les différents types d'acquisitions utilisés en recherche sont le **B-Mode** qui est l'échographie anatomique classique, et le **M-Mode** qui représente un suivi temporel d'une ligne de l'image B-Mode, et le **Power Doppler** pour désigner l'imagerie Doppler. Par exemple, le débit cardiaque (Cardiac Output CO) est calculé via l'imagerie ultrasonore en multipliant le rythme cardiaque (Heart Rate) par le volume d'infarctus (Stroke Volume).

Les analyses de données en échographie s'effectuent directement sur les données brutes avec les logiciels propriétaires fournis avec la machine d'acquisition comme explicité dans la Figure 12, ci-après.

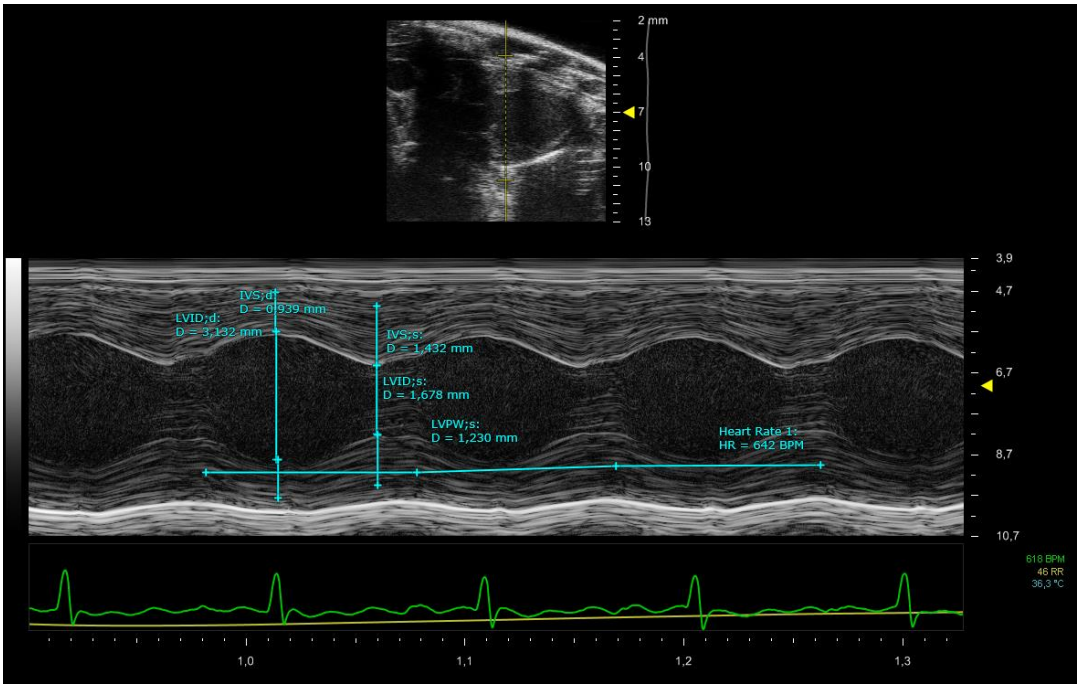


Figure 12 Prise de mesures sur une image M Mode
(source : données d'une étude au laboratoire LRI)

I.2.1.2. Acquisitions in vitro

En recherche biomédicale, l'information moléculaire in vitro est une source d'information riche qui vient compléter les données in vivo. Dans cette section, nous présentons l'histologie, un domaine d'imagerie in vitro pour l'étude des tissus biologiques, et les investigations moléculaires telles que le western-blot pour l'identification des protéines et la spectrométrie de masse pour la protéomique.

I.2.1.2.1. L'histologie

L'histologie est un domaine de la recherche biomédicale qui s'intéresse à l'étude des tissus biologiques. Elle a été fondée au XVII siècle par le médecin italien Marcello Malpighi (Romero-Reverón, 2011). Elle désigne la faculté à observer des tissus fins au microscope et a été appelée autrefois « anatomie

microscopique ». Avec l'arrivée récente des scanners numériques de lames entières, les lames d'histologie peuvent désormais être numérisées et stockées sous forme d'images numériques.

L'histologie est utilisée sur des biopsies et sur des échantillons chirurgicaux pour le diagnostic des maladies telles que le cancer. Ce domaine clinique, aussi appelé anatomopathologie, étudie les anomalies des tissus biologiques ainsi que de cellules présentant une pathologie. Elle représente une discipline d'observation avec des analyses à la fois macroscopiques et microscopiques des échantillons prélevés sur des patients (analyse de biopsies pour reconstituer le micro-environnement du cancer par exemple).

L'anatomopathologie et l'histologie utilisent des équipements spécialisés (des scanners de lames, des caméras, des microscopes, des écrans étalonnés et adaptés...). De point de vue clinique, l'image qui en résulte sert à la demande d'avis d'expert pour le diagnostic, à l'archivage ou l'utilisation des logiciels d'analyse pointus (pour compter le nombre des cellules par exemple).

En préclinique, plusieurs phases préparatoires doivent être effectuées avant l'acquisition de l'image numérique. L'opérateur, après l'euthanasie de l'animal, récupère les organes et les conserves en paraffine (ou autres modes de conservation). Quand un examen histologique est prévu, les organes en paraffine sont coupés en fines lamelles de tissus, fixés avec des solutions spécifiques et positionnés sur une lame du microscope afin de les préparer à la coloration ou au marquage.

Il existe de très nombreuses techniques en histologie : les colorations Rouge Sirius et Trichrome de Masson pour mettre en évidence les fibres de collagènes, le marquage de protéines avec des anticorps primaires et secondaires, l'immunohistochimie (IHC), l'immunofluorescence …etc. Dans la suite de ce paragraphe, nous allons présenter l'immunofluorescence et la coloration au Rouge Sirius.

L'immunofluorescence est le marquage d'une protéine spécifique à l'aide d'anticorps primaires et secondaires couplés à un fluorochrome (émission de lumière quand excité). L'opérateur doit alors effectuer les manipulations nécessaires pour réussir le marquage à l'anticorps primaire et ensuite celui à l'anticorps secondaire. Le but étant de révéler une protéine spécifique dans la cellule par émission de fluorescence. Elle permet donc de déterminer non seulement la présence, ou l'absence d'une protéine, mais aussi de sa localisation dans la cellule ou le tissu analysé. Il est essentiel de marquer les noyaux des cellules. Cette phase s'effectue, entre autres, avec une coloration au fluorochrome DAPI (4',6-diamidino-2-phénylindole) qui se lie fortement à l'ADN présent dans le noyau. Un microscope ou un scanner de lames à fluorescence est nécessaire pour pouvoir détecter les lumières émises par les fluorochromes (par exemple, 455nm est la longueur d'onde pour le DAPI). Sur la Figure 13, ci-après, l'image d'une lame ainsi que les différentes images correspondantes acquises par un scanner de lames à fluorescence.

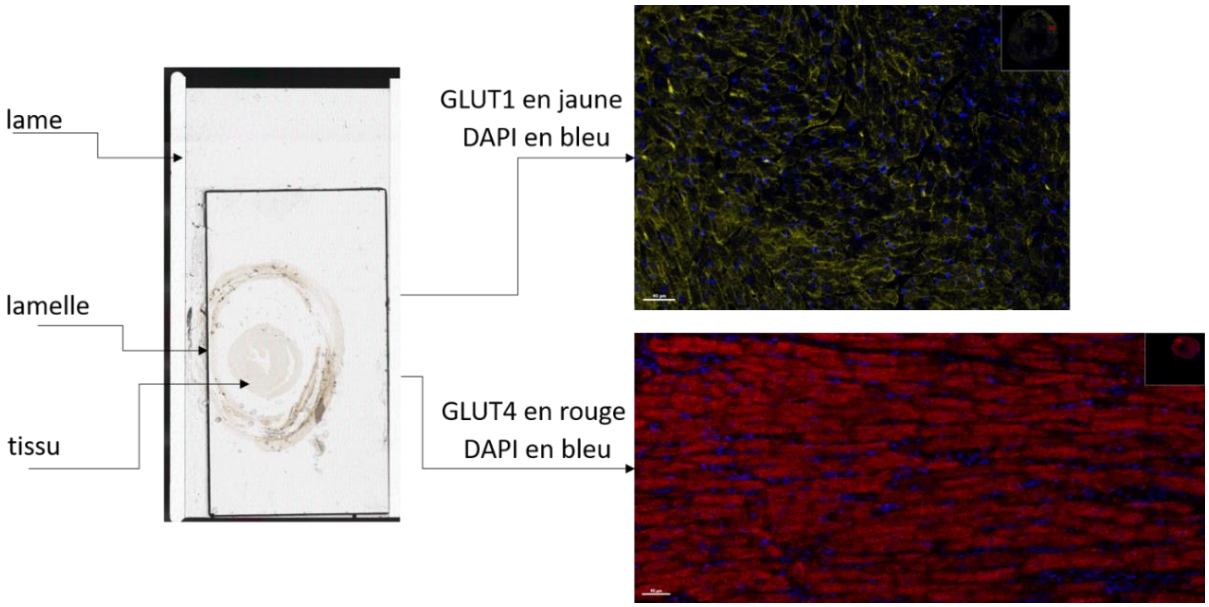


Figure 13 Exemples d'images acquises sur une lame d'histologie avec un scanner à fluorescence (source : données d'une étude au laboratoire LRI)

La coloration au Rouge Sirius (RS) est une coloration de tissu biologique qui peut être observée par un microscope optique sans fluorescence. Elle permet de teinter en rouge les fibres de collagènes et d'examiner le niveau de fibrose « destruction » d'un tissu en conséquence. La figure ci-après montre une coupe transversale du cœur colorée au RS. Elle permet de détecter la fibrose dans le muscle cardiaque.

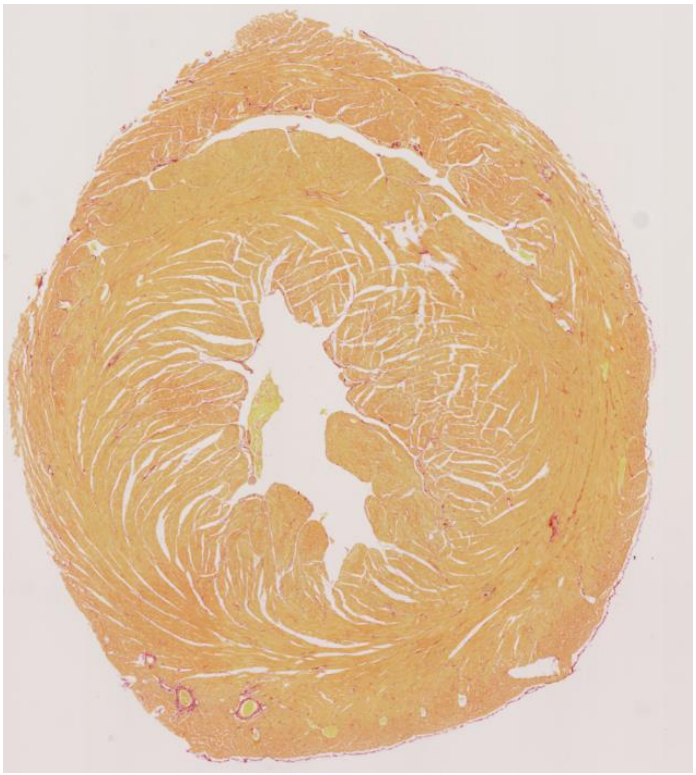


Figure 14 Un scan de lame d'un cœur de souris en coupe transversale colorée au Rouge Sirius (source : données d'une étude au laboratoire LRI)

Les images issues des analyses biologiques peuvent être analysées via des outils logiciels ouverts, propriétaires ou « maison ». Dans le laboratoire LRI, l'analyse d'images biologiques s'effectue

principalement via des applications logicielles Matlab ou occasionnellement via le logiciel ouvert ImageJ. Un potentiel énorme se trouve dans l'utilisation des images numériques histologiques en combinaison avec les technologies d'intelligence artificielle très prometteuses de nos jours. Au laboratoire LRI, un outil se basant sur l'apprentissage permet de distinguer les différentes zones tissulaires dans une série d'images et de quantifier leurs pourcentages.

I.2.1.2.2. Le western-blot

Une des investigations moléculaires utilisées au laboratoire LRI est la technique du western-blot. Elle a est fortement utilisé en biologie moléculaire depuis sa découverte en 1979. L'invention est attribuée aux deux travaux de (Renart et al., 1979) et (Towbin et al., 1979) publiés la même année[11]. Le western-blot permet la détection et la quantification de protéines dans un échantillon biologique en suivant les étapes illustrées en Figure 15:

1. Faire migrer les protéines sur un gel par électrophorèse.
2. Transférer les protéines sur une membrane.
3. Révéler par anticorps la ou les protéines d'intérêt.
4. Passer la membrane sous détecteur de chimioluminescence.

L'image finale se présente comme une matrice (Figure 15 partie 4) où les lignes sont les différents échantillons déposés dans la cuve à électrophorèse et les colonnes représentent les différents niveaux correspondants à la migration des protéines dans le gel. Chaque niveau correspond à un poids moléculaire. La quantification des protéines via l'image en western-blot s'effectue en comparant les intensités des bandes protéiques qui s'affichent après avoir noté leurs positions dans la membrane, elle est une technique de densitométrie. La position peut révéler des informations importantes sur la protéine d'intérêt compte tenu des informations sur son poids moléculaire. La bande noire au-dessous est celle de référence (la valeur 100% de quantité y est attribuée) et les autres bandes sont quantifiées relativement par rapport à elle. Les outils logiciels ImageJ ou Matlab peuvent être utilisés pour effectuer la quantification.

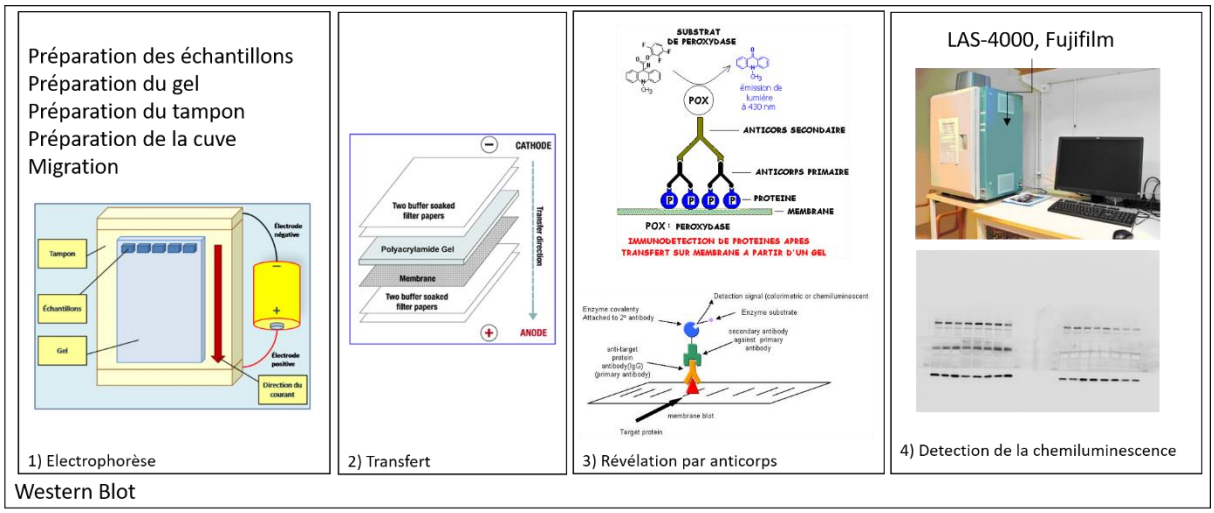


Figure 15 Étapes illustrées de l'expérimentation western-blot

I.2.1.2.3. La protéomique

La deuxième modalité d'investigations moléculaires est la protéomique. Elle est une discipline qui a émergé il y a une trentaine d'années étudiant l'ensemble des protéines (protéome) présentes dans un échantillon biologique à un instant t. Elle s'inscrit dans un champ disciplinaire plus large, celui des

[11] https://earlycareervoice.professional.heart.org/the-history-of-the-western-blot/

« omiques ». Les techniques « omiques » sont des techniques d'exploration biologique à des échelles allant de la protéine à l'ADN en passant par l'ARN et les peptides. Elles nécessitent un broyat de cellules, un sérum biologique ou un prélèvement de sang, objet d'étude afin de caractériser et quantifier sa composition. En recherche, nous pouvons citer la génomique, la protéomique, la transcriptomique, le métabolomique, la fluxomique, etc. Le nom de la modalité « omique » dépend de son élément biologique d'intérêt. La protéomique étudie le protéome, la génomique le génome ou l'ADN, la transcriptomique, le transcriptome ou l'ARN messager, etc.

La Chromatographie en phase Liquide-Spectrométrie de Masse (LC-MS), est la technique de protéomique la plus utilisée par le Laboratoire LRI, mais elle est « sous-traitée » à la plateforme 3P5 à l'hôpital Cochin (Paris). La Figure 16 ci-après, présente les étapes entreprises dans une expérimentation LC-MS. Après préparation des échantillons par chromatographie liquide (LC), ils seront ionisés et passés dans le spectromètre de masse (MS) pour détecter leur spectre en fonction de leurs masses.

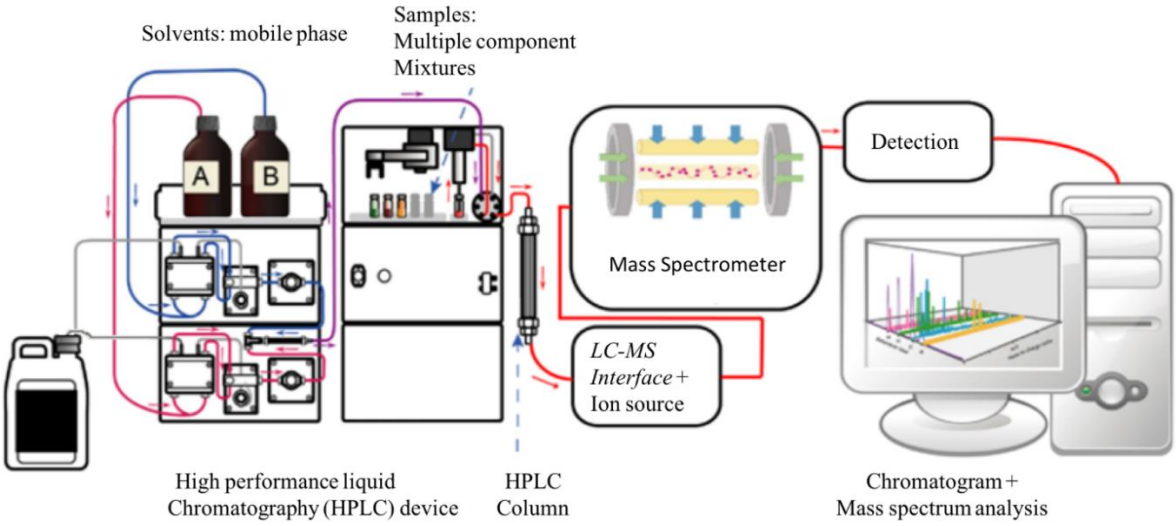


Figure 16 Différentes phases d'un examen de spectrométrie de masse (source : Wikipédia)

La MS est une technique d'analyse d'échantillons biologiques qui mesure le ratio masse/charge des ions (des peptides ionisés). Le résultat est un spectre de masse qui, après analyse, indique la présence et la quantité des peptides dans un échantillon biologique. Cette information, une fois couplée avec les données des voies biologiques, permet d'identifier les protéines et les circuits métaboliques actifs dans le tissu (Niessen, 2006).

Après la préparation des échantillons (broyat, lavage, etc.), il y a une phase de marquage ou labélisation des peptides avec des atomes plus lourds afin de mieux les détecter par MS. Cette phase est facultative, elle dépend du but de l'examen MS. Après ce marquage, une procédure de séparation des peptides ou chromatographie est effectuée (LC). Ensuite, les peptides marqués et séparés sont tous ionisés par spectrométrie de masse (MS). Le résultat des analyses MS est un spectre de masse qui révèle la liste des peptides détectés dans l'échantillon de départ. Ensuite, la séquence de chaque peptide présent dans l'échantillon est récupérée grâce aux bases de données de protéines comme UniProt[12].

Dans une deuxième phase, avec des outils logiciels spécialisés (Maxquant, Perseus, Inguinuity, Pathway Studio), les séquences des peptides sont analysées afin d'identifier et quantifier les protéines présentes dans l'échantillon. Ceci est accompagné d'analyses statistiques permettant d'identifier les protéines les plus significativement présentes. L'information sur la présence et la quantité des différentes protéines

[12] https://www.uniprot.org/

est utilisée pour identifier des circuits métaboliques et conclure sur la présence d'éventuelles pathologies.

I.2.1.3. Le suivi journalier

Un autre type d'acquisition de données que nous avons identifié est ce que nous avons appelé « le suivi journalier ». Il peut s'agir d'entretiens périodiques avec des patients, ou de suivi régulier de la taille d'une tumeur, ou aussi d'un suivi quotidien du régime alimentaire d'une population d'animaux. Par ailleurs, dans un laboratoire de recherche, chaque chercheur doit tenir un cahier de laboratoire (papier ou électronique) afin de noter les différents résultats et expérimentations qu'il a pu effectuer dans sa journée. Ce cahier a une valeur légale en cas de conflits, mais aussi une dimension pratique et utile pour les études de recherche.

Il peut être complété par des registres (papier ou électronique) qui permettent de suivre : (1) les entrées-sorties des animaux, leurs paramètres physiologiques de tous les jours (glycémie, ECG, etc.), le suivi de leur alimentation, ou les horaires de prise de médicaments, de nettoyage de cage, etc. (2) Les expérimentations effectuées sur les machines d'acquisition de données et les paramètres qui y sont associés : qui, quoi, où, quand, comment, dans quel but ?

Dans un cahier de laboratoire, le chercheur note ses protocoles, ses tests, ses expériences et tout ce qui pourrait être utile pour le lendemain et pour le reste de l'étude en cours. Un cahier de laboratoire reste dans le laboratoire après le départ du chercheur (départ très fréquent d'ailleurs en recherche), ce qui constitue le seul moyen de savoir ce qui a été fait préalablement comme recherche, longtemps après le départ d'un chercheur. L'ensemble des cahiers de laboratoire d'un laboratoire constitue sa « mémoire » et permet la transmission de connaissances des anciens aux récents membres au fils des années.

I.2.2. FORMATS DE DONNÉES EN RECHERCHE PRÉCLINIQUE

Pour chaque type d'image et modalité d'acquisition de données, un standard de référence est en général défini et reconnu au sein de la communauté scientifique. Notamment, DICOM[13] pour les images in vivo et le modèle de données OME[14] pour les images ex vivo, et mzML pour les données protéomiques. D'autres formats sont parfois non compatibles avec les standards de référence, comme les formats propriétaires définis par les constructeurs des machines d'acquisition. Nous présentons aussi un format assez récent qui découle d'une initiative de standardisation à l'ESMI (European Society for Molecular Imaging) pour la multimodalité d'imagerie et qui a été proposé en lien étroit avec les constructeurs des machines. Nous donnons pour finir les formats identifiés pour les données dérivées d'analyses.

I.2.2.1. DICOM : le standard de l'imagerie médicale

Le standard DICOM ou la norme ISO 12052 définit le format des images médicales en radiologie. DICOM est aussi un standard de communication et partage des images entre machines d'acquisition et d'analyse et serveurs de stockage et archivage. Il est utilisé par les serveurs d'archivage d'images médicales PACS (Picture Archiving and Communication Systems). Il est publié et maintenu par le NEMA - National Electrical Manufacturers Association. Apparu en 1993 (Mildenberger et al., 2002), il a été adopté par les constructeurs d'imageurs cliniques ainsi que les radiologues et est devenu le standard de référence en imagerie médicale.

DICOM est une solution de gestion des images numériques utilisées par les radiologues et médecins qui leur permet de récupérer les informations sur le patient, les paramètres d'acquisition, etc. ; et ainsi de croiser les données entre elles. Il représente le résultat d'une prise de conscience collective quant à la nécessité d'avoir une convention de nommage entre les professionnels de l'imagerie clinique afin de

[13] http://dicom.nema.org/
[14] https://docs.openmicroscopy.org/ome-model/6.0.0/

pouvoir partager les données entre eux et ainsi améliorer la qualité du diagnostic par le croisement efficace des données. DICOM est plus qu'une convention de nommage, il est un modèle structuré en plusieurs éléments ayant chacun un rôle bien déterminé de description de données.

La Figure 17 présente la structuration d'un fichier DICOM: préambule, préfixe, en-tête (*Header*) et les données des pixels de l'image. Le préambule est composé de 128 octets généralement mis à zéro. Le préfixe est composé de 4 octets réservés aux lettres D, I, C, et M.

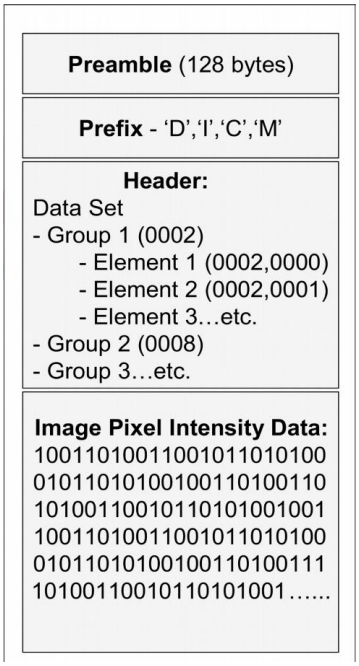


Figure 17 Structure d'un fichier DICOM (Varma, 2012)

L'en-tête DICOM contient des métadonnées décrivant l'image et ses paramètres d'acquisitions labélisés par des couples (Groupe, Élément) de 4 chiffres (0000,0000). L'ensemble des métadonnées DICOM est donc hiérarchisé en Groupes et en Éléments de ce groupe. Même si, par exemple, les attributs concernant le patient ont tous une valeur de Groupe=0010, aucune signification n'est officiellement attribuée aux groupes DICOM. Le Tableau 1 suivant liste quelques couples DICOM ainsi que leurs caractéristiques.

Tableau 1 Exemple de couples (Groupe, Élément) en DICOM

Code Élément DICOM	Tag de l'Élément DICOM	Type DICOM	Description
(0008,0008)	ImageType	Code String (CS)	Caractérisation et identification d'une image
(0008,0020)	StudyDate	Date (DA)	Date de l'examen d'imagerie
(0008,0022)	AcquisitionDate	Date (DA)	Date de l'acquisition
(0008,0031)	SeriesTime	Time (TM)	Heure de prise de la série d'images
(0008,0060)	Modality	Code String (CS)	Modalité d'imagerie : CT, PT ou MRI, etc.
(0008,0090)	ReferringPhysicianName	Person Name (PN)	Nom de la personne référente à consulter
(0008,1050)	PerformingPhysicianName	Person Name (PN)	Nom de la personne qui effectue les examens

(0010,0010)	PatientName	Person Name (PN)	Nom du patient
(0010,1030)	PatientWeight	Decimal String (DS)	Poids du patient
(0018,1020)	SoftwareVersion	Long String (LO)	Version du logiciel d'acquisition
(0018,1030)	ProtocolName	Long String (LO)	Le nom du protocole utilisé dans l'acquisition
(0020,000D)	Study Instance UID	Unique Identifier (UI)	Identifiant unique de l'examen
(0020,0011)	SeriesNumber	Integer String (IS)	Le numéro de la série dans l'examen

Les groupes pairs (à partir de 0008 ;0010 ; etc.) sont des groupes qui appartiennent au standard officiel DICOM, tandis que les groupes impairs (à partir de 0009 ; 0011 ; etc.) sont utilisés pour stocker des informations propriétaires ou non conventionnelles. Or, ces métadonnées ralentissent les processus de conversion ou de conciliation de DICOM avec d'autres types de données et sont problématiques puisqu'ils peuvent même engendrer de la perte d'information. Ceci explique l'importance des groupes de travail cherchant à établir un consensus pour l'extension et l'adoption de DICOM à plus de domaines.

Le standard DICOM a connu plusieurs versions et a été étendu à plusieurs applications et domaines. Il a été adapté notamment à l'imagerie préclinique avec le groupe de travail n°30[15]. Bien qu'utilisé par les constructeurs des machines et scanners en imagerie préclinique, sa mise en œuvre est caractérisée par l'exploitation abondante des champs DICOM propriétaires et donc incompréhensibles par des logiciels de traitement d'images DICOM non fournis par le constructeur.

I.2.2.2. Le modèle de données Open Microscopy Environment (OME)

OME[16] est un consortium d'universités, d'instituts de recherche et d'entreprises privées qui s'intéressent au développement des standards en lien avec les images de microscope. Ils proposent le modèle de données OME sous deux formats distincts : OME-XML et OME-TIFF. Ces formats adaptés aux données de bio-imagerie utilisent des métadonnées pertinentes pour les décrire : informations sur l'acquisition de l'image, informations sur les régions d'intérêt (ROIs), les coordonnées spatiales, etc. la Figure 18 ci-après présente le standard OME avec ses différents éléments de haut niveau (Goldberg et al., 2005).

Au départ, les deux formats (OME-XML et OME-TIFF) étaient destinés à être interchangeables. D'un côté, le format TIFF est un format image par définition adapté pour stocker les données des pixels de l'image. D'un autre côté, le format XML est un format de description par définition plus adapté au stockage des métadonnées de l'image (Date et Instrument d'acquisition, expérimentateur, informations sur l'échantillon visualisé, etc.). La configuration d'origine était que le OME-XML code les pixels de l'image en compression Base64, et que le OME-TIFF contient dans l'en-tête, un bloc de métadonnées OME-XML. Désormais depuis 2016, la recommandation d'OME est de représenter l'image avec deux fichiers : un OME-XML pour les métadonnées uniquement et un OME-TIFF pour les images uniquement[17].

[15] *WG-30 : Small Animal Imaging – DICOM Standard*. https://www.dicomstandard.org/wgs/wg-30/
[16] https://www.openmicroscopy.org/
[17] https://www.openmicroscopy.org/Schemas/Documentation/Generated/OME-2016-06/ome.html

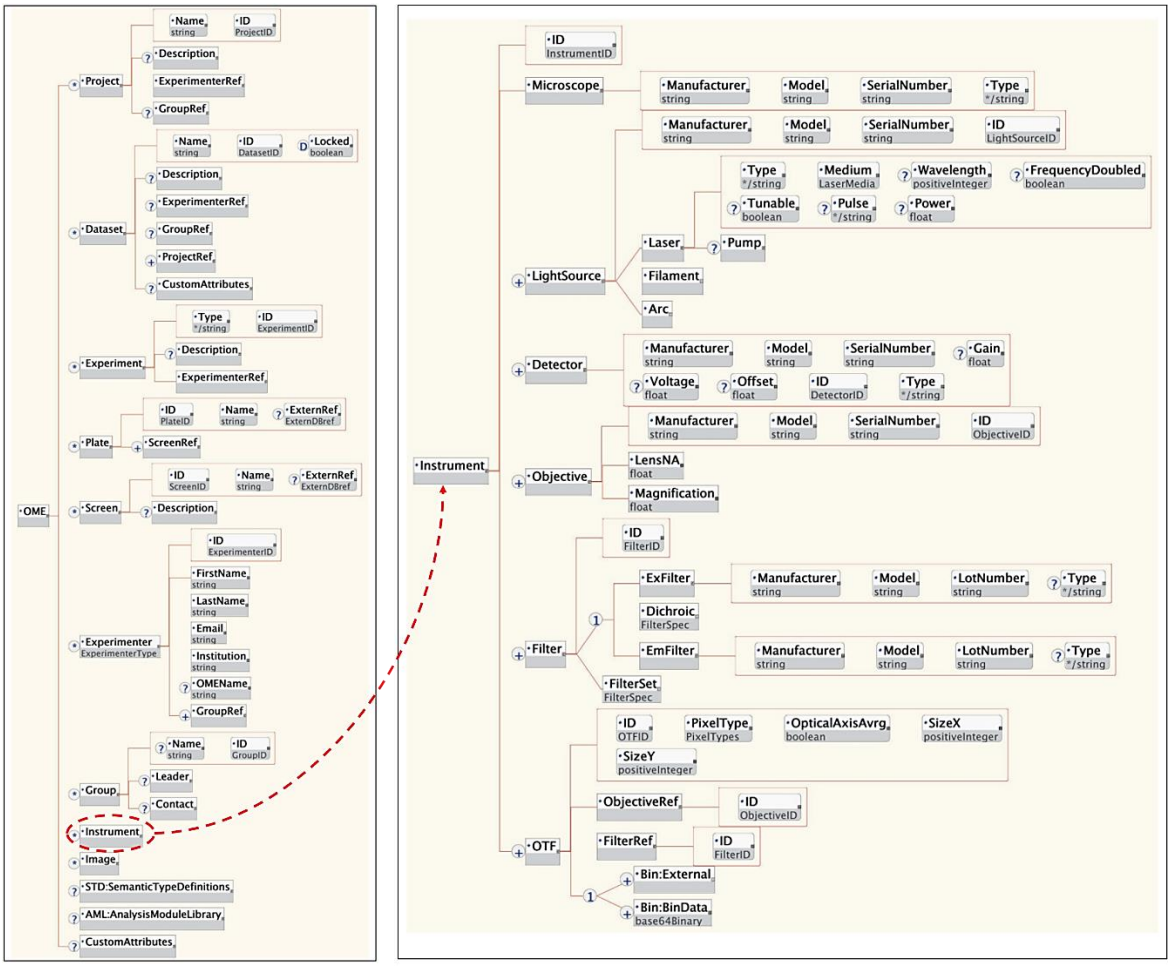


Figure 18 Le modèle de données OME (Goldberg et al., 2005)
les principales briques OME sont à gauche, et une explication de la brique instrument est à droite

Le standard OME est accompagné de la bibliothèque Java Bio-Formats qui permet de lire plus de 150 formats d'images[18] et de les retourner sous format OME-XML et OME-TIFF.

I.2.2.3. Le standard mzML pour les « omique »

Le « Proteomics Standards Initiative (SPI) » du HUmain Proteome Organization (HUPO) (Mayer et al., 2013) propose d'utiliser le format mzML pour unifier les formats des données brutes en protéomique. Ce format (Deutsch, 2010) remplace les formats historiques mzXML et mzDATA en proposant le meilleur des deux formats. Le HUPO SPI développe aussi des normes pour décrire les résultats les processus d'identification et de quantification des protéines, peptides, petites molécules et modifications de protéines, par spectrométrie de masse. Ils proposent le mzTab, un format de fichier texte délimité par des tabulations contenant les résultats de protéomique et de métabolomique.

Plusieurs outils sont basés sur le mzML et le mzTab comme la suite logicielle OpenMS qui permet, entre autres, de convertir des fichiers d'un format propriétaire à un format standard comme mzML. La liste complète des outils propriétaires ou non qui sont compatibles avec le standard mzML peut être trouvée sur le site de la SPI[19]. La Figure 19 montre les différentes sections d'un fichier mzML (Deutsch, 2010). En en-tête, les métadonnées du fichier sont documentées : instrument, logiciel, paramètres d'acquisitions, ensuite le spectre et le chromatogramme sont inclus.

[18] https://docs.openmicroscopy.org/bio-formats/6.5.1/supported-formats.html
[19] http://www.psidev.info/mzML

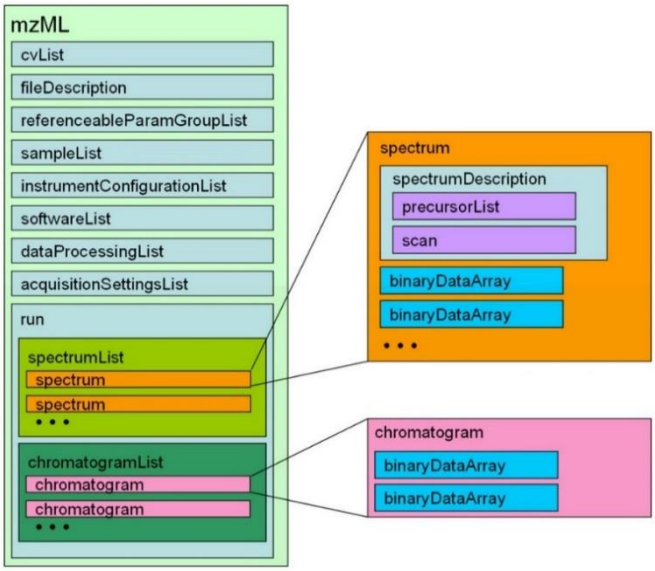


Figure 19 Structure d'un fichier mzML (Deutsch, 2010)

I.2.2.4. Les formats propriétaires

Afin d'acquérir les données en recherche biomédicale, plusieurs équipements sont utilisés : des vieux ou des nouveaux, avec différentes versions de logiciels et différents formats de données proposés. Pour les instruments d'imagerie du petit animal, il y a généralement deux formats proposés : (1) le format du constructeur, qui peut être converti en (2) le format standard. La version du format de données dépend de l'époque où la machine a été construite et de la dernière mise à jour effectuée. Il y a ainsi autant de formats que de constructeurs de machines et autant de variantes que de mises à jour logicielles. Nous citons ici des formats de données rencontrées au laboratoire LRI :

- **Le format Brucker en IRM :** le format par défaut de la machine qui est supporté par tous les outils fournis par l'entreprise Brucker. Ils sont suffisants dans 90% des cas. Le problème se présente quand on souhaite effectuer des analyses avancées dans un contexte de recherche. Dans ce cas précis, l'export en format DICOM peut être utilisé. D'après les chercheurs du laboratoire LRI, ce format DICOM est pauvre en information et n'est pas une traduction complète du format d'origine.
- **Le format Mediso en TEP-TDM :** le scanner TEP-TDM de l'entreprise Mediso fournit un fichier DICOM pour l'acquisition brute TDM et un fichier DICOM pour le TDM reconstitué. Cependant, pour les données brutes TEP, elle propose un format propriétaire dit « list-mode » avec une extension « .data ». Ce dernier représente une liste d'événements enregistrés par les détecteurs de radioactivité du scanner. Pour pouvoir exploiter ce format, l'entreprise Mediso propose un outil de « reconstruction » permettant d'obtenir des images TEP sous format DICOM, exploitables ensuite par d'autres logiciels (voir Figure 20).

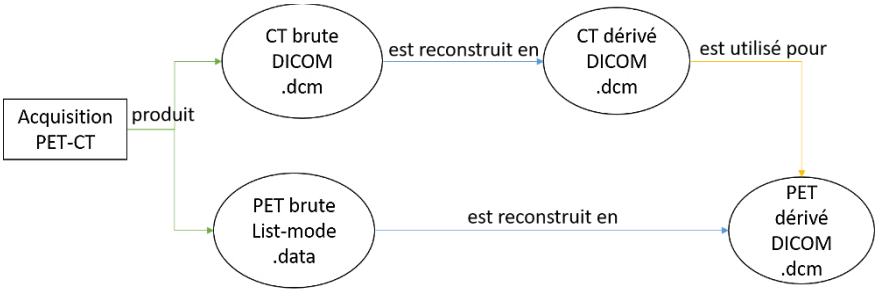


Figure 20 Les différents formats produits par le scanner TEP-TDM Mediso

- **Les formats propriétaires en échographie ultrasonore :** nous avons identifié deux pôles d'imagerie ultrasonore utilisés par le laboratoire LRI. Le premier est la plateforme d'imagerie du petit animal à l'Hôpital Cochin[20], les données brutes y sont sous format propriétaire : format « .bimg » pour l'acquisition B-mode, « .pimg » pour les paramètres physiologiques, « .mimg » pour l'acquisition M-mode et « .dimg » ou « .pwimag » pour l'acquisition Doppler. Quelques fichiers sont encodés en XML, ce qui permet leurs lectures séparément, mais ne donne accès à l'information complète de l'examen que lorsqu'on accède aux données depuis le logiciel « Vevo LAB » du constructeur Visualsonics.
  Le deuxième pôle est une technique en cours de développement qui consiste à coupler un échographe d'imagerie ultrasonore ultrarapide (IUU) avec un scanner TEP-TDM pour une trimodalité appelé PETRUS (PET-US simultanés et TDM pour la correction de l'atténuation du signal). Toutes les données de ce dispositif d'imagerie IUU sont traitées en Matlab.
- **Les formats propriétaires en Histologie :** Au laboratoire LRI, les chercheurs ont la possibilité d'utiliser trois équipements : deux scanners de lames (Hamamatsu et Polaris Vectra) et un microscope à fluorescence (Axioimager Apotome). Les formats que nous avons pu identifier sont le « im3 », « qptiff » et « tif », format des données brutes du scanner de lames Polaris Vectra, qui sont regroupées dans un dossier correspondant au scan d'une lame (voir Figure 21), le format « NDPI » pour le scanner de lames Hamamatsu, et « TIF » pour le microscope Apotome. Tous permettent un export en TIF des images, format ouvert et connu. Les formats propriétaires obligent les opérateurs à utiliser exclusivement les logiciels du constructeur et limitent ainsi les recherches avancées sur les données. Comme mentionné auparavant, la bibliothèque bio-format permet la conversion de 150 formats d'image biologique, dont le format Hamamatsu « NDPI » et Polaris Vectra « im3 » et « qptiff » ainsi que « TIF », en OME-XML + OME-TIFF, les formats standard du modèle de données OME[21].

MSI	18/06/2018 11:45	Dossier de fichiers	
SlideRegistration	18/06/2018 11:45	Dossier de fichiers	
CoverslipMask.tif	14/06/2018 17:17	Fichier TIF	21 Ko
FocusMap.tif	14/06/2018 17:18	Fichier TIF	43 Ko
Label.tif	21/10/2020 12:24	Fichier TIF	2 174 Ko
OverviewBF.tif	14/06/2018 17:17	Fichier TIF	20 173 Ko
OverviewFL.tif	14/06/2018 17:18	Fichier TIF	3 708 Ko
S937_SDHB_TUMEUR_IF_488GLUT1-594CD31_X20_L1_Scan1.qptiff	14/06/2018 17:21	PerkinElmer whole slide scan file	905 246 Ko
S937_SDHB_TUMEUR_IF_488GLUT1-594CD31_X20_L1_Scan1_annotations.xml.lock	09/04/2019 15:35	Fichier LOCK	1 Ko
SampleMask.tif	14/06/2018 17:18	Fichier TIF	15 Ko

Figure 21 Fichiers d'un examen d'histologie effectué avec le scanner de lames Polaris sur une lame

- **Le format ThermoFisher en protéomique :** Les données protéomiques du laboratoire LRI sont produites par le « Dionex U3000 RSLC nanoLC », couplé à « Q-Exactive » ou « LTQ Orbitrap-Velos » pour la spectrométrie de masse Thermo Fisher Scientific de la plateforme 3P5 à l'institut Cochin. Les données directement produites par la machine sont des spectres sous format « .raw ». Elles peuvent être exploitées via l'outil XCalibur fourni par ThermoFisher. Les données dérivées sont issues de logiciels tels que MaxQuant, Perseus, Ingenuity et Pathway Studio. Ils génèrent des résultats d'analyses sous format tableur et image.
  Le standard de référence pour ce type de données est « mzML ». Cependant, ThermoFisher n'offre pas un export directement en « mzML » de leurs données via « XCalibur », mais propose « Proteome Discoverer » qui permet de convertir les fichiers « .raw » en « mzML ». Pour le même but, des outils ouverts existent, tels que RawConverter[22].

[20] http://piv.parisdescartes.fr/modalites-imagerie/echographie-haute-resolution/
[21] https://docs.openmicroscopy.org/bio-formats/5.8.2/supported-formats.html
[22] http://fields.scripps.edu/rawconv/

I.2.2.5. Le format « gff » pour la standardisation de l'imagerie multimodale

Dans (Begley & Ioannidis, 2015), il a été reporté que 75-90% des observations empiriques publiées dans les journaux scientifiques en recherche fondamentale et en recherche préclinique ont une reproductibilité questionnable. Ce qui est un chiffre alarmant. Les auteurs expliquent que tout l'écosystème de la recherche est impliqué notamment : la culture « *publish or perish* », le manque d'application des bonnes pratiques, etc. Ils insistent sur le fait que la responsabilité est partagée entre tout le monde et nécessite la mise en place de plusieurs solutions, et non pas une seule.

Le contexte actuel de la gestion de données de recherche (RDM), l'hétérogénéité des données, ainsi que la faible reproductibilité des résultats en recherche préclinique, rendent la standardisation en imagerie du petit animal et la définition d'un format commun un sujet d'actualité au sein de la communauté scientifique de l'European Society for Molecular Imaging (ESMI) standardisation group, (Mannheim et al., 2018). Ces derniers ont effectué une enquête concernant la standardisation en imagerie du petit animal au niveau des protocoles et méthodes, tels que l'anesthésie, la manipulation des animaux, le relevé des paramètres physiologiques, l'acquisition et l'analyse de données, les machines d'acquisition, etc. Ils ont relevé la nécessité de bonnes pratiques unifiées pour améliorer la reproductibilité des recherches précliniques.

Dans la continuité de cette initiative (Yamoah et al., 2019) ont proposé un format de fichier pour l'imagerie multimodale préclinique et clinique avec l'extension « gff » et permettant de stocker jusqu'à 5 modalités à la fois. Cette structure de données stocke en série les dimensions, les tailles et le type de voxels, les éléments de données de rotation et de translation, les informations de temps et de canal, les informations de compression, ainsi que les métadonnées. Le format prend en compte l'existence des données non structurées (ou plutôt non préalablement définies – telles que les valeurs de SUV, le poids, l'âge, etc.) et permet leur ajout via des pairs de clé-valeur. La Figure 22 ci-après présente les séries d'informations du format « gff ».

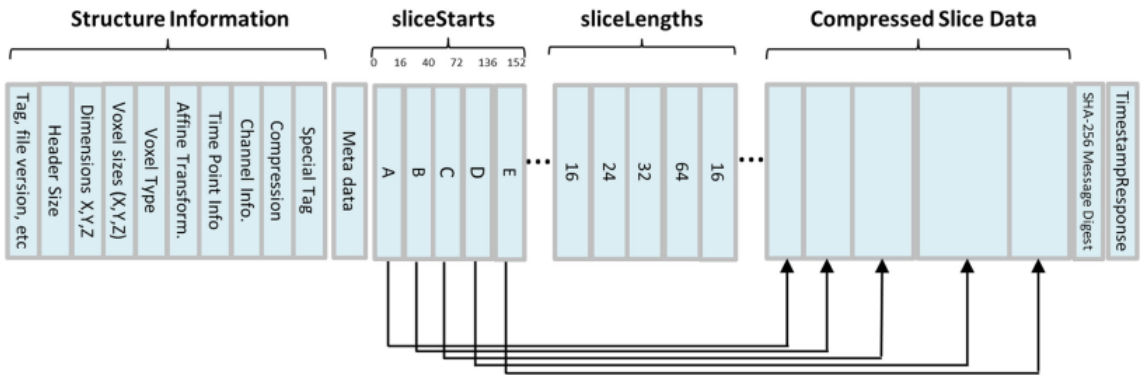


Figure 22 Format de fichier « gff » et ses différents composants en mémoire (Yamoah et al., 2019)

I.2.2.6. Les formats des données issues des analyses

Les données issues d'analyses sont en général des tableaux, des images et des figures ainsi que des fichiers de travail où plusieurs calculs scientifiques et statistiques ont été élaborés. Ils peuvent être aussi des scripts informatiques en Matlab ou autre et des textes d'observations ainsi que des conclusions rédigées.

Tandis que les acquisitions de données au début du cycle de vie sont très variées et diverses, les résultats ou les données d'analyses en fin du cycle de vie d'une étude de recherche se ressemblent dans leurs formats qui dépendent des pratiques communes de valorisation et publication scientifique. Les formats acceptés dans les publications des journaux et dans les conférences sont principalement des images, des tableaux et du texte, ou des présentations de diapositives. Une tendance récente dans le cadre de l'open

science, incite les chercheurs à publier des données atypiques comme les vidéos, les données brutes, etc., mais la réponse reste modeste et l'adhésion des chercheurs pourrait varier en fonction du domaine.

I.2.3. OUTILS UTILISÉS POUR L'ANALYSE DES DONNÉES

Ayant présenté les différentes modalités et formats de données que l'on peut rencontrer en recherche préclinique, nous présentons dans cette section les différents outils communément rencontrés dans un laboratoire de recherche, puis nous détaillons ceux rencontrés au laboratoire LRI. Une journée typique d'un chercheur dans le domaine biomédical est partagée entre la paillasse pour l'expérimentation et l'ordinateur. Pour comprendre les types d'outils utilisés tous les jours en recherche, nous avons listé ci-après les activités de recherche de tous les jours que nous avons identifiées lors des échanges avec des collègues au laboratoire LRI (qui nous a accueilli pendant cette thèse) :

- La recherche bibliographique
- La rédaction de documents scientifiques (rapports, articles ...) ou administratifs (demandes de financement …)
- La mise au point de protocoles d'expérimentation
- La réalisation d'études préliminaires aussi appelées études pilotes
- La préparation de matériels et produits pour les expérimentations
- La gestion des calendriers et des stocks
- L'organisation et la participation à des réunions et des événements scientifiques
- La réalisation des expérimentations
- La collecte et la gestion des données de projets de recherche
- L'interprétation biologique des données et l'énoncé d'hypothèses de recherche
- L'analyse de données qualitativement et quantitativement

I.2.3.1. Types d'outils utilisés pour manipuler les données

Nous avons identifié cinq types d'outils utilisés quotidiennement au laboratoire LRI:

1. Les outils bureautiques : ce sont des outils qui permettent l'édition et la consultation de textes et de tableaux et la préparation de présentations. La plus connue est la suite Microsoft Office.
2. Les outils d'organisation de données : Ce sont des catalogues qui permettent de consulter et situer les données dans leur contexte. Par exemple Osirix[23] en imagerie
3. Les outils spécifiques d'analyse de données : Ce sont des outils spécialisés, qui permettent d'effectuer des analyses adaptées et spécifiques aux types de données en entrée. Par exemple ImageJ en imagerie.
4. Les outils de calcul statistiques : Ce sont des outils qui permettent d'obtenir des résultats statistiques avancés et qui ont en entrée des tableaux et donnent en sortie des graphiques et diagrammes statistiques. Nous pouvons citer R studio[24] ou GraphPad Prism[25].
5. Les codes informatiques faits-maison : Ce sont des outils développés par le laboratoire en Java, Python ou Matlab afin de satisfaire un besoin précis relatif aux projets de recherche en fonction du niveau de connaissances et d'expertises du laboratoire.

Un constat assez clair à ce stade concerne la richesse du panel d'outils utilisés par le chercheur au quotidien. Nous remarquons que surtout au niveau de l'analyse et l'organisation des données, un chercheur peut se retrouver avec autant d'outils manipulés que de types de données produites.

[23] https://www.osirix-viewer.com/
[24] https://rstudio.com/
[25] https://www.graphpad.com/scientific-software/prism/

I.2.3.2. Outils utilisés dans le laboratoire LRI

Afin de réaliser l'état des lieux et caractériser les outils utilisés au laboratoire LRI, nous avons effectué une enquête auprès des membres du LRI en mars-avril 2017. Avant le lancement du questionnaire, un recensement des logiciels utilisés au labo a été effectué. L'enquête collectait la fréquence d'utilisation, et donc l'importance, de chaque outil pour les seize membres du LRI ayant répondu. La liste des outils identifiés figure dans le Tableau 2 ci-après :

Tableau 2 Liste des logiciels identifiés au laboratoire LRI

1	Amira	Logiciel de visualisation, de traitement et d'analyse d'image en 3D et 4D.	Imagerie
2	Pmod	Logiciel de quantification des images produites par des scanners biomédicaux à des fins de recherche. En majorité de l'imagerie avec traceurs TEP.	Imagerie
3	ImageJ	Logiciel libre et open source pour le traitement et l'analyse d'images. Il est développé par le NIH, US et extensible via des macros et plugins Java.	Imagerie
4	Osirix	Logiciel de visualisation et d'archivage d'images médicales sous format DICOM. Il est adapté au système d'exploitation MacOS et utilisé par les radiologues	Organisation
5	Olea Sphere	Logiciel de visualisation avancée et archivage des images IRM et TDM. Il est développé par la société Oléa Medical	Organisation
6	Myrian	Suite logicielle qui permet de visualiser un large panel d'images : IRM, TEP, etc. Elle est extensible en fonction des besoins en visualisation et analyse. Elle est développée par l'entreprise Intrasense	Organisation
7	R	Langage de programmation et un logiciel libre destiné aux statistiques et à la science des données.	Statistiques
8	Excel	Fait partie de la suite Microsoft Office et permet de gérer des tableurs et d'effectuer des statistiques simples	Statistiques
9	XLStat	Extension de statistiques avancées pour Microsoft Excel	Statistiques
10	PrismGraphPad	Logiciel d'analyse statistique sophistiqué et simple d'utilisation	Statistiques
11	OpenOffice	Suite logicielle bureautique libre et gratuite	Office
12	MicrosoftOffice	Suite logicielle bureautique de Microsoft	Office
13	Physiod3D	Utilitaire développé en Matlab par le labo pour l'analyse des données de l'IRM de perfusion DCE	Matlab
14	Lectine_SiriusRed_CD31	Utilitaire développé au laboratoire pour le comptage des cellules marquées dans une image issue de microscope	Matlab
15	Other_Matlabs	Autres outils développés en Matlab	Matlab
16	Other_software	Autres logiciels non cités	Autres

Ensuite, l'enquête a été envoyée à tous les membres du LRI via un lien Foodle[26] (un outil de collecte d'avis de Renater qui a été délaissé fin 2017). Le Tableau 3 ci-après illustre les différents choix possibles pour un (Logiciel XX – ligne 1). Six fréquences sont à sélectionner (ligne 2) et pour chaque fréquence un poids est attribué (ligne 3), les fréquences ont été après convertis en % pour le calcul des résultats.

Tableau 3 Modèle et poids utilisés pour le classement des logiciels au laboratoire LRI

Logiciel XX					
almost_every_day	once_a_week	twice_a_month	once_a_month	rarely	never
20	12	6	3	1	0

Les résultats de l'enquête sont résumés dans le graphique de la Figure 23. Une majorité est attribuée à Matlab et à la suite Microsoft Office avec 23% de l'utilisation globale chacun. Excel qui n'est pas loin avec 22%, en quatrième position, Open Office (9%) et Pmod (8%). Puis Amira avec 4%. Il est à noter

[26] https://groupes.renater.fr/reunion/foodle/A-Survey-on-U970-used-Software-and-Tools-58b73?tab=1

pour finir que Graphpad (1%) n'était pas très utilisé en 2017, mais que son utilisation par le laboratoire a augmenté dans le labo au fil des années.

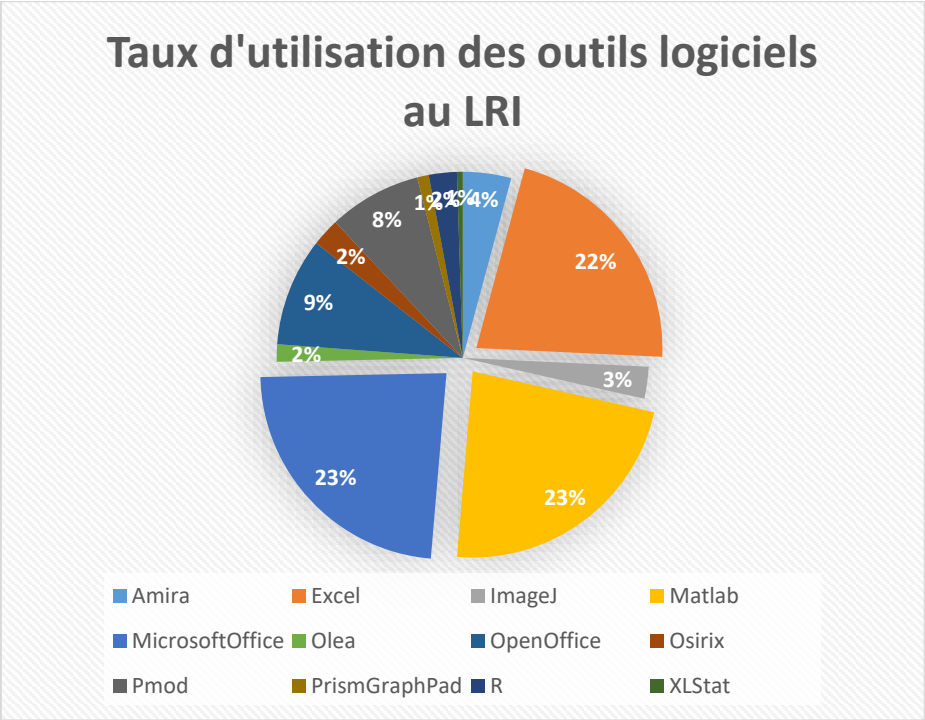


Figure 23 Taux d'utilisations des outils logiciels au LRI

I.2.4. BILAN : DES DONNÉES HÉTÉROGÈNES SUR PLUSIEURS NIVEAUX

Nous avons décrit les données produites et les outils utilisés en recherche préclinique en général et dans le laboratoire LRI en particulier. Le but n'était pas de fournir la liste exhaustive des modalités d'acquisition, mais de mettre l'accent sur leurs points communs et leurs différences afin de mieux les maîtriser pour mieux les gérer.

Compte tenu de tout ce qui a été présenté précédemment, il est clair que l'hétérogénéité des données en recherche préclinique est une hétérogénéité multiple à des niveaux différents. Les plus simples à identifier sont l'hétérogénéité au niveau des procédures d'acquisition des données : avec ou sans agent de contraste, intérêt pour tout le corps ou pour un seul organe, suivi longitudinal en imagerie IRM ou imagerie exploratoire ultrasonore. L'hétérogénéité est aussi observée au niveau des formats électroniques de la donnée générée par ces procédures (.dcm, .xls, .txt, .xml, .mzml, .raw, .tif, .qptiff, .ndpi, .tiff, .gff, .data, .bimg, .ming, brucker, etc.), des outils d'analyse de ces données acquises, des domaines d'expertises biologiques et techniques, de l'étendue d'utilisation des données (notes personnelles vs résultats de recherche publiés) et du rôle de la dimension temporelle qui varie entre un suivi longitudinal et une analyse atemporelle. Nous pouvons aussi noter que les données de recherche préclinique sont multisites, certaines analyses et expérimentations n'étant pas effectuées au laboratoire LRI, mais dans des laboratoires partenaires.

En résumé, les données de recherche préclinique sont souvent multisites, multisources, multiformats et pluridisciplinaires. Elles sont collectées tout au long du cycle de vie d'une étude de recherche et sont acquises avec diverses modalités d'acquisitions et selon différentes temporalités. Leur hétérogénéité freine et alourdit leur partage puisque les données et leurs annotations ainsi que leurs explications respectives doivent toutes être partagées pour assurer la compréhension des données et ils ne sont pas toujours présents lors d'un partage. La réutilisation ultérieure en suivant les recommandations FAIR

(Wilkinson et al., 2016) devient impossible alors. Le schéma de la Figure 24 résume les caractéristiques des données de recherche évoquées jusque-là.

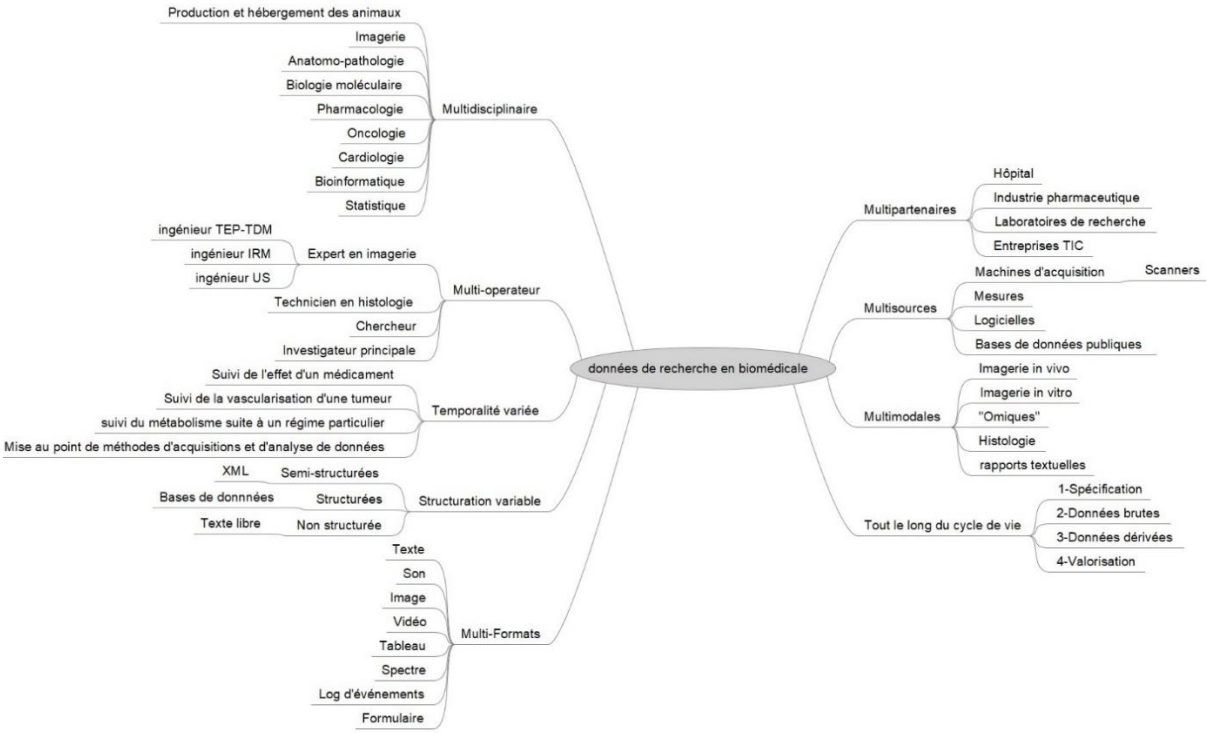


Figure 24 Facteurs d'hétérogénéité en recherche préclinique identifiés suite à l'immersion au Laboratoire LRI

Au laboratoire LRI, plusieurs tâches de gestion de données peuvent être identifiées: préparation des données pour leur passage dans un script, export de données d'un format propriétaire à un format ouvert pour utiliser des outils libres, rapport des expérimentations et des différentes mesures biologiques et physiologiques faites sur l'animal, archivage des données anciennes pour laisser de la place aux nouvelles données récemment acquises, préparation des résultats pour la publication (changement de formats ou autres), etc.

Il faut mettre l'accent sur le fait que de nouvelles questions en recherche émergent au fur et à mesure que l'étude avance. Par conséquent, les solutions de gestion, de stockage, et d'archivage des données existantes au sein du laboratoire LRI sont orientées sur un besoin précis à un instant T d'une personne P, et ne sont pas forcément compatibles avec l'archivage à long terme, ni avec la gestion des données dans le cadre des principes FAIR et les recommandations RDM en France. La gestion des données scientifiques est vue par le chercheur comme une tâche triviale, annexe et peu questionnable. Dans ce contexte, les habitudes et les procédures d'un chercheur l'aident à gagner du temps et à compenser la complexité biologique et informatique qui l'entoure, ce qui n'est pas suffisant sur le long terme, pour lequel il faudrait pouvoir :

- Mettre en place une solution d'archivage robuste des données permettant une conservation intacte de l'information et une réutilisation fiable même sur le long terme.
- Avoir un maximum de traçabilité lors de la collecte, du stockage, du partage et de l'analyse des données.
- Simplifier les procédures complexes et automatiser ou semi automatiser des tâches fastidieuses via l'outil informatique, ce qui permettra de dégager plus du temps pour la recherche.
- Disposer d'outils d'exploration simples de l'ensemble des données du laboratoire afin de permettre de tester des hypothèses de recherche.

- Fournir au laboratoire LRI une gestion des données scientifique au jour le jour compatible avec les standards du domaine mis en jeu et respectant les recommandations FAIR, pour plus de reproductibilité des résultats scientifiques dans un cadre global d'intégrité scientifique.

I.3. CYCLE DE VIE D'UNE ÉTUDE DE RECHERCHE BIOMÉDICALE PRÉCLINIQUE

Plusieurs procédures et modèles de cycle de vie de la recherche biomédicale ont été instaurés par les instituts et par les chercheurs afin de mieux organiser et réglementer le déroulement des études biomédicales.

Le cycle de vie peut être défini comme étant une succession d'étapes, de phases, ou d'états que subissent un objet d'étude qui peut être un organisme vivant, un produit commercialisé, un produit en cours de fabrication, etc. Le mot cycle de vie implique des changements, des événements, des procédures, et le suivi dans le temps depuis le début de la vie de l'objet jusqu'à sa fin d'exploitation pour réexploitation.

Dans cette section, nous nous intéressons tout d'abord aux différents « cycles de vie » identifiés dans la recherche biomédicale afin de mieux identifier le « cycle de vie », objet d'étude de notre thèse. Les cycles de vie reconnus qui sont en lien avec la recherche biomédicale sont principalement liés au dispositif médical, au médicament, aux essais cliniques et à la recherche translationnelle. Dans ce paragraphe, nous les décrivons pour situer le cycle de vie d'une étude biomédicale dans le grand spectre.

I.3.1. LE CYCLE DE VIE DES PRODUITS DE SANTÉ

Un produit de santé est tout objet matériel ou logiciel utilisé dans un contexte de santé. Ils sont répartis selon l'ANSM (Agence National de Sécurité des Médicaments et des produits de santé) en 14 catégories et inclus les médicaments et les dispositifs médicaux[27].

Le cycle de vie des produits de santé est décrit du point de vue réglementaire par l'ANSM en trois phases :

1. **Avant la mise sur le marché** : Pendant cette phase, la recherche biomédicale est exécutée afin de pouvoir concrétiser et réaliser le produit de santé. L'ANSM intervient pour attribuer les autorisations et réaliser diverses inspections pour garantir le respect de la réglementation.
2. **La mise sur le marché** : Lors de cette phase, le produit de santé subit une évaluation interne scientifique, réglementaire et en rapport avec le patient. L'ANSM délivre à la fin de cette phase, les autorisations (ou refus) de mise sur le marché du produit. Le produit est en phase d'utilisation.
3. **Après la mise sur le marché :** Le rôle de l'ANSM est de réévaluer en continu l'Autorisation de Mise sur le Marché (AMM) du produit. Un suivi des effets indésirables du produit est effectué afin de pouvoir reconduire son autorisation ou le retirer si des effets notaires sont détectés (affaire Médiator[28]).

Le cycle de vie décrit par l'ANSM est centré sur la dimension réglementée des produits de santé. Ceci est retrouvé aussi dans le cycle de vie de médicament proposé par son homologue européen l'EMA (European Medicines Agency). Elle a mis en place une procédure centralisée de développement et évaluation des médicaments appelée « le voyage d'un médicament du laboratoire au patient » (EMA, 2019). Il s'agit d'une procédure en 6 étapes (voir Figure 25). La recherche biomédicale couvre les deux premières phases (incluant l'avis scientifique en 02). Après, il y a l'évaluation EMA en 03 et l'Autorisation de Mise sur le Marché (AMM) en 04, qui est naturellement suivi par 05 (accès et

[27] https://www.ansm.sante.fr/L-ANSM/Cycle-de-vie-des-produits-de-sante/Avant-la-mise-sur-le-marche/(offset)/0
[28] https://fr.wikipedia.org/wiki/Affaire_du_Mediator

utilisation du produit). La phase 3 des étapes proposées par l'ANSM est à retrouver en 06 pour l'EMA (pharmacovigilance et évaluation continue).

Figure 25 Phases de développement et évaluation d'un produit selon le (EMA, UE)

La FDA (US Food and Drug Administration)[29] décrit elle aussi un processus comparable du cycle de vie des dispositifs médicaux et des médicaments comme indiqué dans la Figure 26 en 5 étapes.

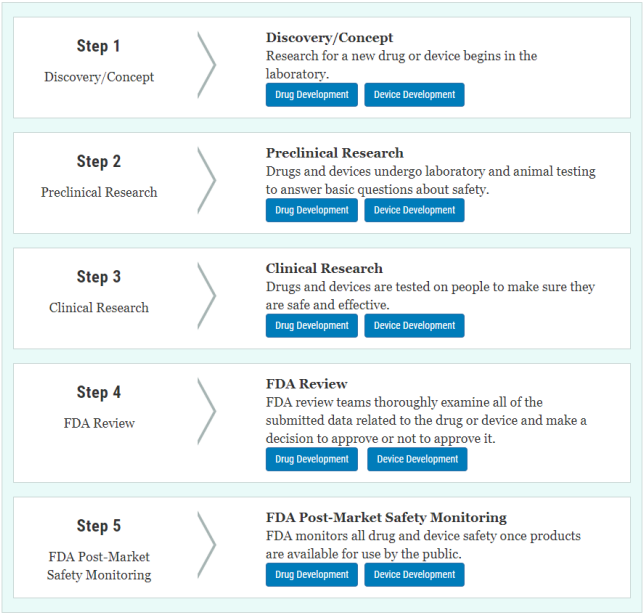


Figure 26 Différentes macro-étapes du développement d'un dispositif médical aux US (source : https://www.fda.gov/patients/learn-about-drug-and-device-approvals)

Premièrement, la découverte du dispositif ou du médicament et sa conceptualisation/conception lors d'une recherche biomédicale. Deuxièmement, le test de son prototype sur des animaux. Cette phase est appelée la recherche préclinique. Troisièmement, et afin d'approuver le dispositif médical ou le médicament, une phase de recherche clinique doit être engagée. Ceci consiste à tester le dispositif médical ou le médicament sur des personnes volontaires et selon un protocole approuvé éthiquement.

[29] https://www.fda.gov/patients/learn-about-drug-and-device-approvals

Passé cette phase, il est aux organismes de référence en santé dans le pays d'approuver ou non le dispositif médical et le médicament : aux US c'est FDA, en Europe, c'est l'UE et en France c'est l'ANSM. La dernière étape est plutôt une étape de suivi de la mise sur le marché du dispositif ou du médicament et de son utilisation afin d'intervenir en cas de dérégulation. Lors de cette étape, des inspections et des contrôles sont assurés par l'organisme concerné.

La Figure 27 résume les différents cycles des produits de santé qui ont été définis de point de vue réglementaire par l'ANSM (France), l'EMA (EU) et la FDA (US). Finalement, ils se croisent et référencent la même chose à quelques différences près.

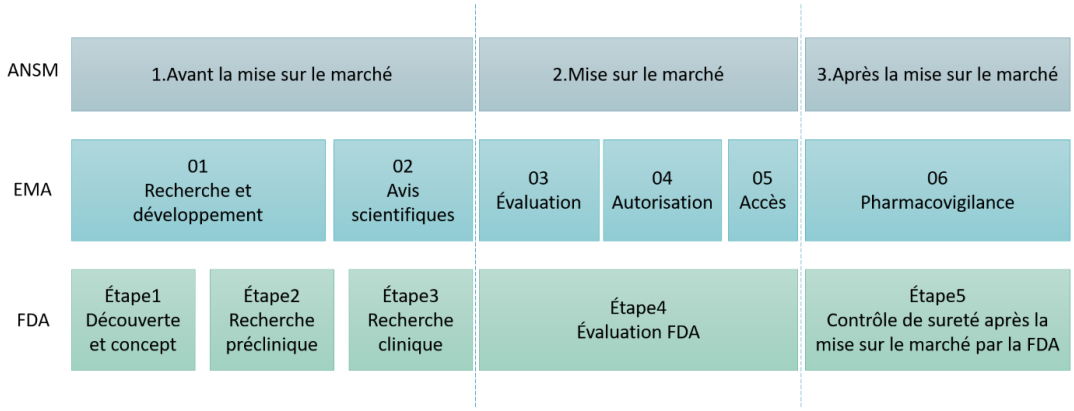


Figure 27 Correspondances entre les différents cycles de vie des produits de santé identifiés

La recherche biomédicale constitue une part importante du cycle de vie des produits de santé identifiés.

I.3.2. LE CYCLE DE VIE D'UNE ÉTUDE DE RECHERCHE BIOMÉDICALE

Il est important à ce stade de bien définir une étude biomédicale. Dans ce manuscrit, une recherche biomédicale contient plusieurs études biomédicales. Pour mieux comprendre cette distinction, prenons l'exemple de la recherche sur la toxicité d'un médicament sur les êtres humains. Cette recherche peut être déclinée en plusieurs études : l'étude de l'effet de ce médicament sur le cœur des souris via des techniques in vivo, une autre étude qui concerne un autre organe, une autre qui concerne des techniques in vitro par exemple ou des techniques d'intelligence artificielle. La définition des différentes études liées à un sujet de recherche s'effectue au fil de l'eau et dépend des moyens à la disposition du chercheur au moment de la définition de l'étude.

Une étude biomédicale est une situation expérimentale dans laquelle une hypothèse liée à la médecine est testée sur un animal (y compris l'Homo sapiens) ou sur un dérivé/prélèvement biologique de ce dernier. Le cycle de vie d'une étude biomédicale est inclus dans le cycle de vie de la recherche biomédicale. Il désigne la succession d'étapes nécessaires à l'exécution de la recherche hormis les étapes liées à la mise en place du projet de recherche. Il est défini dans (Allanic, 2015) en quatre étapes : (1) la spécification de l'étude, (2) l'acquisition des données brutes, (3) l'analyse des données, et (4) la publication des résultats (voir Figure 28).

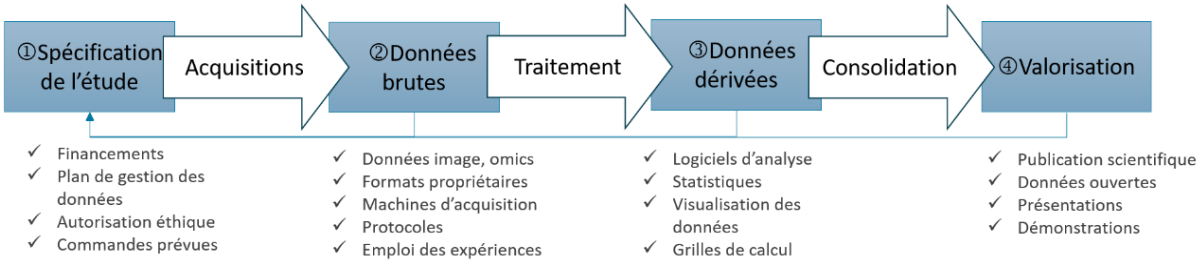


Figure 28 Cycle de vie de la recherche biomédicale, adapté de (Allanic, 2015)

1. **La spécification de l'étude de recherche :** Avant de commencer une étude, une phase de spécification est primordiale. Lors de cette phase, l'investigateur principal de l'étude prépare les documents pour la demande de financement, l'autorisation à l'expérimentation animale, les devis pour les produits commandés, le protocole et le plan d'expérimentation, le Plan de Gestion de Données (PGD ou DMP) et plusieurs autres documents indispensables au lancement de l'étude.
2. **L'acquisition des données brutes :** Une fois l'autorisation acquise, le protocole et le plan d'expérimentation établis, les chercheurs commencent les expérimentations et les acquisitions des données brutes issues des différentes observations, examen et mesures. Ces données sont les premières sources d'informations de l'étude. Ils doivent suivre des règles de qualité, de structuration et de stockage strictes.
3. **L'analyse de données et la consolidation de données dérivées :** Lors de cette étape, qui est celle qui prend le plus de temps et qui génère des données d'une importance capitale pour les résultats de l'étude, les chercheurs utilisent une multitude d'outils logiciels afin d'analyser les données brutes et de produire des résultats statistiques et biologiques intéressants pour la communauté scientifique. De nos jours, il y a de plus en plus d'utilisation de grilles de calcul et de traitement de gros volumes de données.
4. **La valorisation des résultats :** La phase finale d'une étude de recherche est la publication scientifique des résultats de l'étude. La publication peut porter sur un outil de traitement, un résultat biologique ou statistique, un protocole, un jeu de données structurées, etc. Elle référence d'autres articles liés.

Les études et résultats publiés peuvent constituer une base pour les études futures, lançant ainsi un nouveau cycle de vie avec des objectifs raffinés et déduits des précédents résultats. Cette réutilisation est au cœur de nos travaux de recherche qui visent à la faciliter et la favoriser.

Nos recherches portent sur le cycle de vie des études de recherche et de leurs données scientifiques, au bout de la chaîne de tous les autres cycles de vies comme montré par le schéma récapitulatif en fin de cette section (Figure 30).

I.3.3. LE CYCLE DE VIE DES DONNÉES DE RECHERCHE

Les données résultant des recherches scientifiques suscitent de plus en plus l'intérêt des documentalistes et bibliothécaires dans le but d'élargir le concept de la bibliothèque numérique au-delà des livres électroniques afin de permettre la mise à disposition publique des données scientifiques dans le cadre de l'« open data » ou en français l'ouverture des données scientifiques. Le cycle de vie des données est décrit par le « UK Data Archive »[30] , repris par l'INIST[31], en 6 étapes comme le montre la Figure 29: la création des données, le traitement des données, l'analyse des données, la conservation des données, l'ouverture d'accès aux données et la réutilisation des données.

[30] https://www.ukdataservice.ac.uk/manage-data/lifecycle.aspx
[31] https://www.inist.fr/wp-content/uploads/donnees/co/module_Donnees_recherche_7.html

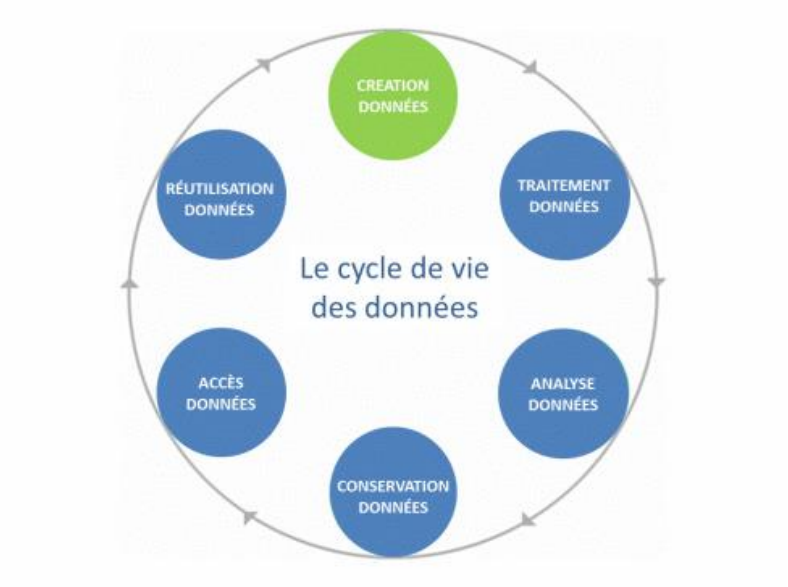


Figure 29 Cycle de vie des données de recherche, adapté de UK Data Archive

I.3.4. CONCLUSION SUR LES DIFFÉRENTS CYCLES DE VIE EXPLORÉS

Les différents cycles de vie présentés dans cette section peuvent être vus comme étant des cycles imbriqués les uns dans les autres comme explicité dans la Figure 30 suivante.

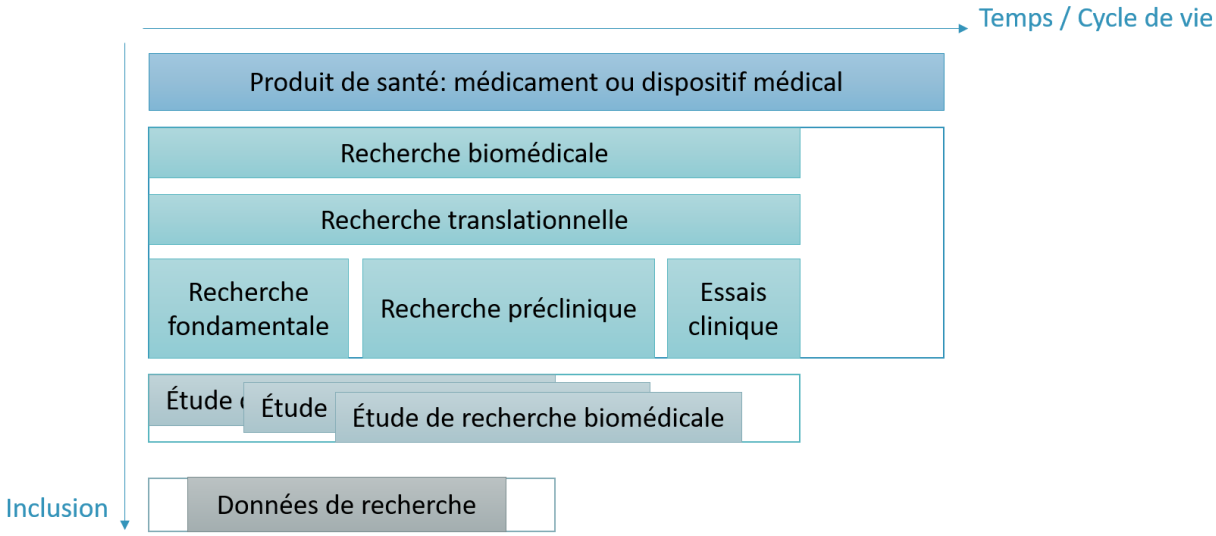


Figure 30 Différents cycles de vie en recherche biomédicale

Nous nous positionnons au niveau du cycle de vie d'une étude de recherche biomédicale et nous prenons compte aussi du cycle de vie des données scientifiques. Nous nous intéressants en effet au cycle de vie des études de recherche préclinique dès la spécification de l'étude à la publication des résultats en passant par l'acquisition et l'analyse des données.

I.4. LA GESTION DE CYCLE DE VIE DES PRODUITS (PLM) APPLIQUÉE À LA RECHERCHE EN NEUROIMAGERIE

Jusqu'ici, nous avons décrit les données hétérogènes en recherche préclinique et nous avons identifié et caractérisé les cycles de vie concernés par nos recherches. Nous avons conclu que les données en

recherche préclinique sont hétérogènes sur plusieurs niveaux et nous nous sommes focalisés sur le cycle de vie d'une étude de recherche biomédicale. Dans ce qui suit, nous présentons les éléments de contexte qui vont nous permettre de les étudier, à savoir : la gestion de cycle de vie des produits (PLM) et son application pour la neuroimagerie réalisée dans le cadre du projet BIOMIST (ANR-13-CORD-0007) et de la thèse de (Allanic, 2015).

I.4.1. LA GESTION DE CYCLE DE VIE DES PRODUITS (PLM)

Nous explicitons dans ce paragraphe les différents atouts, composantes et fonctionnalités des solutions de gestion de cycle de vie des produits ou Product Lifecycle Management (PLM).

I.4.1.1. Historique du PLM

(Terzi et al., 2010) définissent la gestion de cycle de vie des produits (PLM) comme suit :

« PLM can be broadly defined as a product centric – lifecycle-oriented business model, supported by ICT, in which product data are shared among actors, processes and organizations in the different phases of the product lifecycle for achieving desired performances and sustainability for the product and related services. »

Le PLM est une approche intégrée de gestion du flux d'informations des produits, orientée sur le cycle de vie. Il permet la centralisation de l'information, et le partage des connaissances relatives au produit entre collaborateurs au sein de l'entreprise étendue dans le cadre de l'ingénierie collaborative (Belkadi et al., 2010). En effet, de plus en plus de partenaires et d'experts, de différentes localisations géographiques et expertises, travaillent ensemble afin de développer, produire et livrer un produit ou un service. Le cycle de vie d'un produit est composé de processus, d'acteurs, de ressources, d'outils et de méthodes. Il s'agit d'un environnement complexe et fortement collaboratif et compétitif.

Les logiciels PLM ont pour caractéristique la flexibilité de leurs modèles de données et la capacité à implémenter tous types de workflow métier (validation et traitement des données, etc.). Ceci est concrétisé via la conception d'outils logiciels qui intègrent toutes les données créées et échangées tout au long du cycle de vie d'un produit. Le PLM est entre autres utilisé pour la conception 3D et la gestion des fichiers de la CAO (Conception Assistée par Ordinateur) en automobile et aéronautique (Eynard et al., 2004). Le PLM est historiquement très fortement lié aux outils et activités de développement des produits.

I.4.1.2. Les étapes du cycle de vie des produits en industrie

Selon (Terzi et al., 2010), le cycle de vie d'un produit est composé de trois phases détaillées dans la Figure 31 : le « Beginning-Of-Life (BOL) » avec la conception et la fabrication, le « Middle-Of-Life (MOL) » avec la livraison, l'utilisation et le support et le « End-Of-Life (EOL) » avec le retrait de service et le recyclage.

(Jun et al., 2007) considèrent plutôt le « closed-loop PLM » afin d'insister sur la gestion de tous les flux d'information mis en jeu, et réutilisés, tout au long du cycle de vie d'un produit sans contraintes temporelles, ni spatiales. Le « closed-loop PLM » valorise le retour d'expérience (feed-back) et la communication entre les phases du cycle de vie d'un produit. Dans la phase de conception et fabrication (BOL), les experts prennent compte des expériences et connaissances précédentes afin d'améliorer le produit en cours.

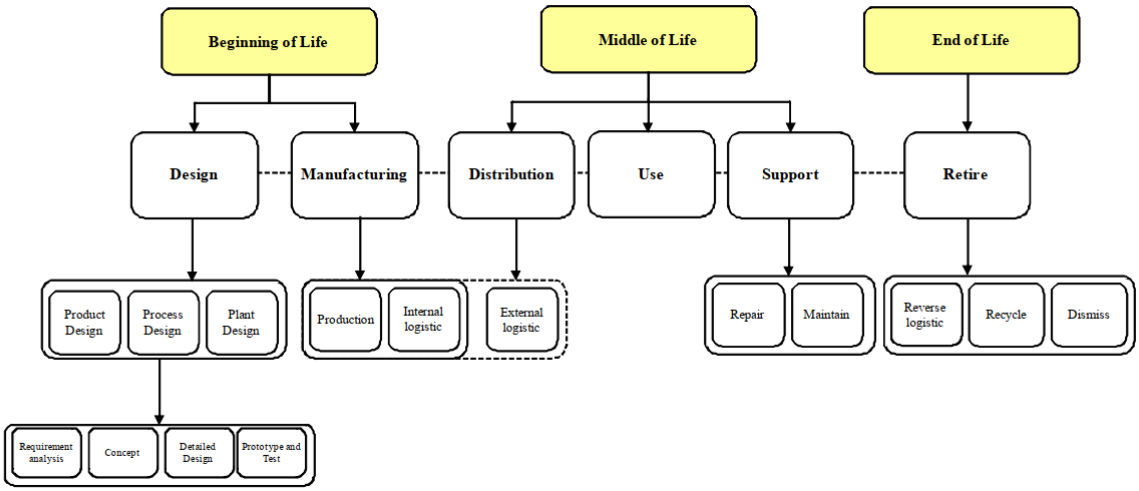


Figure 31 Étapes de cycle de vie des produits (Terzi et al.2010)

I.4.1.3. Les plateformes logicielles en PLM

Afin d'assurer la gestion des données et informations des produits tout au long du cycle de vie, les plateformes PLM ont été développées. Les plateformes PLM sont utilisées dans l'industrie manufacturière depuis plus de vingt ans pour la gestion du cycle de vie des produits complexes et la conception collaborative. Elles ont fait leurs preuves et sont exploitées quotidiennement dans le monde par des millions d'utilisateurs (16 millions de licences actives dans le monde).

Une plateforme PLM est une suite logicielle permettant l'intégration de données multisources, hétérogènes et complexes afin de faciliter la gestion et le partage d'informations entre les différents partenaires et sites. Plusieurs suites logicielles ont été proposées, soit en accès libre telles que OpenPLM ou ARAS, soit en formats propriétaires telles que Teamcenter de Siemens[32], Windchill de PTC[33] et Enovia de Dassault Système[34].

Historiquement, les plateformes PLM s'appuient principalement sur les fonctionnalités des SGDT (Système de Gestion de Données Techniques) ou systèmes PDM (Product Data Management). Ces systèmes permettaient de gérer les documents issus de bureaux d'études afin d'aider les concepteurs à les organiser et à gérer leurs versions. Ceux-ci assuraient la gestion des données des produits sur des sites distants et permettaient de centraliser et livrer l'information au bon moment, mais ils ne considéraient pas l'ensemble du cycle de vie (Tony Liu & William Xu, 2001). Le PLM est un paradigme en perpétuelle évolution et maturation depuis plus de 20 ans.

Une plateforme PLM est destinée à gérer les informations complexes provenant de plusieurs sites et tout au long du cycle de vie des produits. Elle assure la traçabilité des modifications et évolutions du produit et de son environnement dès sa création à son obsolescence. De plus, un produit est géré comme une abstraction assurant la cohérence entre ses différentes représentations : géométrique, cinématique, électromécanique, etc., et son environnement : moyens de production, documentation de fabrication, systèmes industriels, etc. Le PLM permet ainsi de « fournir la bonne information au bon moment à la bonne personne » (Terzi et al., 2010).

[32] https://www.plm.automation.siemens.com/global/fr/products/teamcenter/
[33] https://www.ptc.com/fr/products/plm/plm-products/windchill

[34] https://www.3ds.com/fr/produits-et-services/enovia/

I.4.1.4. Les trois fondamentaux du PLM

Les trois fondamentaux du PLM, explicités dans la Figure 32 de (Terzi et al., 2010), sont : les méthodologies, les processus et les TICs.

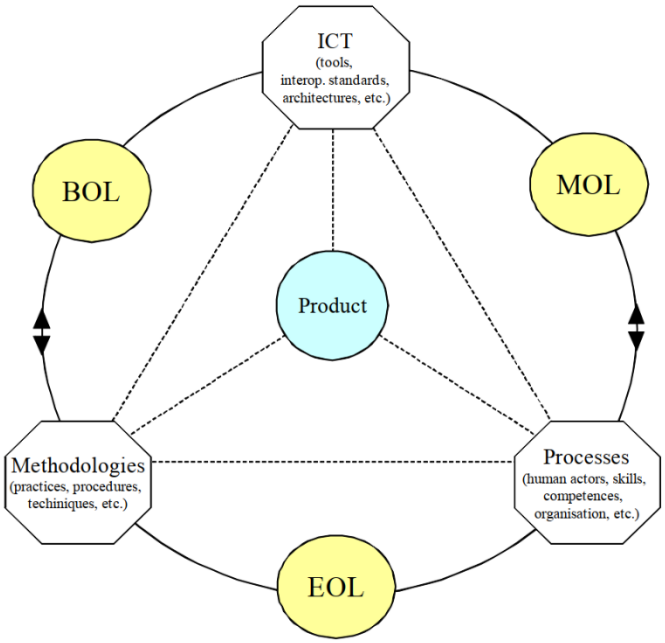


Figure 32 Les trois fondamentaux du PLM (Terzi et al., 2010)
BOL : Beginning of life, MOL : Middle of life, EOL : End of life, ICT : Information and Communication Technology

- **Méthodologies :** Une méthodologie est un ensemble de principes, de procédures et de pratiques communément reconnus dans un domaine. Plusieurs méthodologies différentes doivent être utilisées tout au long du cycle de vie des produits. Le PLM permet de fournir et d'assister la préparation des données indispensables pour la mise en œuvre des méthodologies concernées, d'où son importance.
- **Processus :** Un processus est une liste d'activités coordonnées entre plusieurs acteurs, départements et fonctions au sein d'une entreprise afin de créer une valeur ajoutée ; un produit ou un service. En terminologies PLM, il correspond à un workflow (suite d'activités) collaboratif, une des composantes essentielles des systèmes PLM.
- **TIC ou Technologies de l'Information et de Communication :** un ensemble des technologies issues du mariage informatique, multimédia et télécommunications. La composante logicielle des plateformes PLM repose fortement sur la maturité des TICs. Les TICs fournissent les outils et architectures logiciels, les standards d'échange de données, les mécanismes d'interopérabilité entre applicatifs, etc.

I.4.1.5. Briques fonctionnelles et méthodologiques du PLM

Les plateformes PLM proposent des fonctionnalités logicielles matures, facilitant la gestion des données et processus complexes d'un produit industriel. Elles sont en lien étroit avec les différentes phases de cycle de vie d'un produit et concernent sa conception, son utilisation et sa fabrication (Stark, 2015).

Dans la thèse de (Ducellier, 2008), les composants des plateformes PLM ont été introduits via quatorze briques, comme l'indique la Figure 33. Elles sont réparties sur l'axe X, selon si elles sont des fonctions générales de l'entreprise ou des fonctions relatives aux bureaux études, et sur l'axe Y, selon la problématique traitée : des problématiques d'organisation ou des problématiques de gestion de l'information. Ces briques sont pensées dans le cadre d'un système PLM « classique » orienté sur la conception d'un produit industriel.

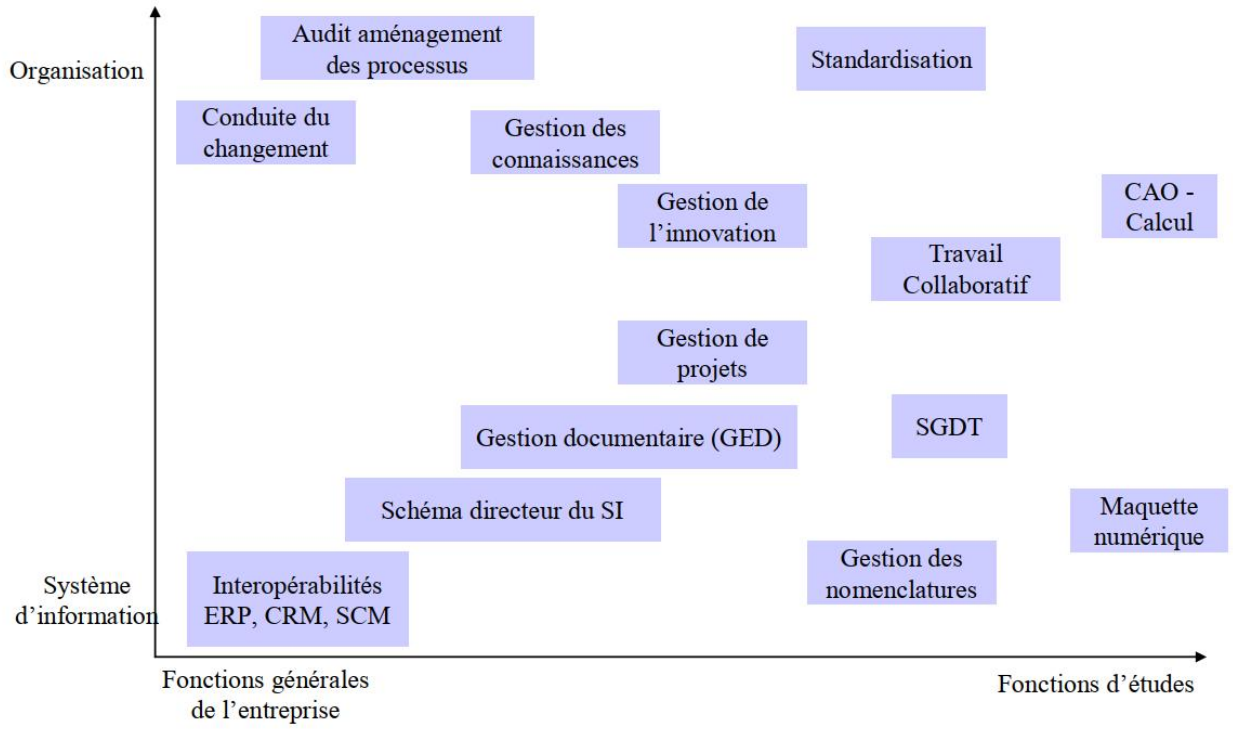


Figure 33 Les quatorze briques d'une plateforme PLM définies par (Ducellier,2008)

La liste des briques PLM introduites ci-dessus se traduit par un ensemble de fonctionnalités logicielles et méthodes de déploiement détaillées en partie dans (Ducellier, 2008) et mis à jour dans ce qui suit.

- **Standardisation :** La standardisation des formats des données et des métadonnées ainsi que des processus, permet une meilleure communication entre collaborateurs et améliore l'efficacité tout au long du cycle de vie d'un produit. La standardisation n'est pas toujours triviale et demande de l'investissement temporel et financier afin d'identifier le standard à suivre ou définir à zéro celui de l'entreprise en question.
- **Gestion des connaissances :** Les plateformes PLM sont considérées comme des systèmes d'information qui gèrent l'échange des données et informations d'un produit tout au long de son cycle de vie. Ces systèmes d'informations sont régis par les connaissances des experts qui travaillent sur le produit en question. La gestion des connaissances est alors étroitement liée au PLM. La gestion des connaissances des experts vise à intégrer les expertises métiers dans toutes les phases du cycle de vie des produits.
- **Gestion documentaire (GED) : U**ne mise à disposition sécurisée de documents techniques à jour à des utilisateurs sur un produit. Les documents sont stockés dans un coffre-fort électronique et ne peuvent être accessibles que via la plateforme PLM. Ceci garantit la cohérence, l'intégrité et la sécurité des données. La gestion documentaire est étroitement liée à la fonctionnalité de gestion des droits et règles d'accès par type d'utilisateur et type de document.
- **SGDT-Système de Gestion de Données Techniques :** Un système PLM est avant tout un SGDT, qui est une suite logicielle centrée sur les données techniques. Il doit inclure toutes les

fonctionnalités primitives d'un SGDT à savoir, la gestion documentaire (GED), la classification et la distribution des données, ainsi que la recherche et la visualisation des composants, afin de permettre une réutilisation efficace des données et composants d'un projet à un autre et d'un site à un autre. Le tout en s'assurant de l'absence de redondance et un maximum de traçabilité concernant les modifications des données techniques (Eynard, 2005).

La visualisation des données techniques s'inscrit dans le cadre de la simplification de la gestion de données pour les experts, mais n'est pas une fonctionnalité inhérente aux plateformes PLM malgré son importance. Elle dépend principalement des outils auteurs de la donnée technique en question. Par exemple, pour visualiser un modèle 3D, une intégration dans la plateforme PLM d'un logiciel de visualisation adapté doit être mise en place (Ducellier, 2008).

- **Gestion de la nomenclature (BOM) :** La gestion de nomenclature correspond à la gestion de la structure arborescente d'un produit, de ses pièces et de leurs différentes configurations et fonctions possibles. Un produit est décomposé récursivement en sous-systèmes jusqu'aux dernières pièces physiques (composants élémentaires indissociables). En terminologie PLM, ceci est appelé BOM (Bill Of Material) ou nomenclature et il y en a plusieurs types (eBOM, mBOM, cadBOM, etc.).
- **Interopérabilité :** L'interopérabilité est la capacité d'intercommunication entre deux systèmes. (IEEE, 1991) la définit comme « La capacité pour plusieurs systèmes ou composants à échanger des informations et à utiliser les informations échangées ». Selon (Danjou, 2015), en s'appuyant sur la norme (ISO 14258, 1998), plusieurs niveaux d'interopérabilité peuvent être identifiés entre les plateformes PLM et les systèmes environnants, ou aussi entre les différentes phases du cycle de vie : niveau sémantique, technique, organisationnel. L'interopérabilité en PLM est un sujet de recherche actif : (Danjou et al., 2013), (Penciuc et al., 2014), (Afoutni et al., 2017),entre autres, témoignent de son importance.
- **Maquette numérique :** Une maquette numérique ou DMU (Digital Mock-Up) est une modélisation en 3D d'un objet avec ses différents composants (généralement assez complexes) et qui visent à simuler son comportement via des calculs et scénarios de simulations variés. Quand une DMU est produite dans le domaine du bâtiment, il est appelé BIM pour « Building Information Modeling ». De nos jours, on parle de jumeau numérique qui permet d'intégrer, en plus de la 3D, la cinématique, la simulation des automates et capteurs, l'utilisation d'un moteur physique, etc. (Gregorio, 2020).
- **CAO-calcul :** La simulation numérique ou IAO (Ingénierie Assistée par Ordinateur) est l'étape de calculs et simulations associés au modèle virtuel issue de la CAO. La gestion des données de ces deux briques est essentielle pour assurer une traçabilité et une maîtrise de complexité de tout le cycle de vie du produit dans un contexte collaboratif. Des travaux de recherche ont été menés afin de pousser encore plus le lien entre la CAO et l'IAO et réduire le triplet coût/délai/qualité de la phase BOL (Assouroko, 2012).
- **Audit et aménagement des processus :** L'audit est en général utilisé pour faire le point sur l'état du système d'information (SI) d'une entreprise. Il peut être global ou concernant un axe en particulier, notamment la sécurité ou la productivité.

  Il passe généralement par des entretiens et des comptes rendus servant à structurer l'information, à identifier et à cerner le besoin et ainsi permettre d'adapter la solution proposée au besoin de l'entreprise audité. Il est régi par plusieurs standards dont (ISACA, 2018), (ISO/IEC 27000, 2018),etc., et fait l'objet de plusieurs rapports et directives notamment (CIGREF, 2019), (CHAI, 2014).
- **Conduite du changement :** La conduite de changement est une discipline à part entière qui permet à une organisation de bien vivre le changement au sein de son organisation, et/ou de son système d'information. Elle peut se définir comme étant « l'ensemble des opérations effectuées au sein d'une organisation pour lui permettre de s'adapter au changement et à l'évolution de l'environnement ». (Stark, 2015) présente les différents acteurs et méthodes dans une démarche de conduite de changement organisationnel dans un environnement PLM.
- **Gestion de projets :** La gestion de projet se définit comme étant l'ensemble des activités visant à organiser le bon déroulement d'un projet et à en atteindre les objectifs. Dans le contexte collaboratif

de la conception, fabrication, et mise sur le marché d'un produit, la gestion de projet est une fonctionnalité essentielle pour les plateformes PLM (Gecevska et al., 2010).

- **Travail collaboratif :** Les fonctionnalités de travail collaboratif dans une plateforme PLM sont principalement : la notification lorsqu'une pièce est disponible pour un contrôle qualité ou une intégration dans un assemblage plus large ; l'échange de documents nécessitant révision et/ou validation, ainsi que le partage via le système des articles (items) préalablement sélectionnés ; les commentaires et la traçabilité des modifications, le système de jeton de réservation d'une ressource pour traiter l'accès concurrent.
  Les plateformes PLM exploitent largement les mécanismes des systèmes CSCW (Computer Supported Collaborative Work) afin de permettre et faciliter la collaboration à distance entre plusieurs sites de conceptions. Le CSCW est un domaine de recherche active en Europe et dans le monde. Il est fortement interdisciplinaire et fait intervenir des chercheurs en Informatique, en Sciences Humaines et Sociales (SHS) et en Sciences et Technologies de l'Information et de la Communication (STIC) (Lewkowicz, 2012).

I.4.2. L'APPLICATION PLM À D'AUTRES DOMAINES

Des travaux de recherche sur l'application du PLM à d'autres domaines en industrie ont vu le jour, comme : l'ingénierie assistée par ordinateur (IAO), la mécatronique, l'architecture et la construction, les services, l'éducation (Fielding et al., 2014) ou encore l'internet des objets. Les éléments présentés dans ce paragraphe sont une extension de l'état de l'art présenté dans (Allanic, 2015).

Dans le domaine biomédical, l'approche PLM a été un bon allié pour la réduction du TTM (Time To Market) des médicaments (Prajapati & Dureja, 2012). L'industrie pharmaceutique est régie par des processus et règles strictes vu les enjeux médicaux et économiques associés à la chaîne d'industrialisation d'un nouveau médicament. Ce contexte a motivé l'utilisation du PLM pour ce domaine.

Les principales fonctionnalités reprises du PLM sont : la spécification préalable du produit, le planning et la définition claire des rôles et des responsabilités, la capitalisation des connaissances et des expertises, la prédisposition du système au changement, notamment le changement des règles de gouvernance et d'organisation. De plus, et avec l'émergence des paradigmes de l'industrie 4.0 qui font évoluer aussi les paradigmes PLM, l'industrie biopharmaceutique peut bénéficier de plus de fonctionnalités technologiques en utilisant le PLM (Branke et al., 2016)

En ingénierie biomédicale, et plus précisément autour des technologies pour les dispositifs médicaux, le PLM est aussi présent comme un cadre de développement prometteur (Lantada, 2013). Des applications pour la fabrication de prothèses personnalisées (Lantada, 2013) ainsi que son implantation (Ngo, 2018) ont été identifiées. En effet, l'analogie est faite entre une prothèse personnalisée, conçue à partir des scans patients reconstitués en 3D (Lantada, 2013) et une conception en 3D d'une pièce de voiture. Dans les deux cas, les besoins en gestion des flux de données et informations dans un environnement collaboratif de CAO sont identifiés. En guise d'exemples, une application PLM à des implants visage personnalisés est décrite en (Ardila-Mejia et al., 2018) et un rapport récent sur l'application du PLM aux dispositifs médicaux et plus précisément aux prothèses des membres inférieurs est donné par (Martínez Gómez et al., 2019).

Au-delà du périmètre de l'ingénierie, (Hervy et al., 2017) a étudié la possibilité d'utiliser un système PLM en muséologie pour la gestion des données et connaissances en lien avec l'histoire. En effet, l'une des fonctionnalités PLM pertinentes est l'archivage des données ainsi que le suivi de leur évolution et leur modification dans le temps. Une plateforme de StoryTelling d'une exhibition a été proposée dans ce contexte (Khundam & Noël, 2017).

I.4.3. L'APPLICATION PLM À LA NEUROIMAGERIE

Dans le cadre de la gestion des données pour la recherche biomédicale, les paradigmes PLM ont été appliqués à la recherche en neuroimagerie lors des travaux de thèse de Marianne Allanic (Allanic, 2015) et à l'issue du projet BIOMIST (ANR-13-CORD-0007). En effet, des similarités entre les deux domaines ont été identifiées, à savoir : compétitivité, pluridisciplinarité, échange sécurisé de données sûres entre sites distants, besoin de réutilisation des données des produits/études préalables. La neuroimagerie est l'étude du fonctionnement du cerveau à l'aide de techniques issues de l'imagerie médicale.

L'application du PLM à la neuroimagerie a été principalement centrée sur les fonctionnalités de visualisation des données complexes, la standardisation ainsi que les fonctionnalités de base d'un SGDT comme l'administration et la gestion documentaire (GED) (voir §I.4.1.5. pour rappel des briques du PLM).

Dans ce qui suit, nous présentons les produits de l'application PLM au domaine de la recherche en neuroimagerie : Le modèle de données BMI-LM et la plateforme BIOMIST, (Allanic et al., 2017) ont pour mission d'améliorer la gestion des données hétérogènes en neuroimagerie via la traçabilité des données et de leur provenance. Ensuite, nous analysons leurs apports et leurs manques par rapport à l'état des lieux en gestion de données hétérogènes en recherche biomédicale en général en recherche préclinique en particulier.

I.4.3.1. Le modèle de données BMI-LM

Le modèle de données BMI-LM (BioMedical Imaging – Lifecycle Management) est composé de 19 classes génériques (voir le Tableau 4 ci-après) qui permettent de structurer les données. Le modèle de données BMI-LM présente trois catégories de classes génériques :

1. **Classes de définition (D) :** elles décrivent comment les classes de résultat doivent être obtenues et peuvent être réutilisées d'une étude à l'autre. Elles constituent un pilier de la stratégie de provenance. Exemple : « Data Unit Definition – DUD » du Tableau 4.
2. **Classes de résultat (R) :** elles contiennent les vrais paramètres d'acquisition et de traitement de données utilisés ainsi que les données, brutes ou dérivées, d'une étude, sous la forme de fichiers ou de métadonnées. Exemple : « Data Unit Result – DUR » du Tableau 4.
3. **Classes ambivalentes (A) :** en fonction du contexte, ces classes peuvent être utilisées comme classes de définition ou classe de résultat. Exemple : « Reference Data– RFD » du Tableau 4.

Tableau 4 Liste des objets du modèle de données BMI-LM à l'issue de la thèse de (Allanic, 2015)

Objet	Sigle	Traduction française	Description
Acquisition	ACQ	Acquisition	Période indivisible d'acquisition de données
Acquisition Definition	ACD	Définition d'une acquisition	Description d'un protocole d'acquisition
Acquisition Device	AQD	Dispositif d'acquisition	Description d'un dispositif utilisé pendant un examen
Bibligraphical Reference	BBR	Référence Bibliographique	Article scientifique
Data Unit Result	DUR	Unité de données	Donnée acquise isolée
Data Unit Definition	DUD	Définition d'une unité de données	Description d'une unité de données
Exam	EXA	Examen	Ligne continue d'acquisitions
Exam Definition	EXD	Définition d'un examen	Description de la chaîne d'acquisitions
Subjects Group	SGP	Groupe de sujets (dans l'étude)	Ensemble de sujets dans l'étude regroupés selon un critère (brut ou dérivé)
Processing Definition	PCD	Définition d'une chaîne de traitement	Description d'une chaîne de traitement
Processing	PCR	Chaîne de traitement	Chaîne de traitement
Processing Parameter	PCP	Paramètres de traitement	Jeu de paramètres utilisés pour un traitement
Processing Unit Definition	PUD	Définition d'une unité de traitement	Description d'une unité de traitement
Processing Unit Result	PUR	Unité de traitement	Traitement effectué sur des données
Reference data	RFD	Donnée de référence	Donnée d'entrée d'un traitement hors du contexte d'une étude
Software Tool	STL	Logiciel	Description d'un logiciel de traitement
Study	STU	Etude	Etude de recherche
Study Subject	SSU	Sujet dans l'étude	Sujet dans le contexte d'une étude
Subject	SUB	Sujet	Sujet unique dans la base

Les classes du modèle de données BMI-LM sont reliées entre eux en utilisant deux types de relations :

- Relation de traçabilité : un lien entre deux objets de Résultat ou deux objets de Définition, elle permet de tracer la provenance des données.
- Relation d'identification : liaison entre un objet Résultat et l'objet de Définition qui lui correspond.

Le schéma UML suivant en Figure 34 résume les différents liens qui existent entre les classes du modèle de données BMI-LM.

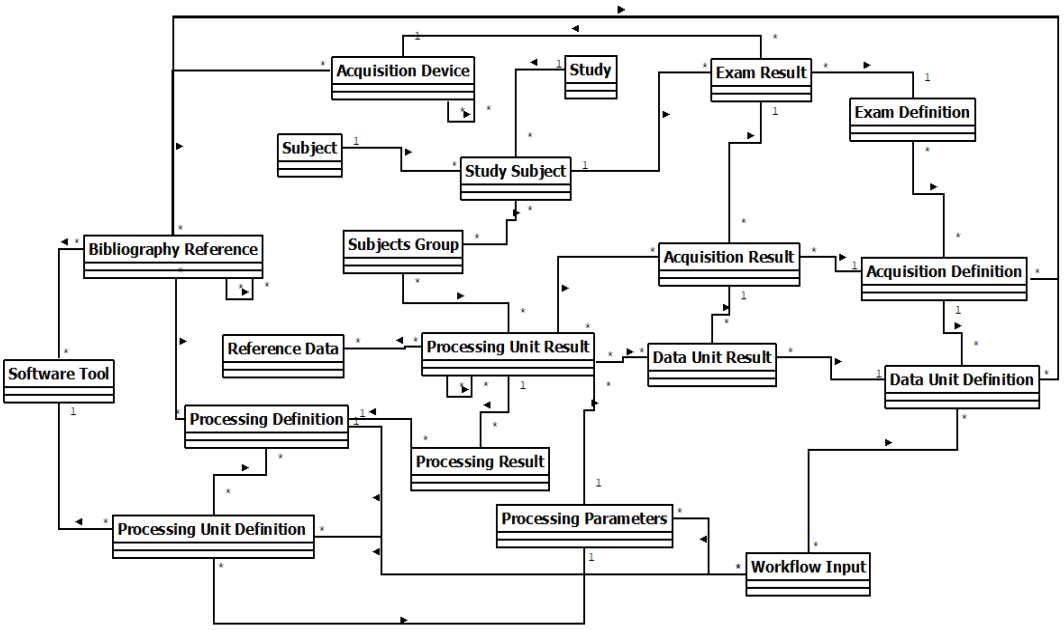


Figure 34 Schéma UML du modèle de données BMI-LM (Allanic, 2015)

Associé aux classes génériques du modèle de données BMI-LM, des classes spécifiques ont été proposées afin d'accompagner les classes génériques dans la description et l'organisation des données de la recherche en neuroimagerie. L'ensemble des classes spécifiques sont rangés dans un arbre hiérarchique appelé « Classification », elle a comme racine les classes présentées dans la Figure 35 suivante. Comme on peut le remarquer, à chaque classe du modèle de données BMI-LM, une branche de la classification est attribuée. Cet arbre est évolutif et permet de spécifier les classes en fonction des disciplines et des types de données mis en jeu.

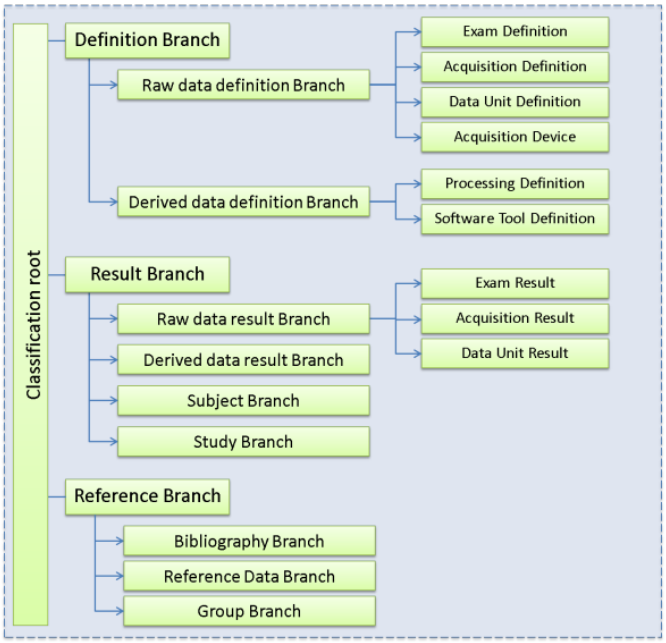


Figure 35 Classes racines d'un arbre de classification associé au modèle BMI-LM (Allanic, 2015)

L'arbre de classification est dépendant du domaine d'application puisqu'il contient des concepts spécifiques. La classification pour la neuroimagerie a été construite en réutilisant des ontologies de

domaine [35] comme QIBO (Quantitive Imaging Biomarker Ontology) (Buckler et al., 2013), OntoNeurolog (ONL-DA Dataset …)[36], NIF[37] et OCRe[38] comme indiqué dans la Figure 36 ci-après.

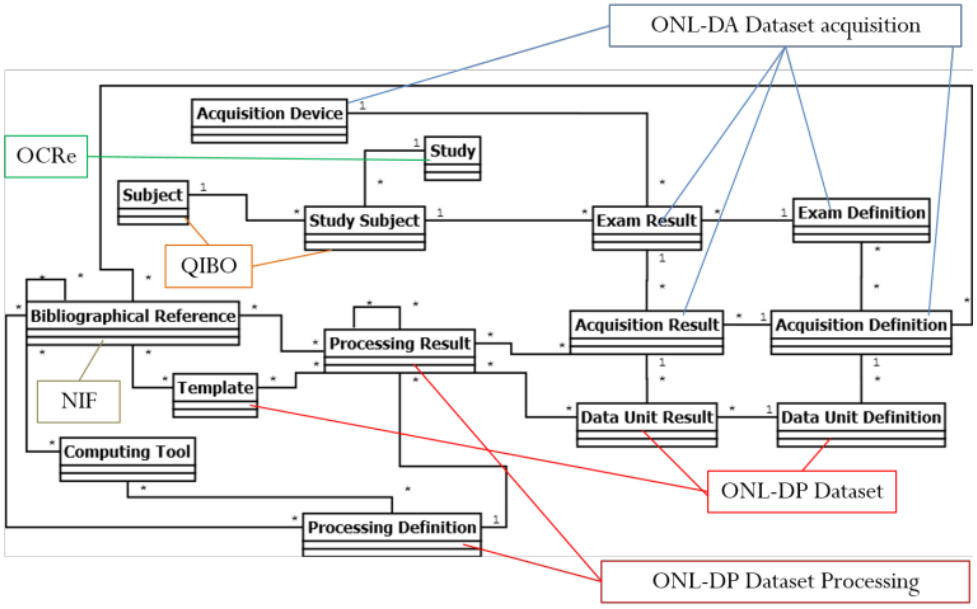


Figure 36 Ontologies et standards utilisés pour la classification pour de la neuroimagerie (Allanic, 2015)

I.4.3.2. La plateforme du projet BIOMIST

À l'issue du projet BIOMIST ANR-13-CORD-0007, une suite logicielle qui a été développée. Elle a été élaborée par l'entreprise Fealinx en collaboration avec le Groupe d'Imagerie Neurofonctionnelle (GIN) de l'Institut des Maladies Neurodégénératives UMR 5293 de l'Université de Bordeaux. Elle est commercialement distribuée par l'entreprise Fealinx sous le nom de SWOMed[39]. Cette plateforme a été testée avec des données en imagerie IRM et avec des données textuelles de questionnaires de repos. Les résultats obtenus étaient prometteurs.

La gestion des données dans la plateforme BIOMIST est supportée par un logiciel de Product Lifecycle Management (PLM), en l'occurrence la solution Teamcenter[40] éditée par Siemens Industries Software. Il s'agit d'une solution logicielle en architecture 4 tiers comme explicité dans la Figure 37.

Du bas en haut de la Figure 37: Le premier niveau est celui des ressources. La plateforme Teamcenter dispose d'une base de données en Microsoft SQL server et d'un serveur de cache pour l'accélération et la sécurisation des fichiers stockés dans le système. Ces fichiers sont cryptés dans des dossiers appelés « volumes » ou « coffre-fort ». Ensuite, pour la couche « entreprise » applicative, il y a la possibilité d'avoir plusieurs serveurs d'applications répartis sur plusieurs sites ou équipes et qui sont synchronisés avec la couche web en se basant sur le « Pool manager ». La couche suivante est la couche web. Le serveur web permet d'avoir des applications clientes via des navigateurs web et des requêtes HTTP.

[35] Vocabulaires communément connus des domaines
[36] http://neurolog.i3s.unice.fr/public_namespace/ontology
[37] https://bioportal.bioontology.org/ontologies/NIFSTD
[38] http://www.ontobee.org/ontology/OCRe
[39] http://www.fealinx-biomedical.com/fr/swomed/
[40] https://www.plm.automation.siemens.com/global/fr/products/teamcenter/

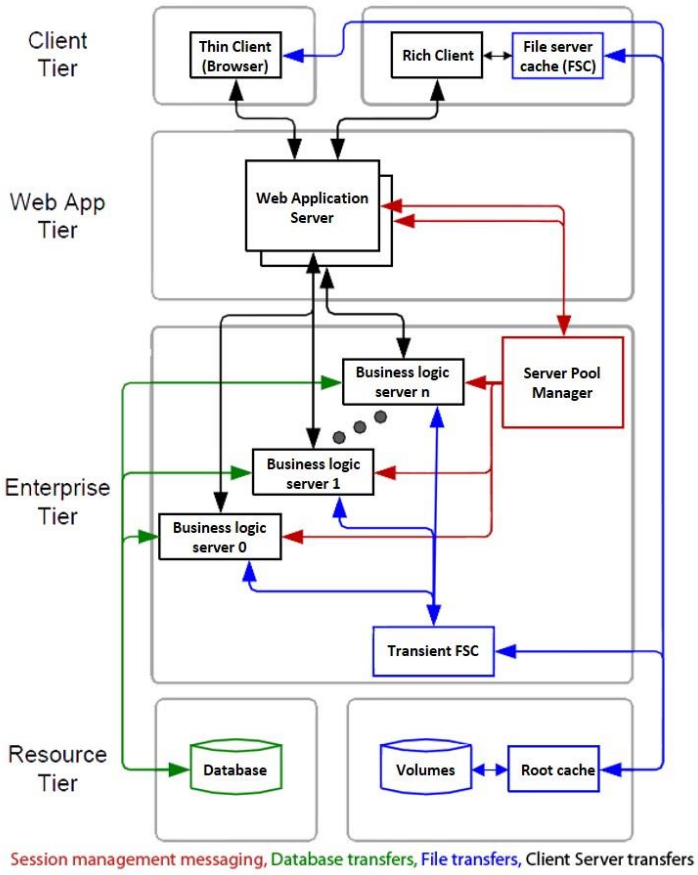


Figure 37 Architecture en 4 tiers de la solution PLM Teamcenter (source : documentation Teamcenter Siemens)

Et enfin, au plus haut, les applications clientes, fournies par le logiciel Teamcenter, sont au nombre de trois :

- Un client bureautique appelé « Client Riche » qui est installé sur une machine spécifique et utilisable par plusieurs personnes sans risque de sécurité.
- Un client web appelé « Client Léger » qui est une transcription du client bureautique mais via le navigateur (ceci a été progressivement délaissé par Siemens)
- Un client web « moderne » appelé AWC (Active Workspace Client ; non présent dans la Figure 37). Il est une version « nouvelles technologies » et qui se veut ergonomique et compatible avec n'importe quels navigateurs et appareils utilisés.

Les différentes réalisations explicitées dans ce manuscrit ont été réalisées principalement via le « Client Riche » et occasionnellement via le « Client Web AWC ». Des captures d'écran de leurs interfaces graphiques respectives sont données et commentées en Annexe A pour plus de clarté.

La plateforme BIOMIST (Allanic et al., 2017) réutilise les nœuds de la plateforme PLM Teamcenter et y ajoute un ensemble de composants et nœuds spécifiques à la gestion de données de recherche en neuroimagerie avec traçabilité de la provenance des données, qui sont :

- Le modèle de données BMI-LM spécifique à la gestion de cycle de vie de la recherche biomédicale en neuroimagerie.
- L'import des données en format DICOM via le protocole ad hoc en utilisant la bibliothèque dcm4che[41].

[41] https://www.dcm4che.org/

- L'envoi de données à des clusters de calculs pour leur analyse (Allanic et al., 2016), et la récupération des résultats en utilisant l'intégrateur de workflow Nipype (Gorgolewski et al., 2011).
- L'import de données de questionnaires de repos (des données spécifiques à la neuroimagerie) via un webservice dans la plateforme BIOMIST.
- L'export des données stockées dans la base de données de la plateforme BIOMIST, via un webservice.
- L'exploration de la base de données via l'outil JMP[42] en utilisant un connecteur ODBC (Open DataBase Connectivity). ODBC est une interface logicielle de connexion standardisée à différents types de bases de données. Elle est compatible avec plusieurs systèmes de gestion de bases de données (SGBD) et est disponible en plusieurs langages de programmation.

Toutes ces adaptations et fonctionnalités sont plus amplement décrites dans (Allanic,2017). Des plateformes sont actuellement déployées chez deux partenaires historiques : le Groupe d'Imagerie Neurofonctionnelle et le PARCC. La solution logicielle SWOMed est en cours de déploiement pour des projets de recherche RHU et PSPC. La Figure 38 résume ses principaux modules et fonctionnalités : import et requêtes des données, intégration des chaînes de traitements, organisation des projets et définition des droits d'accès, stockage en utilisant les volumes de données, gestion via un modèle de données et une classification.

Figure 38 Fonctionnalités de la suite logicielle SWOMed (source : documentation SWOMed version 1.5)

La suite logicielle SWOMed est proposée pour être utilisée tous les jours en routine dans le but d'organiser les données, de tracer l'information et de valoriser les connaissances tout au long du cycle de vie d'une étude de recherche.

I.4.4. CONCLUSION

Les différentes propositions et réalisations lors de l'application PLM à la neuroimagerie ont été validées lors de la thèse (Allanic, 2015) et ont été testées dans le laboratoire GIN, Bordeaux. Ce qui renforce notre hypothèse. Nous continuons à explorer cet axe de recherche en utilisant le modèle de données BMI-LM et la plateforme BIOMIST pour la gestion des données hétérogènes de la recherche

[42] https://www.jmp.com/fr_fr/software.html

préclinique. Nous utiliserons pour cela, les différentes briques des plateformes PLM identifiées dans la section I.I.4.1.5.

I.5. POSITIONNEMENT SCIENTIFIQUE ET QUESTIONS DE RECHERCHE

Même si la neuroimagerie et la recherche préclinique sont régies toutes les deux par des études biomédicales. Ils n'ont pas exactement les mêmes caractéristiques (voir Tableau 5). Ce qui a motivé cette thèse comme continuité des travaux de recherche sur l'application PLM à la recherche biomédicale.

Tableau 5 Comparaison entre la recherche préclinique (au LRI) et la recherche en neuroimagerie (au laboratoire GIN)

	Recherche en neuroimagerie clinique (GIN)	Recherche préclinique (LRI)
Durée	plusieurs années	quelques mois
Expérimentations	principalement in vivo	in vivo et in vitro
Cardinalité des groupes	des milliers de participants	une dizaine d'animaux
Modalités	IRM, examens cliniques neuropsychologiques	IRM, TEP-TDM, US, histologie, omiques, western-blot

En plus des besoins de flexibilité de la recherche biomédicale, la recherche préclinique est marquée par plus de changements que la recherche en neuroimagerie. Les études se réalisent sur une plus courte durée et font intervenir des modalités d'acquisition de données diverses et variées ainsi que des expertises pointues. Dans cette thèse, nous élargissons notre champ de recherche à tout le domaine de la recherche biomédicale tout en exploitant les données de la recherche préclinique, notre domaine d'application.

Dans le Tableau 6, l'analogie entre le cycle de vie d'une étude biomédicale et d'un produit est présentée sur les plans de : collaboration, aspects multisites, pluridisciplinarité, hétérogénéité, complexité des données et des workflows, partage, réutilisation de données, et flexibilité. Ils présentent des critères assez semblables sur tous les plans à part celui de la flexibilité. En effet, la recherche biomédicale est plus évolutive que la conception des produits.

Tableau 6 Analyse comparative de la gestion de cycle de vie des : produits et études biomédicales

Gestion de cycle de vie de	**Produit**	**Étude biomédicale**
Collaboration complexe	Entre les concepteurs CAO, les ingénieurs industriels, le responsable logistique	Entre médecins, radiologues, biologistes, physiciens, informaticiens, ingénieurs de recherche
Multisite	Entre plusieurs sites de productions ou entreprises partenaires ou fournisseurs/clients	Entre plusieurs laboratoires ou plateformes de recherche ainsi que des industriels pharmaceutiques
Pluridisciplinarité	Conception, production, test, support, maintenance, mécanique, électronique, mécatronique, pneumatique, etc.	La neurologie, la cardiologie, la biologie, la physique, l'informatique, la cancérologie…
Hétérogénéité	Utilisation de différents logiciels CAO et donc différents formats	Différentes sources, différents formats, différente sémantique, forte hétérogénéité des données
Complexité des données et des workflows	La complexité du produit qui peut contenir plusieurs milliers de pièces	Le croisement de données fortement hétérogènes est une tâche complexe pour un chercheur en biomédical

	qui évoluent toutes avec des cycles de vie en parallèle	
Partage	À toutes les étapes du cycle de vie. Par exemple, en vue de collaboration sur la conception d'une voiture, un avion	Partage de données pour revue de projet et double analyse et vérification
Réutilisation de données	Conception de pièces et assemblages complexes réutilisables d'un produit à un autre	Réutilisation de données d'une première étude dans une autre étude pour des tests de faisabilité ou de l'économie de ressources
Flexibilité de gestion	Une fois le cahier des charges et les délais fixés, la flexibilité n'est plus un critère de gestion	De nouvelles idées émergent en recherche et doivent être parfois testées rapidement. Le système de gestion doit pouvoir suivre ces innovations sans difficulté.

Le contexte de la recherche biomédicale largement décrit dans ce chapitre introductif conditionne la gestion de données de recherche et fait émerger des questionnements quant à l'aptitude des données scientifiques à être partagé en « données ouvertes » plus tard et à être exploitées et réutilisées. Nous avons identifié deux freins pour la réutilisation des données :

1. La perte de confiance dans les données, due au départ de leur producteur ou à la perte d'annotations (de provenance ou autre) les décrivant.
2. L'incapacité de déchiffrer les annotations des données due à une utilisation de termes vernaculaires et au manque de standardisation.

Nous nous focalisons sur la problématique de la préparation pour partage et réutilisation ultérieurs, de données scientifiques hétérogènes tout au long du cycle de vie des études de recherche. Nous développons nos propositions en étroite collaboration avec les chercheurs du laboratoire LRI en recherche multimodale préclinique. Nos recherches ont été étroitement liées aux domaines suivants :

- La gestion de données de recherche (RDM) dans le cadre de l'ouverture des données de recherche
- L'application PLM à d'autres domaines
- La recherche biomédicale et la recherche préclinique
- L'organisation, la gestion et l'ingénierie des connaissances
- Les systèmes d'information pour la gestion de données

Nous formulons la question générale de recherche suivante :

1. **Comment gérer les données hétérogènes de la recherche biomédicale préclinique en vue de préparation de leur partage et leur réutilisation ?**

Cette question peut être découpée en deux principales sous-questions :

La réutilisation des données implique une confiance suffisante dans sa provenance, ceci a été mis en place en neuroimagerie via le modèle de données BMI-LM et la plateforme BIOMIST. Nous analysons et évaluons l'existant afin de répondre à la question suivante :

   a. **Comment gérer le cycle de vie des données hétérogènes de la recherche préclinique et leur provenance ?**

La réutilisation implique aussi une compréhension des données qu'on réutilise. Cette compréhension est freinée par les termes locaux vernaculaires propres à chaque équipe ou discipline. Nous considérons la

compréhension, dernier niveau de la pyramide de la connaissance DIKW, afin de répondre à la question suivante :

b. Comment assurer la compréhension des données hétérogènes de la recherche biomédicale lors d'une réutilisation ultérieure ?

La Figure 39 résume les articulations des problèmes et des objectifs de recherches de notre thèse. Elle sera reprise au fil des chapitres pour repositionner le contexte et clarifier nos propos.

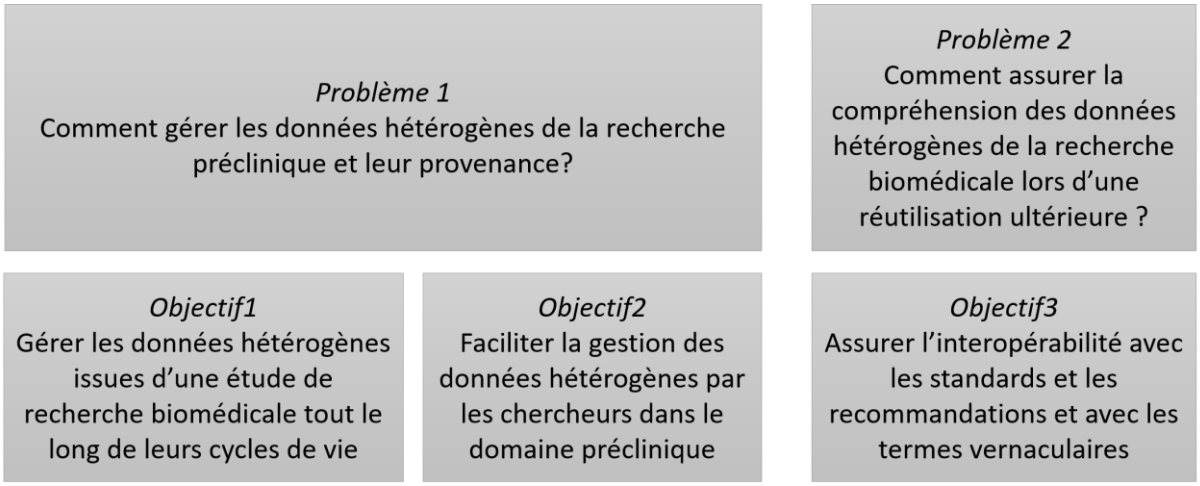


Figure 39 Objectifs de recherche de la thèse

Dans le reste du manuscrit, nous répondons aux questions précédentes en adoptons les étapes suivantes :

2. Immersion dans le quotidien du chercheur via un premier chapitre d'état de l'art sur : l'étude des besoins de la communauté en recherche biomédicale, et en particulier en recherche préclinique, l'analyse des systèmes de gestions de données de recherche identifiés en bibliographie, et l'identification de leviers qui nous permettront de résoudre notre problématique (Chapitre II).
3. Exploration de la compréhension à partir de données via un deuxième chapitre d'état de l'art sur : la pyramide de la connaissance DIKW, la sémiotique, la gestion des connaissances (KM), l'organisation des connaissances (KO), et l'ingénierie des connaissances (KE) (Chapitre III).

À la suite de ces deux chapitres d'état de l'art, nos deux propositions pour la résolution des problèmes de recherche identifiés sont décrites :

4. Proposition du modèle de données et du système BMS-LM pour (BioMedical Study-Lifecycle Management), les successeurs de BMI-LM et de BIOMIST, pour la structuration et la gestion des données de recherche hétérogènes tout au long de leur cycle de vie. Un focus sur les méthodes d'intégration des données hétérogènes brutes et dérivés, et des calculs scientifiques est effectué (Chapitre IV).
5. Proposition d'une représentation ontologique du nouveau modèle de données BMS-LM afin de renforcer la compréhension des données et ainsi leur réutilisabilité. La méthode de mise en place d'une interopérabilité sémantique avec d'autres terminologies et standards du domaine est explicitée (Chapitre V).

Les plans d'expérimentation effectués dans le laboratoire LRI et leurs réalisations respectives seront explicités dans les deux chapitres suivants :

6. Expérimentation et mise en œuvre du système et de l'ontologie BMS-LM pour les données précliniques multimodales au laboratoire LRI. Applications et exemples issus du domaine

préclinique pour la validation des différentes propositions de gestion de cycle de vie des études de recherche biomédicale (BMS-LM) (Chapitre VI).

Un dernier chapitre (Chapitre VII) est dédié à la discussion des résultats et l'annonce des perspectives de recherche.

La structure du manuscrit, de l'introduction aux perspectives est présentée dans la Figure 40.

CONCLUSION DU CHAPITRE I

Dans ce chapitre, nous avons situé l'imagerie et la recherche préclinique dans leurs contextes réglementaires et translationnels. Nous avons montré et explicité l'hétérogénéité des données de recherche ainsi que les outils et formats de données mis en jeu et nous avons présenté l'état des lieux dans le laboratoire LRI. Ensuite, nous avons situé nos recherches sur la gestion du cycle de vie dans le cadre global de la recherche biomédicale. Avant de présenter notre problématique, nos questions de recherches, et la structure détaillée du manuscrit, nous avons présenté le Product Lifecycle Management (PLM), élément de notre première hypothèse. Son application à d'autres domaines et à la neuroimagerie peut être étendue aux études biomédicales précliniques. Les produits de cette dernière sont le modèle de données BMI-LM (BioMedical Imaging - Lifecycle Management) et la plateforme BIOMIST.

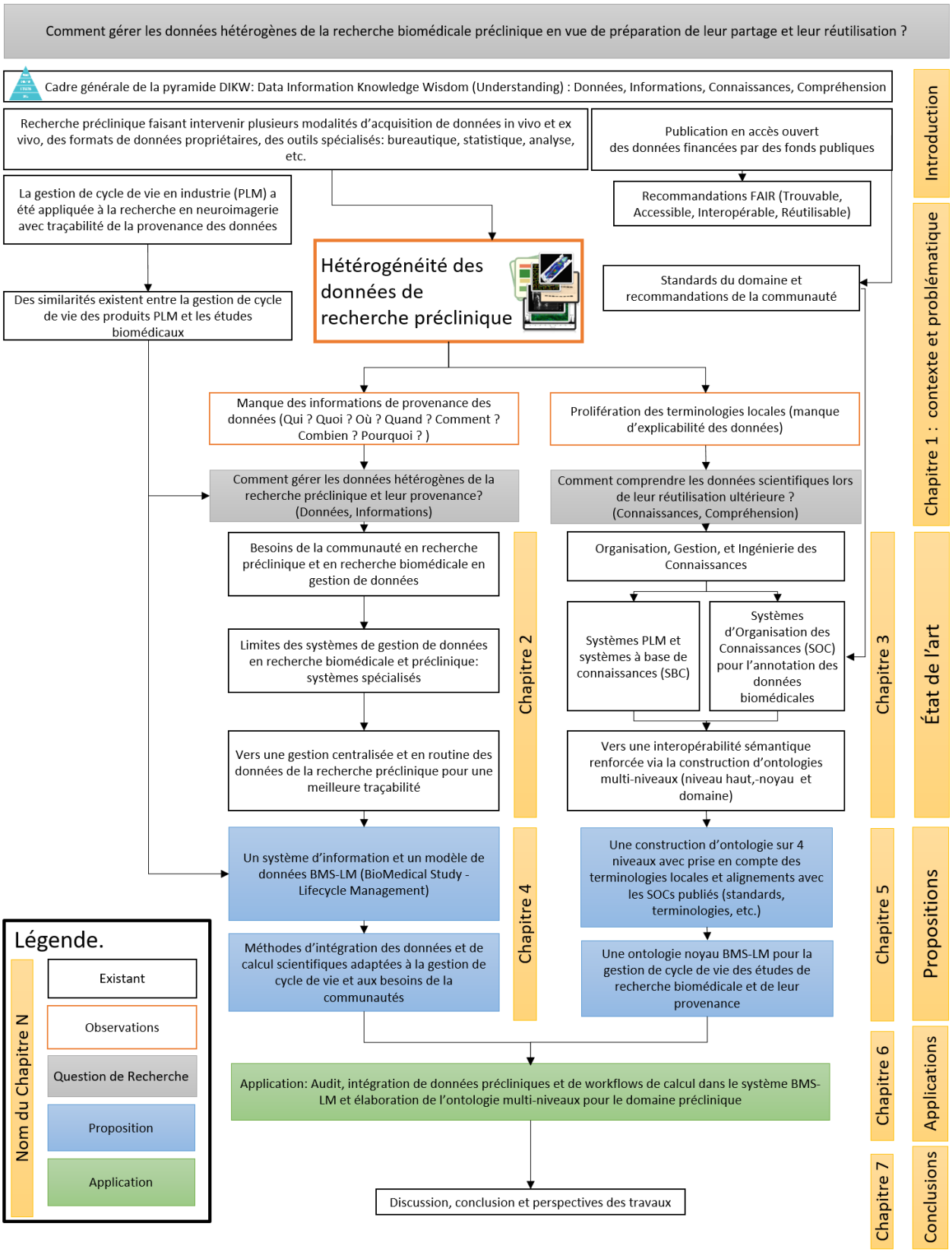


Figure 40 Structure de la thèse

Chapitre II. État de l'art de la gestion des données de recherche biomédicale et préclinique

Au chapitre I, nous avons exploré l'état des lieux en recherche préclinique et au laboratoire LRI et nous avons mis l'accent sur l'hétérogénéité des données à gérer ainsi que la variété des outils mis en œuvre pour leur analyse. La problématique de la gestion de ces données et de leur provenance en vue de leur partage et leur réutilisation a été identifiée (voir Figure 41). En effet, face au contexte global fortement réglementé et face aux données scientifiques lourdement hétérogènes dans le cadre d'une étude de recherche, un chercheur dans le domaine biomédical doit s'équiper d'outils et de procédures lui permettant de bien gérer ses données, d'être cohérent avec les recommandations nationales et internationales, et de gagner du temps.

De nos jours, la recherche biomédicale est orientée sur la publication, vu son caractère compétitif. Ainsi, en informatique pour la recherche, un focus est réalisé sur les outils d'analyse de données et sur l'accès aux ressources de calcul scientifique avancé. Plus spécifiquement, les investissements vont pour les clusters de calcul universitaire pour lancer des scripts de traitement d'image par GPU et pour l'apprentissage automatique. Les demandes de financement pour des nouveaux projets prennent en compte ces coûts en plus pour l'analyse, mais mentionnent rarement le coût correspondant à la gestion de données de recherche (5% sont recommandés par l'Union européenne).

Dans ce chapitre, nous identifions tout d'abord une liste de besoins en gestion de données issues de la bibliographie en recherche biomédicale. Ensuite, nous présentons un état de l'art sur les systèmes de gestion de données en recherche préclinique et biomédicale. Ce qui nous permettra d'identifier leurs limites et d'introduire les différents avantages attendus de l'application des paradigmes du PLM à ce domaine.

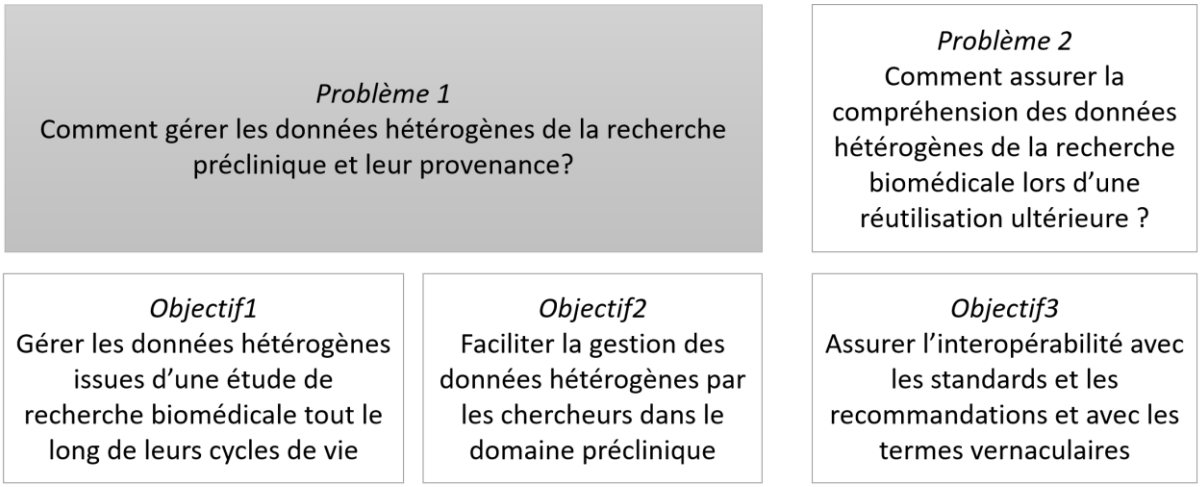


Figure 41 Problème1 : la gestion des données hétérogènes de la recherche préclinique en vue de partage et de réutilisation

II.1. DES BESOINS DE LA COMMUNAUTÉ EN RECHERCHE BIOMÉDICALE

Une liste de besoins de la communauté a été identifiée lors de l'enquête de (Anderson et al., 2007), où 286 chercheurs du Nord-Ouest Pacifique (Université de Washington) ont été interviewés via un questionnaire de 36 questions, et parmi eux 15 candidats ont suivi un entretien semi-structuré. Leurs domaines d'origine étaient assez variés avec une majorité de généticiens, de neuroscientifiques, et de

biologistes cellulaires, mais pas seulement. Le résultat de cette étude est très intéressant et cohérent avec les observations que nous avons effectuées dans le laboratoire LRI.

En effet, une difficulté avec la gestion des données scientifiques a été énoncée : les solutions de gestion existantes sont chères et complexes et les chercheurs ont besoin de solutions sur mesures compte tenu de l'évolution rapide des besoins en recherche. Ces solutions doivent respecter les budgets et le temps alloué à la gestion des données. Ce temps de RDM est supérieur à 10 heures par semaine pour 50% des chercheurs en protéomique et en biologie structurelle, et est, en général, de plus de 5 heures par semaine pour plus de 50% des participants.

En attendant la solution miracle, les chercheurs utilisent des applications logicielles tout publiques qui se caractérisent par leur courte courbe d'apprentissage, leur simplicité et leur abondance, telles que les tableurs, les fichiers textes et des outils simplissimes de partage de documents (via clé USB, un cloud grand public, un serveur local, etc.). De plus, chaque chercheur développe une méthode « maison » de gestion de ses données et qui est adaptée à son besoin immédiat sans se soucier du partage et de la réutilisation ultérieure des données. Dans ce contexte, des difficultés d'indexation, d'annotation, de stockage à long terme, d'extraction et d'organisation de données émergent ; avec le nombre croissant de fichiers à traiter, la taille grandissante des documents et les formats de plus en plus complexes et variés.

Les besoins énoncés dans (Anderson et al., 2007) concernent ainsi deux catégories principales : premièrement, les méthodes de gestion de larges jeux de données qui doivent –selon les répondants- être définies et standardisées à l'échelle du(des) laboratoire(s) et pas à une échelle individuelle, non utilisable ailleurs, et deuxièmement, les méthodes d'analyse et de traitement de ces données qui doivent être améliorées de point de vue ergonomie et interface utilisateur (UXDesign) afin de les inciter à les choisir, au dépend des solutions « faciles » ou « par défaut » décrites dans le paragraphe précédent.

En guise de solution, des répondants à l'enquête proposent une implication beaucoup plus engagée par les instituts de recherche qui doivent- selon eux- investir dans des services d'appui à la recherche, en informatique, bio-informatique et biostatistique, ainsi que des outils intuitifs, adaptés et simples à utiliser pour les laboratoires de recherche. Ceci afin de prendre en charge la composante financière et permettre un gain de temps pour le chercheur qui n'est pas censé passer son temps de recherche à gérer les problématiques d'accès et de traitement des données scientifiques.

À propos des risques, une réticence concernant la collaboration avec les experts en bases de données et sciences de données a été remarquée lors de l'enquête. Ceci s'explique par les contraintes de confidentialité et la culture protectrice fortement présente en recherche scientifique. Qui plus est, la spécificité des recherches et son caractère évolutif risquent de rendre les solutions proposées par les instituts et les experts en gestion de données rapidement obsolètes même si elles ont été conçues dans les règles de l'art expliquées auparavant (standards, adaptés, simples, etc.).

D'autres enquêtes ont été menées afin de caractériser l'état de la gestion de données de recherche (RDM) dans plusieurs domaines : (Poline et al., 2012) pour les données en neuroscience, (X. Chen & Wu, 2017) pour les données en chimie, (Houston et al., 2020) et (Krahe et al., 2020) pour les données cliniques. Ces enquêtes mènent toutes aux mêmes conclusions qu'il y a 13 ans : en plus des besoins en formation pour les chercheurs, et en sensibilisation pour les décideurs des instituts de recherche, une généralisation et standardisation d'outils et de méthodes doit pouvoir s'effectuer afin de commencer à mettre les fondations de ce domaine, i.e. la RDM. Un chemin long, mais indispensable. Ceci permettra une meilleure réutilisation et partage d'un chercheur à un autre. Et quant aux spécificités de chaque domaine de recherche, elle sera prise en compte en collaboration, étroite et continue, entre les data managers et les experts en recherche biomédicale.

Nous avons présenté des besoins et des observations en recherche biomédicale en général faites principalement par (Anderson et al., 2007). Il en résulte les éléments ci-après :

- Le besoin d'avoir un outil robuste et simple d'utilisation fourni par les instituts de recherche pour la gestion des données est omniprésent.
- Le besoin d'amélioration des interfaces utilisateurs pour une meilleure expérience utilisateur et une meilleure prise en main.
- Le besoin de standardisation des structures des données à l'échelle du laboratoire afin de faciliter leur partage et leur réutilisation au sein d'une même équipe et entre institutions partenaires.

Il est aussi important de rappeler à ce stade les observations suivantes tirées aussi de (Anderson et al., 2007) :

- Beaucoup de temps est consacré à la gestion de données de recherche par les chercheurs, considérée comme une tâche fastidieuse.
- Il y a une prolifération de solutions « maisons » qui ne sont pas pérennes et qui bloquent le partage et la réutilisation des données.
- La collaboration informaticien-biologiste doit être travaillée davantage.
- La prise en compte du caractère évolutif de la recherche est primordiale pour la pérennité de l'outil de gestion de données proposé.

Pour clôturer cette section sur les besoins en recherche préclinique, nous souhaitons mettre en lumière un article des années 80, qui n'a pas été identifié jusque-là comme un article de référence en gestion de données préclinique, et qui confirme que ces questionnements et ces besoins sont vraiment très ancrés dans le monde de la recherche préclinique avant même que le mot « préclinique » n'existe.

(Leaders et al., 1980) proposent de décrire les habitudes en gestion et en traitement des données d'un chercheur comme « tried and true » et défend la thèse que : « tant que les nouveaux outils ne présentent pas un intérêt majeur ou ne bénéficient pas d'une impulsion forte, un chercheur ne va pas laisser tomber ses méthodes « tried and true » ».

Ils évoquent aussi le fait qu'il n'y avait que le « Peer Reviewing » comme évaluation réelle du chercheur, et de ce qu'il a effectué tout au long de son étude de recherche, ce qui est bien très minimal comme validation. Aujourd'hui, il y a les comités d'éthique et les financeurs ou même les contrôles légaux, mais ce n'est qu'à un niveau réglementaire qui n'inclut pas, par exemple, la vérification systématique de l'intégrité scientifique des résultats.

Nous aimerions reprendre une citation de cet article qui nous a toujours parue triviale mais que nous n'avons pas réussi à la formuler ainsi :

« *If the biologists are to use a system efficiently, or indeed if they are to be convinced to use it at all, the software must mold the computer capabilities to the day to day operating requirements of the laboratory scientist-not the other way around.* » [43]

De plus, ils ont identifié un constat confirmé 25 ans plus tard par (Anderson et al., 2007) qui énonce que les informaticiens et les biologistes doivent collaborer ensemble dès le début du projet de développement logiciel pour la recherche afin de pouvoir fournir des outils utilisables.

[43] En français : « Pour que les biologistes puissent utiliser un système efficacement, ou même pour les convaincre de l'utiliser, le logiciel doit adapter les capacités de l'ordinateur aux besoins opérationnels quotidiens du scientifique de laboratoire - et non l'inverse »

« Computer scientists are trained to be as precise as a mathematician or a linguist in their thinking. Biologists, on the other hand, have been forced to accept inherent biological variation as a baseline way of life and consequently have allowed their thought processes to accept less precise limits. » [44]

Et pour finir, ils ont proposé une liste de critères qui doivent être présents dans un système d'acquisition, de gestion et d'annotation des données de recherche sur l'animal du point de vue du chercheur, à laquelle nous adhérons parfaitement, 40 ans plus tard :

- Le système informatique doit être « invisible » pour l'utilisateur : un maximum d'automatisation.
- Le système doit être guidé par le protocole : les étapes d'annotation des données via des formulaires doivent être adaptées aux étapes réelles du protocole à exécuter (acquisition TEP-TDM par exemple). Le système doit aussi guider le chercheur dans son travail, garantir la traçabilité étape par étape, et ainsi une meilleure qualité des notes à la fin de l'expérimentation. Ceci demande une flexibilité énorme.
- Le système doit assurer une traçabilité en arrière-plan (en mode log) des différents événements et saisie d'information pour générer des rapports en temps et en heure. Ces rapports doivent être compatibles avec les standards de référence et les recommandations gouvernementales.
- Les données dans le système doivent être sécurisées, et des niveaux de droits d'accès doivent être mis en place
    - Niveau 1 : Le personnel du labo – entrée de donnée uniquement
    - Niveau 2 : Les chargés de qualité de la recherche – accès aux données pour contrôle qualité, changement de données non autorisé
    - Niveau 3 : Statisticiens et autres chercheurs - accès aux données et possibilités de réorganisation et préparation des données pour l'analyse, mais changement de données non autorisé
    - Niveau 4 : Directeur de l'étude - accès aux données et possibilité de changer avec documentation du changement.
- Le système doit être flexible lors de l'import et l'export de données (nécessité d'avoir plusieurs formats en entrée et en sortie)
- Le système doit maximiser l'utilisation des capacités des ordinateurs (via l'automatisation et le Scripting).
    - Le contrôle des valeurs saisies par l'opérateur automatiquement. Par exemple, définir l'intervalle acceptable pour le poids de l'animal.
    - Le contrôle de la cohérence des données. Par exemple, si une souris meurt, elle doit rester morte.
    - Les calculs simples doivent être faits automatiquement. Par exemple, groupement des animaux selon leurs poids
    - La saisie des données doit pouvoir être possible partout à côté de tout équipement utilisé afin de limiter les erreurs.
    - La programmation des calendriers d'examen en optimisant les ressources pour l'équipe de recherche

[44] En français : « Les informaticiens sont formés pour être aussi précis qu'un mathématicien ou un linguiste dans leur réflexion. Les biologistes, en revanche, ont été contraints d'accepter la variation biologique inhérente comme mode de vie de base et ont donc permis à leurs processus de pensée d'accepter des limites moins précises. »

- La compilation d'un registre d'animaux depuis les données qui ont été fournies lors de l'annotation des données pendant et après les expérimentations.

II.1.1. SYNTHÈSE DES BESOINS

Dans cette section, nous avons présenté les différents besoins identifiés dans la bibliographie (pas très abondantes en recherche préclinique). Ci-après deux tableaux (Tableau 7 et Tableau 8) résumant les besoins et leviers identifiés en matière de gestion de données scientifiques hétérogènes. Pour chaque élément des deux tableaux, un mot et un numéro sont choisis pour le désigner. Une brève description est donnée pour chacun afin de dresser le profil du système « idéal » de gestion de données de recherche biomédicale.

Tableau 7 Liste des besoins en gestion de données de recherche

	BESOIN	DESCRIPTION
B1	archivage	Constituer un entrepôt de données hétérogènes pour l'archivage à long terme comme une mémoire du laboratoire
B2	import	Importer tout type de données (hétérogènes) utilisées dans un milieu de recherche
B3	export	Exporter les données selon un groupement particulier dans tout type de formats standard en recherche
B4	requête	Poser des « questions » par le chercheur, sous forme de requêtes à la base de données du système afin de tester et formuler de nouvelles hypothèses
B5	partage	Partager des données (y compris celles volumineuses) et collaborer en toute sécurité, simplicité et traçabilité
B6	analyse	Offrir un panel d'analyses (statistiques ou autres) et de traitement de données intéressantes et pertinentes pour le chercheur
B7	réutilisation	Permettre une réutilisation efficace des anciennes données de recherches du laboratoire
B8	traçabilité	Tracer automatiquement les événements liés aux données hétérogènes et à l'activité de recherche (log) ainsi que leur provenance
B9	simplicité	Pour un scénario d'utilisation donné, les étapes de son exécution doivent être les plus minimales et simples possibles
B10	automatisation	Pour toute tâche de gestion de données fastidieuse, une automatisation, ou à défaut une semi-automatisation, doit être effectuée
B11	standardisation	Les données doivent être standardisées au sein du laboratoire afin de permettre une compréhension mutuelle entre les collaborateurs
B12	ergonomie	Les interfaces graphiques, et toute utilisation des outils de gestion de données doivent être intuitives et simples d'utilisation
B13	efficacité	Les interactions avec le système doivent toutes aboutir à un résultat pertinent pour le chercheur compte tenu du temps investi
B14	évolutivité	Le système doit pouvoir être évolué par l'administrateur en fonction des besoins émergeant de la recherche scientifique et du laboratoire
B15	flexibilité	Exécuter des scénarios spécifiques différents de ceux par défaut du système, mais qui leur ressemble
B16	vérification	Assurer un contrôle qualité des données manuellement saisies par l'utilisateur et une vérification de l'intégrité et la cohérence des données
B17	reporting	Générer des rapports de traçabilité des activités de recherche pour le cahier de laboratoire ou des contrôles d'intégrité scientifique. Générer des rapports de mise en forme des données pour des besoins de conformités aux standards du domaine
B18	sécurité	Garantir une sécurisation des données contre les incidents techniques et les cyberattaques. Garantir aux chercheurs que leurs données confidentielles ne seront jamais visibles aux autres utilisateurs du système et définir des niveaux de droits d'accès spécifiques à tout projet de recherche

B19	suivi	Suivre dans le temps les activités de recherche sur tout le cycle de vie d'une étude

Tableau 8 Liste des leviers pour la gestion de données de recherche

	LEVIER	DESCRIPTION
L1	collaboration bio-info	La collaboration entre biologistes, imageurs et autres disciplines de recherche et les ingénieurs-informaticiens, doit être étroite dès le début de la conception de la solution logicielle de gestion de données
L2	changement	La prise en compte des changements au cours d'une étude est primordiale dès le début de la mise en place de la solution logicielle
L3	adaptation	Même avec prise en compte des changements lors de la conception d'une solution logicielle, son adaptation à la réalité terrain est nécessaire. En recherche, il y a beaucoup de « bricolage » puisque les standards des domaines n'évoluent pas rapidement. Le système doit pouvoir s'adapter aux nouveautés tout en respectant les standards
L4	mirroring	Les saisies d'informations dans le système doivent suivre le workflow de recherche (protocole d'acquisition ou d'analyse) pour permettre une meilleure utilisation du système
L5	intérêt	Toute utilisation du système par un chercheur doit être liée à un intérêt majeur pour son étude de recherche

Les besoins et les leviers expliqués ci-dessus (Tableau 7 et Tableau 8) résument les résultats de nos recherches bibliographiques et nos observations terrain. Nous allons les utiliser tout le long de ce manuscrit afin d'analyser les systèmes existants et valider nos propositions.

Dans la section suivante, nous nous intéressons aux systèmes de gestion de données proposés en recherche biomédicale et préclinique. Nous allons les présenter en identifiant leurs apports et leurs manques.

II.2. SYSTÈMES DE GESTION DE DONNÉES EN RECHERCHE BIOMÉDICALE

Dans cette section, nous présentons des systèmes existants identifiés dans la bibliographie. Ils s'intéressent à la gestion de données de recherche biomédicale préclinique et clinique (en imagerie, en biologie, en suivi journalier). Nous effectuons enfin un bilan de cet état des lieux et des potentielles voies d'amélioration.

II.2.1. DES SYSTÈMES POUR CHAQUE DOMAINE D'ACQUISITION DE DONNÉES

Considérant que les données les plus populaires en recherche biomédicale sont les analyses biologiques, les examens d'imagerie et les rapports de soin, il devient trivial que des systèmes spécifiques pour chaque type de donnée voient le jour.

II.2.1.1. Les systèmes en radiologie, imagerie

L'initiative la plus ancienne en gestion de données numériques de radiologie que nous avons pu identifier est celle de (Fishman et al., 1991). Dans une époque où la RAM était de 2Mo, celle du disque dur de 40 Mo, et les programmes informatiques se partageaient via une disquette, une attention particulière a été portée à la collecte des données des examens des patients en imagerie TDM (des cas uniques, ou intéressants pour la recherche, ou pour le suivi dans le temps), afin de constituer une base

de données consultable et utilisable facilement. Les requêtes à l'époque étaient déjà préconfigurées. (Fishman et al., 1991) est une référence bibliographique précieuse qui liste des besoins en radiologie vis-à-vis de l'outil informatique et qui sont encore d'actualité : La facilité d'utilisation, la facilité d'accès à l'information utile, la proximité physique maximale avec le radiologue (utilisable là où il le veut quand il le veut), l'adaptabilité de la base de données aux nouveaux besoins et demandes, la rapidité pour pouvoir libérer le chercheur afin qu'il se focalise sur ses travaux à forte plus-value. Ils ont développé un outil de requête simple pour les données en imagerie TDM et un outil de saisie présent en Figure 42.

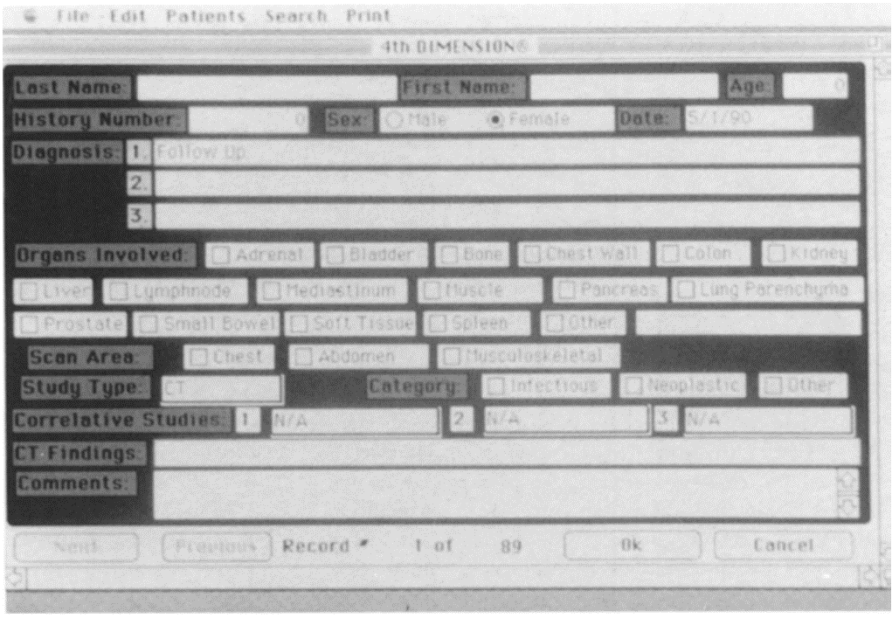


Figure 42 Plus ancienne interface de saisie d'informations identifiée en radiologie (Fishman et al., 1991)

De nos jours, le premier système de référence en radiologie est le PACS (Picture Archiving and Communication System) qui est étroitement lié au standard DICOM. Les premiers travaux identifiés remontent au début du deuxième millénaire (Strickland, 2000) peu après le passage en radiologie du film conventionnel à l'image digitale. La révolution numérique a permis d'avoir des systèmes PACS avancés et qui ont vite adopté le standard DICOM (Clunie, 2000) (Mildenberger et al., 2002). Un système PACS permet de centraliser les données d'un centre de radiologie ou un laboratoire de recherche en imagerie. Le standard DICOM largement adopté, confère au système PACS une force de communication et d'intégration avec les machines d'acquisition et aussi les serveurs de stockage de données. Les données acquises sont instantanément disponibles pour leur visualisation depuis le système PACS. L'archivage à moyen et à long terme, la compression des images, leur annotation automatique par des métadonnées DICOM et plusieurs autres fonctionnalités de base (de type COTS[45]) sont tous disponibles via un système PACS.

Le deuxième système de référence en radiologie est le RIS (Radiology Information System). Il est centré sur les données administratives de gestion d'examens et des résultats d'analyse en radiologie, tandis que le système PACS est centré sur l'image (Hecht, 2008).

La Figure 43 présente une architecture typique dans un centre d'imagerie médicale de nos jours où le système PACS et le RIS sont interconnectés aux différentes modalités d'imagerie (Maxhelaku & Kika, 2020). Les autres éléments de l'architecture sont expliqués plus tard dans cette section.

[45] Commercial off-the-shelf : Désigne tout produit informatique fabriqué en série avec une liste de fonctionnalités génériques, standardisées et prédéfinies par avance.

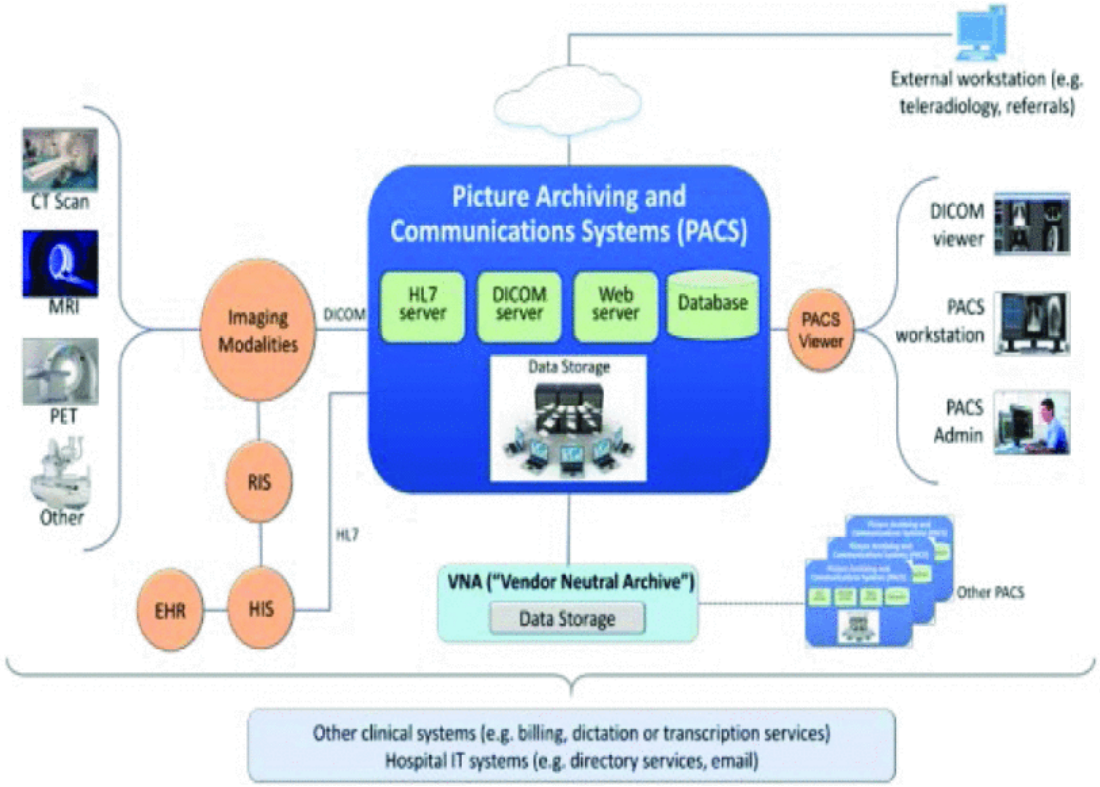


Figure 43 Architecture typique d'utilisation d'un système PACS dans un centre hospitalier (Maxhelaku & Kika, 2020).

La frontière entre la recherche et les pratiques cliniques en milieu hospitalo-universitaire est fine. Par conséquent, un RIS ou un PACS peuvent être utilisés pour stocker et gérer des données de recherche clinique en sus des données d'examens patients ou de prestations de soins.

Un chercheur en imagerie clinique utilise différents systèmes d'informations et de gestion de données pour accéder à l'information pertinente. (Taira et al., 1996) propose l'architecture en Figure 44 qui repose sur une couche « middleware » afin de lui donner accès d'une manière transparente et unifiée, aux données hétérogènes d'imagerie et de soins patient. Ces données sont extraites des systèmes HIS (Hospital Information System), RIS et PACS.

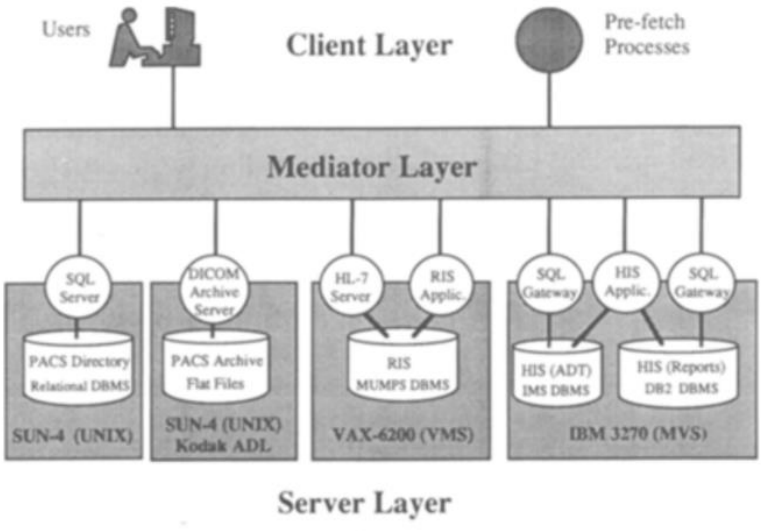


Figure 44 Architecture client-middleware-serveur de l'intégration entre systèmes HIS-RIS-PACS (Taira et al., 1996)

Depuis l'architecture client-middleware-serveur de (Taira et al., 1996) , l'intégration HIS-RIS-PACS a été aussi proposée par : (Mansoori et al., 2012) (Forsberg et al., 2016) (Mongan & Avrin, 2018) (Maxhelaku & Kika, 2020). L'émergence d'autres systèmes l'ont aussi favorisée (Figure 43). Nous pouvons citer l'EHR (Electronic Health Record) et l'EMR (Electronic Medical Record), deux termes qui désignent (à quelques différences près) le dossier patient électronique et qui seront explicité dans la section II.II.2.1.3.

Même à un niveau plus restreint, des hétérogénéités ont été observées entre constructeurs des systèmes PACS, ce qui a alimenté l'initiative d'avoir un système PACS indépendant du constructeur et 100% conforme au DICOM appelé Vendor Neutral Archive (VNA) (Agarwal & Sanjeev, 2012).

Force est de constater que les systèmes les plus connus en imagerie et en radiologie ne sont ni des systèmes purement axés sur la recherche biomédicale ni des systèmes qui prennent en compte le domaine préclinique et l'imagerie du petit animal. En effet, les initiatives dans ce sens et spécifiquement pour la recherche préclinique sont très limitées. Elles font l'objet d'une présentation détaillée à la section II.II.2.2.2.

(Camarasu-Pop et al., 2010) a présenté un ensemble de huit projets qui avaient comme objectif de gérer, à la fois, des données images en recherche clinique et en recherche préclinique. Les auteurs ont souligné la rareté de ce genre d'approche où l'on souhaite lever les barrières entre l'hôpital et le laboratoire, et aussi entre les différents sites et services.

Les systèmes comparés sont :

1) **MammoGrid** pour la recherche clinique en mammographie.

2) **Trencadis** qui entreprend une approche multisite pour le partage des données DICOM entre plusieurs hôpitaux.

3) **Globus Medicus** et 4) **MDM**, qui ressemblent plutôt à un entrepôt de données DICOM (data warehouse) dé-identifiées afin de permettre un stockage et une recherche sécurisés dans une banque de données.

5) **MAGIC-5** et 6) **MediGrid** assurent un traitement des données dé-identifiées et pas que DICOM.

7) **The Virtual Data Grid** est un catalogue d'images (pas que DICOM) qui permet de faciliter le partage des données entre centres de recherche.

8) Le **BIRN** (The Biomedical Informatics Research Network) est une initiative des États-Unis d'Amérique depuis 2008 pour le partage de données entre sites de recherche biomédicale. Les données partagées (DICOM, NIFTI et autres) sont sécurisées en utilisant des accès restreints et un processus de dé-identification. Il s'agit d'un dispositif puissant et permettant en plus du partage sécurisé, d'accéder à des ressources de calcul intéressantes.

Les systèmes étudiés ont été évalués selon quatre critères indispensables pour les chercheurs :

1. La sécurité des données cliniques.
2. L'interopérabilité entre les différents formats d'images.
3. Le partage sécurisé entre l'hôpital et le laboratoire de recherche.
4. L'accès sécurisé et facilité à des grilles de calcul pour l'analyse des données.

Des freins à l'appui de la recherche ont été identifiés, notamment, ceux posés par la sécurisation des données patient et qui limitent le partage interinstitution. Est considéré un frein aussi, l'accès limité à des ressources de calculs avancés. Ce dernier est toujours d'actualité malgré des progrès énormes dans

le domaine du calcul scientifique et le déploiement d'infrastructures de calcul pour la recherche partout en France et dans le monde ces dernières années.

Et pour finir, une conclusion particulièrement intéressante a été reportée concernant les approches identifiées pour faire face à l'hétérogénéité des données : les auteurs ont identifié deux approches en imagerie : Premièrement, choisir une architecture du système basé sur DICOM, le standard par défaut en imagerie, et « dicomiser » les autres images ayant d'autres formats. Cette approche a l'avantage d'être plus acceptable par les radiologistes, cliniciens et chercheurs qui sont habitués à utiliser du DICOM. Cependant, la « dicomisation » n'est pas toujours possible. La deuxième approche est de construire un système générique et de développer pour chaque type de données d'intérêt un driver/connecteur spécifique, comme l'initiative BIRN par exemple. Nous retenons l'approche générique pour la suite de nos travaux.

II.2.1.2. Les systèmes en biologie, histologie

La plateforme de référence en biologie est le LIMS (Laboratory Information Management System) apparu dans les années 80s (Gibbon, 1996). Il est le système qui gère le catalogue des données et des échantillons biologiques dans un laboratoire de recherche ou une plateforme de services en biologie. Un système LIMS permet la traçabilité des lots de produits utilisés en laboratoire, la gestion des résultats d'analyses biologiques, la traçabilité des protocoles et des échantillons, la programmation des analyses, la gestion des coûts et des factures, et le contrôle qualité[46]. Un LIMS est configurable selon les besoins du laboratoire, il peut intégrer des workflows d'analyse scientifique et peut aussi s'interfacer avec d'autres systèmes d'informations si besoin. Plusieurs LIMS commerciaux sont proposés sur le marché dont on peut citer TEEXMA LIMS de Basseti Group[47] ou le LIMS de Thermofisher Sientific[48]. Par ailleurs, les laboratoires d'analyses peuvent choisir de développer un LIMS « maison », spécifiquement adapté à leurs besoins. Par exemple, (Helsens et al., 2010) proposent un LIMS « maison » pour gérer les données issues de la spectrométrie de masse (MS) en protéomique : Les données brutes, ou spectres, sont stockées et analysées en utilisant un workflow automatique défini selon les besoins spécifiques des utilisateurs. De plus, les données dérivées et les données brutes sont archivées et ainsi peuvent être retrouvées via le LIMS.

Un système PACS (voir section II.2.1.1) peut potentiellement être utilisé en bio-imagerie. En effet, les données en biologie peuvent être sous format image ; en anatomopathologie par exemple, des échantillons peuvent être photographiés pour la demande d'un deuxième avis ou pour l'archivage numérique. En histologie, des lames entières sont scannées plusieurs fois à chaque nouvelle coloration ou nouveau marquage. Pour pouvoir stocker ces images proprement, le standard DICOM en radiologie a été repris dans (R. Singh et al., 2011) pour la pathologie numérique ainsi il est désormais possible d'utiliser un système PACS pour la gestion des données en biologie.

Plus récemment, le framework OMERO (Allan et al., 2012) désigne « Open Microscopy Environment Remote Objects » et a été proposé pour unifier les formats et les procédures autour de la donnée biologique. Il s'agit d'un système de gestion de données de microscopie, en biologie expérimentale et en bio-imagerie. Il se base sur le standard OME (OME-XML et OME-TIFF) et supporte plusieurs (140 ou plus) formats de données grâce à sa bibliothèque d'outils logiciels Bio-Format. Il est maintenu par l'Université de Dundee en Écosse et possède une grande communauté d'utilisateurs open source et plusieurs organismes financeurs (universités, centres de recherche, industriels). OMERO offre plusieurs fonctionnalités aux chercheurs pour les données de microscopes telles que : la visualisation, l'analyse, le partage, la gestion des droits d'accès.

[46] https://www.gazettelabo.fr/archives/pratic/1998/25LIMS.htm Février 1998 - n°25
[47] https://www.bassetti-group.com/logiciel-lims-fr/
[48] https://www.thermofisher.com/fr/fr/home/digital-solutions/lab-informatics/lab-information-management-systems-lims.html

Le quatrième système considéré est Cytomine (Marée et al., 2016). Il s'agit d'un outil de la même famille que OMERO avec plus de fonctionnalités en visualisation et analyse de bio-images. Il permet de gérer et partager les données brutes ainsi que leurs analyses. Il offre une gestion de projet et permet d'annoter et enrichir les bio-images. Il est un outil orienté sur l'analyse d'image (voir Figure 45)

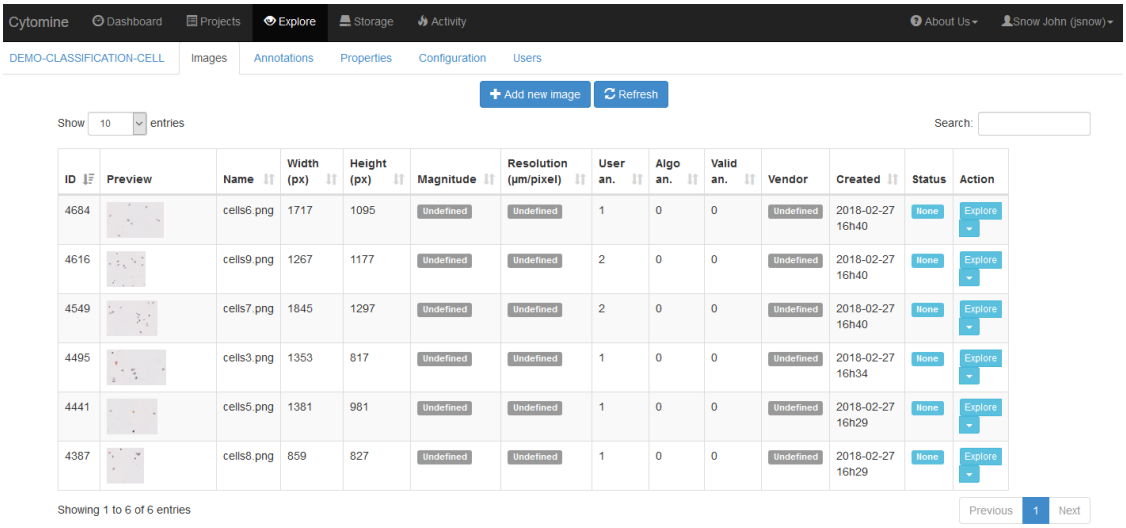


Figure 45 Exploration d'une liste de bio-image dans Cytomine
capture d'écran du site http://pf-01.lab.parisdescartes.fr/#tabs-images-4327

Image Data Ressource (IDR) (Williams et al., 2017), est une plateforme d'intégration et de publication des données images in vitro (voir Figure 46). Il constitue un grand atlas de données publiées en « open access » pour réutilisation. Il contient des données de microscope et de scanners de lames pour des acquisitions sur tissu biologique en histopathologie ou histologie, et sur culture cellulaire en 96 puits. IDR offre un large éventail d'images en biologie, que l'on peut visualiser, annoter et utiliser afin de lancer des scripts de traitement. IDR utilise OMERO comme standard d'échange de données et est en accès public.

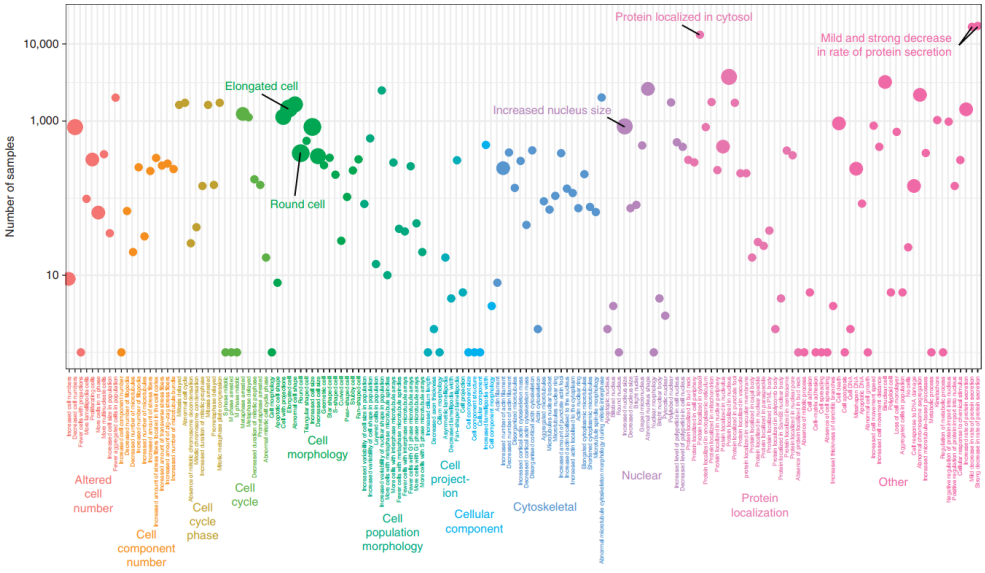


Figure 46 Exploration de la liste des phénotypes issues des échantillons biologiques de la base de données IDR

En biologie, les systèmes identifiés présentent plusieurs points communs notamment en matière de composants : stockage, analyse, visualisation des images. Le LIMS historique est le plus sophistiqué des systèmes présentés puisqu'il gère des données de facturation et de gestion de temps en plus des données hétérogènes d'analyse biologique. Néanmoins, la séparation des systèmes proposés par type de

données est toujours présente. Aussi, la portée de chacun diffère ; il y a ceux open source et utilisable ponctuellement pour des analyses en ligne (Cytomine, IDR, etc.) ou ceux à utilisation interne à chaque équipe et ayant vocation à gérer toutes les données d'un laboratoire (le LIMS, OMERO, PACS via l'extension du standard DICOM à la biologie, etc.)

II.2.1.3. Les systèmes en suivi de soin et de la recherche

La troisième catégorie de données biomédicales largement répandue est le rapport textuel (ou narratif) de soins ou de suivi longitudinal en essais cliniques (questionnaires, observations, mesures, etc.). En recherche préclinique, il s'agit, par exemple, les fiches de suivi journalier des animaux (hébergement, température, poids, état de progression d'un processus biologique, dose administrée d'un produit, etc..), mais aussi, les notes des chercheurs tous les jours dans le cahier de laboratoire électronique ou papier.

Dans ce paragraphe, nous présentons tout d'abord les systèmes de gestion de données patients tels que le dossier de santé électronique (EHR) et le dossier médical électronique (EMR). Ensuite, nous présentons le cahier de laboratoire, seul outil de référence et d'une valeur législative du chercheur par excellence.

Un EHR ou un EMR, deux termes interchangeables par abus d'utilisation, est une collection de données patient qui, jadis, était sous forme de fiches en papier, et qui de nos jours, est devenue numérique. Les fonctionnalités de base d'un EHR sont décrites dans (Jha et al., 2009). Il présente trois ensembles de fonctionnalités allant des plus basiques aux plus étendues. Les fonctionnalités de base sont : les informations sur le patient, l'historique des problèmes rencontrés et des médicaments pris, les rapports de radiologie et d'analyse biologique effectués ainsi que les résultats du diagnostic de routine. D'autres fonctionnalités « étendues » peuvent s'ajouter à la liste comme : la gestion des notes des médecins et des soignants, le stockage des images acquises en plus des rapports d'examens, et des fonctionnalités d'aide à la décision (des recommandations cliniques et des alertes d'allergies à un médicament ou de surdosage par exemple).

En réalité, il y a une différence entre un EMR et un EHR (Garets & Davis, 2006). Un EMR est généralement connu comme étant la version locale et basique de la gestion des données patient par un clinicien, tandis que le EHR est celui avec plus de fonctionnalités et qui est centré sur le parcours de soins du patient.

La centralisation des données patient a conduit à des bases de données riches et intéressantes pour la recherche clinique et à l'identification des cas pertinents pour les essais cliniques. La forte présence des textes donne une opportunité exceptionnelle pour le Data Mining et le Traitement Automatique de Langue (TAL). Ces techniques sont exploitées pour la découverte de phénotypes pour des recherches cliniques : caractérisation des symptômes liés à des maladies rares, identification des réactions de patient à des prises de médicaments, etc. Par exemple, (Jonnalagadda et al., 2017) propose un algorithme en TAL qui permet d'analyser un EHR et d'identifier, avec une sensibilité de valeur égale à 0.95, les patients à recruter pour l'essai clinique « PARAGON » (qui étudie l'insuffisance cardiaque avec fraction d'éjection préservée).

Les données textuelles en recherche biomédicale ne sont pas limitées aux données des rapports patient en recherche clinique. Les chercheurs détiennent aussi un cahier de laboratoire papier nominatif et d'une valeur légale. Le cahier de laboratoire est un outil assez ancien et assez commun à tous les domaines. La plus ancienne référence, que nous avons pu repérer dans la bibliographie, au cahier de laboratoire papier est celle des cahiers de Marie Sklodowska-Curie et de Pierre Curie datant de 1898 (Adloff, 1999).

En France, toute personne faisant de la recherche doit tenir un cahier de laboratoire papier ou, depuis peu, électronique. Le cahier est un registre des activités quotidiennes du chercheur : expérimentations, hypothèses, résultats, et toutes activités et informations en lien avec les projets de recherche en cours et à venir. Chaque note doit être datée, signée et montrée à une personne dite « témoin » qui doit aussi

dater et signer. L'ensemble des cahiers de laboratoire d'une équipe constitue la mémoire des expérimentations réalisées en son sein. Il s'agit d'une traçabilité de la recherche.

Les premières initiatives de Cahier de Laboratoire Électronique (CLÉ) datent de 1990 (Butler, 2005). Dans sa version basique, il s'agit d'une transcription du cahier de laboratoire papier. Le but premier d'un cahier de laboratoire électronique ou papier est d'assurer la traçabilité des expérimentations en cours et d'améliorer la reproductibilité de la recherche. Le CLÉ a évolué, dans une version enrichie, vers un système de gestion des données quotidiennes de la recherche avec plusieurs fonctionnalités telles que : le partage du cahier inter et intra études de recherche, la gestion des équipements et des inventaires, l'intégration avec les autres outils et logicielles utilisés dans le laboratoire, la gestion des agendas et planification des expérimentations, la gestion électronique de documents (GED), la collecte, analyse, stockage et archivage de données (Nelson et al., 2011). La frontière entre un CLÉ et un LIMS est de plus en plus floue puisque le CLÉ, jadis individuel, est désormais partagé au sein de l'équipe de recherche (Tabard et al., 2008) et, donc, permet d'y intégrer facilement les fonctionnalités d'un LIMS. L'apport du CLÉ par rapport à un LIMS standard se définit dans la dimension de traçabilité et de suivi journalier sur tout le cycle de vie d'une étude de recherche. La différence entre les deux systèmes réside dans la portée de l'information : individuelle puis collective, et évolutive dans le temps pour le CLÉ ; collective, administrative, partagée et statique pour le LIMS.

L'INSERM (Institut National de la Santé et de Recherche Médicale) a introduit officiellement le cahier de laboratoire électronique en 2018[49], il a depuis une valeur légale. Il s'agit d'une solution industrielle qui a été adaptée aux besoins de l'INSERM, elle s'appelle LABGURU[50]. Elle est principalement dédiée à la biologie et la biochimie. Effectivement, il est à noter qu'un cahier de laboratoire électronique dans sa version COTS[51] n'est pas adapté à la recherche en imagerie et nécessite un remodelage pour pouvoir intégrer les grands volumes de données d'imagerie et le vocabulaire qui y est lié.

Comme pour le EHR, un cahier de laboratoire électronique offre la possibilité d'effectuer des recherches avancées dans tout CLÉ appartenant au labo, ce qui en fait un outil puissant et une mine d'informations. Des outils d'extraction sémantique peuvent être mis en place pour faciliter la recherche de protocoles ou de données par les chercheurs. Par exemple, (Dianat et al., 2013) ont utilisé l'analyse sémantique de texte présent dans les CLÉs, pour proposer des ressources bibliographiques issues de PubMed aux chercheurs.

En conclusion, les EHRs, CLÉs ou tout autre système de reporting scientifique sont cruciaux pour la traçabilité et la reproductibilité de la recherche en général et de la recherche biomédicale en particulier. Ces systèmes doivent être intégrés avec ceux déjà présents et utilisés dans l'environnement des chercheurs et doivent pouvoir fournir le service adéquat au bon moment et qui s'adapte aux besoins et à la spécificité des membres d'un laboratoire. Une fois construites, ces notes sont une mine d'information importante pour toute recherche avancée et outils d'aide à la décision.

II.2.2. DES SYSTÈMES POUR CHAQUE PHASE DE LA RECHERCHE TRANSLATIONNELLE

Lors de l'analyse bibliographique des systèmes de gestion de données existants, nous avons constaté que hormis la séparation biologie/imagerie/texte, il y a aussi un clivage clinique/préclinique. Nous évoquons, tout d'abord, des systèmes proposés en recherche clinique et les types de données qu'ils gèrent, nous présentons après ceux proposés pour la recherche préclinique.

[49] https://cle.inserm.fr
[50] https://www.labguru.com/
[51] Commercial Off-The-Shelf : Désigne tout produit informatique fabriqué en série avec une liste de fonctionnalités génériques, standardisées et prédéfinies par avance.

II.2.2.1. Des systèmes en recherche clinique

La recherche clinique est vaste. Plusieurs disciplines médicales (oncologie, neurologie, urologie, gastrologie, cardiologue, pneumologue, génécologie, chirurgie ...etc.) et pathologies (cancer, alzheimer, mélanome, infarctus, etc.) entrent en jeu. Par conséquent, les systèmes identifiés dans la bibliographie traitent forcément un sous-ensemble des données en recherche biomédicale clinique.

Pour commencer, l'entrepôt de données cliniques (Clinical Data Warehouse ; CDW) est un système assez répondu, depuis les années 90s, dans les hôpitaux (La référence la plus ancienne trouvée sur Semantic Scholar est celle de (Prather et al., 1997)). Il s'agit en général d'un stockage centralisé des données patient afin de mieux les requêter et les organiser.

Le CDW ouvre la voie vers des analyses avancées et qui sont impossibles autrement. Plus récemment, au centre médical de recherche Vanderbilt (Nachville, US), une architecture d'un entrepôt de données cliniques pour leur réutilisation en recherche a été proposée et déployée (Danciu et al., 2014). L'entrepôt est composé de plusieurs éléments (voir Figure 47) : A) la partie clinique où il y a la collecte des données à stocker dans l'entrepôt. F) la partie recherche où les outils d'analyses de données sont présentés. Et entre les deux, le *Research Data Warehouse* (C+B et D) où les données sont dé-identifiées, quand

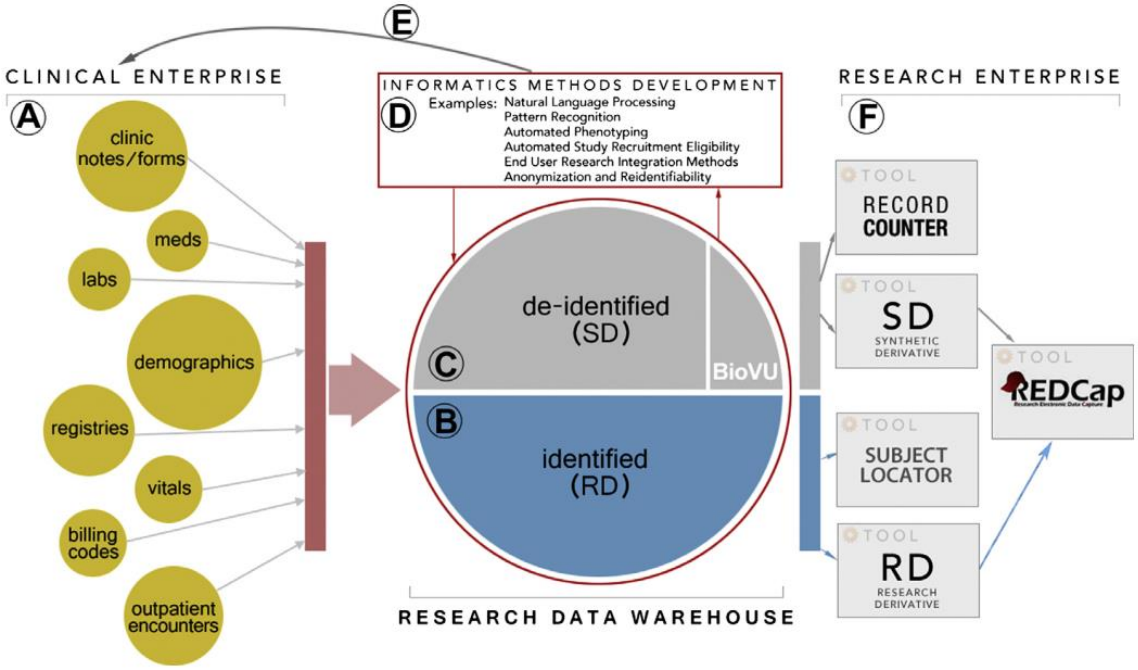


Fig. 1. Vanderbilt clinical and research enterprise.

Figure 47 Architecture de l'entrepôt de données au centre médical Vanderbilt (Nachville, US)

En France, nous avons identifié les entrepôts de données de santé de l'APHP [52], le eHOP du CHU de Rennes (Madec et al., 2019), le clinical data warehouse de l'HEGP (Jannot et al., 2017) (Zapletal et al., 2010), le Dr Warehouse (Garcelon et al., 2018), Semantic Health Data Warehouse de l'université de Rouen (Lelong et al., 2019) et l'historique SNIIRAM-SNDS[53] ainsi que le tout récent Health Data Hub[54] à l'échelle nationale.

La Figure 47 montre l'architecture du CDW de l'université de Rouen. Les données proviennent du système d'information Hospialier (HIS) et sont transformées via une couche intermédiaire pour enfin

[52] https://eds.aphp.fr/eds
[53] https://www.snds.gouv.fr/SNDS/Composantes-du-SNDS
[54] https://www.health-data-hub.fr/

être stockées dans une base de données après dé-identification. Un moteur de terminologies et de traitement sémantique est présent entre les applications de navigation dans les données et la base de données en elle-même. Il se base sur HeTOP[55], une base de terminologies et ontologies de santé développée par le CHU de Rouen.

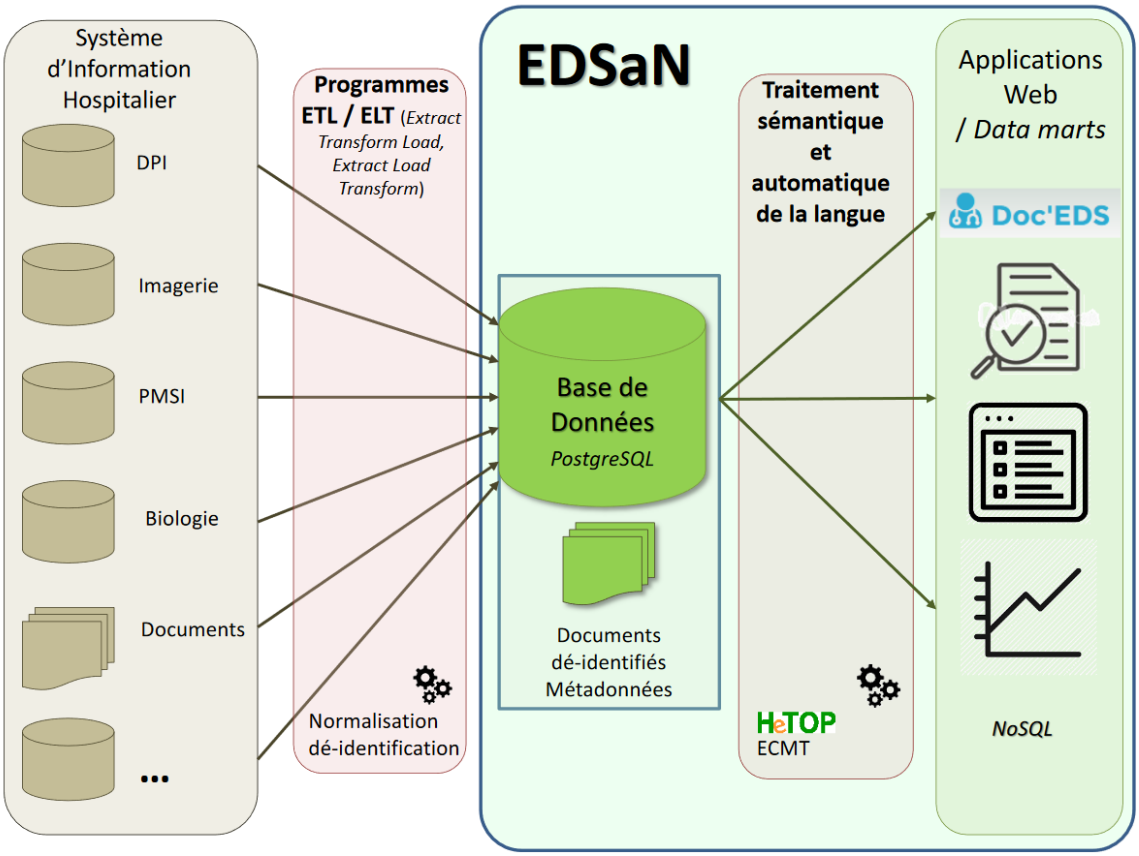


Figure 48 Architecture de l'entrepôt de données de santé de l'université de Rouen
Source : http://www.chu-rouen.fr/cismef/wp/wp-content/uploads/2020/03/EDS%20Rouen%20janvier%202020%20pour%20GSM3.pdf

Nous pouvons aussi citer le clinical data warehouse (Jannot et al., 2017) (Zapletal et al., 2010) qui a été élaboré à l'Hôpital Européen Georges Pompidou (HEGP) à Paris. Il se base sur l'architecture et le modèle en étoile d'i2b2 (PATIENT, PROVIDER, VISIT, CONCEPT et OBSERVATION) (Murphy et al., 2007) tout comme CARPEM (Rance et al., 2016), qui est aussi un entrepôt de données cliniques français qui s'intéresse en particulier aux données de la recherche translationnelle en oncologie dans le cadre du projet CARPEM. Ce CDW est construit en se basant sur i2b2 et Transmart. Il permet de gérer des données d'imagerie de cancer, des tableaux Excel, et des données omiques. À Paris également, le Dr Warehouse (Garcelon et al., 2018) a été proposé. Il est orienté sur les rapports narratifs de soins. Il utilise les technologies de Traitement Automatique de Langue (TAL) et les algorithmes de ressemblance et de détection de la négation. Il permet ainsi une recherche par texte libre dans les documents présents à l'entrepôt des données de l'Hôpital Necker à Paris.

(Gagalova et al., 2020) proposent une analyse pseudosystémique de la bibliographie depuis 2008 et compare les CDW, appelés aussi IDR (Integrated Data Repositories), pour la gestion des données cliniques. Les entrepôts de données comparés se sont avérés principalement dépendants des sources de données, de l'utilisateur final et de l'étendue de l'équipe (interne, multisites, disciplinaire, etc.). La

[55] https://www.hetop.eu/hetop/

Figure 49 ci-après présente les quatre modèles d'architectures identifiés après la revue de plus de 34 articles de référence par (Gagalova et al., 2020).

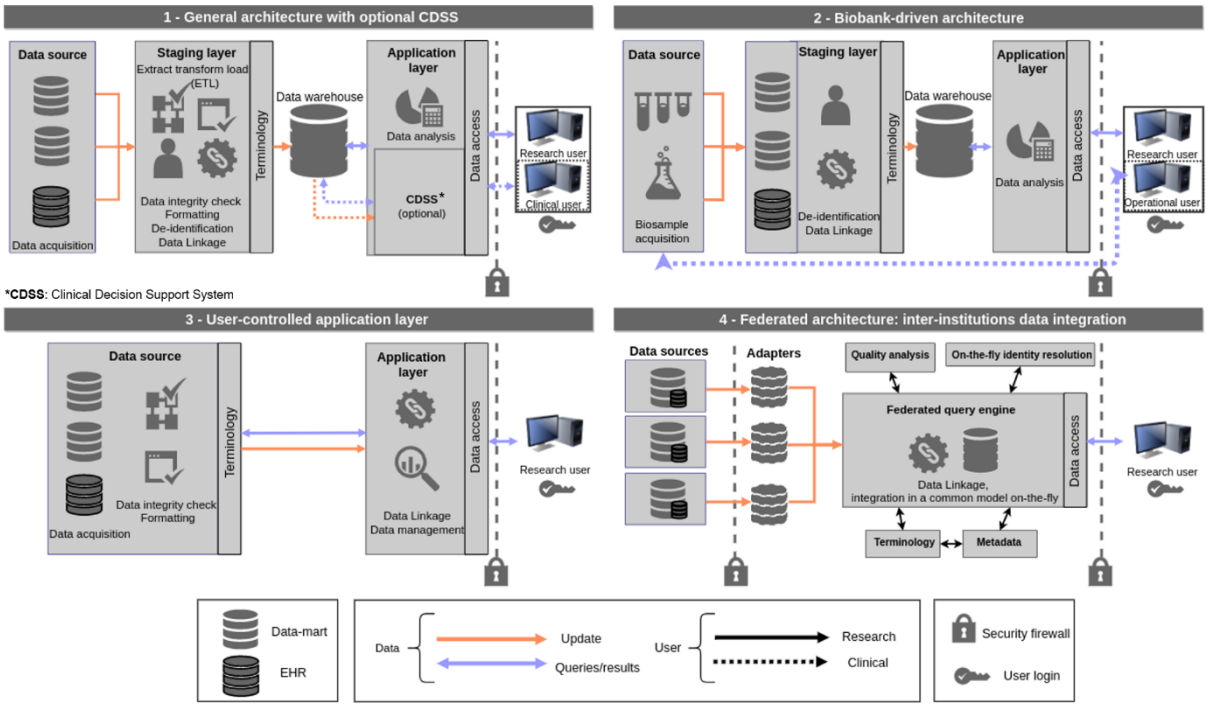


Figure 49 Liste des architectures identifiées pour la gestion des données cliniques hétérogènes (Gagalova et al., 2020)

Dans la Figure 49, les architectures « 1-Générale » et « 3-Utilisateur » sont assez semblables de point de vue « données d'entrée », mais l'architecture 3 est plus minimaliste et est régie par les demandes des utilisateurs et des chercheurs, tandis que l'architecture 1 est plus robuste avec la centralisation des données dans un entrepôt intermédiaire (Data warehouse) qui permet des utilisations plus avancées. L'architecture « 2-Biobank » est axée sur les données biologiques d'analyse et l'architecture « 4-Fédérée » est une architecture axée sur la recherche multisite.

Par ailleurs, (Deserno et al., 2014) propose un système Rare Disease Registry (RDR) pour le stockage et la gestion des données des maladies rares pour la recherche clinique, afin de permettre une meilleure compréhension de ces maladies et un meilleur diagnostic pour le patient. Le système permet une analyse intégrée des images et des signaux. La singularité du RDR est le fait d'être spécialisé dans les maladies rares et d'être adapté aux besoins déjà identifiés dans le domaine par (Bellgard et al., 2013). Selon (Deserno et al., 2014), ces besoins ne peuvent pas être satisfaits par des systèmes axés sur les essais cliniques (clinical trials) et les formulaires de collecte de cas clinique eCRF (electronic Case Report Form), comme, OpenClinica[56] (Cavelaars et al., 2015) et REDCap[57] (Harris et al., 2009).

Archimed (Micard et al., 2016) est un système de gestion des données imagerie en recherche clinique pour le stockage, la visualisation, le contrôle qualité et le traitement des images. Il présente des fonctionnalités plus avancées qu'un système PACS standard. Il est multisite, multiformat, multiorganes étudiés, et centré sur l'étude de recherche. D'autres systèmes existent pour la gestion des données en imagerie pour la recherche clinique tels que Shanoir (Barillot et al., 2016) et CATI (Operto et al., 2016), mais ils sont centrés sur la neuroimagerie.

La liste des systèmes gérant des données cliniques pour la recherche est longue, nous avons présenté des exemples typiques dans ce paragraphe. En effet, les systèmes existants diffèrent principalement par type de données (imagerie, analyse biologique, rapports cliniques, etc.) ou de discipline (cancérologie,

[56] https://www.openclinica.com/
[57] https://www.project-redcap.org/

neurologie, maladies rares, etc.) ou par appartenance à une institution (HEGP, Vanderbilt, etc.) ou un projet (CARPEM, etc.). Pourtant, ils ont tous les mêmes modules (plus ou moins) qui se répètent et qui sont en lien étroit avec des étapes du cycle de vie des données : collecter, traiter, stocker, analyser, partager et réutiliser ; ainsi que la sécurisation des données et le respect de la vie privée des patients.

II.2.2.2. Les systèmes en recherche préclinique

En recherche préclinique, les ressources bibliographiques sont beaucoup plus limitées qu'en recherche clinique. Les premiers efforts que nous avons pu identifier, hormis les recherches sur les données en zoologie et en médecine vétérinaire, traitent plutôt des sujets annexes à la recherche préclinique. Notamment, (Frank et al., 1991) proposent le LAMS (Laboratory Animal Management System) pour la gestion des données d'hébergement des animaux et de leurs expérimentations et (Bruns et al., 1993) propose un LIMS pour le suivi des données de la production des animaux de laboratoire. La première référence que nous avons pu identifier est celle de (Li & Banks, 2006) où ils utilisent le terme « développement préclinique » qui réduit la recherche préclinique à la recherche en biochimie. Dans leurs travaux, ils présentent 3 systèmes de gestion de données : système de données de chromatographie (CDS), système de gestion de données textes (TIMS), et du traditionnel système de gestion de données de laboratoire LIMS. Bien que les besoins en recherche préclinique auraient été exprimés depuis longtemps par (Leaders et al., 1980), la seule référence identifiée et qui s'intéresse à la gestion des données en recherche préclinique est la thèse de (Szymanski, 2008), soit 11 ans après les premiers travaux en recherche clinique. Il s'intéresse aux données produites par les plateformes de recherche préclinique des États-Unis d'Amérique dans des domaines variés : l'imagerie moléculaire, la protéomique, la cytométrie, la génétique, la pathologie, la microscopie, la culture tissulaire et les données expérimentales. Bien que ses travaux aient été intéressants, ils étaient plutôt limités aux données administratives et financières des prestations de recherche ainsi que sur la gestion de temps et de projets pour un chercheur travaillant dans une plateforme de recherche. Le système MIMI (Multimodality, Multiresource, Information Integration) de (Szymanski, 2008) a été appliqué à des données de prestations de service en imagerie et en protéomique.

Plus récemment, nous avons identifié quelques travaux avec une composante forte de gestion (ou management) de données précliniques: ImmunoDB (Rosa et al., 2014) en immunogénicité, (Lapinlampi et al., 2017) pour les données précliniques en épilepsie, Shanoir Small Animal pour les images IRM (Kain et al., 2020), SysWEB (Abidi et al., 2019) pour des données d'histologie, d'imagerie et de la RPE et SABER (Persoon et al., 2019) pour les données de la radiothérapie de précision guidée par l'image.

ImmunoDB est présenté dans (Rosa et al., 2014) comme un outil web pour la gestion des données de recherche en préclinique. Il est néanmoins particulièrement axé sur l'étude d'immunogénicité de peptides. L'immunogénicité est « la capacité qu'a un antigène de provoquer une réponse immunitaire bien spécifique » [58]. Il vise, principalement, à simplifier la présentation et l'analyse de grandes tables de données inexploitables autrement. Il est d'un grand intérêt pour son cas d'application. Il présente des fonctionnalités telles que la gestion de données brutes et dérivées, la possibilité d'ajouter des tags pour enrichir les données, et la visualisation.

EPITARGET (Lapinlampi et al., 2017), un projet européen pour la découverte des biomarqueurs en épilepsie, propose l'harmonisation de la collecte et l'analyse des données précliniques en recherche en épilepsie. Il propose ce que les auteurs nomment « élément de données commun (CDE) » pour stocker les informations précliniques, un dictionnaire de données, des rapports formatés en « CRF » pour les animaux et des recommandations. Il se base sur REDcap (Harris et al., 2009) pour le stockage des données précliniques.

[58] https://www.futura-sciences.com/sante/definitions/medecine-immunogenicite-13234/

Shanoir Small Animal (Kain et al., 2020) (Kain, 2018) est un projet coordonné par l'unité Analyse & Management de l'Information (IAM) de France Life Imaging (FLI)[59]. FLI est un projet français des Programmes Investissements d'Avenir (PIA) « Infrastructure en Biologie et Santé » financé sur 8 ans depuis 2012 et qui a pour vocation de mettre en place un réseau de collaboration académique, industrielle, et clinicienne en imagerie du vivant. Les chercheurs du nœud IAM, ont présenté un outil de gestion des données image pour le domaine préclinique lors du congrès national de l'imagerie du vivant (CNIV2019). Le démonstrateur ne traitait que les images IRM DICOM, mais d'autres modalités sont en cours d'ajout. Il s'agit d'un descendant du Shanoir pour l'imagerie clinique (Barillot et al., 2015). Shanoir-SA permet une gestion d'images sous format DICOM, un partage distant des images, ainsi que leur analyse, visualisation et traitement.

SysWEB est un produit de la société sysNCom qui est utilisé dans le projet Imagerie Du Vivant (IDV) (Abidi et al., 2019). Historiquement, il permet de gérer les images de microscopes et de scanners de lames (imagerie ex vivo). Il a été adapté lors de ce projet à l'imagerie du vivant. Il permet de gérer des images IRM sous format Brucker et des images d'échographie native et aussi des images de RPE (Résonance Paramagnétique Électronique). Il propose les fonctionnalités classiques d'une plateforme d'analyse d'images telles que la visualisation, la gestion des droits d'accès et l'ajout de métadonnées sous forme de tags. Le passage à l'échelle en mode multisite et multimodalité est en cours.

SABER (Small Animal Big-datawarehouse Environment for Research) (Persoon et al., 2019) se présente comme un entrepôt de données scientifiques du petit animal. Elle représente la seule plateforme trouvée qui met en avant la gestion de données sans s'orienter sur l'analyse de données contrairement à la plupart des outils récents. La plateforme comprend trois modules en interaction : la gestion de workflow, la gestion de données, et la gestion de stockage. Les données mises en jeu sont les données de radiothérapie de précision guidée par l'image. Ce sont des données DICOM principalement, mais il y a des références à d'autres types de données.

Les systèmes identifiés en recherche préclinique sont dans la plupart à leur début et sont très domaine-dépendants (épilepsie, immunologie, radiothérapie, IRM). Seul SysWEB présente des données indépendantes du domaine, mais ses fonctionnalités sont limitées à la visualisation et à la gestion documentaire des images. Les données hétérogènes de la recherche préclinique, indépendantes du domaine, et qui sont produites et utilisées tout au long d'une étude scientifique, n'ont pas été traitées jusque-là d'un point de vue centré sur le cycle de vie et en prenant en compte les besoins de la communauté.

II.2.3. SYNTHÈSE SUR LA GESTION DES DONNÉES EN RECHERCHE BIOMÉDICALE

Avant de faire le bilan détaillé des différents systèmes identifiés précédemment, nous présentons celui effectué par (Danciu et al., 2014). En effet, la mise en place de l'entrepôt de données cliniques à l'institut de recherche de Vanderbilt, a permis de mettre l'accent sur des constations critiques lors de la réutilisation des données cliniques en recherche :

1) Les données sont produites en premier pour un objectif de soins et sont donc, hélas, en majorité incomplètes pour une réutilisation en recherche. Réutiliser des données cliniques nécessite préalablement une phase de « nettoyage et préparation » des données.

La qualité des données utilisées en recherche est un sujet d'importance capitale pour la recherche clinique (Houston et al., 2020), mais aussi la recherche préclinique (Steckler et al., 2015).

2) La recherche étant compétitive par nature, réglementée aussi, les chercheurs ont besoin d'outils simples et puissants afin de faciliter rapidement l'accès à l'information utile.

[59] https://project.inria.fr/fli/

Ce besoin très bien explicité dans la section II.1 précédente et est rapporté aussi dans d'autres travaux comme (Anderson et al., 2007) et (Maxhelaku & Kika, 2020).

3) La recherche s'effectue sur le long terme, ce qui représente un défi pour la traçabilité des données scientifiques. (Danciu et al., 2014) préconisent la construction d'une équipe pluridisciplinaire durable maîtrisant le sujet de recherche en question, ce qu'ils reconnaissent être très difficile.

Nous pouvons ajouter que les laboratoires de recherche ont un turn-over très important ce qui engendre une perte de la mémoire du labo. À titre indicatif, l'indice de turn-over au CEA en 2013 était de 38 % par an (Peretti et al., 2015). Autrement dit, dans une équipe de 10 personnes 4 postes sont renouvelés chaque année. Dans un scénario extrême, au bout de trois ans, toute l'équipe est renouvelée. Ainsi, la traçabilité et la provenance des données sont une composante forte à prendre en compte dans toute proposition de système de gestion de données de recherche à long terme.

4) Pour réussir la réutilisation des données de recherche via des outils informatiques, il y a besoin d'une collaboration étroite entre informaticiens et cliniciens, dès le début de projet pour mieux les avertir sur la complexité des données et l'aider à mieux formuler leurs besoins souvent complexes de point de vue des données.

Ceci est rapporté, entre autres, dans (Leaders et al., 1980), 34 ans plus tôt, et dans (Anderson et al., 2007) et a été très bien explicité dans la section II.1 précédente .

Dans cette section ont été présentés des systèmes qui gèrent des données spécialisées en recherche biomédicale. Une constatation majeure est la ségrégation des systèmes de gestion selon le type des données ou le domaine d'application ou l'appartenance à un ensemble de personnes (institut, équipe, consortium).

Le Tableau 9 ci-après présente une vue d'ensemble de l'existant comme il a été présenté dans cette section [(imagerie, biologie, suivi journalier) vs (clinique, préclinique)] en décrivant chaque type de système selon les 19 besoins identifiés à la fin de la section II.1.

Tableau 9 Résumé des systèmes pour la gestion des données en (imagerie, biologie, suivi journalier) selon le domaine (clinique, préclinique)

	Imagerie	Biologie	Suivi journalier
Recherche Clinique	**Système typique : PACS** B1 : archivage présent, B2 et B3 : import et export des données en DICOM, B4 : recherche simple, B5 : partage en DICOM, B6 : analyse limitée, B7 : réutilisation absente, B8 : traçabilité minimale, B9 : simplicité présente, B10 : automatisation DICOM, B11 : standardisation en DICOM, B12 : ergonomie présente, B13 : efficacité présente, B14 : évolutivité absente, B15 : flexibilité absente, B16 : vérification simple, B17 : reporting simple	**Système typique : LIMS** B1 : archivage présent, B2 et B3 : import et export limités des données, B4 : recherche simple, B5 : partage limité, B6 : analyse absente, B7 : réutilisation absente, B8 : traçabilité minimale, B9 : simplicité présente,	**Système typique : EHR** B1 : archivage présent, B2 import présent B3 : export problématique, B4 : recherche avancée, B5 : partage limité, B6 : analyse absente, B7 : réutilisation présente, B8 : traçabilité minimale, B9 : simplicité présente, B10 : automatisation présente, B11 : standardisation présente, B12 : ergonomie présente, B13 : efficacité présente, B14 : évolutivité absente, B15 : flexibilité absente, B16 : vérification simple, B17 : reporting simple B18 : sécurité présente B19 : suivi minimal

	B18 : sécurité présente B19 : suivi absent	B10 : automatisation présente, B11 : standardisation minimale, B12 : ergonomie présente, B13 : efficacité présente, B14 : évolutivité absente, B15 : flexibilité absente, B16 : vérification simple, B17 : reporting simple B18 : sécurité présente B19 : suivi absent	
Recherche Préclinique	**Système typique : *inexistant***		**Système typique : CLE** B1 : archivage présent, B2 import présent B3 : export présent, B4 : recherche avancée, B5 : partage présent, B6 : analyse absente, B7 : réutilisation présente, B8 : traçabilité présente, B9 : simplicité présente, B10 : automatisation présente, B11 : standardisation absente, B12 : ergonomie présente, B13 : efficacité présente, B14 : évolutivité absente, B15 : flexibilité absente, B16 : vérification simple, B17 : reporting simple B18 : sécurité présente B19 : suivi présent

Il est évident, compte tenu de tout ce qui a été présenté auparavant, que le système qui gère les données hétérogènes en recherche biomédicale comme nous les avons définies à la section II.II.1.1 n'existe pas encore, d'autant plus en recherche préclinique.

Qui plus est, les systèmes existants souffrent d'une faiblesse d'évolutivité et de flexibilité (voir Tableau 9 ci-dessus) en regard des demandes des chercheurs et des changements liés à la nature de la recherche scientifique. Ceci peut être expliqué par une collaboration perturbée entre biologistes et informaticiens.

De surcroît, le suivi tout au long du cycle de vie d'une étude scientifique, la réutilisation documentée et le partage sécurisé sont des fonctionnalités limitées dans ces systèmes. Et pour finir, la « **B8** traçabilité » des activités de recherche et de la provenance des données ainsi que le « **L4** mirorring » du workflow de recherche (acquisition et analyse) (voir section II.1.1), sont aussi très limitée.

CONCLUSION DU CHAPITRE II

Pour pouvoir résoudre notre « Problème 1 : Comment gérer les données hétérogènes de la recherche préclinique et leur provenance ? », nous avons exploré dans ce chapitre les besoins de la communauté des chercheurs et les systèmes proposés pour la gestion des données de recherche biomédicale. Les ressources bibliographiques à ce sujet étant limitées pour la recherche préclinique, nous avons élargi notre exploration au domaine clinique. Les tableaux Tableau 7, Tableau 8 et Tableau 9 présentent une synthèse des différents travaux identifiés. Il faut satisfaire les besoins des chercheurs en ce qui concerne la gestion de données hétérogènes de façon à préparer le partage et la réutilisation ultérieure de ces données. Si l'outil logiciel proposé répond aux 19 besoins du Tableau 7, et prend en compte les leviers du Tableau 8, il sera utilisé. La gestion de données sera ainsi plus robuste, les données plus matures et la réutilisation plus efficace. En réalité, les initiatives dans le domaine de la gestion de données précliniques sont modestes et traitent généralement de l'analyse de données. La seule référence bibliographique que nous avons identifiée comme accordant le plus d'importance à l'organisation des données est SABER (Persoon et al., 2019), mais leurs recherches se sont limitées à la radiothérapie.

Dans la section I.4.1.5, nous avons explicité les briques fonctionnelles et méthodologiques du PLM. Ils permettent de gérer les données d'une manière intégrée et sécurisée, tout en assurant leur archivage et leur stockage dans les coffres-forts électroniques (stockage de données). Une plateforme PLM offre la plupart des fonctionnalités demandées par les experts biomédicaux à savoir : archivage (B1), import (B2), export (B3), requête (B4), partage (B5), analyse (B6), réutilisation (B7), traçabilité (B8), automatisation (B10), standardisation (B11), efficacité (B13), évolutivité (B14), flexibilité (B15), reporting (B17), vérification (B16), sécurité (B18), et suivi (B19) des données des produits industriels. L'application PLM à la neuroimagerie dans le cadre du projet BIOMIST a permis de mettre en place des fonctionnalités répondant aux besoins B1-6, B8, B10, B11, B14-18 des chercheurs en neuroimagerie. Elles sont détaillées en Annexe B.

Lors de cette thèse, nous continuons les efforts de recherche pour l'application PLM à la recherche biomédicale en général et à la recherche préclinique en particulier. Nous nous sommes armés des quatre leviers du Tableau 8 et nous proposons d'étudier et satisfaire les besoins identifiés dans le Tableau 7 en utilisant les briques fonctionnelles et méthodologiques du PLM pour la gestion de données scientifiques tout au long du cycle de vie d'une étude de recherche, en vue de partage et réutilisation. Nous présentons le système BMS-LM résultat de nos recherches dans le chapitre IV.

Lors de l'application PLM à la neuroimagerie, il a été reporté que les besoins en ergonomie (B12) et en simplicité (B9) ne peuvent pas être satisfaits avec les plateformes PLM existantes. En effet, ce besoin n'a pas été très demandé en industrie dès le début du paradigme PLM. Les processus industriels étant complexes et avec forte compétitivité, les autres fonctionnalités ont été considérées comme plus prioritaires. Le besoin en ergonomie n'a pas freiné l'utilisation des plateformes PLM puisque l'industrie a des ressources humaines dédiées à la gestion de données. Cependant, de plus en plus, l'intérêt pour ces deux besoins augmente en industrie (Allanic, 2015). Pour cette thèse (application PLM à la recherche biomédicale), nous sommes encore au début de ce domaine de recherche. Nous n'avons pas essayé de résoudre ce problème général des plateformes PLM, mais nous avons veillé à ce que nos propositions, au-delà des interfaces PLM présentées en annexe A, soient les plus simples et ergonomiques possibles. Nous avons priorisé les fonctionnalités répondant aux besoins de l'import (B2), l'analyse (B6), la standardisation (B11), l'évolutivité (B14), et la flexibilité (B15) des données en recherche préclinique afin de satisfaire les besoins en réutilisation (B7), efficacité (B13), et suivi (B19). Nos propositions sont détaillées dans les chapitres 4 et 5 pour la recherche biomédicale en général et leur application pour la recherche préclinique est décrite dans le chapitre VI.

Le chapitre III suivant présente l'état de l'art en organisation, gestion et ingénierie de connaissances. Il détaille l'importance de ces domaines pour la standardisation, la compréhension des données et l'interopérabilité en industrie et en recherche biomédicale. Il permettra d'orienter notre proposition concernant le « Problème 2 : Comment assurer la compréhension des données hétérogènes de la recherche biomédicale lors d'une réutilisation ultérieure ? »

Chapitre III. État de l'art en organisation, gestion, et ingénierie de connaissances et intersections avec les domaines : biomédical et industriel

Dans le chapitre I, nous avons présenté les données hétérogènes de la recherche préclinique et les différents standards et formats qui y sont liés. Nous avons expliqué le besoin du chercheur à utiliser des terminologies locales pour l'annotation de ses données par deux raisons : le retard de la mise à jour des standards par rapport aux avancées scientifiques, et la spécificité de ses recherches qui avancent au jour le jour et qui imposent d'inventer des noms et des concepts. Le sens de ces concepts peut être mal interprété ou incompris lors d'une réutilisation ultérieure des données, ce qui peut rendre les données inutilisables.

Dans ce chapitre, nous explorons l'état de l'art en réponse au problème 2 : « Comment assurer la compréhension des données hétérogènes de la recherche biomédicale lors d'une réutilisation ultérieure ? » (voir Figure 50). Tout d'abord, nous nous situons au niveau connaissance de la pyramide DIKW et nous explorons les domaines d'organisation, de gestion et d'ingénierie de connaissances. Cet axe de recherche est, d'un côté, en lien avec les briques « interopérabilité », « standardisation » et « gestion des connaissances » du paradigme PLM. Nous consacrons la deuxième section pour expliciter leurs liens. D'un autre côté, ce même axe de recherche est étroitement lié aux besoins de standardisation (B11) et aux recommandations RDM en matière d'annotation des données scientifiques avec des terminologies standards et interopérables. Une dernière section lui sera dédiée.

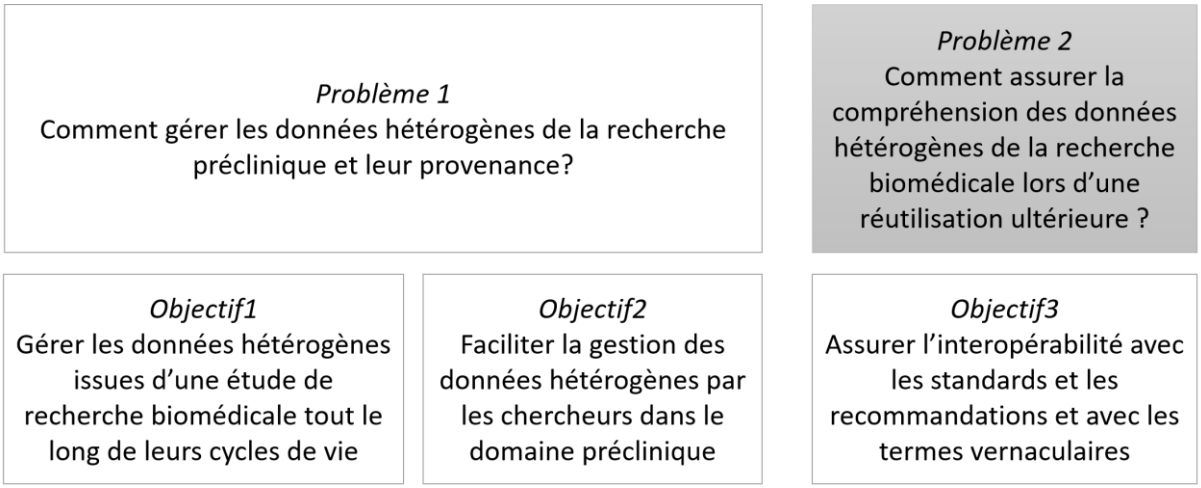


Figure 50 Problème 2 : compréhension des données hétérogènes de la recherche biomédicale lors d'une réutilisation ultérieure

III.1. APPORTS DE L'ORGANISATION, LA GESTION, ET L'INGÉNIERIE DES CONNAISSANCES

L'organisation, gestion, et ingénierie des connaissances accordent une attention particulière au sens des termes employés pour gérer les données. Pour illustrer l'importance de l'interprétation, prenons l'exemple du terme « SUV ». SUV en mécanique est un type de voiture et référence le « Sport Utility Vehicule (SUV) ». Mais SUV est aussi « Standardized Uptake Value (SUV) » en radiologie, et est une mesure importante utilisée en imagerie moléculaire pour quantifier la fixation d'un agent radioactif dans

les organes du corps (Buvat, 2007). Le terme « SUV » sans aucune indication sera interprété par un membre du laboratoire LRI comme mesure en radiologie, tandis qu'il sera interprété par un membre du laboratoire Roberval à l'Université de Technologie de Compiègne (UTC) en tant que type de voiture. Cet exemple met en évidence qu'avoir accès à une donnée (ici le terme « SUV ») ne permet pas à lui seul de « comprendre » cette donnée. Les connaissances et les expertises « dans la tête » de chaque personne conditionnera cette compréhension d'où l'intérêt de s'y intéresser afin de comprendre les données et pouvoir ainsi les gérer.

Dans cette section, tout d'abord, nous présentons la pyramide de la connaissance, DIKW, élément de structuration de notre thèse. Nous expliquons ensuite, l'importance des mots utilisés en présentant l'échelle sémiotique et pour finir nous détaillons, un à un, les domaines de la connaissance et leurs apports respectifs.

III.1.1. LA PYRAMIDE DE LA CONNAISSANCE (DIKW)

La donnée, l'information, la connaissance, et la compréhension ou sagesse forment la pyramide de la connaissance (voir Figure 51). Elle est une représentation connue et référencée dans plusieurs domaines notamment la gestion des données. Elle est souvent désignée par DIKW (Data-Information-Knowledge-Wisdom) (Rowley, 2007) (Ackoff, 1989). Selon (Sharma, 2008), ce modèle tient son origine improbable dans la poésie.

« *The poet T.S. Eliot was the first to mention the "DIKW hierarchy" without even calling it by that name. In 1934 Eliot wrote in "The Rock"*

Where is the Life we have lost in living?

*Where is the **wisdom** we have lost in **knowledge**?*

*Where is the **knowledge** we have lost in **information**?* » (Eliot, 1934)

(Rowley, 2007)

*"An ounce of **information** is worth a pound of **data**.*

*An ounce of **knowledge** is worth a pound of **information**.*

*An ounce of **understanding** is worth a pound of **knowledge**."*

(Ackoff, 1989)

Figure 51 La pyramide de la connaissance DIKW

Au niveau de la base de la pyramide réside les données (plus abondantes, peu utiles en soit) et au-dessus jusqu'au sommet de la pyramide : l'information, la connaissance et la sagesse/compréhension (plus rare, très utile en soit).

Plusieurs définitions et interprétations de la DIKW sont trouvées dans la bibliographie. (Baskarada & Koronios, 2013) résume quelques-unes et propose un cadre sémiotique pour analyser la pyramide ainsi qu'une caractérisation de la qualité à tous ses niveaux. Pour (Baskarada & Koronios, 2013), les données sont les signes physiques (lettres de l'alphabet, support de stockage, fichier numérique) et la qualité des données se définit en matière d'écart entre les caractéristiques du signe physique utilisé et sa spécification. Par exemple, une écriture mal soignée augmente l'écart entre la spécification de la lettre (A) et le signe physique utilisé ().

Par ailleurs, il définit l'information en matière de sens (la sémantique), la connaissance quant à elle, en matière de croyances jugées correctes socialement, et la sagesse en matière des différents jugements corrects et recommandés au sein de la société. Hormis la définition des données, (Baskarada & Koronios,

2013) font appel au mécanisme de la cognition et aux facteurs sociétaux pour définir les éléments de la pyramide. De la même façon, la qualité de l'information, la connaissance et la sagesse est définie par rapport à son adéquation à la finalité désirée par la personne ou la société.

Nous avons cité (Baskarada & Koronios, 2013) afin de montrer l'étendue de la pyramide et l'importance au-delà des TICs de cette distinction D, I, K, W. De notre côté, nous nous centrons sur les définitions qui sont les plus proches de notre contexte de la gestion de données techniques. Nous ne traitons que le niveau connaissance pour comprendre les données, et nous laissons de côté le niveau sagesse. Le Tableau 10 ci-après présente les définitions que nous adoptons en nous basant sur celles présentées dans (Baskarada & Koronios, 2013) :

Tableau 10 Les définitions de Données, Informations, Connaissances retenues

Donnée	Tout enregistrement et observation des choses, événements, activités et transactions qui ne sont ni organisés, ni traités, ni contextualisés	(Awad & Ghaziri, 2004)
Information	Ce qui est délivré par la communication de la donnée. Des données traitées cognitivement générant ainsi du sens et d'utilité pour les êtres humains. Ce qui circule entre les personnes afin d'informer et de s'informer. Il s'agit d'une donnée contextualisée.	(Baskarada & Koronios, 2013) (Laudon & Laudon, 2006)
Connaissance	Un ensemble d'informations qui ont été organisées et travaillées pour illustrer une compréhension, une expérience, un apprentissage accumulé et une expertise précise. Elle est reconnue au sein de la communauté travaillant sur un même sujet. Il s'agit d'une information consolidée et partagée.	(Turban et al., 2004)

III.1.2. LA SÉMIOTIQUE, LE SIGNE ET L'ÉCHELLE SÉMIOTIQUE

La classification des documents et l'annotation des données dans un but d'exploitation et de réutilisation semblent des tâches triviales, mais en leur sein réside un duo philosophique sémiotique ancien entre : le « Signifiant » (le symbole utilisé pour annoter ou classifier) et le « Signifié » (ce que l'on souhaite exprimer vraiment avec ce symbole).

Le « Signifié » est le concept abstrait, élément de connaissance « implicite » que nous souhaitons exprimer et qui est en lien avec le domaine en question. Le « Signifiant » est conditionné par le vécu de chacun et donc l'environnement social de l'acte d'annotation.

La sémiotique est la branche de la linguistique qui étudie les signes en général. Un « Signe » peut être une lettre, une image, un graffiti, on parle plutôt de « Symbole ou Signifiant » (voir Figure 52). Il référence un « Objet réel » ou « Référent » et symbolise l'abstrait cognitif désigné par « Concept ou Signifié ». Ces trois termes « Symbole-Réfèrent-Concept » ou « Signe-Object-Concept » sont connus sous le nom de triangle sémiotique (Ogden & Richards, 1923).

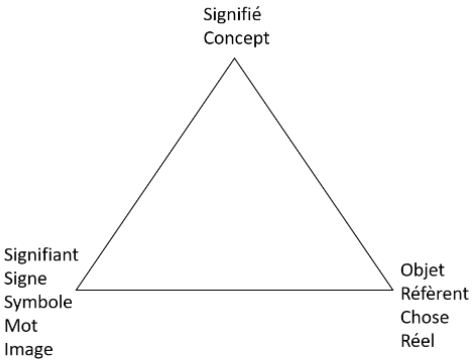


Figure 52 Le triangle sémiotique adapté de (Ogden & Richards, 1923)

(Baskarada & Koronios, 2013) proposent l'échelle sémiotique comme le montre le schéma ci-après (voir figure 47) :

- Niveau Social : actions entreprises et valeurs ajoutées
- Niveau Pragmatique : relation entre les signes (ou signifiant) et les concepts signifiés
- Niveau Sémantique : relation entre les signes et les choses qu'ils désignent (ou leurs sens)
- Niveau Syntaxique : relation structurelle et formelle entre les signes en leur forme physique
- Niveau Empirique : l'encodage des signes physiques pour transmission
- Niveau Physique : la physique et l'ingénierie pour modéliser les signes, les transmettre

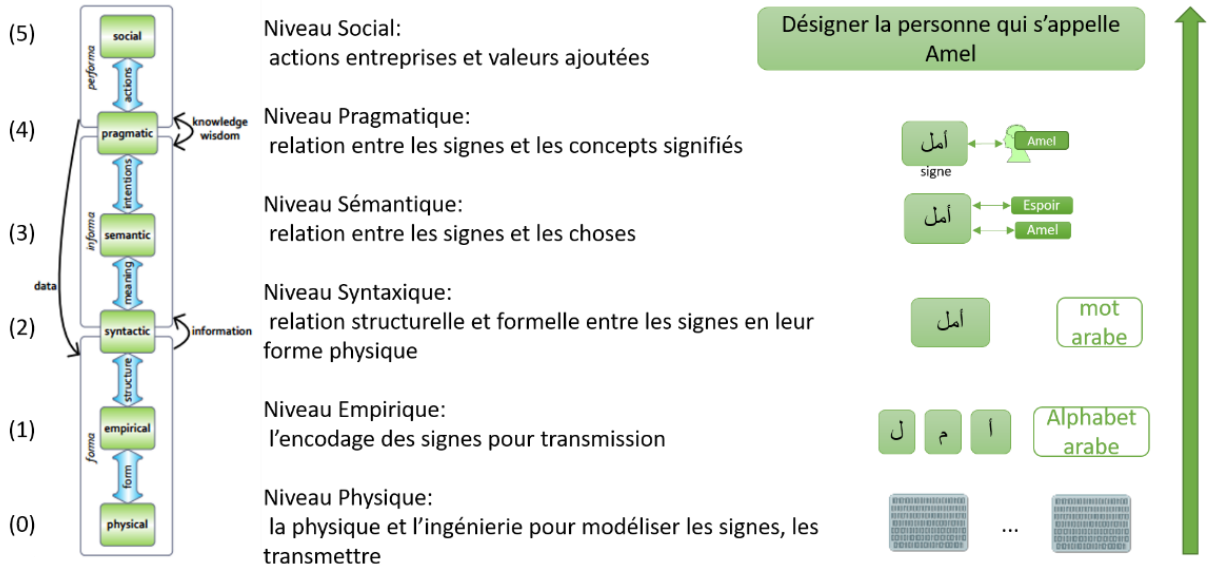


Figure 53 Échelle sémiotique : explication des niveaux du bas en haut et exemple adapté de (Baskarada et Koronios, 2013)

La partie droite du schéma précédent se lit du bas en haut : Les trois signes de départ sont transmis via le support physique numérique (0), encodé en suivant l'alphabet (1) et ensuite structurés en un mot au niveau syntaxique (2). Le niveau sémantique (3) apporte après de la signification et du sens aux trois lettres d'origine. Mais, ce n'est que, au niveau pragmatique (4) que la relation est établie avec le signifié et ainsi, les trois signes de départs compilés via les différents niveaux de l'échelle sémiotique permettent d'entreprendre l'action finale au niveau social (5).

En considérant les différents niveaux de l'échelle sémiotique quand il s'agit de la gestion des connaissances, l'enjeu derrière cette tâche devient de plus en plus important et pas du tout trivial.

III.1.3. LES DIFFÉRENTES DISCIPLINES DE LA CONNAISSANCE : L'ORGANISATION, LA GESTION, ET L'INGÉNIERIE DES CONNAISSANCES

Dans cette section, nous allons expliciter les spécificités de chaque discipline afin d'identifier les différents sujets en lien avec notre proposition. Nous avons identifié trois disciplines dans la bibliographie (Ohly, 2012):

- La gestion des connaissances (KM : Knowledge Management)
- L'organisation des connaissances (KO : Knowledge Organization)
- L'ingénierie des connaissances (KE : Knowledge Engineering)

III.1.3.1. La gestion des connaissances (KM : Knowledge Management)

Dans tout organisme, les données sont considérées comme une mine d'or et constituent la mémoire de l'organisme. En industrie, le terme mémoire d'entreprise est plutôt utilisé. La gestion de cette mémoire est d'un intérêt majeur pour les décideurs. En effet, l'exploitation de cette mine d'or permet de mieux connaître l'entreprise, et de construire des indicateurs de performances et d'aide à la décision. Il s'agit de la gestion des connaissances (KM : Knowledge Management).

Le KM est un domaine qui s'occupe de la capitalisation des connaissances des entreprises afin de permettre une meilleure productivité et une meilleure maîtrise de l'expertise et de la gestion de l'entreprise. L'enjeu managérial et financier est fortement présent. Il est souvent associé à l'informatique décisionnelle (Business Intelligence - BI), à la gestion de la mémoire d'entreprise, et au transfert des expertises.

Le KM est la partie applicative de l'ingénierie des connaissances (KE) et de l'organisation des connaissances (KO). Il donne alors une attention particulière pour la mise en œuvre, l'adoption et l'utilisation des solutions proposées pour la gestion des connaissances. En court, le KM peut être défini comme « *doing what is needed to get the most out of knowledge* »[60] (Becerra-Fernandez & Sabherwal, 2010).

Les collaborateurs d'une entreprise, et par analogie les chercheurs dans un laboratoire de recherche ont aussi un intérêt pour la gestion des connaissances. Ceci permettra de réduire la complexité et l'hétérogénéité des données en facilitant leur compréhension. En effet, l'acquisition des connaissances permet de fournir un référentiel commun aux collaborateurs et facilitera ainsi le dialogue entre eux : « ça permet de nommer un chat un chat et un chien un chien »

III.1.3.1.1. Cycles de vie en KM

Plusieurs Frameworks et cycles de vie de KM ont été proposés dans la bibliographie afin de décerner les frontières, les approches et méthodes en gestion de connaissances (KM). (Shongwe, 2016) présente une synthèse de 20 Frameworks pour la période allant de 1991-2016 et propose après les avoir analysés, son propre Framework.

Selon (Shongwe, 2016), le cycle de vie de la gestion de connaissance comprend les cinq activités suivantes, organisées **selon leur importance** pour les 20 Frameworks analysés:

- ❖ Le partage/transfert de connaissances : le processus de transfert de connaissance d'une personne, endroit, propriétaire à un autre. Dans un contexte d'entreprise, il représente le partage et l'échange de connaissances entre collaborateurs via des réunions, ou en utilisant des outils technologiques

[60] En français: « faire ce qui est nécessaire pour récupérer le maximum de connaissances »

comme les courriels et les groupes de discussion. Une solution technologique de KM doit permettre et faciliter un tel partage.

- ❖ Le stockage/conservation de connaissances : l'activité d'encodage et stockage des connaissances et expertises d'une entreprise pour constituer ce qu'on appelle « La mémoire d'entreprise ». La mémoire d'entreprise permet de collecter toutes les expertises passées et qui sont intéressantes pour le présent et le futur de l'entreprise. L'enjeu est important surtout pour le stockage et la pérennisation des connaissances implicites importantes des experts avant leur départ.
- ❖ L'application/utilisation de connaissances : le processus d'utilisation des connaissances acquises dans la résolution des problèmes, la gestion de l'entreprise, la prise de décision : l'action au niveau pragmatique et social. Dans un projet de gestion des connaissances, l'utilisation de connaissances constitue un indicateur de qualité et d'utilité de la solution technologique adoptée pour le KM.
- ❖ La création de connaissances : le processus de création de connaissances à partir des connaissances existantes. Lorsqu'une entreprise opte pour une solution technologique de gestion des connaissances, elle vise à activer ce processus afin de produire de la valeur ajoutée et de nouvelles connaissances. Par analogie, en ingénierie de connaissances, les moteurs d'inférences ont pour but de créer de la connaissance à partir des connaissances existantes.
- ❖ L'acquisition de connaissances : l'activité d'identification, de capture et de collecte des connaissances depuis l'environnement socio-économique de l'entreprise afin de l'utiliser dans l'entreprise. L'acquisition de connaissances permet d'améliorer et d'enrichir les connaissances stockées de l'entreprise ainsi que de mieux les contextualiser.

(Shongwe, 2016) n'a pas imposé un cycle ou une succession pour le cycle de vie proposé. Cependant, et compte tenu de tous les Frameworks analysés, un ordre s'impose en commençant par l'acquisition des connaissances, puis le stockage, après l'application et l'utilisation, ensuite le partage, et enfin la création de nouvelles connaissances. Le cycle se referme avec l'acquisition des connaissances nouvellement créées. (voir Figure 54).

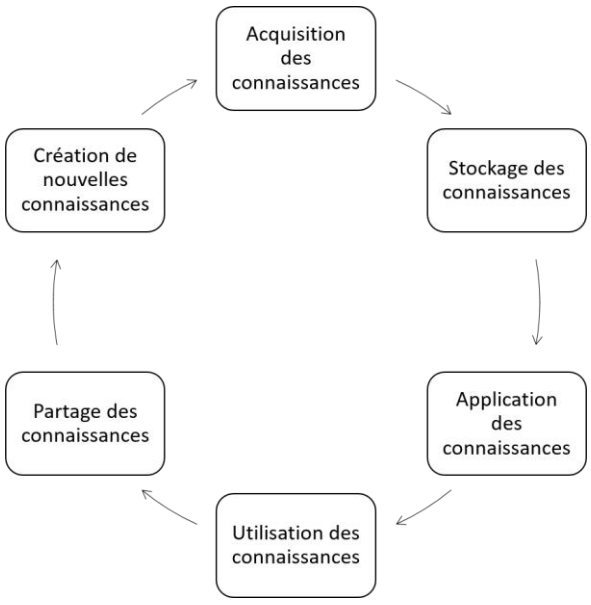


Figure 54 Cycle de vie de la gestion des connaissances

Des statistiques (KPMG, 1998), cités plusieurs fois par les articles en KM notamment dans (Alavi & Leidner, 2001) indiquent que les objectifs premiers des entreprises en s'engageant à une démarche KM sont les suivants (par ordre de pourcentage des voix collectées) :

1. Une meilleure prise de décision (86%)
2. Une réduction de coût (70%)
3. Une réponse rapide à des problématiques clés (67%)

4. Améliorer la productivité (67%)
5. Partager des bonnes pratiques (60%)
6. Créer de nouvelles opportunités de business (58%)
7. Augmenter le profit (53%)
8. Une meilleure interaction entre les collaborateurs (42%)
9. Augmenter la part de marché (42%)

Si l'on classifie les objectifs des entreprises selon le cycle de vie de gestion des connaissances (voir Figure 54) on aura : 1, 3 qui relèvent de l'utilisation de la connaissance ; 2,4,6,7,9 qui relèvent de la création de connaissances ; et 5,8 qui relèvent du partage de la connaissance. Le tout ne peut pas se réaliser sans l'acquisition et le stockage de la connaissance. Ceci nous donne une idée sur le rôle crucial des systèmes de gestion des connaissances d'entreprise (KMS : Knowledge Management System) qui permettent d'effectuer toutes ou quelques étapes du cycle de KM.

III.1.3.1.2. Systèmes de gestion de connaissances (KMS : Knowledge Management System)

(Maier, 2007) montre que divers systèmes et outils ont été utilisés comme KMS dans le cadre d'une enquête effectuée sur un échantillon de 71 entreprises parmi les TOP 500 entreprises allemandes et les TOP 50 banques et assurances. Les KMS identifiés sont en majorité des systèmes « fait-maison » développés pour des besoins spécifiques à l'organisation (18 réponses sur 28). Dans d'autres cas, le KMS a été développé en se basant sur une solution Groupware (logiciel de travail collaboratif) ou une solution de serveur Intranet. Les entreprises restantes ont opté pour d'autres produits commerciaux notamment : "Compass Server" de Netscape, "Knowledge Organizer" de Verity, "Livelink" de Open Text, "Knowledge X" de IBM et d'autres.

Dans la même étude, 48 organisations ont été interrogées concernant le contenu exact de leur KMS et ainsi plus de 50% des répondeurs ont reporté que les connaissances suivantes sont prises en compte dans leur KMS :

- La connaissance à propos des processus et de l'organisation
- La communication interne
- La connaissance à propos des partenaires business
- Les études internes
- La connaissance des produits
- Le répertoire des collaborateurs et leurs compétences

Les technologies de l'information (TIC) jouent un rôle important en KM. Le datawarehouse (DWH), le Management Information System (MIS), le Système à Base de Connaissances (SBC ou KBS), Systèmes Experts (SE) sont tous liés au système de gestion des connaissances (KMS). Avant de construire un KMS, il faut d'abord répondre aux questions énoncées par (Ohly, 2012) afin de maîtriser le cas d'application de l'entreprise bénéficiaire: Qui a une connaissance importante pour l'entreprise? Comment peut-elle être codée ? Pour qui doit-elle être délivrée ?

III.1.3.2. L'organisation des connaissances (KO : Knowledge Organization)

Le domaine de l'organisation des connaissances s'occupe de la classification des ressources documentaires. Le développement récent des bibliothèques numériques lui a permis de s'élargir et s'ouvrir à l'informatique. Des analogies avec le domaine des bases de données, la programmation orientée objet et l'ingénierie des connaissances (KE) s'avèrent très utiles.

En 2008, le journal Knowledge Organization (ISSN 0943-7444)[61] a traité, dans un numéro spécial (McIlwaine & Mitchell, 2008), la question de la recherche dans le domaine de l'organisation des

[61] https://doi.org/10.5771/0943-7444

connaissances et son lien avec les autres domaines. L'organisation des connaissances y est définie comme « *The field of scholarship concerned with the design study, and critique of the processes of organizing and representing documents that societies see as worthy of preserving.* »[62].

(Hjørland, 2008) propose deux définitions de l'organisation des connaissances : une « *narrow* » où il énonce que les activités du KO sont : la description, l'indexation, la classification des documents issus des bibliothèques, des bases de données bibliographiques, des archives, et des autres types de « mémoires d'institutions » par les documentalistes et une « *broader* ».

Bien que cette définition « *narrow* » serait celle de référence dans le domaine, l'organisation des connaissances est plus pertinente dans sa définition « *broader* ». (Hjørland, 2008) la définit comme étant « l'étude de la manière dont les connaissances sont socialement organisées et la manière dont la réalité est organisée ». En d'autres termes, il s'agit de la classification et la structuration des connaissances d'un domaine donné en prenant en considération ses liens avec le social et le réel. (Hjørland, 2008) parle d'un dilemme de la « Connaissance » entre le « Réel » et le « Social » qui nous amène à prendre des décisions d'annotations pragmatiques et contextualisées (cf. La sémiotique §III.1.2).

L'organisation des connaissances (KO) est un domaine de recherche piloté en majorité par les experts en sciences de l'information, de la documentation et des bibliothèques. Elle représente un domaine interdisciplinaire qui fait intervenir des disciplines comme l'informatique, la modélisation et le stockage des données, les sciences humaines et sociales, la philosophie, etc. L'ISKO (International Society for Knowledge Organization) regroupe les chercheurs de ce domaine et édite le journal « Knowledge Organization » depuis 1974.

III.1.3.2.1. Les approches identifiées en KO par les bibliothécaires

Historiquement, les livres sont indexés sur la base de la Classification Décimale de Dewey (CDD, voir Figure 55) en sciences de l'information et de la documentation. Pour assurer une annotation utile, une structuration des connaissances délivrées par ces données (ou ces livres) est indispensable. Plusieurs approches de classification ont été identifiées dans la bibliographie par (Hjørland, 2008) où il a explicité chaque approche et a analysé son apport et ses limites vis-à-vis de l'organisation des connaissances.

L'approche traditionnelle d'organisation des connaissances utilise les systèmes de classification (comme la CDD) et les bases de données bibliographiques (comme MEDLINE (Medical Literature Analysis and Retrieval System Online)). La CDD est la plus utilisée dans les bibliothèques, elle a pour vocation d'indexer les livres avec un code décimal indiquant leur catégorie d'appartenance (voir Figure 55).

[62] En français : « le domaine qui s'intéresse à la conception d'étude et au critique du processus de l'organisation et de la représentation de documents intéressants à préserver pour la société »

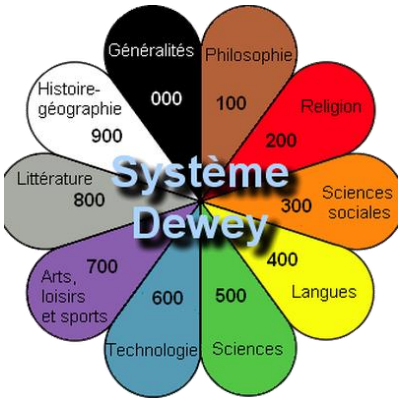


Figure 55 Système de Classification Décimale de Dewey (CDD)[63]

La Figure 56 montre l'exemple du code 611.347 d'un ouvrage, appelé aussi « indice de Dewley ». Pour le comprendre, il faut trouver la classe correspondante dans la CDD : ici, 600, Technologie. Pour continuer, il faut connaître les autres niveaux de la CDD qui sont les divisions (10), les sections (1) et les sous-sections (.3,.04,.007).

611.347
600 TECHNOLOGIE (Sciences appliquées) [classe]
610 Sciences médicales Médecine [division]
611 Anatomie humaine [section]
611.3 Organes de l'appareil digestif [sous-section]
611.34 Intestins
611.347 Gros intestin

Figure 56 Interprétation d'un exemple de code dans la CDD[64]

L'indice de Dewley est suivi de lettres et de chiffres (non présent dans l'exemple), désignant la côte attribuée à l'ouvrage d'une façon unique en tenant compte des informations sur son auteur. Ceci nous rappelle les identifiants uniques des objets dans les bases de données.

L'approche de Dewley a été fortement critiquée par la communauté KO. En effet, plusieurs sont ceux qui la considèrent comme un système sémiotique faible, qui a été construit de point de vue industriel et non pas théorique et scientifique (Frohmann, 1994).

En 1933, S. R. Ranganathan a proposé la formule PMEST (`Personality`, `Matter`, `Energy`, `Space`, `Time`) formée de cinq concepts (ou facettes). Elle permet la classification des sujets (ou documents). Afin d'analyser chaque sujet, une projection dans l'ensemble des concepts PMEST est faite. (Moss, 1964) énonce qu'il s'est peut-être inspiré des catégories d'Aristote (substance, quantité, relation, qualité, lieu, temps, situation ou position). L'analyse à facettes est largement reconnue et utilisée malgré une justification théorique absente. Elle est en l'occurrence utilisée dans la conception de pages web et dans des formats XML d'échange de métadonnées.

L'approche par extraction d'informations, en anglais « Information Retrieval », a été fondée dans les années 50. Il s'agit une méthode basée sur l'analyse systémique du texte afin d'extraire les informations d'intérêt à la suite d'une requête utilisateur. Afin d'évaluer la qualité des termes identifiés et documents proposés, une validation par expert est effectuée. Les métriques d'expérimentation de Cranfield

[63] Source : www.memrise.com
[64]Source : site chercher pour trouver https://www.ebsi.umontreal.ca/jetrouve/biblio/

« précision » et « rappel »[65] sont utilisées. La principale critique de la méthode est qu'elle n'est pas épistémologiquement fondée et trouve son fondement dans les statistiques. Aussi, cette approche ne tient pas compte du contexte ni des différents points de vue avec lesquels un sujet peut être traité. Elle impose, selon (Hjørland, 2008), un seul point de vue, celui de l'expert qui a évalué l'approche, à tout autre type d'utilisateur.

L'approche bibliométrique analyse les réseaux d'articles et utilise le nombre de citations et de co-citations afin de définir les classes de classification, et les liens entre eux. Cette approche utilise des sources fiables (publication scientifique) et des concepts bien contextualisés et identifiés. Cependant, le nombre de citations, critère de base de cette approche, est potentiellement biaisé par le facteur social (i.e. les citations sont très influencées par le réseau des chercheurs). La robustesse intellectuelle de la classification résultante reste questionnable.

L'approche analytique d'un domaine est celle qui s'occupe d'un domaine particulier d'un point de vue social et épistémologique. Elle vise à fournir une classification des documents cohérente avec le besoin d'un groupe d'utilisateurs et servant un objectif idéal. Le but n'est pas de fournir une classification « *one fits all* », mais de fournir une classification adaptée à la réalisation d'une tâche précise. Afin d'avoir une indexation cohérente des documents, une connaissance minimale du domaine et de ses différents enjeux doit être assurée par l'expert en KO.

Un cercle vicieux peut se construire facilement lorsque l'on s'intéresse à un domaine particulier, en effet pour identifier les terminologies d'un domaine, il faut comprendre le domaine, et pour comprendre le domaine, il faut connaître les terminologies. Cette approche doit effectivement être itérative. Par conséquent, la classification, qui en résulte, est en continuelle évolution et raffinement en fonction des avancées dans le domaine, des objectifs de classifications et des besoins des utilisateurs.

Les aspects spécifiques aux « approches orientées utilisateur », selon (Hjørland, 2008), sont notamment l'aspect « *user-friendly* », l'aspect orienté marché, les études empiriques des utilisateurs, ou encore la classification réalisée par les utilisateurs. L'évaluation de cette approche se fait avec les métriques de Cranfield « *précision* » et « *rappel* ».

Des critiques peuvent être faites sur l'idée même de l'utilisation des métriques de « *précision* » et « *rappel* » : est-ce que l'utilisateur ne veut vraiment avoir que les documents strictement en lien avec ce qu'il cherche ? Est-ce qu'on peut faire confiance aux connaissances de celui qui cherche pour qu'il puisse trouver ce qui l'intéresse ? Par exemple, un constat de tous les jours, parfois nous cherchons des choses dont on ne connaît pas le nom (le signe).

Les approches en KO ont beaucoup évolué depuis la CDD (voir Figure 55). Elles se croisent de plus en plus avec des approches en KE et en KM. Dans le cadre de nos recherches, nous retenons les deux approches « bibliométriques » et « analytique ». La force de la première est de fournir une information confirmée par la communauté scientifique et contextualisée. Tandis que, la force de l'approche analytique réside dans l'immersion au sein du domaine étudié et son adaptation aux besoins d'un groupe d'utilisateurs et à la réalisation d'une tâche précise. Elle est aussi en continuelle évolution ce qui est bien adapté au contexte de la recherche biomédicale et préclinique.

III.1.3.2.2. Les Systèmes d'Organisation des Connaissances (SOC ou KOS)

Le produit de l'organisation des connaissances est connu sous le nom de Système d'Organisation des connaissances (KOS). Il représente une structuration de termes et de significations permettant de

[65] « précision » est le pourcentage des résultats positifs retournés par une requête par rapport à l'ensemble des résultats retournés et « rappel » est le pourcentage des résultats positifs retournés par rapport à l'ensemble des résultats positifs.

modéliser la connaissance (pour un jeu de donnée par exemple, ou pour un domaine particulier, ou aussi d'un point de vue général).

Les KOS sont définis selon (Hodge, 2000) comme : « *The term knowledge organization system is intended to encompass all types of schemes for organizing information and promoting knowledge management*.» [66]. KOS est un terme global qui inclut les classifications, les terminologies, les vocabulaires structurés, les glossaires, les réseaux sémantiques, les ontologies, etc. Dans le domaine des bases de données, un KOS peut être rapproché à un schéma SQL de base de données, un modèle de données en UML ou une ontologie en ingénierie de connaissances (KE).

(Souza et al., 2012) ont proposé une taxonomie de KOS (voir Figure 57) en 4 grandes familles du moins structuré au plus structuré :

- Texte non structuré
- Termes et/ou listes de concepts
- Concepts, relations et disposition structurée
- Concepts et structures de relations

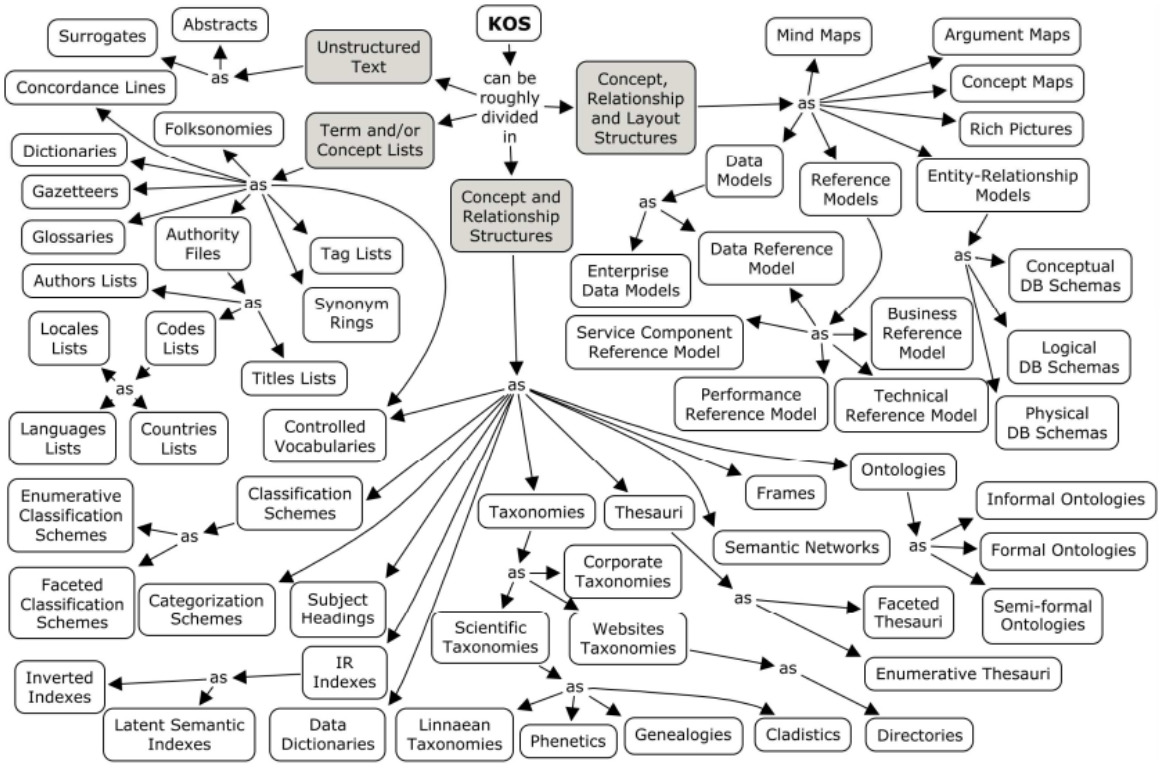


Figure 57 Taxonomie des KOS selon (Souza et al., 2012)

Dans la Figure 57, un dictionnaire est classé comme « Termes et/ou listes de concepts » et des images riches « *Rich Pictures* » sont classés comme « Concepts, relations et disposition structurée » : La différence est énorme, mais le lien est clair : tous les deux constituent une façon de présenter les connaissances, en tant qu'information, via des données en tant que signes ; alphanumérique pour le premier, graphique pour le deuxième.

Les Systèmes d'Organisation de Connaissances peuvent être classés selon plusieurs critères : le degré de structuration, de formalisation vis-à-vis à un moteur de raisonnement, et l'expressivité sémantique. Par exemple, les glossaires en HTML ont une faible expressivité sémantique contrairement aux

[66] En français : « tout type de schéma d'organisation d'information et de gestion des connaissances »

ontologies OWL (Bergman, 2007). Aussi, une liste de concepts est moins structurée qu'un modèle de données ou une ontologie.

III.1.3.3. L'ingénierie des connaissances (KE : Knowledge Engineering)

L'ingénierie des connaissances (KE) est le domaine qui s'occupe de modéliser les connaissances et le raisonnement humain afin de les exploiter dans des systèmes informatiques autonomes et intelligents. Le KE fait intervenir plusieurs domaines de recherche en informatique. Il fait partie des disciplines de l'intelligence artificielle (IA), définie au sens large par « l'externalisation des facultés humaines dans un support indépendant de l'homme ». Elle représente un domaine de l'IA connu sous le nom de l'« IA symbolique ».

En France, une conférence annuelle dédiée à l'ingénierie des connaissances tient lieu tous les ans et regroupent les experts francophones du KE. Elle est pilotée par le collège SIC (Science de l'Ingénierie des Connaissances) de l'AFIA (Association Française pour l'Intelligence Artificielle).

Selon (Charlet, 2002), l'ingénierie des connaissances est au carrefour de plusieurs disciplines : la linguistique; la terminologie et la recherche sur les ontologies; la psychologie; la logique avec les méthodes formelles; l'informatique; l'ergonomie; les sciences de gestion pour l'étude de l'environnement organisationnel des connaissances, etc.

Différentes méthodes et techniques relèvent de l'ingénierie des connaissances et couvrent tout le cycle de vie en KM (voir Figure 54). Pour commencer, l'acquisition de connaissances à partir de texte, d'images, de données semi- ou non-structurées, est réalisée avec des techniques de fouille de données (Data mining), de traitement automatique de langue (TAL), ou d'analyse statistiques pour l'extraction des concepts pertinents. Il s'agit de la première étape du cycle de vie.

Ensuite, la représentation des connaissances via des modèles conceptuels est réalisée via l'ingénierie ontologique (Ontology engineering) (Corcho et al., 2006), un sous-domaine du KE qui s'intéresse aux KOS, principalement les ontologies, et leurs méthodes de conceptions, construction, alignement, et intégration. Une attention particulière est portée à l'interopérabilité sémantique entre ces KOS.

Afin de représenter les KOS, les standards du web sont utilisés. En effet, les technologies du web sémantique ont contribué à la structuration et stabilisation du langage de représentation d'ontologie via l'apparition du langage de description OWL (*Ontology Web Language*) et ses différents niveaux de formalisation logique (OWL-Lite, OWL-DL, OWL-Full), des annotations SKOS (Simple Knowledge Organization System), ainsi que des bases de connaissances ou « triple store » RDF (*Ressource Description Framework*) et le schéma RDFS associé.

Lors des étapes suivantes, d'application et utilisation des connaissances, le KE s'intéresse à l'utilisation des ontologies et KOS, dans l'annotation sémantique des données, la conception de systèmes à base de connaissances (KBS), la conception d'Interfaces Homme Machine (IHM) ainsi que la recherche d'informations en utilisant les connaissances acquises préalablement.

Le raisonnement logique via des moteurs d'inférence ainsi que les différentes techniques d'apprentissage en « IA symbolique », des spécificités de l'ingénierie des connaissances, permettent la création de nouvelles connaissances, dernière étape du cycle de vie en KM avant l'acquisition des connaissances nouvellement créées. L'atout par rapport à la gestion des bases de données est que la machine gère les concepts et la sémantique qui relèvent de la compréhension humaine (i.e. les moteurs d'inférences et de raisonnements). Ces mécanismes sémantiques permettent de remonter la pyramide DIKW des données vers les connaissances.

Les ontologies, les systèmes experts (SE) et les systèmes à base de connaissances (KBS) sont les produits phares de ce domaine.

III.1.3.3.1. Les systèmes experts et les systèmes à base de connaissances (SBC ou KBS)

Un système expert (SE) est un système qui est conçu pour effectuer une tâche pointue en se basant sur une base de connaissances construite en collaboration avec les experts de cette tâche. Les systèmes à base de connaissances (KBS) sont une version plus élargie des systèmes experts. Un KBS est tout système composé d'une base de connaissances (régie par un KOS), d'un moteur de règles ou d'inférences modélisant le raisonnement humain et d'une interface homme-machine via laquelle le système interagit avec l'utilisateur (voir Figure 58) (Umar, 2015).

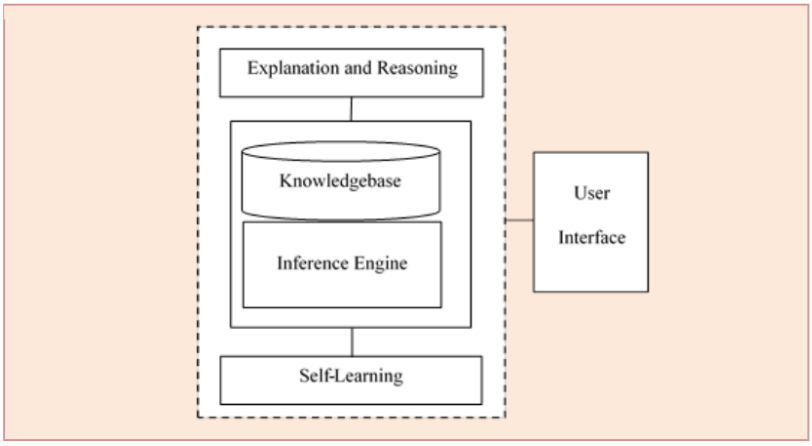


Figure 58 Architecture typique d'un KBS (Umar, 2015)

(Umar, 2015) explique que la capitalisation des connaissances via la base de connaissances du KBS passe par les trois étapes suivantes.

- L'acquisition des connaissances : elle comprend la collecte de connaissances par différentes méthodes, comme l'observation, les entretiens, les questionnaires, etc.
- La vérification des connaissances : lorsque les ingénieurs ou les « ontologistes » appliquent différentes règles sémantiques, afin de formaliser les connaissances collectées précédemment.
- La représentation des connaissances : le format livré à l'utilisateur des connaissances formalisées. Elle doit être compréhensible et adaptée au domaine ou problème en question.

Une méthode assez connue dans ce cadre est la méthode CommonKADS (Schreiber et al., 2000). Elle est utilisée pour la modélisation des connaissances et le développement des KBS. Elle est une méthode de référence dans le domaine à l'échelle Européenne. Elle représente le produit d'une série de projets financés par le programme Européen ESPRIT, dont le premier remonte à 1983. Elle est structurée en six modèles comme dans le schéma de la figure 53 ci-après :

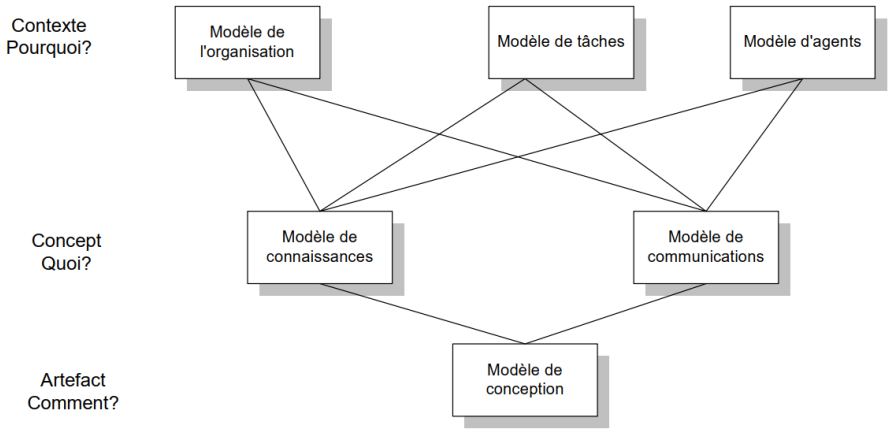


Figure 59 Les six modèles de la méthode CommonKADS (Schreiber et al., 2000)

Selon (Abi-Zeid & Lamontagne, 2000), les trois premiers modèles permettent l'analyse de l'environnement organisationnel afin d'étudier les risques et les leviers derrière le déploiement d'un KBS dans l'organisation. Les deux modèles suivants (de connaissances et de communications) fournissent la description conceptuelle des expertises et des raisonnements pour la résolution des problèmes qui doivent être fournis par le système. Le modèle de conception, réponse à la question « comment ? », est la spécification technique du système livré, basé sur les cinq premiers modèles. Les six modèles ne doivent pas toujours être présents dans un KBS, ceci dépend du contexte d'application.

III.1.3.3.2. L'ingénierie ontologique

L'ingénierie ontologique désigne la branche en informatique en lien avec la construction d'ontologies afin de modéliser les connaissances et les expertises d'un domaine. Une ontologie est définie par (Studer et al., 1998) comme « *An ontology is a formal, explicit specification of a shared conceptualization* » (adapté de (Gruber, 1993) « *explicit specification of a conceptualization* » et (Borst, 1997) « *formal specifications of shared conceptualizations* ») [67].

De nos jours, l'outil d'édition d'ontologie par excellence est « Protégé » de l'Université de Stanford (Gennari et al., 2003). D'autres outils ont été proposés au fil de temps comme OILEd ou WebODE (A. Singh & Anand, 2013), mais « Protégé » reste le plus utilisé. Récemment, une version web appelée « webprotégé » a été proposée pour plus de flexibilité dans l'édition collaborative des ontologies. Il est présenté comme un environnement de développement des KBS.

Différents types d'ontologies sont reconnus par la communauté (Roussey et al., 2011) : ontologies de haut niveau, ontologies noyau, ontologies générales, ontologies de domaine, ontologies de tâches et ontologies locales ou d'application (voir Figure 60).

[67] En français : « spécification formelle et explicite d'une conceptualisation partagée »

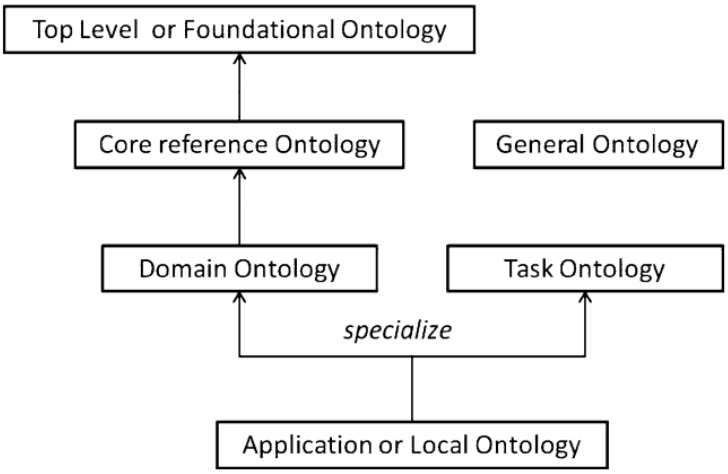


Figure 60 Les différents types d'ontologies communément référencés (Roussey et al., 2011)

- **Les ontologies locales ou d'application** : Il s'agit d'ontologies faites maison, orientées vers les applications, qui dépendent du contexte spécifique et ne sont pas partageables ailleurs.
- **Les ontologies de domaine** : Une ontologie qui explicite tous les termes d'un domaine très spécifique et qui reflète une convention entre un groupe d'experts du domaine. Par exemple, la dose de radioactivité injectée à une souris lors d'un examen d'imagerie TEP-TDM.
- **Les ontologies de tâche** : Cette ontologie contient les connaissances nécessaires à la réalisation d'une tâche ou activité.
- **Les ontologies de référence /noyau** : Une ontologie qui représente les concepts génériques de base d'un grand domaine et qui peut intégrer d'autres ontologies de domaine plus spécifiques. Par exemple, dans une étude biomédicale, on peut trouver examen, protocole, plan, etc.
  Des ontologies noyau de référence ont été proposées dans différents domaines tels que la production de films (Chakravarthy et al., 2009) et l'analyse des processus d'entreprise (Pedrinaci et al., 2008). Une ontologie noyau est essentielle pour établir un engagement ontologique minimal pour un grand domaine (Gruber, 1995). Elle garantit la cohérence entre les KOS qui la réutilisent.
- **Les ontologies générales :** Cette ontologie représente un groupe de concepts qui ne sont pas liés à un domaine particulier, mais qui contiennent des connaissances générales d'un vaste champ disciplinaire.
- **Les ontologies de haut niveau** : Ce sont des ontologies qui « modélisent le monde » en référence aux origines en philosophie analytique du mot « ontologie ». Une ontologie de haut niveau est une ontologie indépendante d'un contexte ou d'un domaine particulier (Roussey et al., 2011). Elle reflète une vision cohérente et formelle du monde, également appelée engagement ontologique (Guarino et al., 1994) et (Gruber, 1995). Les ontologies de haut niveau représentent le très haut niveau de connaissance. Elles sont peuplées de termes abstraits tels que : entité, objet, processus, etc.
  Dans le domaine biomédical, l'ontologie DOLCE (Descriptive Ontology for Linguistic and Cognitive Engineering) (Gangemi et al., 2002) et BFO (Basic Formal Ontology) (B. P. Smith et al., 2005) sont les ontologies de haut niveau les plus connues. Elles décrivent des concepts généraux tels que `entité`, `concept`, `perdurant` (DOLCE), `continuant` (BFO), `entité physique`, `entité immatérielle`, `processus`, etc.

III.1.3.4. Synthèse des KM, KO et KE

Les trois domaines décrits dans cette section ont émergé chacun d'un contexte différent et entretiennent des points de vue différents. Pourtant, en fin de compte, ils se complètent et se nourrissent mutuellement. Pour le KM, le contexte concurrentiel en entreprise prône pour l'efficacité et le rendement et utilise les KMS dans le cadre de ce but premier. Le KO, quant à lui est un domaine qui mise sur la sémantique et

la sémiotique afin de structurer les connaissances sous forme de KOS. Ces derniers servent à l'annotation des documents des bibliothèques numériques. Pour le KE, le contexte n'est ni l'annotation en bibliothèques ni l'efficacité en entreprise (pourtant ce sont des applications potentielles). Le focus en KE est sur la modélisation de la connaissance, en utilisant la logique formelle en informatique, via une ontologie (i.e. un KOS), et ensuite, son exploitation en entrée à un moteur d'inférence, dans le cadre d'un KBS. L'ingénierie ontologique (conception, alignement, etc.) est au cœur du KE et sert à la fois le KM et le KO. De la même façon qu'il y a plusieurs types de KOS, il y a plusieurs types d'ontologies (locale, domaine, noyau, de tâche, générale, haut niveau). Dans le paragraphe suivant, nous présentons les éléments bibliographiques concernant l'interopérabilité sémantique entre KOS.

III.1.4. L'INTEROPÉRABILITÉ SÉMANTIQUE ENTRE LES SYSTÈMES D'ORGANISATION DES CONNAISSANCES (KOS)

L'interopérabilité donne le « I » aux recommandations FAIR (Wilkinson et al., 2016). Elle est aussi l'une des briques fonctionnelles et méthodologiques du PLM (voir §I.4.1.5.) qui sera détaillée dans la section III.3.2. Mais dans ce paragraphe, nous nous focalisons sur la bibliographie en ce qui concerne l'interopérabilité entre KOS.

Selon (Patel et al., 2005), l'interopérabilité peut être définie comme « la capacité de différents KOS à communiquer de manière à préserver le sens ou la « signification voulue » ». Il définit deux types d'interopérabilité :

- Interopérabilité syntaxique : elle permet l'échange de données, d'informations et de connaissances en mettant l'accent sur le format des données, les codages, les propriétés, les valeurs et les types.
- Interopérabilité sémantique : elle permet une compréhension correcte du langage, de la terminologie et des métadonnées utilisés.

L'interopérabilité syntaxique est au niveau physique de l'échange de données et elle est étroitement liée aux technologies d'implémentation des systèmes sources de données et aux langages de description des données. Elle vient après l'interopérabilité sémantique.

Les ontologies sont couramment utilisées comme une étape vers l'interopérabilité sémantique. En effet, l'interopérabilité est inhérente à leur définition « spécification formelle et explicite d'une conceptualisation partagée ». Le premier avantage de l'ontologie se trouve dans le mot-clé « formelle ». En effet une ontologie suit une logique formelle lors de sa conception et son développement. Elle est formée de classes, de propriétés et d'axiomes. Cela permet d'avoir des ontologies bien formées. Le deuxième avantage de l'ontologie est dans le mot-clé « explicite ». En effet, des ontologies bien formées permettent d'expliciter la sémantique des données dans un contexte donné et assure ainsi une compréhension précise. Le troisième avantage se trouve dans le mot-clé « partagée », puisque l'ontologie ne regroupe que des terminologies et des règles communes pour un sujet d'intérêt, ce qui facilite l'intercommunication et ainsi l'interopérabilité.

En ingénierie ontologique, plusieurs initiatives existent pour la mise en place d'alignements et interopérabilité entre ontologies. Nous citons les outils OntoAnimal[68] et les principes XOD (eXtensible Ontology Development) (He et al., 2018). Ils ont été proposés afin de faciliter la réutilisation partielle des ontologies publiées lors du développement de nouvelles ontologies pour les rendre interopérables. Par ailleurs, une méthodologie d'alignement composée a été proposée par (Oliveira & Pesquita, 2018) pour stimuler l'intercommunication entre les ontologies (i.e. l'interopérabilité). Sur un autre axe de recherche, le centre CEDAR (Center for Extended Data Annotation and Retrieval) (Martínez-Romero et al., 2017) fournit des outils ergonomiques aux chercheurs pour l'édition collaborative et la

[68] http://www.hegroup.org/ontozoo/

standardisation des métadonnées, ce qui permet de standardiser les KOS utilisés pour l'annotation et les aligner avec les standards du domaine (i.e. les rendre interopérables). Il s'agit d'une méthode basée sur les outils ergonomiques et l'engagement des utilisateurs finaux.

Plus sur le plan de construction d'ontologies, (Gangemi et al., 2002) encouragent l'utilisation d'une ontologie de haut niveau pour concevoir une ontologie, conceptuellement plus rigoureuse, cognitivement transparente et efficacement exploitable, donc une ontologie plus interopérable sur le plan sémantique. En outre, (Patel et al., 2005) estiment que l'interopérabilité sémantique est renforcée par l'utilisation d'une approche de construction d'ontologie à trois niveaux (top/core/domain ou haut/noyau/domaine).

L'ontologie de haut niveau fournit le cadre général et l'engagement ontologique « représentation du monde », qui est partagé entre plusieurs ontologies la réutilisant. Le niveau noyau fournit les concepts généraux du large domaine en question et prépare le cadre sémantique pour le niveau suivant. Le niveau domaine est le niveau le plus riche sémantiquement et en nombre de concepts puisqu'il représente les connaissances du domaine précis d'un contexte donné. (Patel et al., 2005) donnent également deux exemples dans le domaine du multimédia, d'une ontologie noyau (Chakravarthy et al., 2009) et d'un Framework (Hunter, 2003) qui ont été proposés selon l'approche haut /noyau/domaine, et ont été évalués positivement pour l'amélioration de l'interopérabilité sémantique.

CONCLUSION INTERMÉDIAIRE

Dans cette section III.1, nous avons présenté la pyramide de la connaissance et les liens entre les données et la compréhension, ainsi que ceux entre les signes utilisés et le concept signifié via le triangle sémiotique. Nous avons présenté l'échelle sémiotique pour montrer les différents niveaux mis en jeu lors de l'interprétation des signes physiques ainsi que l'importance des niveaux : social et pragmatique. Nous avons exploré les trois domaines KM, KO, et KE et leurs apports en matière de : (1) modélisation des connaissances via des KOS et des ontologies, et (2) exploitations des connaissances formalisées via les systèmes KBS et KMS. Nous avons, enfin, étudié les facteurs de la mise en place de l'interopérabilité sémantique dans ce contexte et nous avons retenu la méthode de construction d'ontologie en trois niveaux haut/noyau/domaine et la mise en place d'alignement entre ontologies comme solution pour l'interopérabilité sémantique. Les deux prochaines sections III.2 et III.3, seront consacrées au lien entre la gestion des connaissances et les deux autres domaines principaux de la thèse à savoir : la recherche biomédicale et la gestion de cycle de vie des produits (PLM).

Le domaine biomédical est plus avancé par rapport au domaine du PLM dans l'utilisation et le développement des terminologies, standards, vocabulaires, ontologies et plus globalement la mise en œuvre de KOS pour l'annotation des données. Ceci est dû en partie à l'histoire des deux domaines puisque le PLM est assez récent en industrie. Aussi, le domaine biomédical a été toujours riche en terminologies complexes ce qui a motivé assez tôt la mise en place de la standardisation entre professionnels en biologie et en médecine.

III.2. GESTION DES CONNAISSANCES ET RECHERCHE BIOMÉDICALE

Nous allons décrire dans cette section les différents KOS en lien avec notre contexte à savoir « la gestion des données de recherche biomédicale et de leur provenance tout au long du cycle de vie d'une étude de recherche ». Nous allons présenter tout d'abord les types de KOS que nous avons identifié. Ensuite, nous allons explorer les KOS qui traitent spécifiquement la provenance et ceux qui traitent spécifiquement la gestion des données de recherche. Ce dernier est un domaine étendu, les KOS explorés sont plutôt des KOS noyau (cf. III.1.3.3.2).

Deux portails en accès web constituent une référence en ce qui concerne les KOS pour la recherche biomédicale.

- Le portail web BioPortal (Whetzel et al., 2011) qui fournit un accès à une bibliothèque numérique de KOS biomédicaux via les web services du Centre National des Ontologies Biomédicales (NCBO) financé par le NIH (National Institute of Health). Il est accessible à cette adresse : http://bioportal.bioontology.org.
- Le portail BioSharing renommé récemment en FAIRsharing (Sansone et al., 2019) est un effort de mise en catalogue des ressources terminologiques, de bases de données partagées, et de guidelines en science du vivant pour la standardisation et l'annotation des données dans le but de les partager et les réutiliser. Il a été lancé en 2011, en collaboration avec le groupe de travail « *Research Data Alliance RDA/Force11* » et en collaboration avec l'« *International Society for Biocuration* ». Il est accessible à cette adresse : https://fairsharing.org/

III.2.1. KOS DE DOMAINE POUR L'ANNOTATION DES DONNÉES DE RECHERCHE BIOMÉDICALE

Deux grandes catégories de KOS sont utilisées pour annoter les données en recherche biomédicale : Les KOS publiés et les KOS locaux. En effet, comme les études de recherche biomédicale sont en constante évolution, des KOS locaux sont souvent utilisés en plus des KOS publiés. Effectivement, les terminologies locales sont référencées à plusieurs reprises dans la bibliographie (Vreeman & McDonald, 2005) (Daniel-Le Bozec et al., 2007) (Zunner et al., 2012).

III.2.1.1. KOS publiés

Les KOS publiés sont ceux reconnus par la communauté d'experts d'un domaine. Par exemple en imagerie médicale, le standard DICOM est le KOS de référence pour l'annotation des images biomédicales. Le Tableau 11 ci-après liste quelques KOS publiés et reconnus, leur type et les spécialités en recherche biomédicale qu'ils couvrent.

Tableau 11 Liste de KOS pour l'annotation de données en recherche biomédicale

Sigle du KOS	Nom du KOS	Type	Domaine
QIBO	Quantitative Imaging Biomarker Ontology	Ontologie	Biomarqueurs d'Imagerie Biomédicale
RadLex	Radiology Lexicon	Lexique	Radiologie
DICOM	Digital Imaging and Communications in Medicine	Taxonomie	Imagerie médicale
OME	Open Microscopy Environment	Modèle de données	Imagerie biologique
SNOMED-CT	Systematized Nomenclature of Human and Veterinary Medicine - Clinical Terms	Terminologie	Médecine
CIM ou ICD 10	Classification Internationale des maladies	Terminologie	Maladies
HPO	Human Phenotype Ontology	Ontologie	Phénotype des maladies
LOINC	Logical Observation Identifiers Names & Codes	Terminologie	Examens et soins cliniques
BioLOINC	La partie bio de LOINC	Terminologie	Analyses de laboratoire
IOBC	Interlinking Ontology for Biological Concepts	Ontologie	Biologie
GO	Gene Ontology	Ontologie	Génétique
MS	Mass Spectrometry Ontology	Ontologie	Protéomique
PRO	Protein ontology	Ontologie	Protéomique
FBbi	Biological Imaging Methods Ontology	Ontologie	Imagerie biologique
NCIT	National Cancer Institute Thesaurus	Thesaurus	Recherche sur le Cancer
MESH	Medical Subject Headings	Vocabulaire contrôlé	Publications en sciences du vivant

OBI	Ontology for Biomedical Investigation	Ontologie	Expérimentation biomédicale
HL7	Health Level 7	Standard, norme	Système d'information hospitalier
ontoVIP	Virtual Imaging Platform ontology	Ontologie	Imagerie médicale
NMR-CV	Nuclear Magnetic Resonance-Controlled Vocabulary	Vocabulaire contrôlé	Imagerie IRM
NIFSTD	Neuroscience Information Framework Standard ontology	Ontologie	Neuroscience
BIRNLEX	Biomedical Informatics Research Network Project Lexicon	Lexique	Bio-informatique
NDDO	Neurodegenerative Disease Data Ontology	Ontologie	Neurologie
DDI	Ontology for Drug Discovery Investigations	Ontologie	Médicaments
EGO	EpiGenome Ontology	Ontologie	Génétique
MedRA	Medical Dictionary for Regulatory Activities	Dictionnaire	Pratique médicale
EDAM	bioinformatics operations, types of data, data formats, identifiers, and topics	Ontologie	Bio-informatique et biologie
HUPSON	Human Physiology Simulation Ontology	Ontologie	Physiologie
ERO	eagle-i resource ontology	Ontologie	Ressources pour la recherche
ONTOAD	Bilingual Ontology of Alzheimer's Disease and Related Diseases	Ontologie	Neurologie
LABO	clinical LABoratory Ontology	Ontologie	Tests de laboratoire
SIO	Semanticscience Integrated Ontology	Ontologie	Recherche scientifique
ONL_MR_DA	Magnetic Resonance Dataset Acquisition Ontology	Ontologie	Acquisition de données IRM

Toutes ces KOS sont consultables sur le portail BioPortal. Nous avons listé celles que nous avons identifiées lors de nos recherches de standards pour notre contexte d'application : la recherche préclinique.

III.2.1.2. KOS locaux

Les KOS locaux vernaculaires sont de différentes formes : une norme qui a été modifiée pour répondre aux besoins d'une étude, une liste de vocabulaire contrôlée partagée entre collègues de laboratoire via des fichiers Excel, des termes d'une ontologie de domaine associés à d'autres termes maison et liés au contexte, etc.

Dans la bibliographie, plusieurs noms sont attribués au « KOS locaux » : terminologies locales, terminologies de laboratoire, terminologies d'interface, etc. (Daniel-Le Bozec et al., 2007) défend l'idée que les KOS locaux ou terminologies locales sont irremplaçables par les KOS publiés ou terminologies de référence. Les KOS locaux selon lui fournissent de la matière pour les terminologies d'interface, terminologies qui sont principalement utilisées pour les interfaces graphiques (GUIs) ou les Interfaces Homme Machine (IHMs) des outils du laboratoire et de l'hôpital. Les terminologies d'interface constituent un sous-ensemble des terminologies locales. Tous doivent être mappés aux terminologies de référence (ou KOS publiés) afin de pouvoir profiter mutuellement de leurs avantages.

(Rosenbloom et al., 2006) souligne le rôle des terminologies d'interface en étant un lien entre les termes locaux utilisés dans les rapports de soins, et les terminologies de référence codée en une structure bien définie dans les programmes et outils informatiques. Elles sont entre autres appelées terminologies jargonneuses, terminologies d'application et terminologies d'entrée.

(Schulz et al., 2017) parle d'écosystème de terminologies où l'on trouve des terminologies d'interfaces, ceux de références et ceux d'agrégation. Les terminologies d'interface sont celles que nous avons appelé KOS locaux. Ceux de référence et ceux d'agrégation sont les KOS publiés. La différence entre les deux consiste à avoir la composante « règles et contraintes » en plus, chez les terminologies d'agrégation. Les

auteurs citent SNOMED-CT comme exemple de terminologie de référence et ICD-9, 10, comme exemple de terminologie d'agrégation. Les auteurs soulignent entre autres l'importance des efforts terminologiques locaux et « *bottom-up* ». Nous retenons l'importance des KOS locaux qui permettent d'enrichir les terminologies de référence ou d'agrégation.

III.2.2. KOS POUR L'ANNOTATION DE LA PROVENANCE DES DONNÉES SCIENTIFIQUES

Le modèle de données PROV-DM (PROV-DM) (Belhajjame et al., 2013) et son ontologie correspondante PROV-O (Lebo et al., 2013) sont deux KOS de provenance proposés par le consortium World Wide Web (W3). Selon PROV-DM, la provenance est définie « comme un enregistrement qui décrit les personnes, les institutions, les entités et les activités impliquées dans la production d'une donnée ou d'une chose dans le monde »(Belhajjame et al., 2013). Une « entité » peut être physique, numérique ou conceptuelle. Une « activité » se déroule au cours d'une période de temps et agit sur une ou plusieurs entités. Un « agent » est responsable de l'exécution d'une activité. Les entités, les activités et les agents sont les trois principaux concepts de PROV-DM et sont modélisés par un ensemble de relations. PROV-O (Lebo et al., 2013) est l'ontologie associée à PROV-DM et est actuellement identifiée comme l'ontologie de provenance de référence sur le web. D'autres ontologies existent, telles que PAV (Ciccarese et al., 2013), mais elle est plus adaptée à la création de documents et au versionnage.

La provenance est définie dans (Simmhan et al., 2005) comme l'information sur la nature d'une donnée ; quand, où et comment a-t-elle été produite ; pourquoi et pour qui a-t-elle été exécutée. Les informations sur la provenance sont nécessaires pour pouvoir reproduire, réutiliser et partager les résultats scientifiques. Pour les collecter, les questions en cinq points « *Five Ws* : *Who ? do What ?, Where ?, When ? and Why ?* » ont été proposées comme outil efficace pour la collecte d'informations sur la provenance (Sahoo et al., 2008) et (Ding et al., 2010). QQOQCCP est leur équivalent français : Qui ? Quoi ? Où ? Quand ? Comment ? Combien ? Pourquoi ?

La provenance des études biomédicales est complexe (Allanic et al., 2017). Les informations sur la provenance doivent être associées aux données générées au cours d'une étude, de (1) sa spécification à (4) ses résultats publiés. La provenance de (2) l'acquisition de données et (3) de leur analyse comprend les dispositifs et leur configuration, les algorithmes utilisés, les paramètres, les outils, etc.

III.2.3. KOS NOYAU POUR L'ANNOTATION DES DONNÉES DES ÉTUDES BIOMÉDICALES

Dans la bibliographie, plusieurs Frameworks pour l'annotation et la gestion des données de recherche biomédicale ont été proposés. Dans cette section, huit d'entre eux sont présentés : le modèle conceptuel de la suite logicielle ISA [46], les listes minimales du projet MIBBI (Taylor et al., 2008), l'ontologie pour les investigations biomédicales OBI (the Ontology for Biomedical Investigation), et quatre ontologies noyau pour la description des expériences scientifiques ; CSMO (Brahaj et al., 2012), EXACT (Soldatova et al., 2008), SMART Protocols (SP) (Giraldo et al., 2017) et SECO (Aloulen et al., 2019).

La suite logicielle ISA (Rocca-Serra et al., 2010) propose un modèle à trois concepts (`Investigation`, `Study`, `Assay`) ainsi qu'une variété d'outils (ISACreator, ISAConfigurator, ISAValidator...) pour le l'annotation des investigations biologiques et moléculaires par les chercheurs. Récemment, le format de référence d'ISA, ISA-Tab, a été converti en RDF selon l'approche LinkedISA en utilisant les standards du web sémantique (González-Beltrán et al., 2014).

Le projet MIBBI (Taylor et al., 2008) est un autre effort de gestion de données de recherche biologique et biomédicale. Il vise à enrichir les métadonnées avec des listes d' « information minimale » et à former les chercheurs pour mieux les impliquer dans les initiatives de gestion des données de recherche (Sampaio et al., 2019).

L'ontologie pour les investigations biomédicales (OBI) est une ontologie de domaine qui spécialise l'ontologie fondamentale BFO. Elle a été proposée pour couvrir la sémantique des expériences et des tests biologiques. L'OBI couvre toutes les phases d'un processus d'investigation, telles que la planification, l'exécution et la rédaction de rapports.

(Brahaj et al., 2012) ont proposé le Core of Scientific Metadata Ontology (CSMO) pour l'annotation les données scientifiques par des informations contextuelles telles que le matériel, l'institution, les personnes, les logiciels, l'étude, la recherche ou l'expérience, les données résultantes et la publication correspondante.

L'ontologie EXACT (EXperiment ACTions) a été proposée (Soldatova et al., 2008), après un examen des protocoles publiés, pour décrire efficacement et sans ambiguïté les protocoles expérimentaux, et pour renforcer leur fiabilité et leur reproductibilité. Des travaux récents fournissent un Framework pour la conversion des protocoles du format de langage naturel au format défini sémantiquement (Soldatova et al., 2014), afin de permettre la réutilisation des protocoles par les humains et les machines.

Un effort similaire, pour simplifier et rendre explicite la sémantique des protocoles, est l'ontologie des protocoles SMART (SP) (Giraldo et al., 2017) et son modèle SIRO (`Sample`, `Instrument`, `Reagent`, et `Objective`) d' « information minimale » associé.

Récemment, (Aloulen et al., 2019) ont proposé une méthodologie simple et indépendante du domaine et une ontologie noyau appelée SECO (Scientific Experiments Core Ontology) pour l'annotation des données du LIMS (Laboratory Information Management System). Elle réutilise les ontologies de domaine et est basée sur l'ontologie fondamentale BFO.

Il est intéressant de constater qu'au niveau sémantique, il y a des efforts de structuration et gestion des connaissances indépendamment des domaines : chose que nous n'avons pas trouvée aux niveaux systèmes de gestion de données existants (présentés en section II.2.) Les KOS identifiés ne couvrent pas toutefois tout le cycle de vie d'une étude de recherche et n'ont pas tous une considération pour la provenance des données et pour l'interopérabilité sémantique.

III.3. APPLICATIONS DE LA GESTION DES CONNAISSANCES DANS LES SYSTÈMES PLM

La gestion des connaissances s'inscrit dans la continuité de la gestion des données et informations des produits. Elle est une des briques méthodologiques du PLM présenté dans la section I.4.1.5. En effet, les connaissances représentent l'expertise métier qui, une fois acquise, permet de réduire les coûts d'une action et d'améliorer les performances globales. La gestion des connaissances d'un produit dans une démarche PLM couvre tout ou partie du cycle de vie : analyse des besoins, conception, industrialisation, fabrication et assemblage, logistique, utilisation et fin de vie (recyclage). Elle se traduit par la proposition de modèles et de métamodèles de données et ontologies pour la structuration des données et informations des produits (Le Duigou, 2010). Ces modèles et ontologies servent à façonner les systèmes PLM et les faire évoluer vers des Systèmes à Base de Connaissances (KBS).

III.3.1. KOS ET ONTOLOGIES POUR LA GESTION DU CYCLE DE VIE DES PRODUITS

(Kadiri & Kiritsis, 2015) proposent un état de l'art explicitant les principaux rôles des ontologies dans les systèmes PLM et qui a été repris dans (Le Duigou, 2017). Une ontologie peut être utilisée pour collecter le savoir et savoir-faire, commun et vérifié, d'un ensemble d'acteurs (humain ou machine) du cycle de vie des produits. Elle permet aussi de faire le lien entre différents domaines et d'assurer l'interopérabilité entre systèmes. L'ontologie facilite aussi la recherche contextuelle.

Plusieurs ontologies ont été proposées dans le cadre de la gestion de cycle de vie des produits : L'ontologie noyau « closed-loop PLM », l'ontologie produit (PO), l'Engineering Design Ontology (EDO), le Semantic Object Model (SOM), le Core Product Model (CPM), ONTO-PDM, Onto-STEP, OntoSTEP-NC. Dans ce qui suit, une présentation générale est donnée pour chacune : étape couverte du cycle de vie, niveau de généralité ou de spécificité, les concepts principaux, et les cas de validations/applications utilisés.

- **Ontologie noyau « Closed-loop PLM »** : (Jun et al., 2007) proposent une ontologie noyau pour représenter les données méta des produits (indépendamment du domaine d'application) tout au long du cycle de vie et plus précisément dans un contexte de closed-loop PLM (§I.4.1.2). Cette ontologie est basée sur une classification des connaissances du cycle de vie selon quatre points de vue : selon le niveau d'abstraction (des données de domaine vs des données méta) ; selon le contenu des données (`Produit`, `Processus`, `Ressource`, `Agent`, `Temps`) ; selon la phase du cycle de vie (BOL, MOL, EOL) ; et selon la faculté de changement (des données dynamiques ou statiques). L'ontologie a été appliquée à la maintenance d'une automobile. Dans la même logique de recherche, (Fortineau et al., 2013) proposent 5 concepts de base représentant tous les éléments entourant un produit tout au long de son cycle de vie : `Produit`, `Processus`, `Resource`, `Règles` et `Business`. Ces concepts représentent la base d'une ontologie noyau, indépendante du domaine d'application. Ils ont été appliqués au métamodèle d'une centrale nucléaire (le produit).
- **L'ontologie produit (PO)** : (Lee & Suh, 2007) proposent l'ontologie produit (PO) et un cadre de développement ontologique à trois niveaux : méta-, générique et spécifique. Ceci correspond, en terminologie ontologique, aux trois niveaux : haut, noyau et domaine. L'ontologie PO ne couvre pas tout le cycle de vie d'un produit et est limitée à la phase BOL. Elle a été appliquée à la conception d'un système mécanique de purification d'air. Ses concepts noyau sont répartis selon les deux concepts fondamentaux : `Objet` et `Processus`, détaillés dans la Figure 61.

Concepts in meta-PO	Primitive concepts in generic PO
Object	Requirement, Function, Failure, Cause, Part, Form, Energy, Information, Solid Geometry, Geometry Feature, Sketch, Material, Manufacturing Machine, Manufacturing Machine Tool.
Process	Method, Behavior, Operation, State, Transition, Precondition, Manufacturing Process.

Figure 61 Les concepts primitifs de l'ontologie Product Ontology (PO) (Lee & Suh, 2007)

- **EDO (Engineering Design Ontology)** : (Zhang & Yin, 2008) ont proposé d'appliquer les ontologies dans un environnement de conception distribuée à l'aide des systèmes multiagents[69]. Il en résulte une architecture d'un KBS qui intègre les connaissances métier dès leur acquisition jusqu'à leur réutilisation. Dans ce contexte, ils proposent un cadre modulaire de conception d'ontologies pour la conception produit (voir Figure 62). Du haut en bas, les modules vont du général (generic product design ontology) au spécifique : (material ontology), (Quality standard ontology), etc.

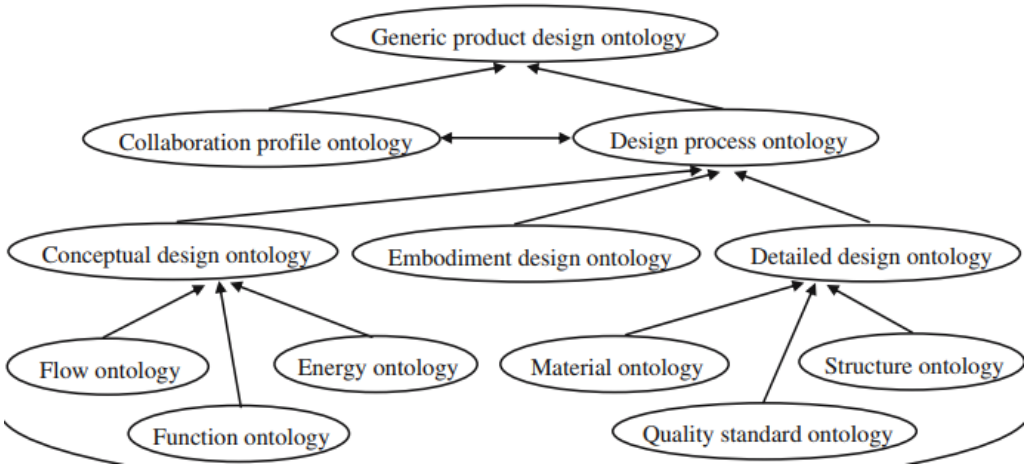


Figure 62 Les différents modules d'ontologies de la EDO (Engineering Design Ontology) (Zhang & Yin, 2008)

- **SOM (Semantic Object Model)** : Dans le cadre du projet Européen PROMISE FP6-IST (PROduct lifecycle Management and Information tracking using Smart Embedded systems) (Cassina et al., 2009), le modèle de données SOM (Semantic Object Model) a été proposé. Il a été conçu en UML et contient 26 classes génériques (indépendantes de l'application). Ces classes sont réparties principalement en classes « as-designed » de description du produit à son début de vie (BOL), et des classes qui concernent les autres phases du cycle de vie et les différentes activités et données qui y sont associées. Le projet PROMISE comprend dix scénarios d'application qui couvrent chacun, tout ou partie du cycle de vie. (Cassina et al., 2009) ont présenté une mise en œuvre dans les cycles de vie des réfrigérateurs. Plus récemment, le modèle de données SOM a été transformé en une ontologie noyau pour le PLM qui a été appliquée à l'automobile en milieu de vie (Matsokis & Kiritsis, 2010).
- **Core Product Model** : Le CPM (Core Product Model) (Fenves et al., 2008) et son extension OAM (Open Assembly Model) forment un modèle générique de données qui permet de modéliser les informations produit de tout le cycle de vie. Il a son origine au NIST (National Institute of Standards

[69] « Un système multiagent est un système composé d'un ensemble d'agents (un processus, un robot, un être humain, etc.), qui interagissent entre eux. Un agent est une entité autonome ou partiellement autonome »

and Technology). Il est basé sur trois vues principales : la fonction, la forme et le comportement d'un produit. (Fiorentini et al., 2007) propose la version ontologie du CPM et son extension OAM.

- **ONTO-PDM** : ONTO-PDM ou « Product ontology » (Panetto et al., 2012) est une ontologie compatible avec les normes (ISO 10303, 1994) et (IEC 62264-1, 2003). Elle est axée sur l'interopérabilité sémantique. Elle réutilise les 8 concepts principaux de (IEC 62264-1, 2003), à savoir : `Définition du produit`, `matériel`, `équipement`, `personnel`, `segment de processus`, `calendrier de production`, `capacité de production` et `performance de production`**.** Elle est plutôt proposée pour la phase de l'industrialisation, la fabrication et l'assemblage d'un produit sur des sites distants et en présence de données hétérogènes. Sa validation a été faite via une simulation des activités distribuées, nécessaires à la fabrication d'un prototype de produit simple.
- **Onto-STEP** : STEP, Standard for the Exchange for Product model data, (ISO 10303, 1994), est le standard le plus largement utilisé pour l'échange de données tout au long du cycle de vie du produit, voir Figure 63 (Rachuri et al., 2008). Onto-STEP, l'ontologie associée, a été proposée par (Barbau et al., 2012). Elle se base sur des modèles de données de la norme STEP (AP203, 214 et 239), mais aussi CPM et OAM du NIST. Elle étend aussi STEP avec des informations non géométriques à savoir les fonctions, les exigences, et le comportement.

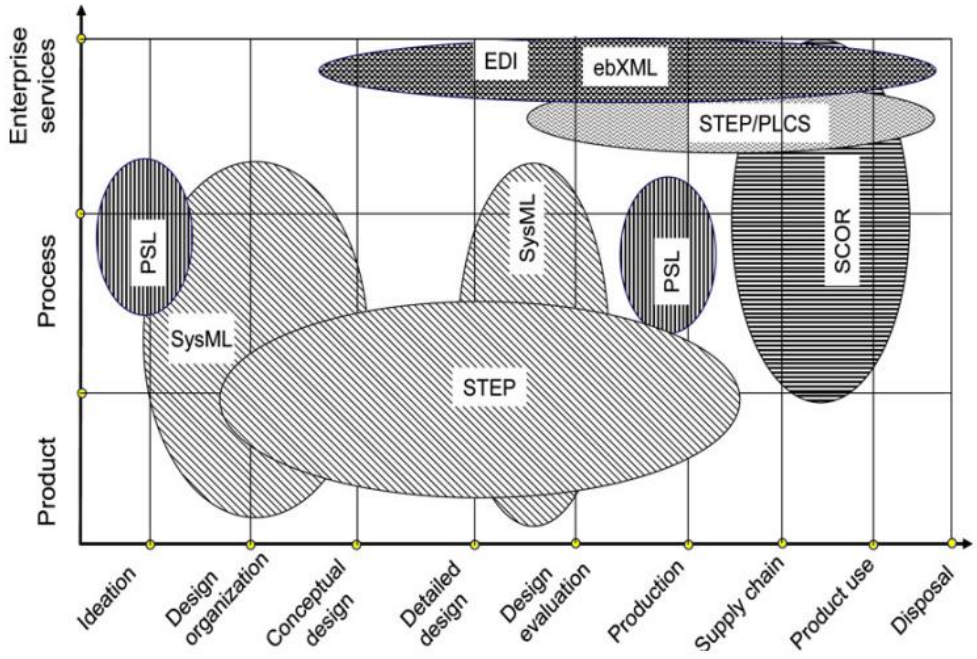


Figure 63 Liste des standards PLM juste après la sortie de STEP et avant l'arrivée des ontologies (Rachuri et al., 2008)

- **OntoSTEP-NC** : OntoSTEP-NC est l'ontologie associée au standard STEP-NC(Danjou et al., 2017). STEP-NC est l'application des méthodes proposées par le standard STEP à l'usinage (NC machines). STEP-NC est l'abréviation de « STEP Data Model for Computerized Numerical Controllers ». OntoSTEP-NC permet de mettre en lien les informations de fabrication depuis les machines d'usinage jusqu'aux systèmes de CAO (Conception Assistée par Ordinateur) et FAO (Fabrication Assistée par Ordinateur) et inversement. Il s'agit du « closed-loop manufacturing ».

En synthèse, les ontologies que nous avons identifiées sont principalement des ontologies noyau. Ceci s'explique par le fait que la gestion de cycle de vie est une notion fondatrice appliquée à plusieurs domaines en industrie (cf. § III.1.3.3.2). Les ontologies identifiées aussi ne se basent pas sur une ontologie de haut niveau et sont quelques fois centrées sur une étape du cycle de vie : la fabrication, le BOL, etc. Nous notons aussi l'existence d'efforts de mise en place d'une interopérabilité sémantique

entre standards et modèles comme Onto-STEP et ONTO-PDM et l'adoption d'une approche modulaire comme dans EDO.

III.3.2. INTEROPÉRABILITÉ DES SYSTÈMES PLM

Ce paragraphe est une description étendue de la brique interopérabilité de la section I.4.1.5. L'interopérabilité est la capacité d'intercommunication entre deux systèmes. (IEEE, 1991) l'a défini comme « La capacité pour plusieurs systèmes ou composants à échanger des informations et à utiliser les informations échangées ».

Plusieurs niveaux d'interopérabilité peuvent être identifiés entre les plateformes PLM et les systèmes environnants, ou aussi entre les différentes phases du cycle de vie, selon (Danjou, 2015), en s'appuyant sur la norme (ISO 14258, 1998):

- Le niveau sémantique : par exemple, l'utilisation unifiée d'un standard comme STEP ou STEP-NC dans sa version ontologique pour permettre les échanges de connaissances entre les systèmes (Danjou, 2015).
- Le niveau technique : par exemple, l'intégration des outils logiciels externes dans la plateforme PLM ou le développement de connecteurs entre les logiciels CAO classiques et les plateformes PLM, ou encore, des connecteurs génériques (Penciuc et al., 2014)
- Le niveau organisationnel : par exemple, la mise en place de méthodologies, processus sous forme de workflow collaboratif d'échange d'informations entre les différentes équipes au sein d'une même entreprise. Il s'agit d'un niveau peu traité dans la littérature, puisque dépendant du facteur humain (D. Chen et al., 2008).

L'interopérabilité dans le contexte de l'entreprise étendue rencontre plusieurs barrières et fait ainsi l'objet d'étude de plusieurs Frameworks de modélisation. L'EIF (Entreprise Interoperability Framework) propose une modélisation de trois dimensions qui est centrée sur les freins pour l'interopérabilité (voir Figure 64). L'EIF est le résultat des travaux du réseau d'excellence INTEROP de 2003 à 2007[70].

La Figure 64 résume les trois composantes de l'interopérabilité à savoir : les barrières à l'interopérabilité (Conceptuelle, Technologique et Organisationnelle), les niveaux concernés par l'interopérabilité dans l'entreprise (Business, Processus, Service et Donnée) et les approches de résolution abordée (Intégrée, Unifiée, et Fédérée).

[70] http://interop-vlab.eu/interop/

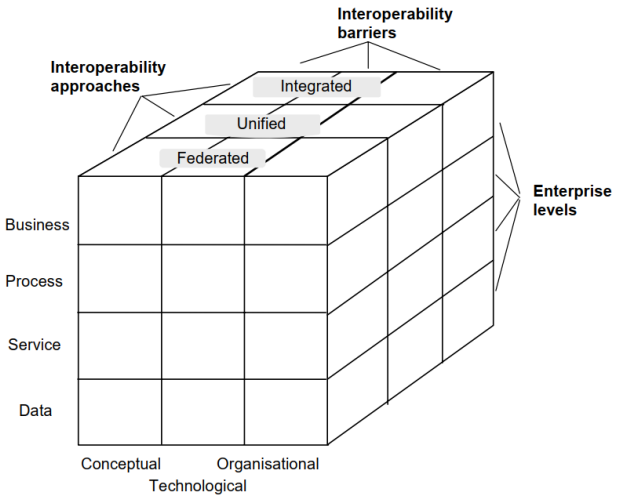


Figure 64 Les trois dimensions du cadre d'interopérabilité en entreprise ou EIF (source : https://en.wikipedia.org/wiki/Enterprise_interoperability_framework)

Le framework EIF ainsi que la norme (ISO 14258, 1998) présentent trois différentes approches d'interopérabilité en industrie : (1) l'intégration (2) l'unification et (3) la fédération.

- **L'intégration** permet de modéliser toutes les données via un seul format de données et de les manipuler via ce seul format commun. Ceci masque l'hétérogénéité des données et facilite un traitement transparent pour l'utilisateur. Le résultat donne des systèmes fortement interdépendants.
- **L'unification** est une démarche plus flexible et propose d'unifier le modèle à un haut niveau en laissant libre le choix du modèle pour chaque système inter-opérant. L'utilisation des standards à haut niveau fournit une des implémentations possibles de cette approche.
- **La fédération** exclut le recours au format commun et permet à chaque système d'avoir son propre modèle. L'interopérabilité s'effectue alors en utilisant les ontologies et les technologies du web sémantique. En effet, une ontologie en web sémantique utilise les standards OWL ou RDFS, qui sont interopérables au niveau technique et encodage. Ils permettent ainsi une aisance de mise en correspondance entre les différentes ontologies représentant les connaissances d'un système.

Dans le cadre de l'intérêt croissant pour de la mise en œuvre de l'interopérabilité sémantique dans le domaine industriel. L'initiative « Industrial Ontologies Foundry (IOF) » a vu le jour le 6 décembre 2016 via la création d'un groupe de travail pour le pilotage de ses activités et un workshop annuel pour faire avancer les discussions et ainsi créer des ponts entre la sémantique et l'industrie (https://www.industrialontologies.org/)

III.3.3. Systèmes à Base de Connaissances (KBS) et systèmes PLM

Les approches PLM à base d'ontologies et plus généralement avec la brique gestion de connaissances sont très variées et couvrent différentes étapes et tâches du cycle de vie produit. Ces travaux sont nombreux et nous avons présenté ici les plus représentatifs en relation avec nos objectifs de recherche. Dans cette section, nous citons quelques exemples qui sont centrés sur l'amélioration de l'échange de données, d'informations et de connaissances. Il est à noter que les KBS dans le domaine du PLM sont différents de ceux définis dans le domaine de l'ingénierie de connaissances. Parfois, KBS désigne un système à base d'ontologie qui n'est pas équipé d'un moteur de règles (cf. III.1.3.3.1)

(Ducellier, 2008) a proposé un système expert (SE) à base de règles pour enrichir les plateformes PLM. Il permet ainsi l'intégration des connaissances d'experts en conception, et leur échange à un niveau transversal entre les différentes phases du cycle de vie. La Figure 65 présente les éléments de cette architecture. Les connaissances sont modélisées dans « les règles liant les paramètres » et « la base de

données de paramètres ». Il est à noter que ce système est assez ancien et représente le début des applications des SE aux systèmes PLM.

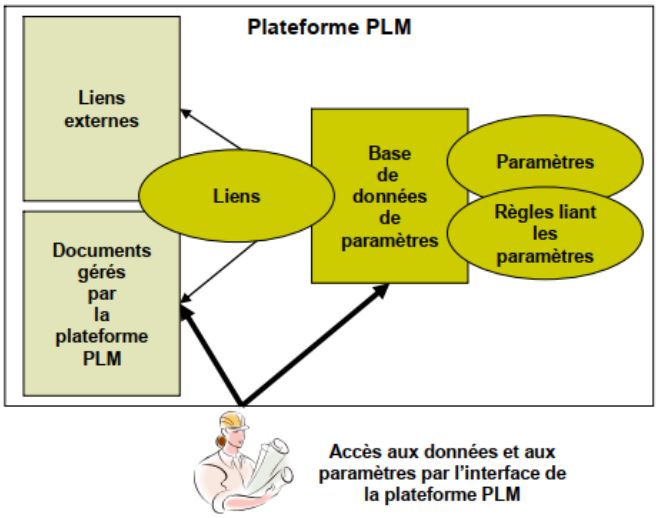


Figure 65 Plateforme PLM augmentée par les connaissances des experts (Ducellier, 2008)

(Sriti, 2008) a proposé la plateforme I-Semantec pour permettre une gestion, une modélisation et un partage de connaissances tout au long du cycle de vie d'un produit. Elle est basée sur les technologies du web sémantique et est décrite dans la Figure 66. Les principaux éléments de son architecture sont : la base de connaissances en RDF, les instances d'installations PLM sources de données, les différents extracteurs, ainsi que le réconciliateur d'ontologie et de modèle de données. La plateforme a été testée avec des données structurelles de nomenclature (BOM) extraites depuis le système PLM industriel Teamcenter de Siemens.

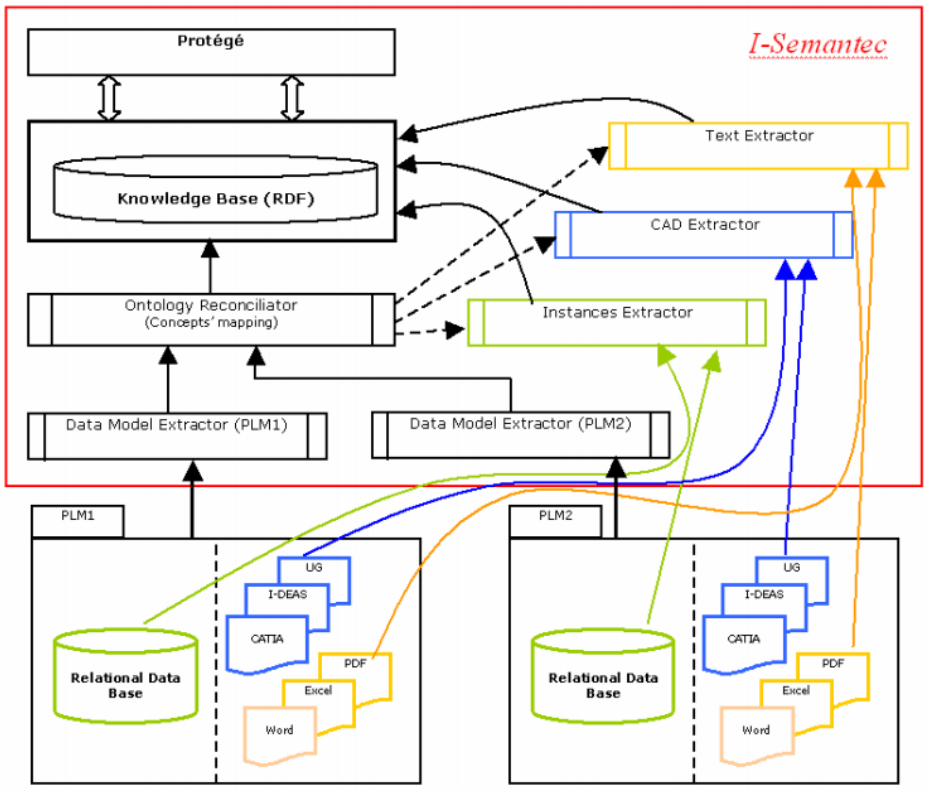


Figure 66 Plateforme I-Semantec de couplage entre PLM et connaissances (Sriti, 2008)

(Assouroko, 2012) a proposé un cadre méthodologique et des outils de visualisation des relations sémantiques (GdR : Gestionnaire de Relation) qui se basent sur les ontologies pour un meilleur déploiement des exigences relatives au produit. L'objectif est de capitaliser et partager l'ensemble des données, des informations, et des connaissances dans le cadre de processus de développement collaboratif des produits, principalement la conception mécanique et la simulation numérique. L'approche proposée a été appliquée au moteur thermique à 4 temps Wankel. La Figure 67 présente les différentes briques de ce cadre méthodologique. On y trouve l'interface graphique, la base de données qui est structurée à partir d'une ontologie des exigences métiers, et la couche serveur qui permet d'explorer les données en exploitant la base de connaissances fournie. Nous pouvons dire que cette couche, qui contient les traitements nécessaires au fonctionnement du système, se substitue au moteur d'inférence, troisième élément dans la définition d'une architecture KBS après l'interface graphique et la base de connaissances.

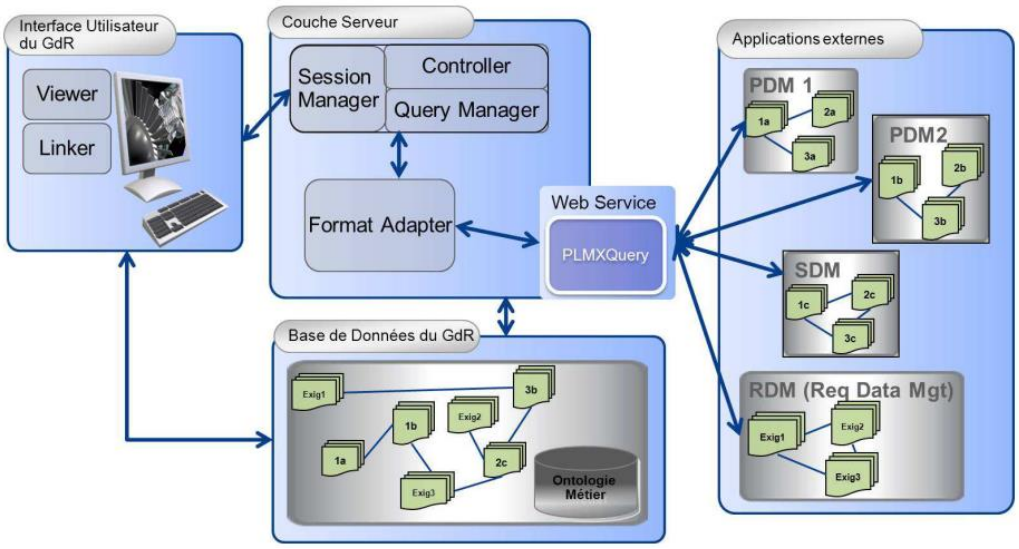


Figure 67 Architecture de la plateforme GdR (Assouroko, 2012)

(Y.-J. Chen et al., 2009) proposent un Framework basé sur les ontologies pour l'intégration des connaissances et expertise métier tout au long du cycle de vie des produits. L'objectif est de gérer le partage des données hétérogènes et multisites. Il est un système à base de connaissances qui permet la capitalisation des connaissances tout au long du cycle de vie dans un but de partage, d'exploration et de transfert de connaissances. Encore une fois, dans cet exemple aussi, le moteur d'inférence est remplacé par une couche d'intégration d'ontologies réalisée par un administrateur d'ontologies (voir Figure 68).

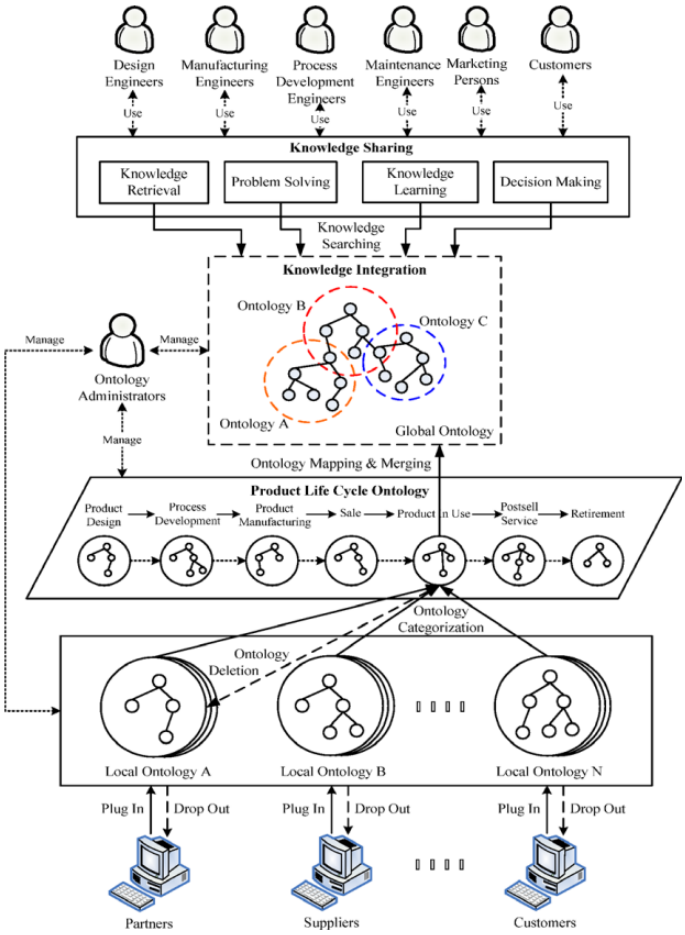


Figure 68 Framework pour la capitalisation des connaissances tout au long du cycle de vie d'un produit (Y.-J. Chen et al., 2009)

CONCLUSION DU CHAPITRE III

Dans ce chapitre, nous avons exploré la gestion des connaissances selon plusieurs angles de vue. Nous avons rappelé les fondements théoriques avec le DIKW, le triangle sémiotique et l'échelle sémiotique. La clarification des apports des domaines de KM, KO, et KE nous ont permis d'identifier : une dimension « pratique » de la gestion des connaissances avec les enjeux organisationnels au sein d'une entreprise, une dimension « conceptuelle » avec la modélisation de la connaissance via les KOS et en particulier les ontologies et une dimension « technologique » avec les méthodes et les systèmes (KMS, KBS) qui peuvent être utilisés et mise en place pour la gestion des connaissances.

Dans le reste du manuscrit, nous détaillons nos propositions dans chacune de ces dimensions (technologique, conceptuelle et pratique) pour la gestion des données, informations et connaissances tout au long d'une étude de recherche :

- La dimension « technologique » : au chapitre IV, nous décrivons nos méthodes pour la conception et la mise en œuvre (à partir de la plateforme BIOMIST) d'un système de gestion de données hétérogènes et de leur provenance dans le cadre d'une démarche BMS-LM (BioMedical Study-Lifecycle Management ; par analogie au PLM).
- La dimension « conceptuelle » : dans le chapitre V, nous expliquons nos méthodes de modélisation de connaissances pour une meilleure interopérabilité sémantique entre KOS locaux et KOS publiés. L'ontologie multi-niveaux qui en résulte est aussi présentée.

- La dimension « pratique » : au chapitre VI, nous détaillons la mise en œuvre du système de gestion y compris l'ontologie multi-niveaux pour la recherche préclinique via des cas d'applications et des exemples.

Chapitre IV. Proposition et méthodes de mise en œuvre du système BMS-LM pour la gestion de cycle de vie des études de recherche biomédicales

Dans le Chapitre II, une étude de plusieurs systèmes existants en gestion de données de recherche biomédicale a été présentée et a donné lieu au constat suivant : les systèmes de gestion de données sont conçus pour un type de données ou un domaine spécifique et ne sont pas généralisables à d'autres typologies de données et domaines. De plus, aucun système proposé dans la bibliographie ne permet de partager et réutiliser les données scientifiques dans le cadre de la « science ouverte ». Notre hypothèse est que la traçabilité de la provenance des données tout au long du cycle de vie est une étape importante pour leur partage et leur réutilisation ultérieure en renforçant la confiance dans les données. Elle est en lien avec nos objectifs de recherche 1 et 2 (voir Figure 69).

Ce chapitre présente notre proposition pour la mise en œuvre d'un système de gestion de données hétérogènes et de leur provenance dans le cadre d'une démarche de « gestion de cycle de vie des études biomédicales » ou BMS-LM (BioMedical Study-Lifecycle Management ; par analogie au PLM). Notre point de départ est la plateforme BIOMIST dans la continuité des travaux de (Allanic, 2015) et (Allanic et al., 2017). Nous avons proposé de faire évoluer le modèle BMI-LM d'origine en une deuxième version. Nous présentons l'architecture générale de ce système BMS-LM y compris le nouveau modèle de données. Nous détaillons ensuite les méthodes d'intégration de données et de calculs scientifiques que nous avons proposées pour sa mise en œuvre. Les expérimentations et résultats issus de notre proposition sont détaillés dans le Chapitre VI.

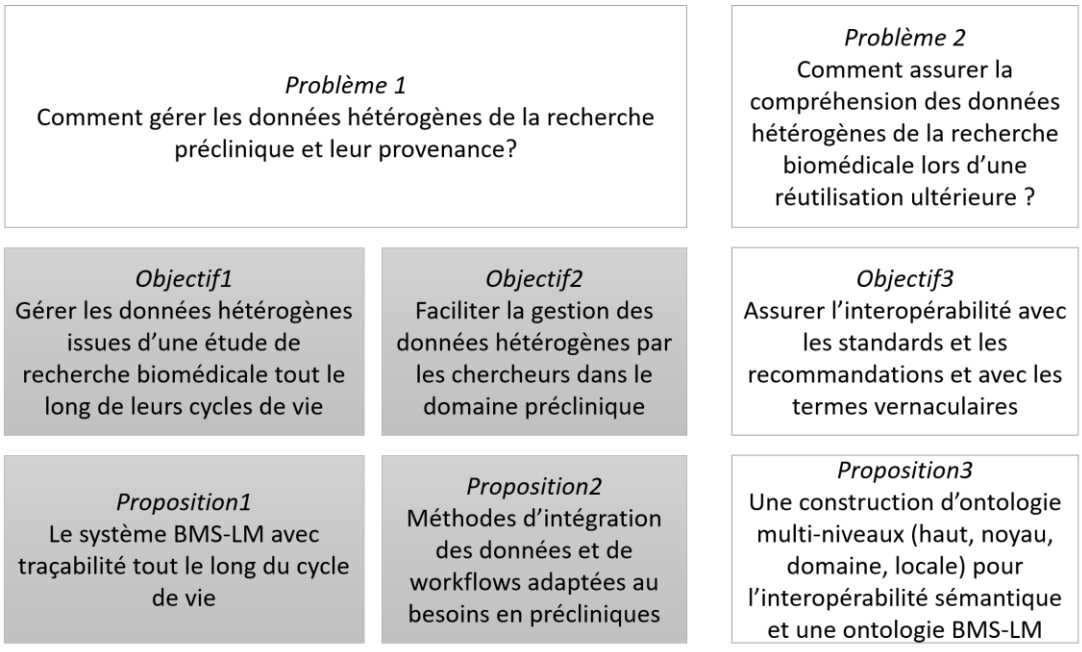


Figure 69 Objectifs de recherche pour la gestion des données hétérogènes

IV.1. BIOMEDICAL STUDY – LIFECYCLE MANAGEMENT (BMS-LM)

Le BMS-LM BioMedical Study – Lifecycle Management ou « gestion de cycle de vie des études de recherche » est un paradigme que nous avons introduit dans cette thèse et qui représente les résultats de l'application du PLM pour le domaine biomédical. Il constitue un parallèle du PLM pour la recherche biomédicale. Les travaux de (Allanic, 2015) et de (Pham, 2017) représentent les premiers efforts dans ce cadre. Notre thèse s'inscrit dans la continuité de ces travaux.

La « gestion de cycle de vie des études de recherche biomédicale » se définit comme :

Une démarche intégrée de gestion de données, informations, connaissances et de leurs provenances avec traçabilité de toutes les étapes d'une étude de recherche biomédicale (i.e. cycle de vie) : (1) la spécification, (2) l'acquisition de données brutes, (3) la production de données dérivées, et (4) la valorisation scientifique des résultats qui serviront à (1) la spécification d'autres études.

Dans cette section, nous présentons l'architecture du système BMS-LM que nous avons proposée ainsi que ses différents composants. Nous détaillons ensuite le modèle de données correspondant.

IV.1.1. SYSTÈME BMS-LM

Le système BMS-LM que nous avons proposé pour la gestion de données et de leur provenance est un système d'information (SI) de gestion de cycle de vie. Il consiste en une plateforme PLM qui a été adaptée à la gestion de données et d'informations du domaine biomédical. Les données sont intégrées dans ce système et gérées via ses modules dès leur création, jusqu'à leur publication et leur réutilisation. La traçabilité des données est mise en place à chaque utilisation du SI. Il s'agit d'une fonctionnalité inhérente au PLM.

À la suite de notre analyse bibliographique, nous avons identifié une liste de 19 besoins des chercheurs dans le domaine biomédical (voir Tableau 7 section II.1.1)) et de 4 leviers (voir Tableau 8 section II.1.1). Le schéma de la Figure 70 présente les différentes briques du système BMS-LM réparties sur tout le cycle de vie d'une étude de recherche qui inclut celui de la donnée scientifique. Ces briques seront présentées brièvement dans les paragraphes suivants en précisant chaque fois le(s) besoin(s) en lien avec chacune d'elle, une présentation plus détaillée se trouve en Annexe B :

1. Gestion documentaire (GED) et de Données Techniques (SGDT) : les fonctionnalités de base des plateformes PLM sont réutilisées pour le système BMS-LM : édition, versionnage, stockage dans les coffres-forts électroniques à long terme, interdiction de suppression, protection des données, réservation de ressources pour modification dans le cadre d'un accès concurrent, traçabilité des modifications, recherche de documents, intégration des logiciels de visualisation des contenus d'un document ou d'un fichier.
   Besoins : (B1) archivage, (B4) requête, (B8) traçabilité, (B10) automatisation, (B18) sécurité
2. Gestion des utilisateurs, des droits d'accès et des sites distants : ce sont des fonctionnalités inhérentes aux plateformes PLM. Cette brique permet un partage sécurisé entre les différents sites de recherche.
   Besoins : (B5) partage, (B7) réutilisation, (B18) sécurité,
3. Traçabilité : à chaque utilisation du système, les actions réalisées sont enregistrées sur le serveur et dans des fichiers de « log ».
   Besoins : (B8) traçabilité, (B5) partage, (B7) réutilisation,
4. Annotation des données par les informations de provenance : cette fonctionnalité est spécifique au BMS-LM. La provenance est une composante essentielle pour l'annotation des données scientifiques afin de renforcer la confiance dans les données. Afin de pouvoir l'utiliser, le système BMS-LM impose d'annoter les données avec les informations de provenance à toute étape du cycle de vie.

Besoins : (B5) partage, (B7) réutilisation, (B8) traçabilité, (B11) standardisation, (B17) reporting,

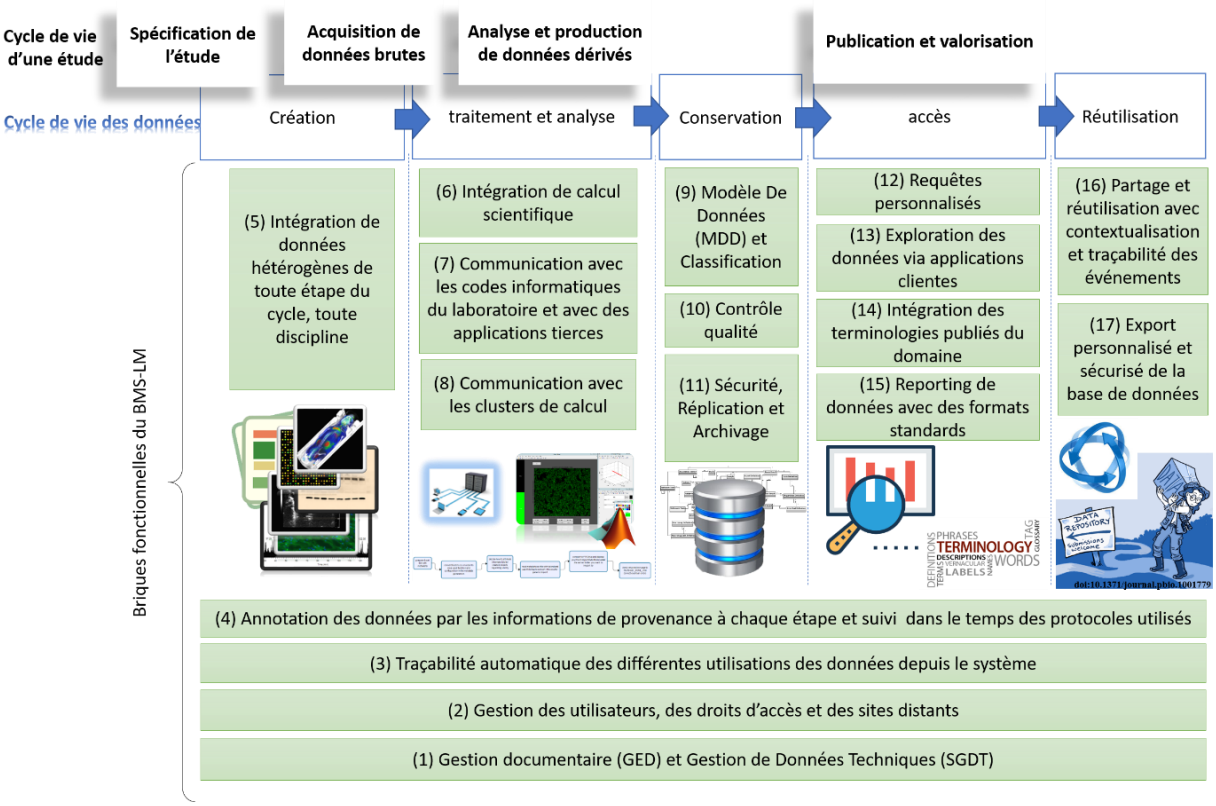


Figure 70 Briques fonctionnelles du système BMS-LM réparties sur tout le cycle de vie

5. Intégration de données hétérogènes : le système BMS-LM a été équipé d'un module d'import de données hétérogènes issues de la recherche biomédicale. Nous avons détaillé cette intégration dans la section IV.2.1
   Besoins : (B1) archivage, (B2) import, (B10) automatisation, (B11) standardisation, (B15) flexibilité, (B8) traçabilité, (B16) vérification,
6. Intégration de calculs scientifiques : le système BMS-LM propose deux méthodes d'intégration de calcul scientifique. Elles sont détaillées dans la section IV.2.2.
   Besoins : (B6) analyse, (B7) réutilisation, (B8) traçabilité, (B10) automatisation, (B11) standardisation, (B15) flexibilité,
7. Communication avec les codes informatiques et les applications tierces : le système BMS-LM a été équipé d'une API REST (Application Programming Interface) qui facilite sa communication avec les programmes informatiques « maison » pour le transfert de données et informations entre les deux.
   Besoins : (B6) analyse, (B7) réutilisation, (B8) traçabilité, (B10) automatisation, (B11) standardisation, (B15) flexibilité, (B4) requête,
8. Communication avec les clusters de calcul : le système BMS-LM dans sa version antérieure était déjà équipé d'un module de communication avec des machines distantes pour lancer des calculs scientifiques (des clusters ou autres).
   Besoins : (B6) analyse, (B7) réutilisation, (B8) traçabilité, (B10) automatisation, (B11) standardisation, (B15) flexibilité,
9. Modèle De Données (MDD) et Classification : les données manipulées via le système BMS-LM suivent un schéma de données représenté et géré avec le duo « MDD+Classification ». Leurs versions antérieures ont été présentées dans la section I.4.3.1. Nous les avons fait évoluer pour le système BMS-LM. Ce travail est expliqué en section IV.1.2.

Besoins : (B1) archivage, (B2) import, (B3) export, (B4) requête, (B5) partage, (B8) traçabilité, (B11) standardisation, (B14) évolutivité, (B15) flexibilité, (B17) reporting, (B19) suivi,

10. Contrôle qualité : des statuts ont été définis pour décrire la qualité des données brutes et dérivées du système BMS-LM. Ils peuvent être attribués manuellement via les interfaces graphiques de l'outil (Annexe A) ou automatiquement lors d'un import de données. Les statuts actuels sont : « complet » ou non, « valide » ou non, pour les données brutes, et « calculé » ou non, « utilisable » ou non, pour les données dérivées, et « erreur import » ou non. (Voir Annexe B).
    Besoins : (B16) vérification, (B5) partage, (B7) réutilisation,
11. Sécurité, réplication et archivage : ce sont des fonctionnalités classiques des plateformes PLM en général. Elles gèrent la sécurité, la réplication et l'archivage des données lors des accès distants et multisites avec un système de cache mémoire avancé et de manière transparente pour les utilisateurs finaux.
    Besoins : (B1) archivage, (B18) sécurité,
12. Requêtes personnalisées : deux méthodes sont en lien avec cette fonctionnalité. La première est la construction par un data manager des requêtes personnalisées à la demande et la deuxième est l'utilisation de termes et de graphes pour la formulation de la requête désirée par l'utilisateur lui-même (voir Annexe B).
    Besoins : (B4) requête, (B7) réutilisation, efficacité
13. Exploration des données via des applications clientes : un module d'exploration de données via Excel, ou via tout autre outil ayant un connecteur de bases de données « ODBC », a été mis en place pour la plateforme BIOMIST. Ce module a été repris pour le système BMS-LM. Par exemple, l'outil PowerQuery[71] d'Excel permet d'accéder à un export en SQL de la base de données du BMS-LM. D'autres scénarios d'explorations de la base sont possibles via l'utilisation de l'API du système BMS-LM ; un exemple est donné en Annexe B.
    Besoins : (B4) requête, (B3) export, (B9) simplicité, (B12) ergonomie, (B13) efficacité,
14. Intégration des terminologies publiés du domaine : le système BMS-LM utilise des terminologies standards du domaine pour l'annotation des données qu'il manipule. Cette fonctionnalité est étroitement liée à l'interopérabilité sémantique, objet du chapitre V.
    Besoins : (B11) standardisation, (B14) évolutivité, (B15) flexibilité, (B5) partage, (B7) réutilisation,
15. Reporting (génération de rapports) de données avec des formats standards : la génération des rapports est une fonctionnalité inhérente aux plateformes PLM. Ce reporting peut être adapté aux données de recherche biomédicale selon les besoins des utilisateurs du système BMS-LM. De plus, des rapports peuvent être générés à la demande pour le CLÉ.
    Besoins : (B17) reporting, (B19) suivi, (B10) automatisation, (B11) standardisation, (B8) traçabilité,
16. Partage et réutilisation avec contextualisation et traçabilité des événements : les plateformes PLM sont avant tout proposées pour la collaboration multisite et la gestion des versions des fichiers partagés. Des liens de références entre instances dans le système permettent de tracer l'historique de la donnée : qui l'a réutilisée ? pour quelle autre instance ? quand ? etc. Ces fonctionnalités sont réutilisées pour le partage des données scientifiques entre collaborateurs au sein d'un même projet de recherche et aussi pour la réutilisation ultérieure de ces données entre projets de recherche (réutilisation de données datant de plus de 20 ans possible et traçable via une plateforme PLM)
    Besoins : (B5) partage, (B7) réutilisation, (B8) traçabilité,
17. Export personnalisé et sécurisé de la base de données : Un export de la base de données du système BMS-LM est déjà mis en place en utilisant un Microsoft SQL Server. D'autres types d'exports peuvent facilement être mis en place (via API, via outil de génération de rapport, ou

[71] Un outil d'exploration, filtre, analyse rapide des données intégrées dans Excel.

autre) afin de permettre au chercheur de déposer ses données, gérées par le système BMS-LM, en accès libre pour leur réutilisation par d'autres chercheurs dans le cadre de la science ouverte. *Besoins : (B5) partage, (B7) réutilisation, (B11) standardisation,*

Nous avons résumé dans les paragraphes ci-dessus les différentes briques fonctionnelles du système BMS-LM et les besoins auxquels elles répondent. Une présentation plus détaillée est donnée en Annexe B. Dans cette thèse, nous nous focalisons en particulier sur les briques d'intégration de données et d'intégration des calculs scientifiques dans le système BMS-LM (briques 5 et 6 de la Figure 70) associées aux besoins suivants : *(B1) archivage, (B2) import, (B6) analyse, (B7) réutilisation, (B8) traçabilité, (B10) automatisation, (B11) standardisation, (B15) flexibilité, (B16) vérification*. Nous avons fait évoluer l'ancien modèle de données BMI-LM en un nouveau modèle dans ce cadre (brique 9 de la Figure 70) et nous avons mis en place l'architecture logicielle « *4-tiers* » en quatre niveaux du système BMS-LM explicitée dans la Figure 71 ci-après. Elle présente les différentes briques logicielles implémentées ainsi que leurs interactions avec les éléments externes : sources de données, clusters de calculs, applications tierces et logiciels maison.

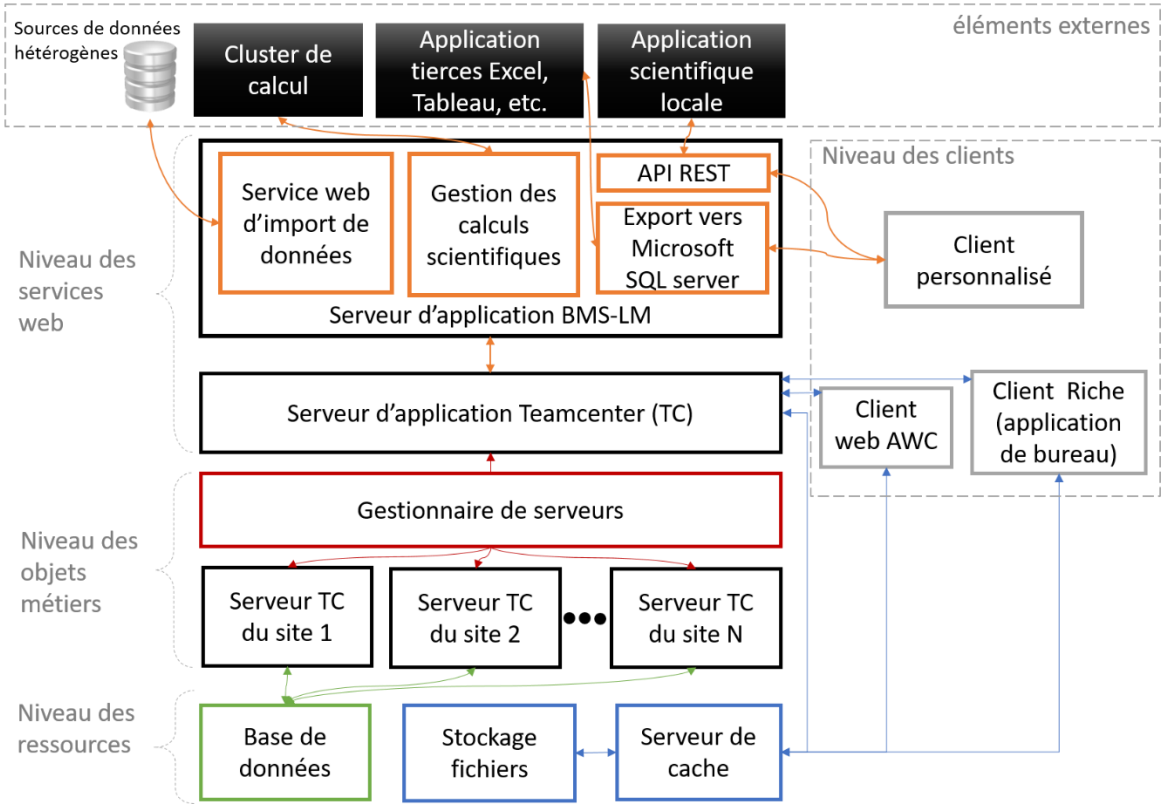


Figure 71 Architecture 4-tiers du système BMS-LM proposé

Les « *4-tiers* » de l'architecture de la plateforme PLM Teamcenter ont été réutilisés. Un « Serveur web d'application BMS-LM » a été ajouté afin d'implémenter les différentes fonctionnalités spécifiques à l'intégration de données et de calculs scientifiques. Les différentes briques internes au serveur d'application BMS-LM vont être détaillées au fur et à mesure dans le présent chapitre.

IV.1.2. Modèle de données (MDD) BMS-LM

Face à l'hétérogénéité multiple des données en recherche biomédicale (explicitée en §I.2.4), nous avons fait évoluer le modèle de données BMI-LM (proposé lors de la thèse de (Allanic, 2015) en neuroimagerie, en un nouveau modèle pour le système BMS-LM que nous appelons « MDD BMS-LM ». Principalement, nous avons ajouté des concepts génériques modélisant les produits « agent de contraste » ou tout autre produit « *Agent* », les « échantillons biologiques - *Sample* », et les

« interventions - *Intervention* », chirurgicales ou autre, ainsi que les branches de la Classification qui leur corresponds :

- *Agent* : tout produit de laboratoire actif dans une acquisition de données, une préparation d'échantillon, un examen, etc. Par exemple, le ^{18}FDG un agent de contraste utilisé en imagerie TEP, ou l'anticorps secondaire utilisé pour la révélation en immunofluorescence.
- *Sample* : tout échantillon biologique prélevé sur un organisme vivant, et utilisé pour une expérimentation dans le cadre d'une étude de recherche.
- *Intervention* : toute action entreprise sur un sujet d'étude dans le but de le transformer pour l'étudier : opération chirurgicale, régime alimentaire, traitement médicamenteux, etc.

Le Tableau 12 liste les différentes classes du MDD BMS-LM, leurs noms officiels, leurs trigrammes et rôles, ainsi que les icônes utilisées pour les désigner. Pour chaque nouveau concept, deux classes sont ajoutées au modèle de données : une classe de Définition, désigné par (1:D) qui représente le protocole de l'expérimentation et une classe de Résultat (2:R) qui représente les données résultantes des expérimentations et la description de leurs déroulements réels. Pour un rappel sur les types Définition (1:D), Résultats (2:R), Ambivalent (3:A), voir section I.4.3.1.

Tableau 12 Liste des classes du MDD BMS-LM après ajout des concepts « Agent », « Sample » et « Intervention »

Classe générique	Trigramme	Rôle : modéliser les informations sur …	Type	Icône
Acquisition Result	ACQ	…les différentes modalités d'acquisition de données utilisées dans un examen et leurs paramètres.	2 : R	
Acquisition Definition	ACD	…le protocole prédéfini de l'acquisition.	1 : D	
Agent Result	AGR	…les produits et leurs doses ainsi que leur utilisation réelle lors d'une expérimentation.	2 : R	
Agent Definition	AGD	…le protocole prédéfini pour l'administration et l'utilisation de l'agent en question.	1 : D	
Device	AQD/DVC	…l'appareil d'acquisition ou la machine d'analyse de données, ses configurations et versions.	1 : D	
Bibliographical reference	BBR	…tout document référencé par une étude de recherche : article de journal, de conférence, etc.	3 : A	
Data Unit Result	DUR	…tout jeu de données produit lors d'une acquisition de données.	2 : R	
Data Unit Definition	DUD	…les jeux de données attendus d'une acquisition de données.	1 : D	
Exam Result	EXA	…les paramètres et configurations d'un examen réalisé sur un sujet d'étude (humain, animal, etc.)	2 : R	
Exam Definition	EXD	…le protocole prédéfini de l'examen.	1 : D	
Intervention Result	ITR	…le déroulement d'une intervention sur un sujet d'étude.	2 : R	
Intervention Definition	ITD	…le protocole de l'intervention.	1 : D	
Processing Result	PCR	…le résultat d'exécution d'une chaîne de traitement, composé de plusieurs PURs.	2 : R	
Processing Definition	PCD	…le protocole d'exécution d'une chaîne de traitement, composé de plusieurs PUDs.	1 : D	
Processing Unit Result	PUR	…le résultat de l'exécution avec les paramètres (PCP) d'une unité de traitement (PUD)	2 : R	
Processing Unit Definition	PUD	…la spécification d'une unité de traitement : son algorithme, ses types d'entrées et ses types de sorties	1 : D	

Processing Parameters	PCP	…le (ou les) paramétrage(s) possible(s) d'une unité de traitement (PUD) utilisé(s) pour produire ses résultats (PUR) lors d'une exécution.	1 : D	
Reference Data	RFD	…les données utilisées comme référence dans une étude ou publiées comme référence pour d'autres études : atlas du cerveau, liste de protéines, etc.	3 : A	
Sample Result	SAR	…les échantillons biologiques prélevés sur des sujets d'étude et le déroulement du prélèvement.	2 : R	
Sample Definition	SAD	…le protocole du prélèvement d'échantillons biologiques.	1 : D	
Software Tool	STL	…l'outil logiciel utilisé dans le cadre de l'étude pour l'analyse de données.	1 : D	
Study	STU	…l'étude de recherche.	2 : R	
Study Subject	SSU	…le sujet dans le cadre de l'étude : souris, lapin, être humain (les informations sont collectées après consentement de la personne).	2 : R	
Subject	SUB	…le sujet unique dans la base de données, ses liens de parenté avec les autres sujets et les SSU qui lui sont liés.	1 : D	
Subject Group	SGP	…le groupe de sujets dans l'étude et son rôle (contrôle, traité, placebo, etc.).	3 : A	
Workflow Input	WFI	…les données, l'algorithme, le logiciel et la machine nécessaires pour l'exécution d'une chaîne de traitement (PCD).	1 : D	

Par la suite, nous optons pour les deux appellations suivantes : classes de Provenance (pour désigner les classes de Définition y compris ceux Ambivalents) et classe de Résultat (pour désigner les classes de Résultats y compris ceux Ambivalents). Les Trigrammes indiqués précédemment seront utilisés tout au long du manuscrit. Il est conseillé d'utiliser le tiré-à-part du Tableau 12 (voir liste des abréviations) afin de s'y référencer si besoin.

Le diagramme UML de la Figure 72 suivante montre les différents liens relatifs aux nouvelles classes du MDD BMS-LM.

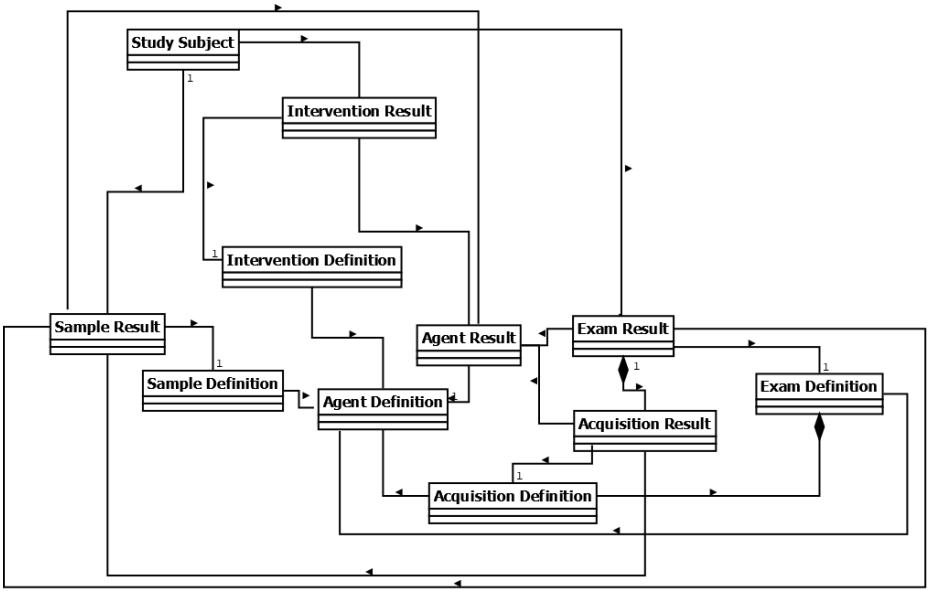


Figure 72 Diagramme UML des nouvelles classes du MDD BMS-LM

Chaque classe de résultat référence sa classe de définition. Un « Study Subject (SSU) » référence ses résultats de « Intervention Result (ITR) », « Sample Result (SAR) » et « Exam Result (EXA) ». Un « Exam Result (EXA) » est composé de « Acquisition Result (ACQ) », de même pour les définitions. Une intervention (ITR), un examen (EXA), une acquisition (ACQ), ou un échantillon (SAR), utilisant un agent doivent le référencer. Un examen (EXA) ou une acquisition (ACQ) utilisant un échantillon (SAR) doivent le référencer.

IV.1.3. CLASSIFICATION LIÉE AU MDD BMS-LM

Le système BMS-LM est régi par le MDD BMS-LM mais aussi par une « Classification » qui spécialise chaque classe générique du MDD (voir Figure 73 ci-après). La « Classification » se présente sous la forme d'une arborescence de classes de domaine qui s'adapte aux disciplines mises en jeu. Pour éviter toute ambiguïté, la notion de « classe » sera utilisée ci-après pour désigner les classes de la « Classification », et la notion d'« objets » sera utilisée pour désigner les classes génériques du MDD.

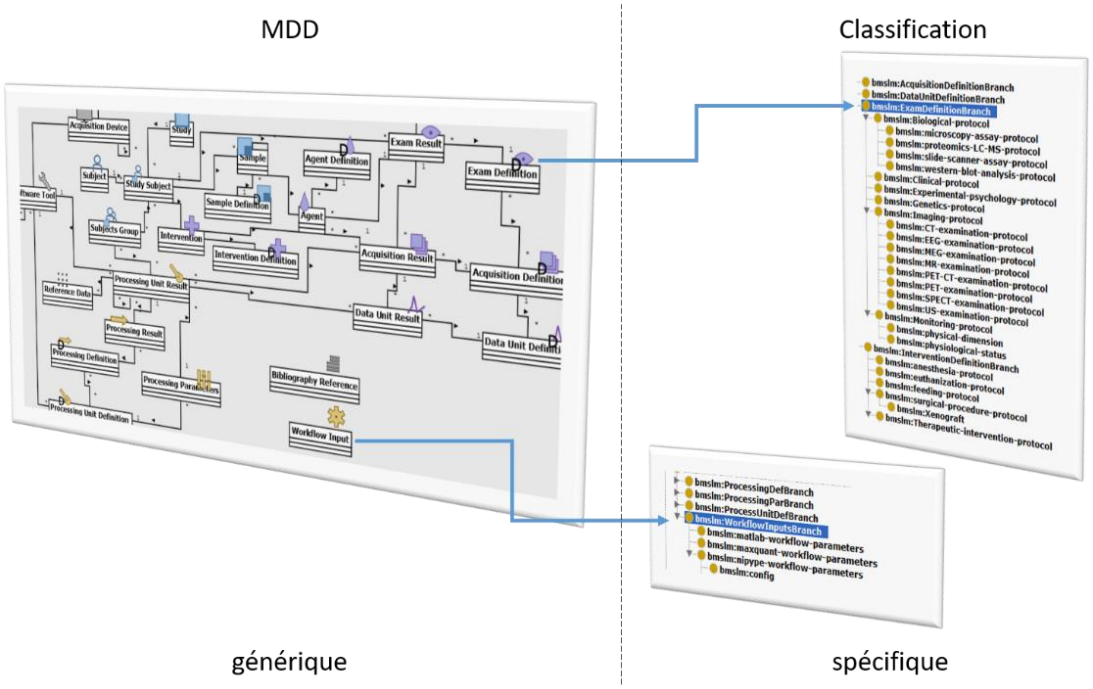


Figure 73 Les objets génériques du MDD spécialisés via la « Classification » dans le système BMS-LM

Le duo « MDD+Classification » est un choix d'implémentation de la modélisation des données de recherche qui a été conditionné par la plateforme PLM utilisée (Teamcenter). Cette dernière présente des contraintes fortes sur la conception et l'implémentation des schémas de données en son sein. Ces contraintes sont liées aux exigences de performance sur les requêtes exécutées par la plateforme Teamcenter dont l'infrastructure profonde est une base de données relationnelle. En effet, bien qu'ayant une approche de modélisation de données orientée objet, la plateforme Teamcenter tente de transformer les objets du MDD le plus directement possible dans l'implémentation relationnelle. Pour maintenir une performance acceptable, il est souhaitable que l'ensemble et le nombre d'objets principaux utilisés soient raisonnablement figés. Cela facilite également les développements logiciels faits autour du système (outils de qualité, traitements scientifiques, rapports, extractions…) et leur maintenance.

Les objets génériques du MDD sont alors figés au maximum pour éviter de redéployer fréquemment le modèle de données. Il s'agit d'une tâche coûteuse et doit être minimisée au maximum. Nos travaux sur le MDD BMS-LM ont eu alors deux portées : éviter les déploiements multiples du MDD qui sont accompagnés d'une période d'indisponibilité du système BMS-LM pour les utilisateurs, et modéliser tout type d'étude biomédicale avec les objets génériques du MDD. Ces objets seront spécifiés à l'aide

de la « Classification » qui consiste en une description sémantique spécifique et facilement évolutive (pas d'arrêt système). Chaque branche de l'arbre de classification correspond à un objet générique du MDD afin de le décrire spécifiquement.

Après l'ajout des trois concepts « Agent », « Sample » et « Intervention », nous avons fait évoluer les branches de classification comme dans la Figure 74.

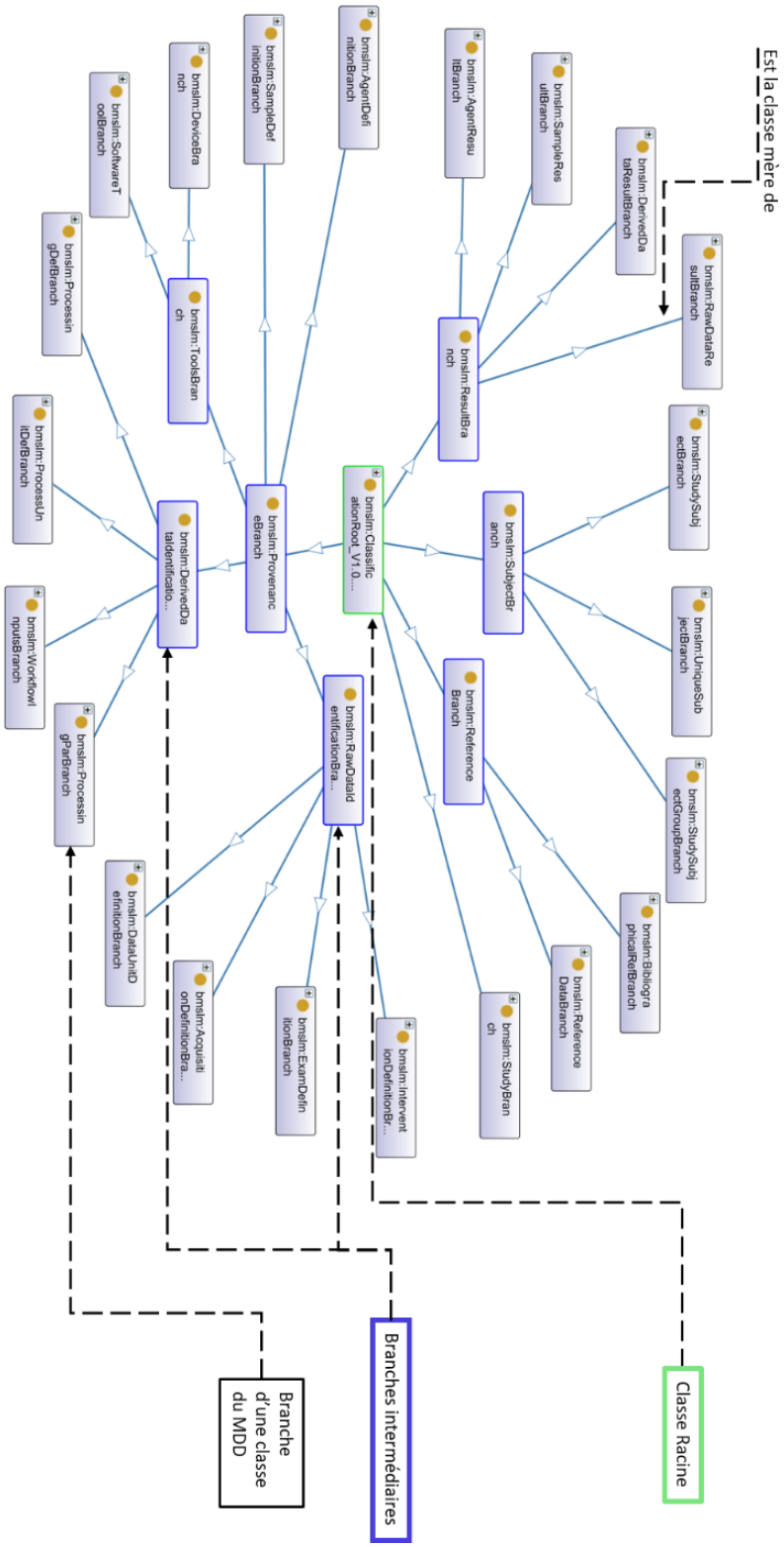


Figure 74 Branches racines de la « Classification » du système BMS-LM

La classe racine « ClassificationRoot » est entourée de classes intermédiaires (« Provenance Branch », « Raw Data Identification Branch », etc.) qui regroupent les branches de chaque objet du MDD BMS-LM. Sous la branche « SampleDefinitionBranch », sont associés les deux éléments de description d'un « Sample » à savoir le « *sample preparation procedure* » et le « *sample identification* ». De même pour la branche « AgentDefinitionBranch », nous avons séparé « l'administration de l'agent » de son « identification ». Pour l'intervention, nous l'avons défini comme sous-classe de « Raw Data Identification Branch ».

IV.2. MÉTHODES D'INTÉGRATION DE DONNÉES ET DE CALCUL SCIENTIFIQUE

Pour récolter les avantages du système BMS-LM, il faut pouvoir gérer les données via ses outils logiciels. Ceci n'est possible qu'après l'intégration des données dans le système. Nous consacrons la première partie de cette section à la présentation de notre méthode « générique » d'intégration de données hétérogènes. Une fois l'intégration des données réalisée, un utilisateur clé a la possibilité d'explorer ses données et d'exécuter des requêtes via les différents clients du système BMS-LM (Annexe A). Une deuxième étape du cycle de vie de l'étude commence : l'analyse des données. Afin d'appliquer la traçabilité aussi au niveau des résultats d'analyses, il faut intégrer les calculs scientifiques utilisant ces données dans le système BMS-LM. La deuxième partie de cette section présentera nos méthodes dans ce cadre.

IV.2.1. MÉTHODE « GÉNÉRIQUE » D'INTÉGRATION DE DONNÉES DANS LE SYSTÈME BMS-LM

Nous nous positionnons dans le contexte global de la gestion de données de recherche biomédicale, hétérogènes par nature, issues de différentes sources et de niveaux de granularité et portées variés. Nous proposons de les centraliser, via une méthode d'intégration dans le système BMS-LM afin de permettre leur suivi et leur gestion avec un maximum de traçabilité. Nous l'avons appelé « Méthode générique d'intégration ».

Les besoins en lien avec cette proposition sont : *(B1) archivage, (B2) import, (B10) automatisation, (B11) standardisation, (B15) flexibilité, (B8) traçabilité, (B16) vérification*. Nous allons expliciter notre méthode d'intégration de données sur deux niveaux : le niveau fonctionnel préparatoire et le niveau technique exécutoire.

IV.2.1.1. Préparation fonctionnelle des données et du système BMS-LM

Pour plus de clarté, nous avons séparé les préparatifs des données de ceux du système BMS-LM.

IV.2.1.1.1. Préparation des données pour l'intégration

Les jeux de données à intégrer sont annotés via des terminologies locales du laboratoire ou du domaine en question. Ils sont aussi relativement « plats », dans le sens où ils n'ont pas de structure qui correspond avec la structure hiérarchique du modèle de données BMS-LM. Afin de mieux connaître et décrire les données et leur provenance, nous avons utilisé les réponses aux questions suivantes :

- À quelle phase du cycle de vie ces données appartiennent-elles ? Est-ce qu'il s'agit de données dérivées ? De données brutes ? De documents de spécification ? De données à valoriser ? etc.
- Qui ? Quoi ? Où ? Quand ? Comment ? Combien ? Pourquoi ? => les QQOQCCP pour la provenance
- Avec quels logiciel et/ou matériel sont acquises ces données? Quelle(s) version(s) ? Quelles configurations ?
- Comment ces données sont-elles acquises, produites ? Les protocoles associés ?

- Dans quel domaine ces données peuvent-elles être décrites ? Quelles sont les KOS de domaine publiés que nous pouvons utiliser pour les décrire ?

Passée cette étape d'annotations des données, les informations sont rangées dans un tableau Excel. Le fichier est préparé par le chercheur lui-même en utilisant son jargon spécifique local. Elle représente l'entrée du premier nœud de notre méthode d'intégration de données. Il contient des informations descriptives des données avec des termes locaux ou de domaine, des informations sur la provenance des données : protocoles suivis, machines et logiciels utilisés. Plus précisément, la version du logiciel, la configuration de la machine, l'ID du protocole et l'ID du projet de recherche.

Ensuite, nous avons attribué une structure au jeu de données à importer. Celle-ci est réalisée en étroite collaboration entre le data manager et le producteur des données. Il s'agit de les modéliser via les concepts du MDD. Par exemple, pour un examen TEP-TDM identifié par T001080, produit par le « Scanner Mediso NanoScan PC » et dans le cadre du projet « Vesicules », où une souris mâle a passé un examen TDM, et une acquisition TEP dynamique, l'examen est représenté comme indiqué dans la Figure 75. Les objets du MDD sont représentés via leurs icônes et leurs noms, et les classes de la « Classification » sont décrites via les attributs en encadré jaune (Pour comprendre les trigrammes cf. le tiré-à-part du Tableau 12 (voir liste des abréviations)).

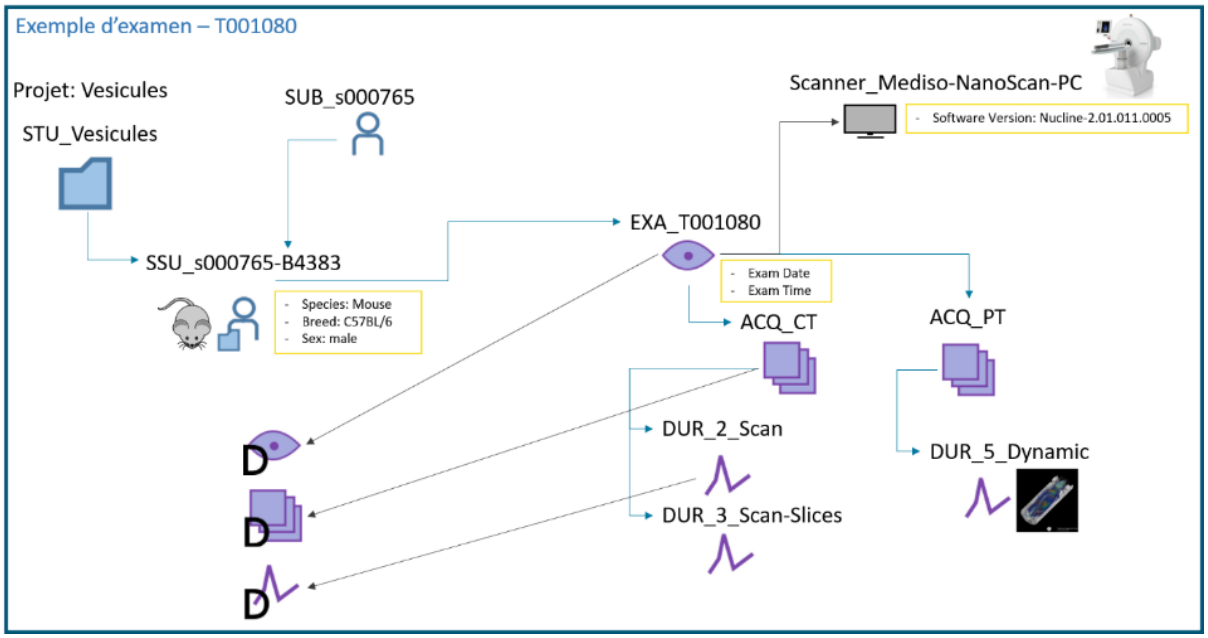


Figure 75 Exemple d'examen TEP-TDM représenté avec les objets du MDD et la « Classification » du système BMS-LM

Cette présentation des données d'entrée dans une structure arborescente nécessite une transcription intermédiaire des données et leurs descripteurs dans un fichier XML. L'écriture de ce fichier XML est accompagnée d'un certain nombre de modifications par rapport au tableau d'entrée. Ces modifications sont réalisées par le data manager et validées par le producteur des données. Elles incluent :

- La structuration des données en arborescence comme dans l'exemple précédent.
- La séparation et contextualisation des métadonnées. Par exemple, la « version du logiciel » doit être attachée à l'objet « logiciel » ou « matériel » et la « souche de la souris » doit être attachée au « sujet de l'étude ».
- L'attribution d'identifiants uniques pour les objets à intégrer. Les identifiants doivent être uniques et faciles à retenir pour les utilisateurs. En fonction du contexte d'application, une politique de génération des identifiants doit être définie avant l'exploitation du système BMS-LM, afin d'éviter les dupliquas des IDs. Une règle de départ a été proposée comme suit dans le laboratoire LRI :

- Pour les sujets dans l'étude
  - SSU_[id projet]_[id sujet dans l'étude]
- Pour les données résultats
  - EXA_[id projet]_[id sujet dans l'étude]_[id exam]
  - ACQ_[id projet]_[id sujet dans l'étude]_[id exam]_[id acquisition]
  - DUR_[id projet]_[id sujet dans l'étude]_[id exam]_[id acquisition]_[id données]
  - PCR_[id projet]_[id sujet dans l'étude]_[id processing]
- Pour les objets de provenance
  - XXD_[id labo]_[type d'examen]_[espèce cible]_[protocole]_[id version]
  - DVC_[id labo]_[id constructeur]_[id machine]

Le schéma de la Figure 76 ci-après résume les étapes de préparation des données. En haut de la figure, est indiqué le nom de l'étape, et en bas ce qu'elle utilise. Les réponses aux questions QQOQCCP permettent de dresser un tableau descriptif des données. Ce tableau est converti en une arborescence XML décrivant les instances à importer dans le système BMS-LM. Les liens entre les instances et les fichiers sont aussi décrits dans le fichier XML et s'intègrent au niveau du système BMS-LM qui lui aussi doit être préparé pour l'accueil de nouveaux types de données.

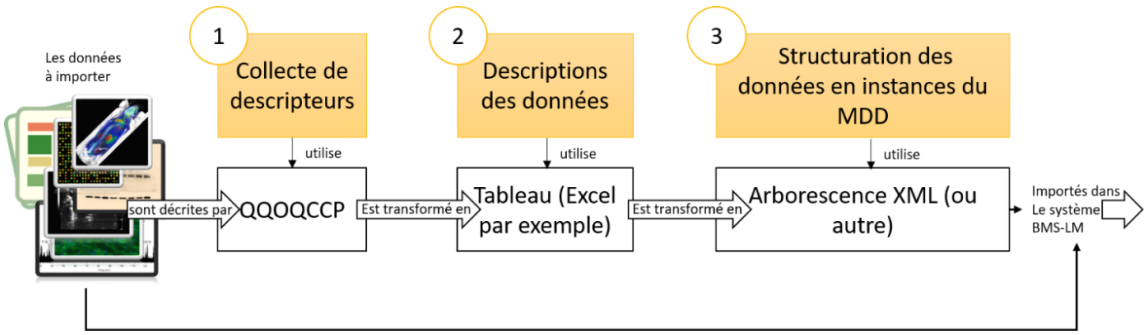


Figure 76 Étapes de préparation de données pour leur intégration « générique » dans le système BMS-LM

IV.2.1.1.2. Préparation du système BMS-LM

En entrée de cette étape, nous disposons d'un fichier XML structuré selon les objets génériques du MDD (Résultats et Provenance) et contenant les descripteurs du jeu de données à intégrer (voir Figure 76). Le fichier XML permet la création des instances du MDD décrivant ces données dans le système BMS-LM. Les descripteurs utilisés sont généralement issus des terminologies locales provenant de l'environnement de production des données. Le rôle de cette étape est de préparer le système BMS-LM à la réception du nouveau lot de données par l'ajout des instances d'objets MDD de Provenance, des classes de domaines à la « Classification » et des alignements entre ces classes et les terminologies locales. Nous expliquons chacune de ces étapes dans les paragraphes suivants. Elles sont réalisées en amont de l'import de données par le data manager.

❖ Ajout des instances d'objets MDD de Provenance : les informations de provenance doivent être collectées lors de la phase QQOQCCP (voir Figure 76). Les instances de provenance doivent être ajoutées au système BMS-LM, en parallèle à la préparation du fichier XML, s'ils ne sont pas déjà dans le système. Le fichier XML décrivant les instances des objets du MDD et les liens entre eux, référence aussi les instances de provenance ajoutées à cette étape. Par exemple, chaque examen (EXA) doit être lié à la version de la machine (DVC) qui a été utilisée pour produire cet examen, et au protocole (EXD) qui a été utilisé pour paramétrer son déroulement. Ces informations sont importantes (versions des protocoles, machines, etc.). Elles doivent être intégrées dans le système BMS-LM et validées par un expert du domaine avant l'import massif de données via les fichiers XML. Les instances de Provenance sont principalement les protocoles (EXD, ACD, SAD, AGD, PUD, PCD), les logiciels STL, les machines DVC, ainsi que les configurations de traitement PCP

et WFI (voir tiré-à-part du Tableau 12). La création des instances de Provenance s'effectue en utilisant un des clients de la plateforme Teamcenter (voir Annexe A).

Jusqu'à présent, il s'agissait d'ajouter des instances du MDD : instances de Résultats spécifiées par les fichiers XML et instances de Provenance référencées par ces mêmes fichiers XML. La structure du fichier XML est détaillée dans la partie technique (§IV.2.1.2). Les instances du modèle BMS-LM sont génériques. À un niveau plus spécifique, la « Classification » du système BMS-LM modélise les classes de domaine associées aux données. Elle est ajustée à l'intégration de chaque nouveau type de données, en réutilisant les termes des KOS publiés, pour une annotation plus standard des données. Les étapes suivantes de cette phase de préparation du système BMS-LM sont alors : l'ajout de classes de domaines dans la « Classification », ainsi que l'ajout d'alignements avec les termes locaux trouvés dans le fichier XML d'import. Ces étapes sont détaillées ci-après.

- L'ajout de classes de domaines : Afin d'identifier les concepts de domaine à ajouter (mesures, type d'acquisition, paramètres vitaux, etc.), il faut chercher dans les KOS publiés des équivalents aux termes locaux utilisés dans le fichier XML. À l'issue de cette phase, la « Classification » du système BMS-LM sera enrichie par des classes publiées, plus standardisées et plus ouvertes, en lien avec le jeu de données d'entrée.
- L'ajout d'alignements entre les différents éléments (terminologies locales du fichier XML, classes de domaines de la « Classification », objets génériques du MDD) : lorsqu'un « concept de domaine » est trouvé, il est aligné avec « le terme local d'origine » si besoin et avec « le terme parent » au niveau MDD. Ces liens sont principalement : l'héritage entre classes ou la description via une liste d'attributs. Dans le Tableau 13 ci-après, nous présentons une liste de termes locaux ainsi que leurs équivalents dans des KOS publiés et comment ils ont été intégrés à la « Classification » du système BMS-LM.

Tableau 13 Liste de termes locaux, leurs équivalents de domaine dans les KOS publiés ainsi que leurs liens avec le MDD et la Classification BMS-LM
Pour connaître le nom complet du KOS, il faut consulter la liste des abréviations

	Terme local	**Décrit quel objet générique du MDD ?**	**Équivalent dans KOS publiés ?**	**Comment le représenter dans la Classification**	**Notes**
1	Animal number	SSU : sujet dans l'étude	DCM : Patient Name	Attribut	Le numéro attribué à un animal constitue son « nom »
2	Project referent	STU : étude	DCM : Referring Physicians Name	Classe	L'ajout des rôles est accompagné d'une définition de droits d'accès pour ce rôle dans l'étude de recherche
3	Operator	EXA : examen	DCM : Operator Name	Classe	//
4	Study ID	EXA : examen	DCM : StudyID	Attribut	« Study ID » dans le standard DICOM désigne l'identifiant de l'examen => donne confusion => remplacer par « ExamID »
5	Slice thickness	DUR : unité de données	DCM : Slice Thickness	Attribut	
6	Series number	DUR : unité de données	DCM : Series Number	Attribut	
7	Number of frame	DUR : unité de données	DCM : Number Of Frames	Attribut	
8	KVP	DUR : unité de données	NCIT : Kilovolt Peak DCM : KVP	Attribut	
9	Q-Exactive	DVC : machine	MS : Q Exactive	Classe	
10	Slide	DVC : machine	OBI : microscope slide	Classe	
11	Souris (mouse)	SSU : sujet dans l'étude	IOBC : mouse	Classe	

12	Immuno	SAR : préparation d'échantillon	PATIT : Immunofluorescence	Classe	
13	SCX	SAR : préparation d'échantillon	NCIT : Strong-Cation-Exchange Chromatography	Classe	
14	Staining	SAR : préparation d'échantillon	OBI : Staining OBI : IHC slide staining OBI : H&E slide staining	Classe	Dans ce cas, nous avons pu spécifier le « Staining » avec des classes encore plus spécifiques (IHC, H&E)
15	Trypsin	AGT : identification d'agent	PRO : trypsin	Classe	
16	Lectine	AGT : identification d'agent	MESH : isolectin B4-binding glycoprotein, mouse	Classe	
17	FDG	AGT : identification d'agent	QIBO : 18-Fluorodeoxyglucose	Classe	
18	DAPI	AGT : identification d'agent	FBbi : 4',6-diamidino-2-phenylindole (DAPI)	Classe	
19	FOV	ACQ : acquisition de données	NCIT : Field of View	Classe	
20	VOI	ACQ : acquisition de données	NCIT : Volume of Interest	Classe	
21	ROI	ACQ : acquisition de données	NCIT : Imaging Region of Interest	Classe	

IV.2.1.2. Implémentation technique de l'intégration et ses composants

Le schéma de la Figure 77 ci-après résume les différents modules techniques de l'intégration des données dans le système BMS-LM. Ils sont numérotés de 1 à 4. Ils vont être décrits un à un : transformation des données via ETL, Mise en correspondance, Moteur d'alignement, et Moteur d'import.

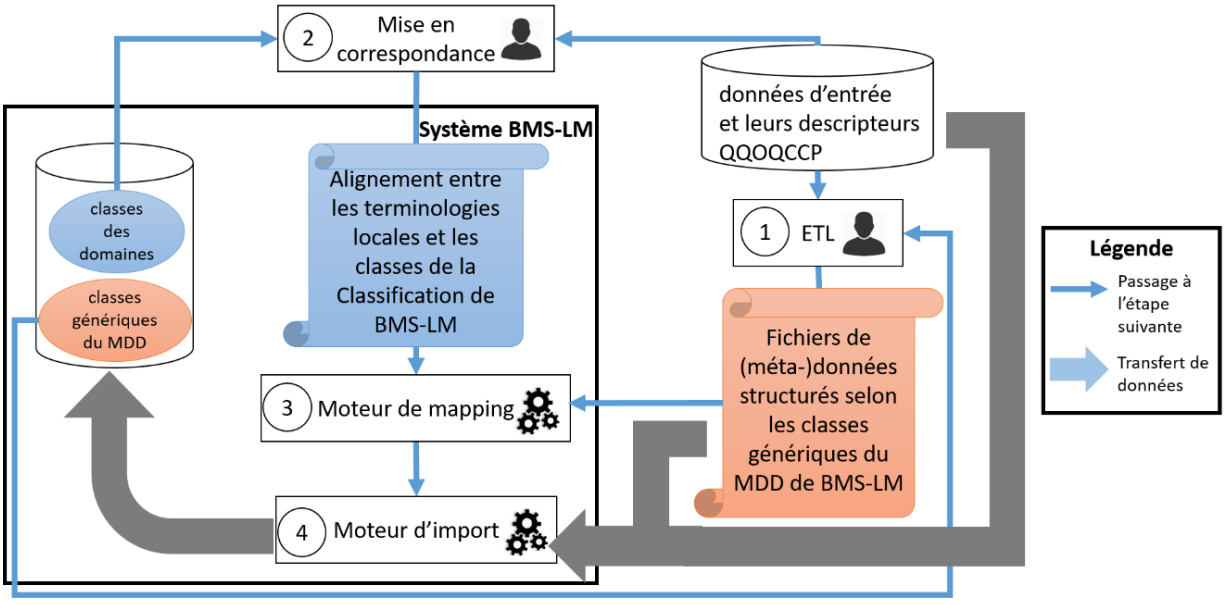


Figure 77 Méthode d'intégration des données dans BMS-LM

IV.2.1.2.1. Extract - Transform – Load (ETL)

Le premier module de l'implémentation de notre méthode est un ETL construit via le logiciel Talend[72]. Il permet d'effectuer la transformation des données d'entrée en arborescence XML prête à être importée dans BMS-LM. L'ETL est réalisée par les experts en gestion de données et pas par le chercheur qui a produit les données. L'ETL Talend est constitué des nœuds présentés dans la Figure 78. Le tableau de métadonnées préalablement fourni par le chercheur décrit le jeu de données à importer. Il est lu dans le premier nœud « metadata ». La liste de ses enregistrements (ou lignes) est réorganisée via le nœud « tSortRow » afin de les grouper en fonction des niveaux de l'arborescence XML de sortie. Ensuite vient le nœud central « tXMLMap » qui a pour rôle de transformer les « données linéaires » en entrée, en « données en arborescence » en sortie. Les deux nœuds qui suivent (tFileOutputXML) permettent l'écriture sur le disque des fichiers XML. Ces fichiers guideront l'intégration des données par la suite.

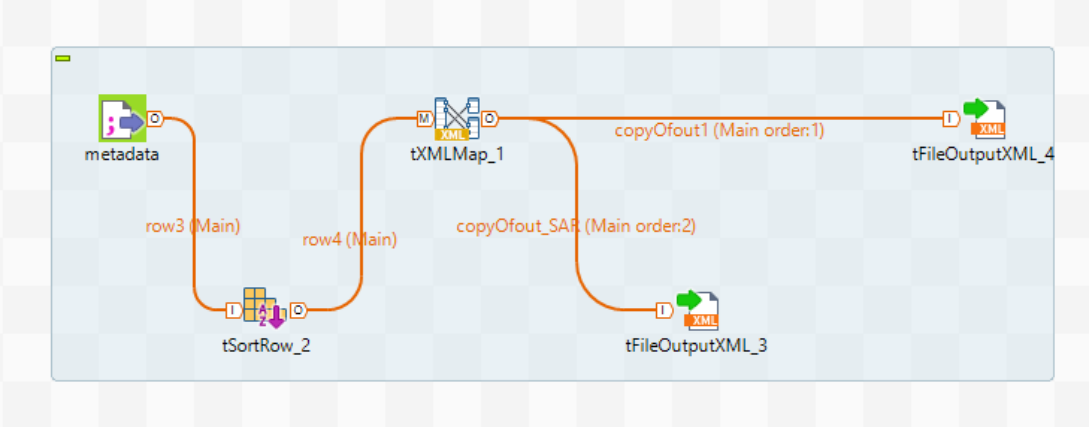


Figure 78 Nœuds d'un ETL Talend pour la transformation des données d'entrée en XML

IV.2.1.3. L'arborescence XML générique

L'arborescence XML de sortie se présente comme dans la capture d'écran de la Figure 79. La balise <import/> contient les éléments à intégrer par le service web d'import de données. Le type d'intégration, qui doit être renseigné, est « generic ». Pour chaque sujet dans l'étude, il faut renseigner ses examens <EXAimas> et ses échantillons <SARs>. Le but du fichier XML est d'attribuer au jeu de données en entrée, une structure dans l'espace du MDD, et une série d'annotations sous forme de liste d'attributs de la « Classification ».

Dans la Figure 79, pour un SSU (sujet dans l'étude), une liste d'examens est spécifiée. Pour chaque EXA (examen), une liste d'ACQ (acquisition) est décrite, et pour chaque ACQ, une liste de DUR (unité de données) est définie. Les ACQs référencent les SARs (échantillons biologiques) utilisés pour l'acquisition. Les chemins d'accès aux fichiers correspondants sont renseignés dans la balise <dataset/> sous DUR. Vers la fin du fichiers XML, les SARs sont listés pour chaque SSU. À chaque niveau où des annotations descriptives peuvent être attachées, elles sont ajoutées via la balise XML <attributes>. À chaque niveau de l'arborescence, les objets de provenance sont référencés, par exemple, EXD (protocole d'examen) pour un EXA (examen).

[72] https://www.talend.com/fr/ dans sa version 6.4

```
<import importRule="generic">
        <SSU id="InternalID" project="projet" label="SSU_nom">
            <STU id="projet"/>
            <EXAimas>
                <EXA id="InternalID" label="EXAima_nom" generateThumbnails="false" checkImport="false">
                    <EXD id="InternalID" />
                    <AQD id="InternalID" revision="A" />
                    <attributes></attributes>
                    <ACQs>
                        <ACQ id="InternalID" label="ACQbio_nom">
                            <ACD id="InternalID" />
                            <attributes></attributes>
                            <DURs>
                                <DUR id="InternalID" label="DURbio_nom">
                                    <DUD id="InternalID" />
                                    <attributes></attributes>
                                    <datasets>
                                        <dataset name="name" type="GIN4_type">path</dataset>
                                    </datasets>
                                </DUR>
                            </DURs>
                            <SARs>
                                <SAR id="InternalID" />
                            </SARs>
                        </ACQ>
                    </ACQs>
                </EXA>
            </EXAimas>
            <EXAbios>
            </EXAbios>
            <SARs>
                <SAR id="InternalID" label="SAR_nom">
                  <SAD id=""/>
                  <SARs>
                    <SAR id ="InternalID" label="SAR_nom">
                        <SAD id="InternalID" />
                        <attributes></attributes>
                    </SAR>
                  </SARs>
                </SAR>
            </SARs>
        </SSU>
</import>
```

Figure 79 Arborescence typique d'un fichier XML décrivant des données brutes

Avec cette structuration générique, nous pouvons formater n'importe quel type de données brutes d'une façon unifiée en garantissant une description minimale et suffisante des données ainsi qu'une information sur leur provenance. La même logique est applicable aux données dérivées. Dans la capture d'écran Figure 80 ci-après, un résultat de calcul scientifique (PCR) est décrit via les balises XML. Il est lié à son SSU (sujet dans l'étude), le groupe correspondant (SGP), les données d'entrée (décrites via un WFI) et les différents résultats des unités de traitement <PURs> qui le composent. Chaque PUR est lié à ses paramètres de calcul <PCP> et enfin le chemin vers le fichier résultant est référencé via la balise <dataset>. Les WFI, SGP et PCP référencés par leurs identifiants dans le fichier XML doivent être déjà instanciés dans le système BMS-LM avant le lancement de l'import.

```
<Import importRule="generic">
  <PCR id="InternalID" label="PCR_nom">
    <SSU id="SSU_InternalID"/>
    <SGP id="SGP_InternalID"/>
    <WFI id="WFI_InternalID"/>
    <attributes/>
    <PURs>
      <PUR id="PUR_InternalID" label="PUR_nom">
        <PCP id="PCP_InternalID"/>
        <attributes/>
        <datasets>
          <dataset name="name" type="Text">path</dataset>
        </datasets>
      </PUR>
      <PUR id="PUR_InternalID" label="PUR_nom">
        <PCP id="PCP_InternalID"/>
        <attributes/>
        <datasets>
          <dataset name="name" type="Text">path</dataset>
        </datasets>
      </PUR>
    </PURs>
  </PCR>
</Import>
```

Figure 80 Éléments d'arborescence XML pour intégration des données dérivées

IV.2.1.3.1. Transformation des données

Le nœud central « tXMLMap » est le nœud où la transformation de données est réalisée. Il est divisé en trois compartiments (voir Figure 81 suivante). Le premier compartiment présente le schéma du tableau en entrée, le deuxième des règles de transformation de données et le dernier les arborescences XML de sortie. Dans le compartiment du centre, il y a une liste de variables intermédiaires utilisées pour transformer les données en entrée. Comme le tableau Excel d'entrée est préparé par les producteurs des données, des adaptations s'imposent avant de les utiliser dans l'arborescence XML de sortie.

Par exemple, les variables « SSU_label » et « SSU_id » (compartiment du centre) sont utilisées pour nommer et attribuer un identifiant unique à l'instance du SSU à créer dans le système BMS-LM (voir troisième compartiment). Ces deux variables sont formées en utilisant les éléments « mouse_ID » et « project_ID » du schéma d'entrée (voir premier compartiment). Les flèches en jaune et gris de la capture d'écran Figure 81 retracent les réutilisations des éléments en entrées pour le « remplissage » de l'arborescence en sortie.

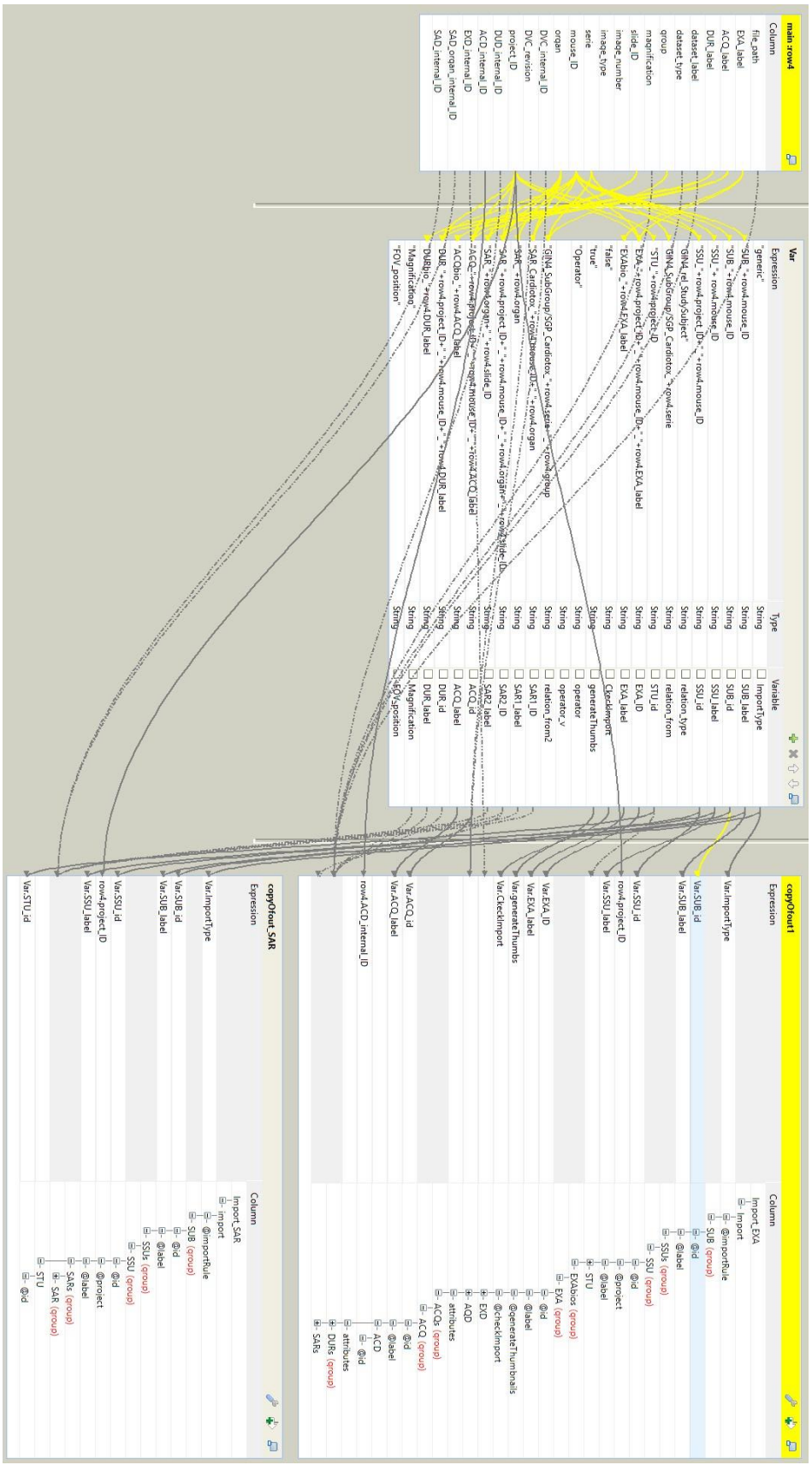


Figure 81 Capture d'écran montrant un nœud « tXMLMap » de transformation de données

IV.2.1.4. La version exécutable de l'ETL

Une fois l'ETL mis au point et testé en fonction du besoin des utilisateurs, il est stabilisé et passe à l'étape de production. Ceci veut dire qu'il sera exporté en un fichier de script « .bat ». Il sera exploité par l'utilisateur, ou le cas échéant par le data manager, afin de convertir son tableau d'entrée en un fichier XML. Le choix du « .bat » est dû au fait qu'au laboratoire LRI, il est plus familier que le « .jar » ou le « .sh ». Dans un autre contexte, il faut trouver la meilleure façon de livrer l'outil d'import de données. Il faut qu'il soit le plus simple possible pour les utilisateurs du système BMS-LM. L'outil produit par l'ETL Talend a été appelé « CSV-2-XML » et est tracé via Subversion (SVN) pour suivi de versions (voir Figure 82 suivante).

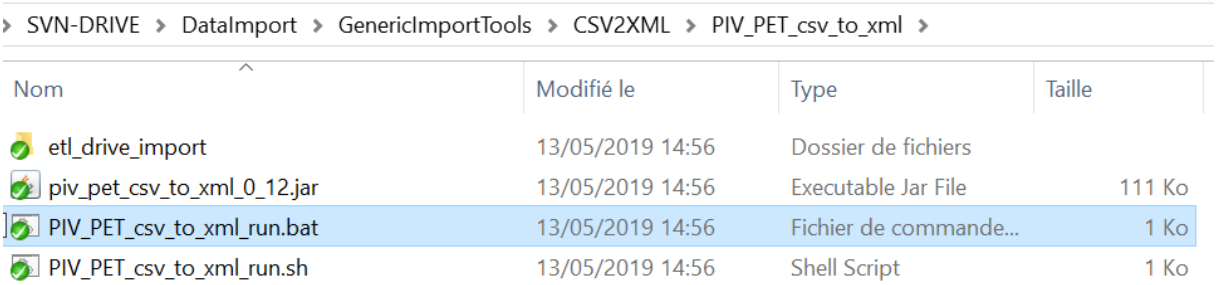


Figure 82 L'exécutable de l'ETL versionné dans SVN, il convertit d'un tableau à un fichier XML

Pour que l'ETL fonctionne, il faut que le schéma du tableau d'entrée soit cohérent avec le schéma défini dans l'ETL. Pour l'utilisateur, l'outil « CSV-2-XML.bat » ne fait que convertir un tableau décrivant des données scientifiques en un fichier XML, qu'il utilisera pour importer ses données. Par conséquent, l'ETL décrit dans cette section a été adapté à chaque type de données afin de répondre au maximum aux demandes des utilisateurs. Par exemple, l'ETL de la capture d'écran Figure 82 est celui dédié aux données TEP-TDM du laboratoire. Son utilisation au laboratoire est décrite en détail dans le chapitre VI. À la fin de l'exécution de l'ETL, les utilisateurs obtiennent les (ou des) fichier(s) XML indispensable(s) à l'intégration des données.

IV.2.1.4.1. Mise en correspondance

Nous avons décrit la mise en correspondance sur le plan fonctionnel en début de cette section (voir Tableau 13 §IV.2.1.1.2). Dans ce paragraphe, nous nous focalisons sur la mise en correspondance entre les données et le fichier XML d'une part, et le système BMS-LM de l'autre, ce qui inclut le MDD et la « Classification ». Il y a plusieurs niveaux pour la mise en correspondance :

1. La correspondance au niveau MDD entre un élément de provenance déjà présent dans le système BMS-LM (machine, protocole, etc.) et la donnée qui lui correspond dans le fichier XML
2. La correspondance entre un terme local utilisé dans le fichier XML et un terme de la « Classification » (issue des KOS de domaine)
3. La correspondance entre objets du MDD et classe de la « Classification »

Toutes ces mises en correspondance sont regroupées dans un fichier d'alignements, qui est fourni au système BMS-LM. Par ailleurs, cela permet au data manager de superviser ses différentes versions. La Figure 83 ci-après donne un exemple de fichier d'alignements utilisé pour l'intégration de données d'histologie. Il a été généré avec un outil interne à l'entreprise Fealinx et il permet d'effectuer les mises en correspondances techniques décrites précédemment. À titre d'exemple, le mot du vocabulaire local « Operator » est mis en correspondance avec « l'attribut d'identifiant 6611 » de la « Classification ». Cet attribut est utilisé pour décrire la classe « slides visualization perfomed protocol » (Classe 1, Figure 83) qui correspond à la réalisation du protocole de « microscopy assay protocol » (Classe de provenance, Figure 83). L'objet générique du MDD qui correspond à ces classes est, EXA ou « Exam Result »

(Classe MDD, Figure 83). Pour information, l'« attribut 6611 » correspond à « Performing Physician's Name » et est issu d'un KOS publié. Il correspond au tag (0008,1050) du standard DICOM.

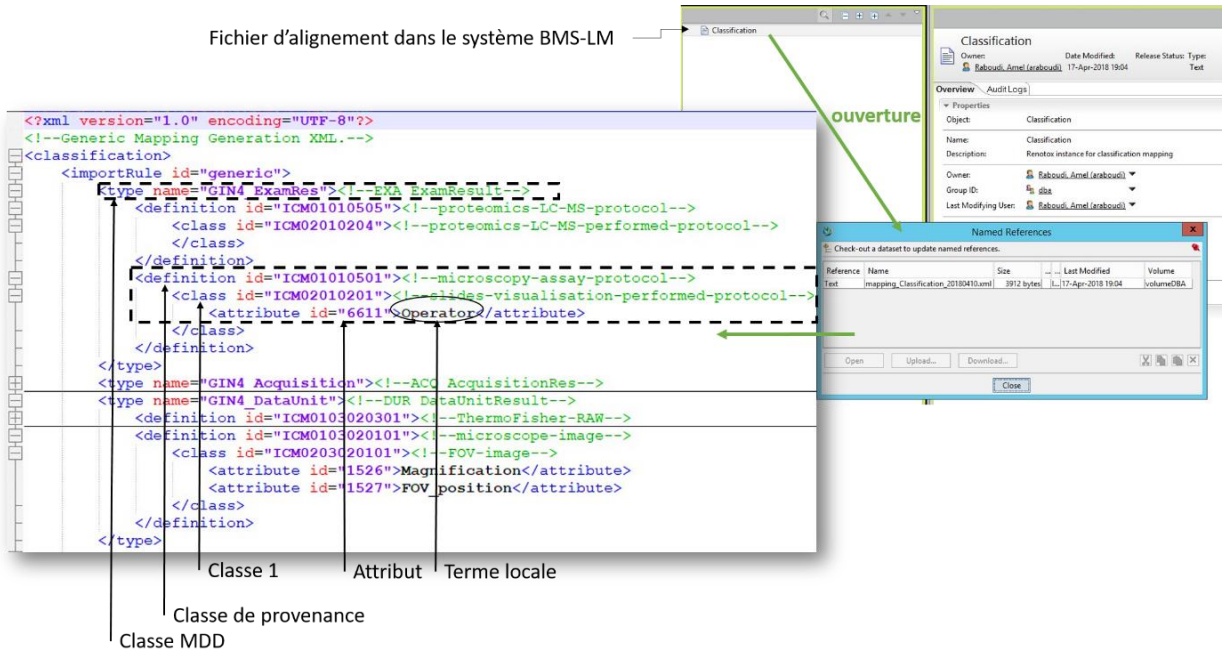


Figure 83 Traçabilité du fichier d'alignement dans le système BMS-LM et contenu du fichier

Le fichier d'alignements généré lors de cette phase sert d'entrée au moteur de *mapping* afin qu'il puisse instancier les bons objets et classes lors de l'intégration de données.

IV.2.1.4.2. Moteur de mapping et moteur d'import

À l'entrée de cette étape, nous disposons d'un fichier XML d'import et du fichier d'alignement. Le moteur de « mapping » effectue la lecture et la vérification de ces fichiers afin de préparer le terrain à l'instanciation des classes dans le système BMS-LM ainsi qu'au transfert des fichiers de données.

Le moteur d'import lit les instructions du moteur de « mapping » et effectue l'intégration de données. Un web service dédié s'en charge.

```
http://<<adresse_ip_du_webservice_d'import>>:8080/spike/swomed/import/generic?path=<<chemin_vers_vos_fichiers_XMLs_d'import>>
```

L'adresse suivante a été utilisée, par exemple, pour l'intégration des données au LRI :

```
http://193.51.82.101:1335/spike/swomed/import/generic?path=/ftp_prod/fichier_import.xml
```

Ce mode d'intégration est adapté à une utilisation par un data manager mais n'est pas *user-friendly* pour un chercheur non initié ; pour cela un *watch folder* que nous avons appelé « Import-Auto » a été mis en place pour le laboratoire LRI. Ce dossier est consulté périodiquement par le service web d'intégration de données. La détection des fichiers XML d'imports déclenche l'intégration de données.

L'utilisateur, voulant commencer une intégration de données, dépose ses données dans le *watch folder* « Import-Auto » ainsi que la conversion en XML du tableau qu'il a préparé pour les décrire (l'outil csv-2-xml.bat a été mis à sa disposition pour cela). Il recevra un courriel quand les données seront remontées dans le système BMS-LM.

IV.2.1.5. Schéma de synthèse

Toutes les manipulations conceptuelles et techniques, présentées dans cette section, sont réalisées en amont de l'import et d'une façon transparente pour l'utilisateur. Elles permettent ensuite une utilisation autonome par les chercheurs dûment formés à l'utilisation de l'outil d'import.

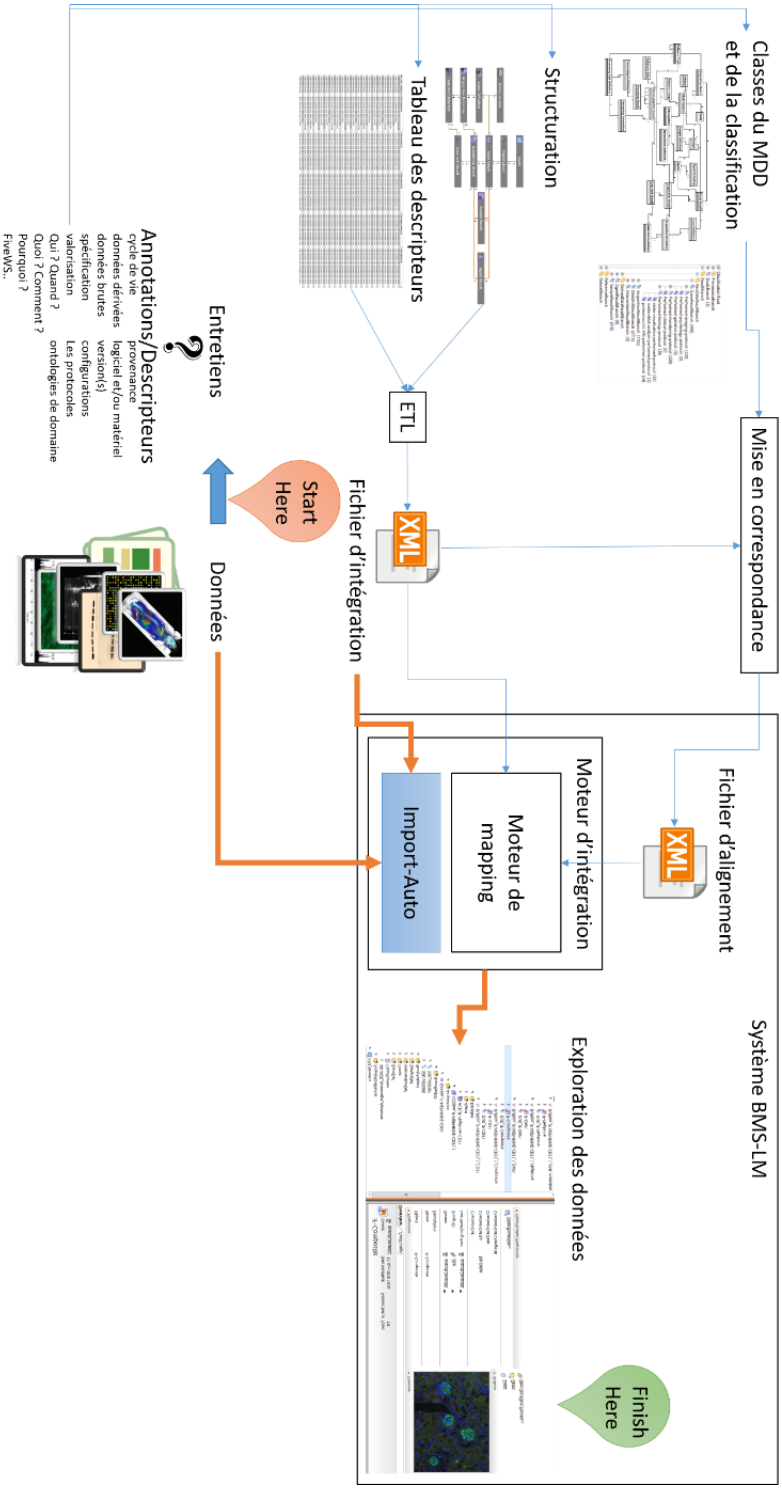


Figure 84 Schéma de synthèse des différents composants de l'intégration de données dans le système BMS-LM

La lecture du schéma doit commencer au niveau des données d'entrée en bas (voir Start Here). La collecte de descripteurs sur les données via des entretiens donne lieu au tableau d'entrée. Ce tableau ainsi que les règles de structuration de données sont l'entrée de la transformation de données via ETL.

Une fois transformé en fichier XML, le fichier sera déposé avec les données dans un dossier « Import-Auto » pour qu'il soit traité par le moteur d'intégration. Ainsi les données seront importées dans le système BMS-LM où elles seront explorées via les interfaces du système décrites en Annexe A. En parallèle à la transformation des données, un enrichissement des classes de la « Classification » est effectué. Cet enrichissement est accompagné d'un alignement entre les termes locaux, les classes de la « Classification » ajoutées et les classes anciennes. Des exemples d'application de cette méthode sont donnés dans le chapitre VI.

IV.2.2. MÉTHODE D'INTÉGRATION DE CALCUL SCIENTIFIQUE

Nous expliquons dans ce paragraphe nos approches d'intégration de chaînes de traitement, qui permettent d'intégrer et gérer des calculs scientifiques via le système BMS-LM. L'intégration de chaînes de traitement signifie dans ce contexte le fait de rendre possible le lancement d'un flux d'activités (*Workflow*) d'analyse de données et la récupération de ses résultats, via, ou avec, l'assistance du système BMS-LM. Le but est de tracer les données produites dans toutes les étapes d'une chaîne de traitement ainsi que leur provenance.

Nous avons testé une première méthode d'intégration et d'automatisation de calculs scientifiques dans le système BMS-LM que nous avons appelé méthode d'intégration « totale ». Cette méthode est issue des travaux antérieurs (Allanic et al., 2016). Elle vise à passer en production un calcul scientifique mature, automatisable à 100% et utilisé en routine. Cependant, la réalité en recherche impose d'avoir des calculs scientifiques moins matures et moins automatisables, notamment, lorsque le calcul demande une assistance utilisateur via une interface graphique. Pour cela, et pour permettre la traçabilité d'un maximum de calculs scientifiques utilisés dans une étude de recherche, nous introduisons la méthode d'intégration « partielle » en plus de celle « totale ».

IV.2.2.1. Méthode d'intégration « totale » des calculs scientifiques

Nous présentons tout d'abord les différents éléments techniques mis en œuvre pour la réalisation de l'intégration totale. Ensuite, nous expliquons les étapes fonctionnelles à entreprendre afin d'intégrer un calcul scientifique de zéro.

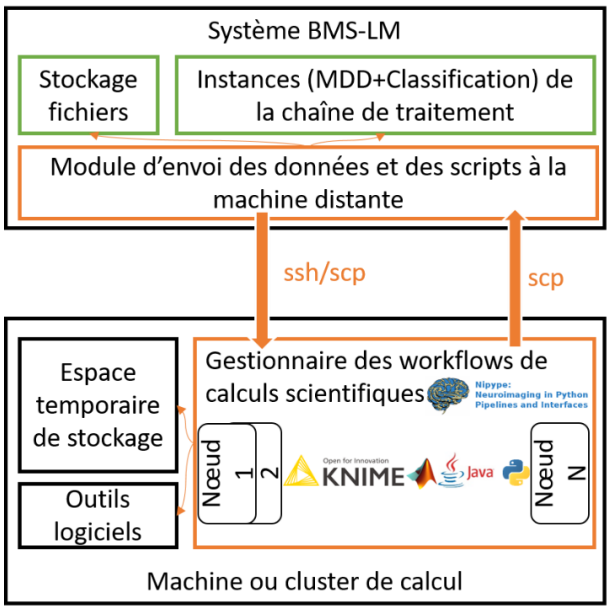


Figure 85 Composants techniques de l'intégration des calculs scientifiques dans le système BMS-LM

La Figure 85 présente les différents éléments logiciels impliqués dans l'intégration des calculs scientifiques. Tout d'abord, une chaîne de traitement est spécifiée dans le système BMS-LM via les

instances du MDD et de la Classification. Les informations, données d'entrée et scripts, qui y sont liés, sont transférés ensuite à la machine de calcul via un module dédié du système BMS-LM. La communication entre ce module et le gestionnaire de workflow de la machine distante s'effectue via le protocole de communication sécurisé SSH et la copie de fichiers via SCP. La machine de calcul doit être équipée d'outils logiciels permettant l'exécution des scripts utilisés dans la chaîne de traitement. Un stockage temporaire est utilisé pour recevoir les données à analyser. Les scripts de tout langage (ou outil) de programmation (Java, Python, Matlab, Knime, etc.) peuvent être enveloppés dans des nœuds Nipype (Gorgolewski et al., 2011) afin de permettre leur gestion et leur exécution automatique. Le gestionnaire de workflow Nipype est la brique centrale permettant l'ordonnancement et l'exécution des calculs scientifiques.

Lors de cette intégration, le système BMS-LM se charge de la traçabilité de l'exécution d'une chaîne de traitement du début jusqu'à la fin, tandis que le gestionnaire de workflow Nipype se charge de son exécution et de la remontée de données vers le système BMS-LM. Pour rappel, les objets du MDD modélisant les chaînes de traitement sont : les PCD, PUD, PCR, PUR, WFI, et PCP. Elles sont rappelées dans la Figure 86 ainsi que les objets qui y sont liés. Dans le cadre d'une « étude (STU) », un « sujet de l'étude (SSU) » ou un « groupe de sujets (SGP) » référence les « résultats de traitement (PCR) » qui lui sont associés (voir Figure 86). Chaque PCR référence « les configurations et données d'entrée (WFI) » de « la chaîne de traitement (PCD ») qui l'a engendré. Chaque PCR est composé de plusieurs « résultats d'unités de traitement (PUR »). Chaque PUR référence « les paramètres de calcul (PCP) » qui ont été utilisés pour exécuter le PUD (unité de traitement à exécuter). La provenance du PUR est déterminée par des liens avec : la version logicielle (STL) utilisée pour le générer, le PUR prédécesseur dans la chaîne de traitement, d'éventuelles « données de références (RFD) », les données brutes (ACQ, DUR) qui ont été analysées en entrée pour le générer en sortie.

La classe « Workflow Input (WFI) » du MDD est une classe centrale dans l'intégration des données scientifiques. Elle détermine les différents éléments en entrée nécessaires au lancement d'une chaîne de traitement depuis le système BMS-LM. Le WFI référence les types de données d'entrée (ACD, DUD) ainsi que la machine de calcul sur laquelle le traitement doit être lancé. Il référence aussi les RFD nécessaires à l'exécution du calcul scientifique sans oublier la définition de la chaîne de traitement (via les PCDs composés de PUD ; voir Figure 86)

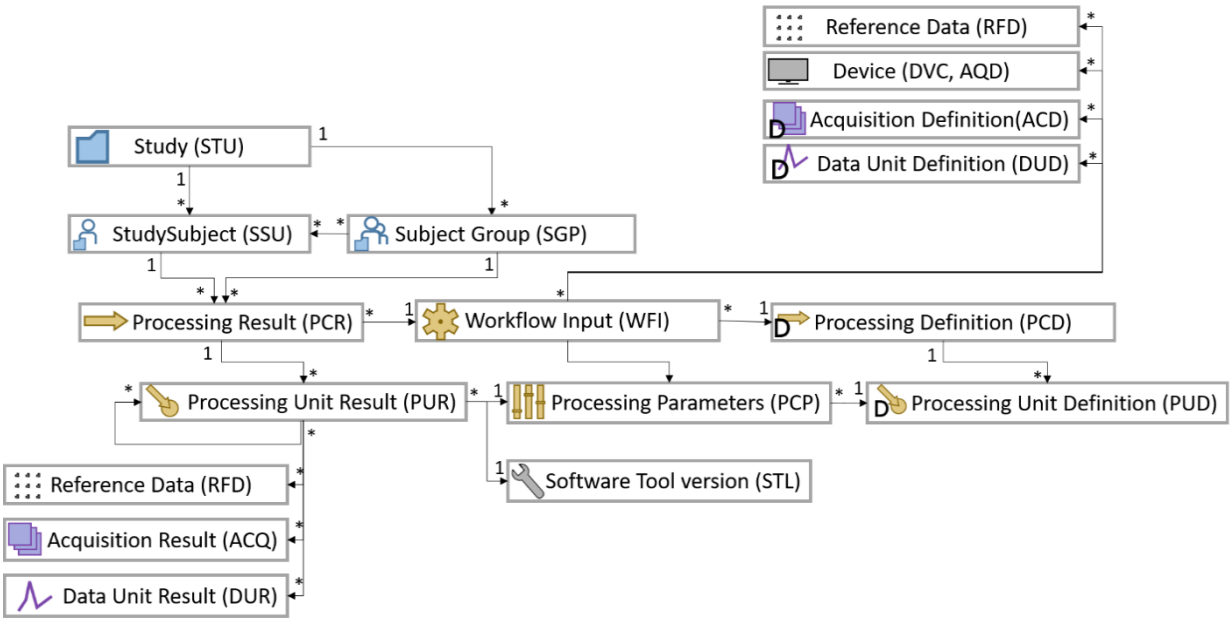


Figure 86 Les objets du MDD modélisant une chaîne de traitement et leurs liens

Pour lancer une chaîne de traitement préconfigurée, l'utilisateur prépare un « groupe de sujets (SGP) » qui va fournir les données d'entrée, ainsi qu'un « workflow input (WFI) » qui décrit les différents éléments d'entrée nécessaires à l'exécution de la chaîne de traitement. Le WFI et le SGP sont envoyés au module de gestion des calculs scientifiques du système BMS-LM via un raccourci clavier (CTRL+P). En plus du WFI et du SGP qui doivent être préparés par l'utilisateur du workflow, la machine qui exécutera le calcul scientifique doit être équipée d'un nœud dédié à son environnement logiciel et le cas échéant, une interface logicielle dédiée au calcul scientifique. Cette interface explicite la connexion entre le système BMS-LM et la machine de calcul pour le bon déroulement de l'exécution de la chaîne de traitement. Elle est préparée par une personne experte en intégration de workflow. Pour résumer, voici les étapes d'exécution d'un calcul scientifique depuis un système BMS-LM :

- La remontée de données au cluster de calcul ou « *Stagging* » :
    - La copie des fichiers d'entrée du serveur BMS-LM au serveur distant de calcul dans un espace de stockage que nous appelons « espace temporaire de stockage ». L'identification des fichiers s'effectue via une jointure entre le WFI et le SGP.
    - La copie des scripts à exécuter ainsi que de leurs paramètres d'entrée grâce aux objets du MDD : PCD, PCP et PUD.
- L'exécution des scripts :
    - L'exécution s'effectue via le gestionnaire de workflow Nipype (Gorgolewski et al., 2011) installé sur la machine de calcul et qui est équipé par un nœud dédié au type de scripts en entrée : Matlab, Knime[73] ,Python, Java, etc.
- Le rapatriement des données résultantes :
    - Une fois les traitements terminés, le gestionnaire de workflow renvoie les données de sortie au système BMS-LM afin qu'il instancie les objets du MDD (PCR, PUR), ainsi que leurs « Classification ».

Comme pour la méthode d'intégration de données, des étapes préparatoires du système BMS-LM, de la machine (ou cluster) de calcul, et du calcul scientifique lui-même s'imposent. La préparation du calcul scientifique consiste à restructurer et modéliser son algorithme en nœuds indépendants avec différentes entrées sorties pour assurer la traçabilité inter-étapes, il devient une chaîne de traitement (PCD). La préparation du système BMS-LM consiste à ajouter les instances des objets du MDD décrivant la provenance du calcul scientifique, ainsi que les classes et attributs de la « Classification » le décrivant spécifiquement. La préparation de la machine qui exécutera le calcul scientifique s'effectue en ajoutant les éléments logiciels permettant d'exécuter la chaîne de traitement depuis le gestionnaire des workflows Nipype (Gorgolewski et al., 2011). Nipype doit être installé sur la machine de calcul et communiquer avec le système BMS-LM. Pour intégrer un workflow de calcul scientifique, il faut donc :

- Restructurer le calcul scientifique à intégrer en une chaîne traitement afin de pouvoir le modéliser en nœuds indépendants avec différentes entrées sorties, pour assurer la traçabilité inter-étapes.
- Préparer les instances correspondantes dans le système BMS-LM : PCD, PUD, PCP, WFI. Il faut aussi préparer les classes de domaine correspondantes dans la Classification.
- Préparer un nœud ou une interface Nipype décrivant les étapes du workflow, leurs entrées, leurs sorties et leurs paramètres.
- Équiper la machine de calcul d'outils spécifiques et nécessaires à l'exécution du workflow. Par exemple, pour un code Matlab, il faut s'assurer que la bonne version du logiciel est installée sur la machine.
- Lancer le workflow depuis le système BMS-LM et récupérer les instances de PCR et PUR depuis la machine de calcul.

[73] https://www.knime.com/knime-server

Une application de cette méthode pour un calcul scientifique en Matlab au laboratoire LRI est présentée dans le chapitre VI.

IV.2.2.2. Méthode d'intégration partielle des calculs scientifiques

La méthode d'intégration « partielle » est adaptée aux applications logicielles « maison » ou « locales » pour le calcul scientifique avec interaction utilisateur via des interfaces graphiques. Lors de l'intégration partielle de ce calcul, le système BMS-LM est utilisé comme un serveur de base de données où l'application locale cherche les données d'entrée et dépose les données intermédiaires et celles de sortie. Le traitement est ainsi lancé et exécuté hors du système BMS-LM, dans un environnement client plus adapté et plus familier pour l'utilisateur : i.e. sa machine locale, une station dédiée au calcul scientifique, etc. Le schéma Figure 87 ci-après représente les différents éléments techniques mis en jeu lors de cette intégration. L'intégration partielle est rendue possible grâce à l'API REST du système BMS-LM et aux librairies de fonctions associées qui sont fournies par l'entreprise Fealinx dans plusieurs langages de programmation (Matlab, Python, etc.).

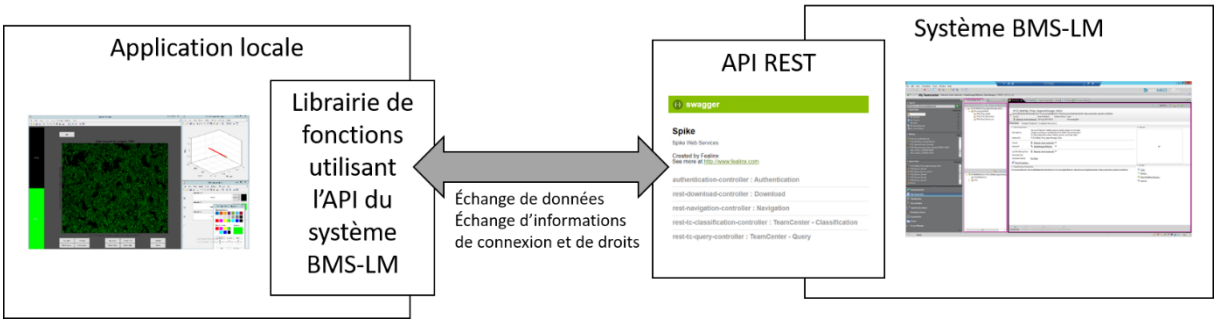


Figure 87 Intégration partielle des calculs scientifiques avec interaction utilisateur

La différence avec l'intégration totale est que l'exécution de la chaîne de traitement s'effectue indépendamment de Nipype et d'un cluster de calcul préconfiguré. Le lien avec le système BMS-LM est un lien de traçabilité uniquement. Les mêmes objets du MDD présentés Figure 86 sont mis en jeu lors de cette intégration. Les étapes de l'intégration partielle d'un calcul scientifique avec interaction utilisateur sont les suivantes :

- La restructuration et la modularisation de l'application logicielle locale afin de permettre la traçabilité inter-étapes.
- La préconfiguration du système BMS-LM pour que la complexité des liens de provenance soit transparente pour l'utilisateur : Ajout des WFIs, PCPs, PUD, PCD avant l'utilisation de l'application logicielle.
- La définition des règles de communication entre l'application locale et le système BMS-LM ainsi qu'un espace temporaire de stockage pour permettre la bonne exécution de la chaîne de traitement et la remontée automatique des résultats d'analyse vers le système BMS-LM.
- L'ajout d'interfaces graphiques pour :
    - La connexion au système BMS-LM via identifiant/mot de passe
    - La récupération des données à analyser depuis le système BMS-LM
- L'ajout d'un module de remontée de données dérivées vers le système BMS-LM depuis l'application cliente.

Une application de cette méthode pour un calcul scientifique en Matlab avec interaction utilisateur a fait l'objet d'un encadrement de stage de Master 1 et est présentée dans le chapitre VI.

CONCLUSION DU CHAPITRE IV

Dans ce chapitre, nous avons présenté les briques fonctionnelles du système BMS-LM que nous proposons. Ses fonctionnalités permettent de gérer les données de recherche tout au long du cycle de vie d'une étude de recherche biomédicale et répondent aux 19 besoins de la communauté des chercheurs dans le domaine biomédical. Nous avons expliqué l'architecture technique de la solution actuelle implémentée en « *4-tiers* » et nous avons analysé les différents niveaux de réponses apportés aux besoins de la communauté en Annexe B. Nous avons présenté les différentes évolutions effectuées sur le modèle MDD générique du BMS-LM, et de la « Classification » spécifique correspondante. Ce MDD a été utilisé dans notre méthode « générique » d'intégration de données et nos deux méthodes « partielle » et « totale » d'intégration de calcul scientifique dans un contexte d'une hétérogénéité multiple de données. Les cas d'application de nos propositions pour la recherche préclinique sont présentés dans le chapitre VI.

Le système BMS-LM peut être considéré comme un système pour la gestion de données scientifiques régi par le duo « MDD+Classification », son KOS (Système d'Organisation des Connaissances). Dans cette configuration, le MDD BMS-LM représente le niveau générique-noyau, et la « Classification » représente le niveau spécifique-domaine.

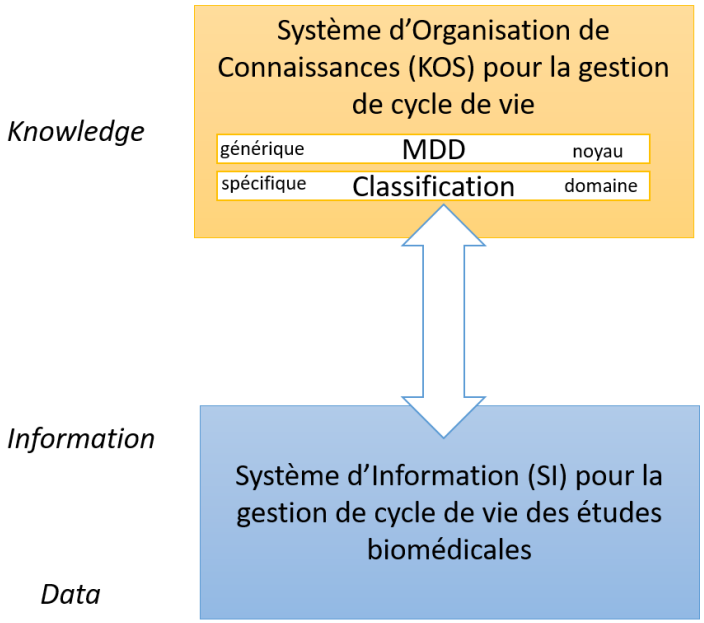


Figure 88 Les deux composants essentiels de l'architecture du système BMS-LM

Dans le cadre de l'ouverture de données de recherche, l'interopérabilité sémantique de ce KOS avec les KOS publiés du domaine est un levier pour faciliter la compréhension de données et ainsi leur partage et leur réutilisation. De même, son interopérabilité avec les KOS locaux des équipes de recherche renforce son utilité pour la gestion au jour le jour des données. Nous nous intéressons dans le chapitre suivant à cette problématique afin d'apporter une réponse conceptuelle dans le domaine de l'organisation des connaissances en exploitant les avancées en ingénierie ontologique.

Chapitre V. Ontologie BMS-LM : Une construction multi-niveaux pour l'interopérabilité sémantique entre KOS

Dans le chapitre III, nous avons effectué un état de l'art sur la gestion, l'organisation, et l'ingénierie des connaissances (KM, KO, KE) afin d'explorer l'existant en réponse à notre deuxième question de recherche « Problème n°2 : Comment assurer la compréhension des données hétérogènes de la recherche biomédicale lors d'une réutilisation ultérieure ? ». L'annotation de ces données à l'aide des KOS locaux est fréquente en recherche, ce qui est une nécessité à cause, entre autres, du décalage entre l'évolution des terminologies scientifiques et la mise à jour qui leur correspond dans les KOS publiés. Cependant, les KOS locaux ne sont pas partageables ailleurs et nécessitent une projection dans l'espace des KOS publiés, pour que les données soient compréhensibles par un tiers. La mise en place de ce processus d'intercommunication entre systèmes d'organisation de connaissances (KOS) relève de l'interopérabilité sémantique, notre objectif de recherche n°3 (voir Figure 89).

Nous avons retenu au chapitre III, que la modélisation des connaissances à l'aide des ontologies, ainsi que la méthode de construction d'ontologies « haut/domaine/local », améliore l'interopérabilité sémantique des données dont elle modélise les connaissances. Dans ce chapitre, nous présentons notre méthode de construction d'ontologie multi-niveaux pour la mise en œuvre de l'interopérabilité sémantique des données scientifiques et l'ontologie qui en résulte.

Au chapitre précédent, nous avons présenté le système BMS-LM avec ses briques fonctionnelles et techniques. Ce système est régi par le KOS composé d'un modèle de données (MDD) générique et une « Classification » spécifique. Dans le cadre de la mise en place de l'interopérabilité sémantique entre KOS de données hétérogènes, nous proposons une évolution du KOS du système BMS-LM (MDD+Classification) en une ontologie multi-niveaux : l'ontologie BMS-LM. Ce chapitre est consacré à sa définition et à l'explication des méthodes utilisées pour sa construction.

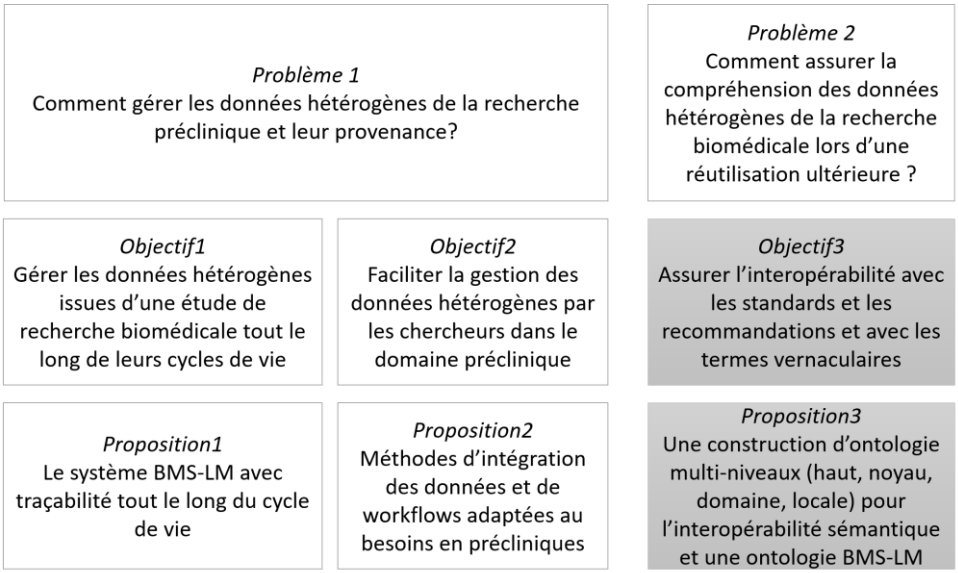


Figure 89 Objectifs de recherche pour assurer la compréhension des données hétérogènes lors d'une réutilisation ultérieure

V.1. MÉTHODE DE CONSTRUCTION DE L'ONTOLOGIE MULTI-NIVEAUX

Dans cette section, la méthode de construction d'ontologies « haut/noyau/domaine » de (Patel et al., 2005) a été reprise et adaptée au contexte de l'hétérogénéité et de l'évolutivité en recherche biomédicale. Un niveau pour les terminologies locales vernaculaires a été ajouté dans ce sens. Dans les paragraphes suivants, nous présentons notre démarche pour la transformation du KOS du système BMS-LM (MDD + Classification) en une ontologie multi-niveaux (haut, noyau, domaine et local) et les différents choix et méthodes que nous avons utilisés. La Figure 90 décrit l'état de départ de cette construction.

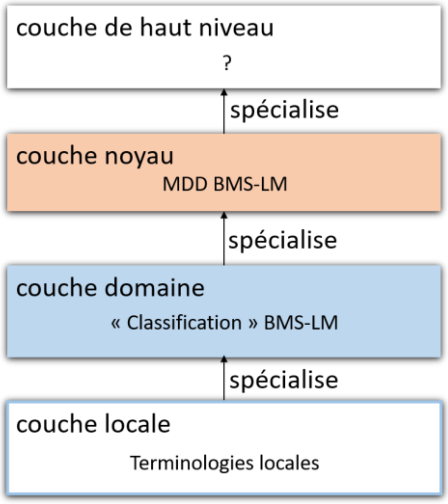


Figure 90 Les quatre niveaux au départ de la construction de l'ontologie multi-niveaux BMS-LM

V.1.1. CHOIX DE L'ONTOLOGIE DE HAUT NIVEAU

Une liste de critères a été constituée pour choisir l'ontologie de haut niveau parmi celles reconnues par la communauté dans le domaine biomédical : DOLCE ou BFO (voir Tableau 14). Ces critères sont : l'activité de la communauté et le nombre de citations et de réutilisations de l'ontologie dans le domaine biomédical. En particulier, ont été pris en compte le nombre de réutilisations dans BioPortal (Whetzel et al., 2011), le nombre d'articles publiés sur Pubmed[74], et les citations de Semantic Scholar[75] et Google Scholar[76].

Tableau 14 Le choix d'une ontologie de haut niveau : résultats de la comparaison entre les ontologies de haut niveau DOLCE et BFO selon le nombre de réutilisations de l'ontologie, le nombre de citations et l'activité de la communauté

	CRITÈRE	DOLCE	BFO	SOURCE
1	nombre de réutilisations d'ontologies biomédicales	34 modules De l'ontologie OntoNeuroLOG (ONLOG)	220 ontologies dans la OBO Foundry	Les ressources officielles respectives http://www.obofoundry.org/ http://neurolog.unice.fr/ontoneurolog/v3.0/Documentation_OntoNeuroLOGv3.pdf
2	nombre de réutilisations dans des ontologies BioPortal	4 ontologies (ONLOG : physical object)	246 ontologies (BFO : material entity)	BioPortal calculant le nombre de SAME_URI dans l'onglet "mappings > colonne source" de respectivement (ONLOG : physical object) et (BFO : material entity).
3	Nombre d'articles dans Pubmed	13 (ontologie DOLCE)	41 (ontologie BFO)	Pubmed à l'aide de l'outil de recherche avec les mots-clés (DOLCE ontology) et (BFO ontology) au 31/10/2019
4	activité de la communauté	non maintenue	maintenue	FAIRsharing.org (Sansone et al., 2019) visité le 31/10/2019

[74] https://pubmed.ncbi.nlm.nih.gov/
[75] https://www.semanticscholar.org/
[76] https://scholar.google.fr/

5	nombre de citations	1263 (DOLCE+ WonderWeb)	42(BFO) 2137 (OBO Foundary)	Google Scholar pour les articles suivants : wonderweb (Masolo et al., 2003), DOLCE (Gangemi et al., 2002), BFO (B. P. Smith et al., 2005) OBO Foundary (B. Smith et al., 2007) au 31/10/2019
6	nombre d'articles en sémantique	44 résultats (ontologie de haut niveau DOLCE)	17 résultats (ontologie de haut niveau BFO)	Semantic scholar utilisant l'outil de recherche avec les mots-clés (DOLCE upper level ontology) et (BFO upper level ontology) au 23/02/2020

Selon les résultats du tableau comparatif, l'ontologie BFO et le consortium correspondant « Open Biological and Biomedical Ontology (OBO Foundry) » ont été choisis, pour les raisons suivantes :

1. L'ontologie BFO est réutilisée par plus d'ontologies biomédicales que DOLCE (voir ligne 1 et 2 du Tableau 14)
2. L'ontologie BFO est citée dans plus d'articles que DOLCE dans le domaine biomédical (voir lignes 3 du Tableau 14). Ce n'est cependant pas le cas en général (voir lignes 5 et 6 du Tableau 14). Dans l'ensemble, l'indicateur du domaine biomédical était plus pertinent pour l'ontologie BMS-LM.
3. La communauté BFO est plus active dans le domaine biomédical que celle de DOLCE (voir ligne 4 du Tableau 14).

Le choix de l'ontologie BFO comme ontologie de haut niveau, impose de suivre son engagement ontologique et ses 11 principes de construction d'ontologie[77] à savoir : la « Disponibilité » et la « Licence » de l'ontologie, la « Réutilisation d'autres ontologies : crédit et annotations », l'utilisation d'un « Format commun »,et d' « Identifiant dans un format <IDSPACE>_<NUMBER> », l' « Utilisation de versions », un « Champ d'application clair », le recours aux « Définitions textuelles », la réutilisation de l'ontologie de relations (RO) au niveau « Relations », la « Documentation », la documentation de la « Pluralité d'utilisateurs », l' « Engagement à Collaboration », et enfin le « Locus d'Autorité ». Ces principes sont appliqués et expliqués dans la section V.2.3 pour la construction de l'ontologie multi-niveaux BMS-LM.

V.1.2. CONSTRUCTION DU NIVEAU NOYAU

Pour rappel, une ontologie noyau est une ontologie qui représente les concepts génériques de base d'un grand domaine et qui peut intégrer d'autres ontologies de domaine plus spécifiques. La « gestion de cycle de vie des études de recherche biomédicale et de leur provenance » est un domaine large qui a des sous-domaines et qui diffère de la « modélisation du monde », rôle des ontologies de haut niveau. Nous avons identifié le MDD générique du système BMS-LM comme un KOS noyau et nous avons procédé à sa transformation en ontologie.

Tout d'abord, nous avons renommé les classes du MDD en cohérence avec les règles de nommage de BFO et « OBO Foundry », pour obtenir les concepts de l'ontologie noyau. Nous avons ensuite construit les relations entre ces concepts en utilisant l'ontologie de relation (RO) du consortium OBO et l'ontologie de Provenance PROV-O. Les liens de spécifications entre les deux niveaux (haut et noyau) ont été construits après analyse des ontologies publiées sur BioPortal. En effet, nous nous sommes appuyés sur des ontologies publiées sur Bioportal et réutilisant BFO, pour déterminer si par exemple, un « sujet de l'étude SSU » est communément défini comme étant un `BFO:role` ou un `BFO:material entity`.

Nous avons ainsi établi un « script1 » de recherche dans Bioportal pour la réalisation de cette tâche, détaillées en Annexe C. Ceci nous a permis d'étudier les KOS publiés et leurs concepts existants dans

[77] http://www.obofoundry.org/principles/fp-000-summary.html

le but de construire le niveau noyau en cohérence avec les travaux de la communauté scientifique. L'ontologie noyau BMS-LM, la troisième version[78] du KOS du système BMS-LM a été construite en utilisant ces méthodes. Elle est détaillée en §V.2.1.

V.1.3. NIVEAU DOMAINE ET INTEROPÉRABILITÉ AVEC LES KOS PUBLIÉS

La règle générale que nous avons respectée est de réutiliser au maximum les concepts des KOS publiés afin d'améliorer l'interopérabilité sémantique. Le choix de BFO comme ontologie de haut niveau permet d'avoir un terrain d'interopérabilité sémantique avec plus de 220 ontologies du consortium OBO Foundry. Pour construire la couche domaine, il faut identifier les concepts et ontologies de domaine à réutiliser. Pour cela, nous avons exploré les KOS biomédicaux publiés et nous avons utilisé un « script2 » pour la recherche dans Bioportal (expliqué en Annexe C). Le processus ressemble à celui utilisé pour construire la « Classification » du système BMS-LM. La différence est qu'avec les outils en ingénierie ontologique : Protégé, OWL, SKOS, etc., l'alignement entre les concepts de la couche domaine et ceux des ontologies réutilisées est explicité en utilisant des annotations comme `skos:narrowMatch`, `skos:broadMatch`, `IAO:imported from` et `OWL:equivalentClass` (voir §V.2.5). Ainsi, l'interopérabilité sémantique est activée sur le plan logique informatique en plus du plan conceptuel.

Le niveau domaine ne consiste pas en une seule ontologie comme pour le niveau haut et noyau. En effet, pour chaque nouveau contexte d'application du système BMS-LM et de son KOS (recherche en neuroimagerie, recherche préclinique, recherche en psychologie, etc.), une ontologie de domaine est ajoutée comme module sous le niveau noyau. Pour chaque nouvelle ontologie de domaine utilisant l'ontologie noyau BMS-LM, il faut définir un espace de noms (« *namespace* ») correspondant et un fichier OWL séparé pour bien définir les modules relatifs à chaque contexte. Par exemple, pour notre contexte d'application, nous avons défini le « namespace DRIVE (nom du projet de recherche) » dans lequel sont définis les niveaux « domaine » et « local » de l'ontologie multi-niveaux, ajoutées spécifiquement pour le projet.

Pour identifier les concepts du domaine d'application, il est nécessaire d'effectuer une enquête de terrain. II faut collecter des jeux de données typiques et minimaux pour chaque sous-domaine d'intérêt dans le contexte d'application. Par exemple, les données d'un examen d'IRM choisi par l'opérateur, les rapports d'expérimentations, les images brutes, leurs analyses, les données statistiques correspondantes, la publication scientifique qui les références, etc. Tous les termes qui peuvent être identifiés dans ces documents d'exemple doivent être ajoutés à ce qu'on appelle le « lot initial des termes ». Ce lot de termes nous permettra d'élaborer une recherche plus poussée afin d'identifier les concepts et les ontologies qui s'y rapportent, et juger de leur pertinence.

À titre d'exemple, en IRM, nous pouvons noter le terme « Gadolinium ». Une recherche sur BioPortal nous confirme qu'il est bien référencé et décrit par les ontologies biomédicales. Cette recherche peut aussi nous aider à trouver la meilleure méthode d'intégration de ce terme dans la couche domaine. Par exemple, nous trouverons des ontologies qui définissent l'atome de gadolinium de point de vue chimique comme CHEBI,[79] mais ceci ne nous intéresse pas en imagerie. Le « Gadolinium » nous intéresse dans l'agent de contraste au Gadolinium utilisé dans un examen IRM. Cela nous amène à plutôt considérer la mention du « Gadolinium » dans le NCIT comme indiqué dans la Figure 147 ci-après. L'exploration des concepts voisins dans NCIT nous donne accès à d'autres types d'acquisitions (ou séquences) IRM, qui sont potentiellement intéressantes pour l'ontologie du domaine imagerie.

[78] La première version est le : MDD BMI-LM, la deuxième version est le : MDD BMS-LM
[79]http://purl.obolibrary.org/obo/CHEBI_50161

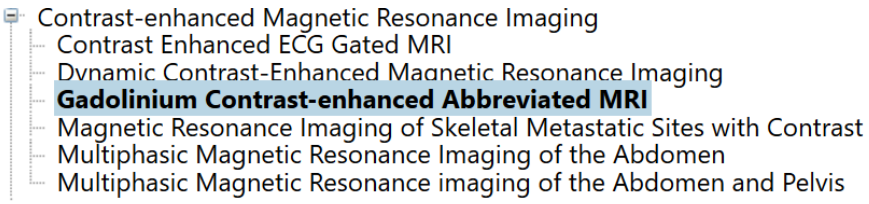


Figure 91 Mention du Gadolinium dans NCIT comme « agent de contrast »

Par ailleurs, le « lot initial de termes » doit être validé par le personnel concerné. Ils sont généralement les utilisateurs clés du système d'information. Ils doivent donc être à minima avertis de l'intérêt de ce travail ontologique et de l'importance de la gestion des données de recherche en cohérence avec les standards. La liste de questions ci-après peut guider la demande de validation d'un terme « T » :

- Que signifie le terme « T » ?
    - Si la personne référente répond avec « confiance » et donne des détails sur le terme, ceci est probablement un terme à retenir.
- « T » est-il indispensable pour décrire une expérimentation/une analyse ?
    - Il faut se concentrer à ce stade sur la « longueur » de la réponse. Si la personne donne une réponse « riche », ceci est probablement un terme à retenir.
- Est-ce qu'il y a des alternatives à « T » ?
    - Cette question permet d'élargir le spectre et d'identifier potentiellement d'autres termes intéressants. Encore une fois, la réponse « certaine » nous confirme que le terme est à retenir.
- Si « T » est une valeur numérique, on peut poser la question sur les différentes valeurs possibles de T afin d'ajouter ceci comme contrainte à intégrer plus tard dans l'ontologie, mais aussi de confirmer que le terme est bien à retenir.

Dans les questions précédentes, « T » peut être remplacé par un ensemble de termes (expression) ou une liste de valeurs. La liste précédente est une liste "antisèche" pour amorcer la discussion. Elle permet d'éviter les questions vagues et directes, qui ne sont pas utiles dans ce cas de figure, telles que celle-ci :

- Est-ce que tu valides que « T » est intéressant pour l'ontologie d'application ?

La validation serait alors indirecte, en fonction du « ton » et de la « richesse » de la réponse reçue pour chaque terme ou lot de termes « T ». Les utilisateurs du SI, sont spécialisés dans la pratique et l'application de leur cœur de métier, et connaissent peu le domaine de la modélisation ontologique. Ils ne peuvent donc pas répondre seuls à ce type de question qui demande une passerelle entre leur champ disciplinaire et le champ de la gestion de l'information.

Parmi le « lot initial de termes », il y a ceux qui seront validés par les utilisateurs comme décrit précédemment et ceux qui seront validés par « voie systématique ». Cette dernière se base sur le critère de « La redondance » comme critère d'inclusion des termes, i.e. si un terme a été identifié dans plusieurs ontologies de domaine (publiées sur BioPortal, FAIRsharing ou autres), il obtient une légitimité de la part de la communauté de ce domaine et il sera intégré à l'ontologie en cours de construction. Les termes identifiés par la « voie systématique » doivent faire l'objet d'une analyse au cas par cas en vue de vérifier s'ils sont cohérents avec le domaine d'application ; aujourd'hui et sur le long terme. Les différents cas de figure pour l'intégration des termes au « lot initial » sont résumés dans le Tableau 15 ci-après.

Dans le cas où la validation utilisateur est obtenue, mais, sans que le terme « T » soit identifié dans aucune ontologie de domaine, le terme est appelé « terme local ». Il s'agit d'un terme essentiel pour l'annotation et l'organisation des données dans le contexte d'application, mais qui n'est pas encore publié.

Tableau 15 Décisions d'inclusion des termes initiaux en fonction des résultats de leurs validations

Validation utilisateur	Validation systématique via les ontologies de domaine : Oui	Validation systématique via les ontologies de domaine : Non
Oui	À intégrer dans le niveau domaine	À intégrer dans le « niveau local »
Non	À décider au cas par cas en fonction de l'utilité du terme dans le futur	À exclure

V.1.4. INTÉGRATION DES KOS LOCAUX DANS UN NIVEAU LOCAL

Le recours aux termes locaux pourrait être expliqué par l'évolution croissante des domaines de recherches d'un côté, et le temps long pour la publication des standards de référence d'un domaine, de l'autre côté. Il pourrait aussi être expliqué par les domaines « niches » qui ne seront jamais publiés en tant que standard de référence puisque les termes qui les constituent sont tous des termes spécifiques à ce domaine « niche ». Il peut aussi s'agir de raccourcis pratiques permettant des notations rapides. Dans d'autres configurations, les termes locaux peuvent représenter des termes techniquement intéressants pour la gestion interne des données au sein d'une équipe de recherche, dans le contexte précis du domaine d'application, mais qui ne sont pas utiles à partager avec la large communauté de ce domaine.

Les intérêts d'utilisation des termes locaux sont donc multiples. Par conséquent, nous avons ajouté aux trois niveaux « haut/noyau/domaine », un quatrième niveau, que nous avons appelé : « niveau local ». Celui-ci regroupe les termes d'application, les termes de fonctionnement technique de la base de données, les termes qui relèvent plutôt du jargon de recherche, mais qui ont une importance capitale dans le contexte précis d'application. Ils sont conformes à l'activité réelle des producteurs de données. Tous les concepts publiés du domaine appartiennent donc à la couche domaine, et tous les termes locaux du domaine « niche » ou du contexte d'application appartiennent à la couche locale. L'ontologie résultante de cette approche est une ontologie à 4 niveaux comme résumé dans le schéma de la Figure 92 ci-après.

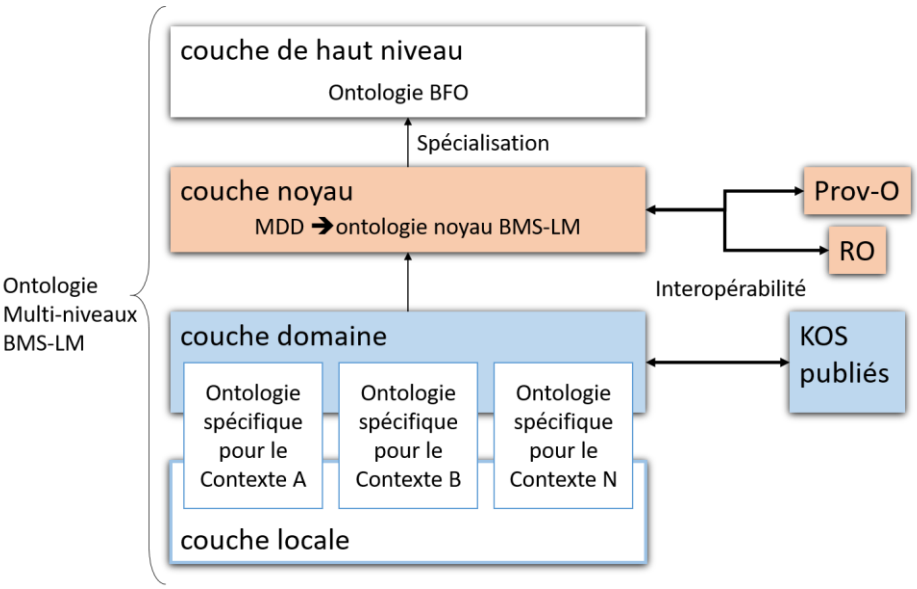


Figure 92 Les quatre couches de l'ontologie multi-niveaux BMS-LM

V.2. ONTOLOGIE MULTI-NIVEAUX BMS-LM

Dans cette section, nous détaillons le résultat de l'application de la méthode de construction d'ontologie à 4-niveaux « haut/noyau/domaine/local » pour le BMS-LM. Tout d'abord sont présentés les concepts et les relations de l'ontologie noyau BMS-LM, ainsi que leurs parents dans l'ontologie BFO. Ensuite, nous détaillons les règles de constructions du point de vue de l'ingénierie ontologique et les différents choix d'encodage des propriétés que nous avons été amenés à réaliser. Enfin, nous présentons des exemples de la couche domaine pour le laboratoire d'accueil. La mise en correspondance de l'ontologie multi-niveaux BMS-LM avec le MDD et la « Classification » du système BMS-LM est explicitée au fur et à mesure.

V.2.1. ONTOLOGIE NOYAU BMS-LM

Le choix du niveau haut étant effectué, nous présentons dans cette section le niveau noyau de l'ontologie BMS-LM.

V.2.1.1. Les concepts noyau de l'ontologie BMS-LM

Les concepts de l'ontologie noyau BMS-LM sont au nombre de 29. Ils sont répartis sur tout le cycle de vie d'une étude de recherche biomédicale. Ces concepts sont des concepts génériques qui constituent le schéma global des données et de leurs provenances. L'ontologie noyau BMS-LM représente une évolution sémantique du MDD générique BMS-LM. Dans le Tableau 16 ci-après, chaque concept de l'ontologie BMS-LM est présenté via son nom anglais, sa définition OBO-compatible, son parent dans BFO ainsi que l'étape du cycle de vie où il est défini et/ou utilisé.

Tableau 16 Liste des concepts de l'ontologie BMS-LM avec leurs définitions et leurs parents respectifs

	Nom de concept	*Parent dans BFO*	*Définition*	*Étape du cycle de vie*
1	***biomedical research study***	*planned process*	Un processus planifié* qui se déroule en plusieurs étapes, de la spécification de l'étude à la publication des résultats. source : ontologie HUPSON	1-spécification de l'étude
2	***studied organism***	*material entity*	Une entité matérielle qui est un système vivant individuel, tel qu'un animal, une plante, une bactérie ou un virus, qui est capable de se reproduire, de croître et de se maintenir dans le bon environnement. Un organisme peut être unicellulaire ou composé, comme les humains, de plusieurs milliards de cellules divisées en tissus et organes spécialisés. source : OBO	1-spécification de l'étude
3	***biological sample preparation***	*planned process*	Un processus planifié qui consiste à transformer un objet biologique en un échantillon biologique prêt pour son examen	2-acquisition des données brutes
4	***biological sample***	*object*	Un objet qui a été prélevé à un organisme vivant : organe, tissu biologique, cellules, broyat, etc.	2-acquisition des données brutes

5	***sample preparation protocol***	*protocol*	Un protocole pour décrire les étapes nécessaires à la préparation des échantillons biologiques.	1-spécification de l'étude 2-acquisition des données brutes
6	***agent administration***	*planned process*	Un processus planifié qui consiste à introduire un produit pharmaceutique qui agit sur le tissu biologique dans un organisme vivant, ou dans un échantillon biologique.	2-acquisition des données brutes
7	***agent product***	*object*	Un objet pharmaceutique qui est produit pour agir sur des organismes vivants ou des échantillons biologiques	2-acquisition des données brutes
8	***agent administration protocol***	*protocol*	Un protocole pour décrire les étapes nécessaires à l'administration de l'agent (produit qui agit)	1-spécification de l'étude 2-acquisition des données brutes
9	***device***	*processed material*	Une entité matérielle qui est conçue pour remplir une fonction dans une expérimentation scientifique (examen, traitement, calcul scientifique), mais qui n'est pas un réactif. source : OBO	1-spécification de l'étude 2-acquisition des données brutes 3-production des données dérivées
10	***examination***	*planned process*	Un processus planifié qui consiste à acquérir via différents moyens des informations sur un objet d'intérêt afin de le décrire et caractériser à un instant T d'un point de vue biologique et médical.	2-acquisition des données brutes
11	***examination protocol***	*protocol*	Un protocole décrivant les étapes nécessaires pour l'examen d'un objet d'intérêt de point de vue biologique ou médical.	1-spécification de l'étude 2-acquisition des données brutes
12	***data acquisition***	*planned process*	Un processus planifié qui consiste à acquérir, d'une façon homogène et dans une période de temps précise et indivisible, des données dans le cadre d'un examen.	2-acquisition des données brutes
13	***data acquisition protocol***	*protocol*	Un protocole pour décrire les étapes nécessaires à l'acquisition des données	1-spécification de l'étude 2-acquisition des données brutes
14	***raw data unit***	*information content entity (i-c-e)*	Un i-c-e, résultat d'un processus d'acquisition de données brutes d'une étude de recherche	2-acquisition des données brutes
15	***expected data unit***	*information content entity (i-c-e)*	Un i-c-e, qui décrit ce que l'on doit avoir comme données résultat d'un processus avant de produire les vraies données résultats de ce processus	1-spécification de l'étude 2-acquisition des données brutes 3-production des données dérivées
16	***data processing***	*planned process*	Un processus planifié qui consiste à traiter et analyser les données issues des différents examens afin de délivrer un résultat d'analyse. Il est composé d'une ou plusieurs unités de	3-production des données dérivées

			traitement (« data processing unit »)	
17	***data processing protocol***	*protocol*	Un protocole pour décrire les étapes nécessaires pour entreprendre une analyse ou un traitement de données	1-spécification de l'étude 3-production des données dérivées
18	***derived data unit***	*information content entity (i-c-e)*	Un i-c-e, résultat d'un processus d'analyse des données d'une étude de recherche	3-production des données dérivées
19	***data processing unit (d-p-u)***	*planned process*	Un processus planifié qui consiste à traiter ou analyser des données en entrée pour produire des résultats en sortie qui sont une entrée pour un d-p-u suivant. Cet enchaînement compose un processus de « data processing »	3-production des données dérivées
20	***processing unit protocol***	*protocol*	Un protocole pour décrire les étapes nécessaires pour l'exécution d'un d-p-u	1-spécification de l'étude 3-production des données dérivées
21	***workflow input profile***	*process profile*	Un profil de processus qui permet de "méta-décrire" un processus, en précisant les types d'entrées indispensables à son lancement dans le cadre d'un workflow (ou flux de traitement)	1-spécification de l'étude 3-production des données dérivées
22	***processing parameters profile***	*process profile*	Un profil de processus qui permet de métadécrire un processus, en précisant les paramètres indispensables à son exécution	3-production des données dérivées
23	***published content***	*information content entity (i-c-e)*	Un i-c-e qui représente tout contenu publié et référencé, utilisé ou publié dans le cadre d'une étude de recherche (hormis les données)	4-valorisation et publication
24	***biomedical intervention***	*planned process*	Un processus planifié qui consiste à mener une action qui modifie et influence un objet biologique dans le cadre d'une étude de recherche	2-acquisition des données brutes
25	***biomedical intervention protocol***	*protocol*	Un protocole pour décrire les étapes nécessaires pour mener une intervention biomédicale	1-spécification de l'étude 2-acquisition des données brutes
26	***software tool***	*plan specification*	Une spécification de plan composée d'une série d'instructions qui peuvent être interprétées par une unité de traitement ou directement exécutées par celle-ci pour analyser les données de l'étude. source : OBO	1-spécification de l'étude 2-acquisition des données brutes 3-production des données dérivées
27	***study subject***	*role*	Un rôle joué par un « studied organism » dans une étude biomédicale en participant à l'étude ou en donnant des échantillons pour l'étude	1-spécification de l'étude 2-acquisition des données brutes 3-production des données dérivées

28	***study subject group***	*object aggregate*	Une agrégation ou un ensemble de rôles de « study subject »	1-spécification de l'étude 2-acquisition des données brutes 3-production des données dérivées
29	***reference data***	*information content entity (i-c-e)*	Un i-c-e qui représente toute donnée utilisée comme référence dans une étude ou publiée en tant qu'ensemble de données référencées dans une étude	3-production des données dérivées 4-valorisation et publication

* : processus planifié ou « *planned process : A processual entity that realizes a plan which is the concretization of a plan specification. »*[80]

A ces 29 concepts, cinq autres ont été ajoutés pour représenter les étapes du cycle de vie d'une étude de recherche (voir Tableau 17). Il s'agit d'une évolution majeure par rapport au modèle de données BMS-LM qui n'avait pas de concepts explicites sur le cycle de vie. Nous avons effectué des recherches dans les ontologies du domaine industriel sur le PLM et les ontologies du domaine biomédicales et nous avons constaté que les étapes du cycle de vie ne sont pas modélisées dans les deux domaines. La seule référence pertinente identifiée est celle de l'ontologie UBERON du consortium OBO. Elle définit l'étape du cycle de vie d'un organisme comme suit « UBERON:0000105 - `life cycle stage` -: *A spatiotemporal region encompassing some part of the life cycle of an organism* »[81].La dimension spatio-temporelle est pertinente pour un organisme vivant, car il a une existence matérielle et physique dans l'espace, alors que, pour une étude, seule la dimension temporelle semble pertinente. Ainsi, le concept BMS-LM - `study lifecycle phase` - a été défini comme une sous-classe de `BFO:one-dimensional temporal region`. Les cinq concepts ajoutés sont détaillés dans le Tableau 17.

Tableau 17 Les concepts de cycle de vie de l'ontologie noyau BMS-LM

Concept	*Définition*	*Parent*
study lifecycle phase	Une région temporelle qui représente une étape bien distincte d'une étude de recherche	*one-dimensional temporal region*
1 study specification	La première phase d'une étude de recherche biomédicale	*study lifecycle phase*
2 study data collection	La deuxième phase d'une étude de recherche biomédicale	*study lifecycle phase*
3 study data analysis	La troisième phase d'une étude de recherche biomédicale	*study lifecycle phase*
4 study results publishing	La quatrième phase d'une étude de recherche biomédicale	*study lifecycle phase*

Tous ces concepts sont liés par des relations qui sont détaillées en §V.2.1.2. Dans la Figure 93 ci-après, on peut observer les regroupements de concepts en fonction de leurs parents dans BFO. Les concepts noyau de BMS-LM couvrent un spectre large de concepts de haut niveau (des entités matérielles, des processus, protocoles, entités contenant des informations, des rôles, des profils de processus), preuve de leur généricité. Des rôles joués par les chercheurs dans le cadre d'une étude biomédicale ont été ajoutés à l'ontologie BMS-LM et sont aussi présents dans la Figure 93.

Les fichiers produits dans le cadre d'une étude sont contenus dans les concepts `raw` et `derived data units`. La description de leurs conditions de production est répartie sur les autres concepts : le processus (et ses profiles), le protocole suivi, les rôles mis en jeu, les entités matérielles utilisées, ainsi que la phase du cycle de vie à laquelle ils appartiennent.

[80] Source : l'ontologie BFO. En français : Entité processuelle qui réalise un plan qui est la concrétisation d'une spécification de plan

[81] En français : Une région spatio-temporelle englobant une partie du cycle de vie d'un organisme

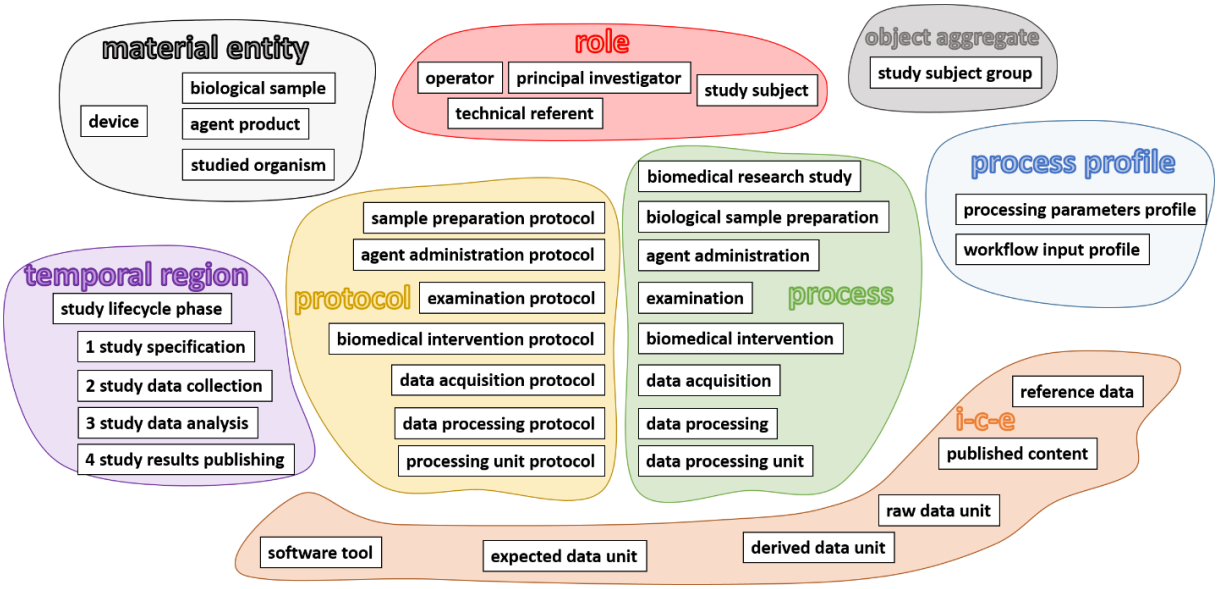


Figure 93 Les différents concepts de l'ontologie noyau BMS-LM et leurs parents respectifs dans BFO (i-c-e = information content entity)

L'ontologie BMS-LM a été conçue pour être interopérable avec l'ontologie PROV-O (expliqué en §III.2.2). Cela rend explicite la sémantique de la provenance dans BMS-LM et permet une interopérabilité sémantique avec les autres ontologies compatibles avec PROV-O. En pratique, PROV-O dispose, dans sa version initiale, d'une liste de trois concepts principaux : `Agent (PROV-O)`, `Entité` et `Activité` comme explicité en section III.2.2. Le Tableau 18 suivant présente la classification des concepts noyaux BMS-LM en fonction des concepts de PROV-O (`Agent (PROV-O)`, `Entité` et `Activité`).

Tableau 18 Répartition des concepts de l'ontologie BMS-LM selon leur rôle de provenance dans PROV-O

Agent (PROV-O)	Entité	Activité	non compatible
software tool	expected data unit	examination	
agent product	derived data unit	data acquisition	Protocols
study subject	raw data unit	data processing	
device	reference data	data processing unit	
	published content	agent administration	Process profiles
principal investigator	biological sample	biomedical intervention	
operator	study subject group	biological sample preparation	
technical referent	studied organism		
		biomedical research study	

Il est à noter que PROV-O ne couvre pas les aspects protocoles et profiles de processus, i.e. la provenance en lien avec la « conception de l'étude » (voir Tableau 18). La raison réside dans la généricité de l'ontologie PROV-O qui n'est pas une ontologie noyau, mais une ontologie générale (cf.§III.1.3.3.2). Dans l'ontologie noyau BMS-LM ces deux groupes de concepts sont bien considérés comme éléments de provenance d'une étude de recherche.

V.2.1.2. Les relations entre concepts de l'ontologie BMS-LM

Au niveau modèle de données du système BMS-LM, les relations sont limitées aux seules relations de l'identification et de la traçabilité (voir §I.4.3.1). En pratique, cependant, les relations entre les concepts de BMS-LM sont plus riches et véhiculent un sens plus spécifique : composition, agrégation, etc. Ainsi, pour faciliter leur compréhension, leur sémantique a été rendue explicite avec l'utilisation de l'ontologie des relations (RO), l'ontologie BFO, et l'ontologie OBI, trois ontologies membres du consortium OBO

ainsi que l'ontologie PROV-O pour les relations de provenance. La liste des relations de l'ontologie BMS-LM est explicitée dans le Tableau 19 ci-après.

Tableau 19 Liste des relations de l'ontologie BMS-LM ainsi que leurs ontologies sources, leurs relations inverses, leurs catégories et des exemples de leur utilisation
(→*à remplacer par le nom de la relation* <u>→</u>*à remplacer par l'inverse de la relation*)

	relation	*ontologie source*	*son inverse*	*exemple d'utilisation*	*catégorie*
1	has member	RO PROV-O	RO :member of	**study subject group** → min 1 **study subject**	agrégation
2	has part	BFO	part of	**examination** → min 1 **data acquisition** **data processing unit** <u>→</u> exactly 1 **data processing**	composition
3	has participant	RO	participates in	**biomedical research study** → *some* **study subject** **agent product** <u>→</u> *some* **agent administration**	contribution
4	has role	RO	role of	**studied organism** → *some* **study subject**	qualification
5	has specified input	OBI	is specified input of	**agent administration** → min 1 **agent product** **raw data unit** <u>→</u> *some* **data processing unit**	ingestion
6	has specified output	OBI	is specified output of	**data processing unit** → *some* **derived data unit** **biological sample** <u>→</u> exactly 1 **biological sample preparation**	production
7	realized in	BFO	realizes	**study subject** → *some* **biomedical research study**	réalisation
8	derives from	RO	xxxx	**biological sample** → *some* **studied organism**	transformation
9	acts on behalf of	PROV-O	xxxx	**operator** → min1 **operator**	provenance
11	is associated with	PROV-O	xxxx	**examination** → min 1 **study subject**	provenance
12	is attributed to	PROV-O	xxxx	**examination protocol** → some **operator**	provenance
13	is derived from	PROV-O	xxxx	**biological sample** → some **studied organism**	provenance
14	is generated by	PROV-O	xxxx	**biological sample** → exactly 1 **biological sample** **raw data unit** → some **data acquisition**	provenance
15	is informed by	PROV-O	xxxx	**data processing unit** → some **data processing**	provenance
16	uses	PROV-O	xxxx	**examination** → min1 **device**	provenance
17	is influenced by	PROV-O	xxxx	concept abstrait non instanciable	provenance
18	has for protocol	BMS-LM	xxxx	**examination** → exactly1 **examination protocol**	provenance
19	has for process profile	BMS-LM	xxxx	***data processing*** → some **workflow input profile**	provenance
20	has for parameters	BMS-LM	xxxx	**data processing unit** → some **processing parameters profile**	provenance

Dans le Tableau 19, les relations ont été classées par catégorie pour expliciter leur rôle dans l'ontologie noyau BMS-LM. L' « agrégation » et la « composition » correspondent aux relations UML classiques modélisées respectivement avec◇et◆. La « contribution » intervient quand un BFO:`continuant` participe à un BFO:`occurrent`. La « production », et sa catégorie inverse « ingestion », désignent respectivement la sortie, et l'entrée, d'un `BFO:process`. La « transformation » est lorsqu'un

BFO:continuant est un résultat modifié d'un autre BFO:continuant d'une manière à changer sa nature. La « qualification » intervient quand un concept décrit une propriété inhérente d'un BFO:continuant. La « réalisation » est utilisée lorsqu'une BFO:realizeable entity est rendue réelle dans un BFO:process. La « provenance » intervient lorsqu'un concept décrit les sources et les conditions de génération d'une donnée ou d'une information dans un autre concept.

V.2.1.3. La compatibilité avec le MDD BMS-LM

Afin de permettre une compatibilité avec le système BMS-LM, une correspondance a été établie entre le modèle de données BMS-LM et l'ontologie BMS-LM (voir Tableau 20). Il est à noter que pour trois anciens concepts du modèle de données BMS-LM, la correspondance n'est pas triviale et constitue une évolution importante du modèle : « Sample Result (SAR) » a été décomposé en deux concepts : un pour le processus (préparation de l'échantillon) et un pour l'objet (échantillon), « Agent Result » a été décomposé selon le même principe. « Processing Unit Result (SAR)» a également été décomposé en deux concepts : un premier pour les données produites du PUR et un second pour l'exécution du script du PUR (voir éléments en **gras** Tableau 20).

Tableau 20 Tableau de correspondances entre le MDD générique BMS-LM et l'ontologie noyau BMS-LM

BMI-LM previous name	*BMS-LM new name*
Study	biomedical research study
Subject	studied organism
Sample Result	**biological sample preparation**
Sample Result	**biological sample**
Sample Definition	sample preparation protocol
Agent Result	**agent administration**
Agent Result	**agent product**
Agent Definition	agent administration protocol
Acquisition Device	device
Exam Result	examination
Exam Definition	examination protocol
Acquisition Result	data acquisition
Acquisition Definition	data acquisition protocol
Data Unit Result	raw data unit
Data Unit Definition	expected data unit
Processing Results	data processing
Processing Definition	data processing protocol
Processing Unit Result	**derived data unit**
Processing Unit Result	**data processing unit**
Processing Unit Definition	processing unit protocol
Workflow Input	workflow input profile
Processing Parameters	processing parameters profile
Bibliographic Reference	published content
Intervention Result	biomedical intervention
Intervention Definition	biomedical intervention protocol
Software Tool	software tool
Study Subject	study subject
Study Subject Group	study subject group
Reference Data	reference data

D'un point de vue technique, les annotations obo:alternative term et owl:backwardCompatibleWith ont été utilisées pour référencer les noms des concepts du modèle de données BMS-LM comme indiqué sur la Figure 94: le nom court ou "trigramme," le nom officiel et le nom technique (voir Figure 94).

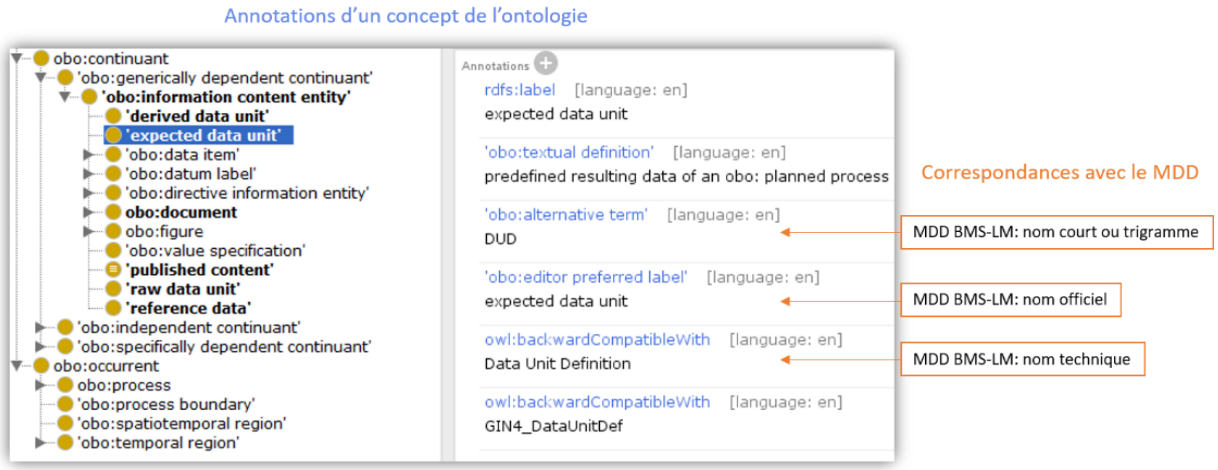


Figure 94 Exemple d'implémentation des correspondances entre l'ontologie noyau et le MDD BMS-LM

V.2.2. LES RÈGLES DE CONSTRUCTION ET ALIGNEMENT AUTOUR DE BMS-LM

L'ontologie BMS-LM a été développée à l'aide du logiciel Protégé 5.1.0 (http://protege.stanford.edu/) et du OWL-DL (Ontology Web Language - Description Logic of the W3 Consortium). Elle a été conçue avec une réutilisation maximale des KOS publiés, et en collaboration avec des experts en gestion de données et des experts en recherche biomédicale de l'entreprise Fealinx et du laboratoire LRI. Nous avons choisi l'ontologie BFO comme ontologie de haut niveau. Son engagement ontologique est à respecter pour garantir la cohérence des niveaux noyau, domaine, et local avec le haut niveau, même s'il est transparent par rapport à l'utilisateur final.

V.2.3. Les 11 principes OBO Foundry et les annotations indispensables dans BMS-LM

L'ontologie BMS-LM a été construite et mise en correspondance avec l'ontologie de haut niveau BFO et son ontologie de relations RO (Relation Ontology). Comme BFO fait partie du consortium OBO pour développer des ontologies interopérables, l'ontologie BMS-LM a été développée en conformité avec les 11 principes OBO Foundry. Ceci est expliqué dans le Tableau 21 suivant. La liste des principes est donnée et leurs applications actuelles et futures dans l'ontologie BMS-LM sont décrites.

Tableau 21 Règles de construction de l'ontologie noyau BMS-LM : la conformité aux principes de l'OBO Foundry.

	PRINCIPE OBO	APPLICATION ACTUELLE DANS BMS-LM	DANS LA FEUILLE DE ROUTE
1.1	Disponibilité	Disponible à l'adresse suivante http://www.fealinx-biomedical.com/ontologies/bmslm/bmslm.owl	Publier dans OBO
1.2	Licences	sous licence CC-BY 3.0	
1.3	Réutilisation d'autres ontologies : crédit et annotations	BMS-LM utilise les annotations MIREOT (Minimum Information to Reference an External Ontology Term) (Courtot et al., 2009)	Utilisation de XOD (eXtensible Ontology Development) (He et al., 2018)
2	Format commun	Le release officiel suit le format RDF/XML	
3	Identifiant dans <IDSPACE>_<NOMBRE> format	Tous les éléments de l'ontologie BMS-LM sont identifiés selon un format BMSLM_0000000	

4	Utilisation de versions	Les versions de BMS-LM suivent le format M.m.(Major.minor). Le suivi du processus de développement se fait en utilisant subversion (svn)	Maintenir un dépôt public de Git
5	Un champ d'application clair	Le champ d'application de l'ontologie BMS-LM est ajouté en tant qu'annotation OWL de l'ontologie. Elle indique qu'il s'agit d'une ontologie noyau, qui couvre le cycle de vie des études biomédicales et la gestion de la provenance avec une réutilisation maximale des ontologies existantes	Un wiki public
6	Définition textuelle	Tous les concepts de base sont définis à l'aide de l'annotation `<IAO : textual definition>`.	
7	Relations : réutilisation de l'ontologie de relations (RO)	Les relations de l'ontologie RO sont toutes réutilisées pour l'ontologie BMS-LM.	
8	Documentation	La documentation de l'ontologie BMS-LM se fait par des commentaires intégrés in situ si nécessaire.	Un wiki public
9	Pluralité d'utilisateurs documentée	L'ontologie BMS-LM est utilisée dans le projet DRIVE-SPC et sera utilisée dans les projets PACIFIC et Psycare de l'entreprise Fealinx. Des informations sur la réutilisation peuvent être trouvées dans les annotations de l'ontologie.	
10	Engagement à Collaboration	La construction de l'ontologie BMS-LM est principalement basée sur une réutilisation maximale et une correspondance avec les ontologies existantes. Cela renforce la collaboration entre les KOS biomédicaux	
11	Locus de l' Autorité	Dans les annotations de l'ontologie BMS-LM, le nom et l'adresse électronique de l'auteur correspondant sont ajoutés. Elle est la personne de référence pour l'ontologie	

Pour nous assurer que tous les concepts de l'ontologie BMS-LM sont bien annotés selon les recommandations OBO Foundry, un modèle « *template* » d'annotations pour les classes de BMS-LM a été défini. Il permet de guider l'ajout de nouveaux concepts dans l'éditeur d'ontologies « Protégé » (voir Figure 95). Il est à noter que les deux annotations `imported from` et `in branch` ne sont pas utilisées pour un concept interne à l'ontologie BMS-LM. Elles sont cependant utilisées pour les concepts importés selon les principes MIREOT (Courtot et al., 2009). L'exemple dans la Figure 95 ci-après donne les annotations actuelles (v1.1) pour le concept `BMS-LM:biomedical intervention` ainsi que pour son équivalent ERO : `intervention`. Les informations à renseigner impérativement sont : la définition textuelle et sa source (si elle est externe) ainsi que les annotations `imported from` et `in branch` si le concept est externe.

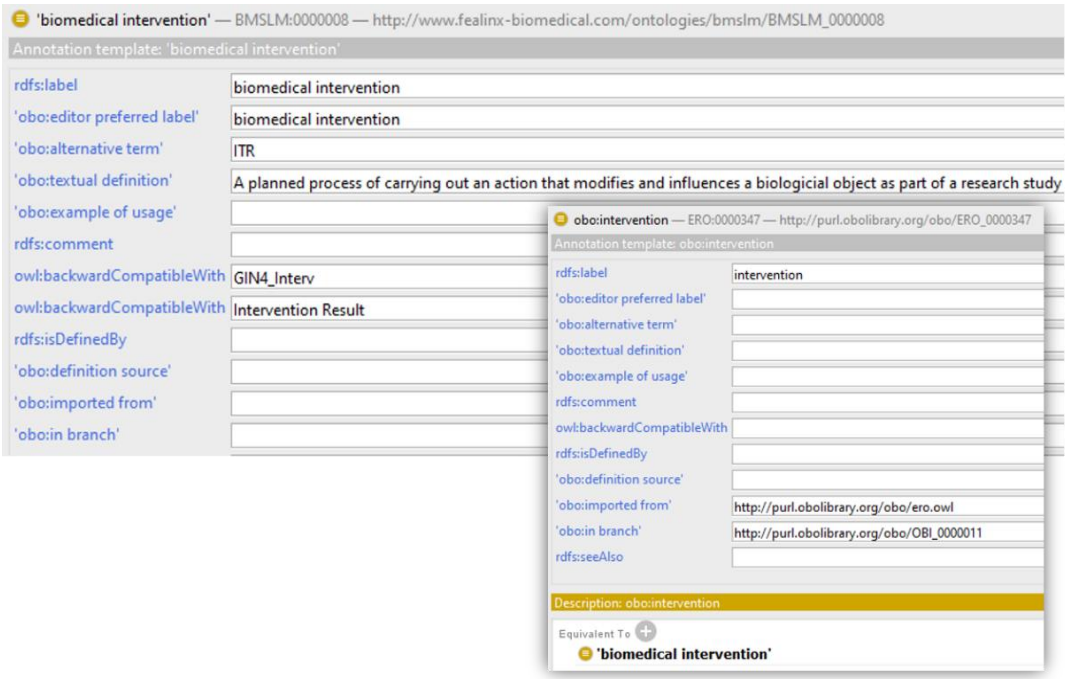


Figure 95 Exemple d'utilisation du « template » d'annotations BMS-LM

V.2.4. Encodage OBO des propriétés de l'ontologie multi-niveaux BMS-LM

Une méthode en deux phases a été adoptée pour l'enrichissement sémantique des concepts de l'ontologie BMS-LM en utilisant des propriétés OWL pertinentes. Cette méthode est utilisée au niveau noyau de BMS-LM, mais elle est aussi préconisée lors de l'utilisation de BMS-LM pour des domaines spécifiques (imagerie, biologie, etc.). Dans la suite de ce paragraphe, nous allons décrire comment pour un concept donné, nous avons identifié et implémenté les descripteurs dans l'ontologie BMS-LM. Le concept décrit est identifié par la lettre $\mathscr{C}$ et un terme le décrivant est appelé un descripteur δ.

La première étape consiste à constituer une collection de descripteurs candidats pour le concept $\mathscr{C}$. Pour ce faire, il faut répondre aux questions bien connues QQOQCCP (Qui ? Quoi ? Où ? Quand ? Comment ? Combien ? Pourquoi ?) sur chaque concept. QQOQCCP est un outil de référence pour la collecte d'informations sur la provenance (Sahoo et al., 2008). Par conséquent, les descripteurs de provenance sont aussi recueillis à l'aide de QQOQCCP. Le résultat est une liste de termes et de mots qui décrivent les principaux concepts d'intérêt.

La deuxième étape est l'encodage des descripteurs qui doit suivre l'encodage OWL utilisé dans les ontologies « OBO Foundry ». Pour ce faire, une analyse des codages OWL dans les ontologies OBI et IAO a été effectuée (voir Annexe C pour une explication des exemples identifiés dans ces ontologies). La liste des possibilités qui en résultent pour encoder les descripteurs de manière conforme aux règles d'OBO est présentée dans la Figure 96. Nous l'avons appelée **ADQIV** (**A**nnotation property, **D**ata property, **Q**uality, data **I**tem, and **V**alue specification).

Huit différentes options d'encodage OWL ont été identifiées : 1) l'utilisation de `owl:AnnotationProperty`, 2) l'utilisation de `owl:DataProperty`, 3) l'utilisation de `BFO:Quality`, 4) l'utilisation de `IAO:data item`, 5) l'utilisation de `owl:ObjectProperty`. Lors de l'utilisation d'un `IAO:data item`, trois autres options sont possibles : A) l'utilisation d'un data type, B) l'utilisation de l'équivalence à une liste d'individus et C) l'utilisation d'une liste de concepts de valeur (voir Figure 96).

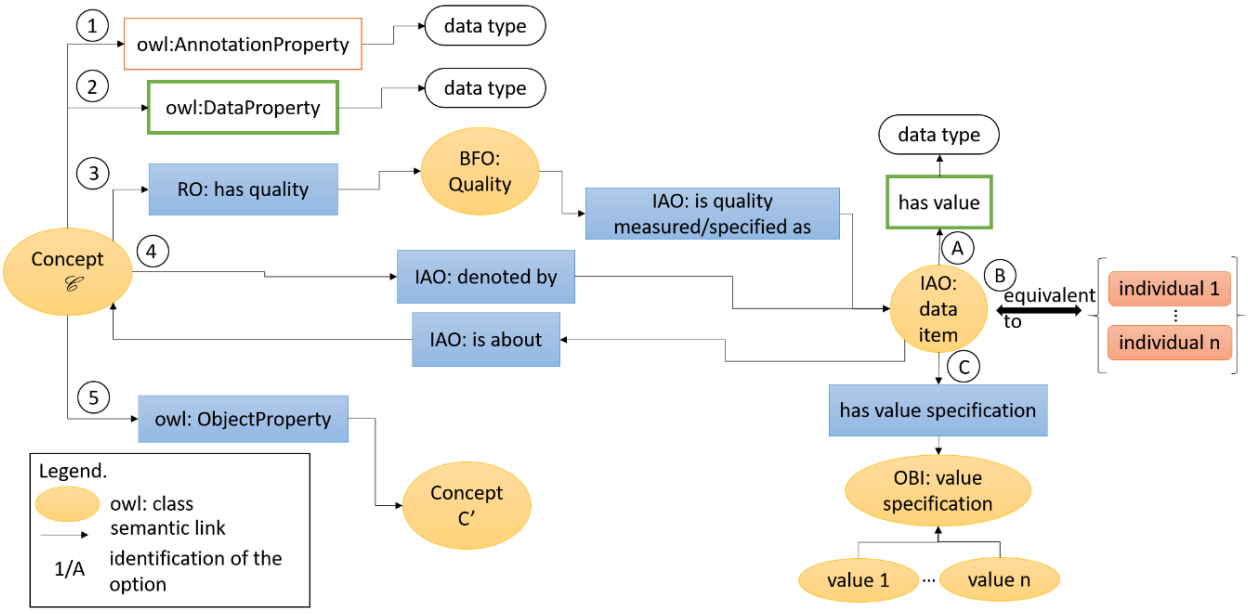


Figure 96 Les règles d'encodage de l'ontologie BMS-LM : la liste des possibilités d'encodage des descripteurs des concepts en conformité avec OBO.

Avec un choix aussi large, une stratégie pour un encodage cohérent en utilisant la liste **ADQIV** était nécessaire. L'arbre de décision de la Figure 97 explique comment choisir si un descripteur δ pour un concept $\mathscr{C}$ doit être encodé en tant que `OWL:AnnotationProperty`, `OWL:DataProperty`,

BFO:quality ou IAO:data item. Les principaux critères de décision sont les suivants : (1) la spécificité de la relation entre le descripteur δ et le concept 𝒞 (2) la nature du premier parent BFO du concept 𝒞 (un BFO:continuant ou un BFO:occurrent), (3) si une propriété OWL existe déjà, elle pourrait être réutilisée et (4) si le descripteur δ est « inhérent » ou non au concept 𝒞.

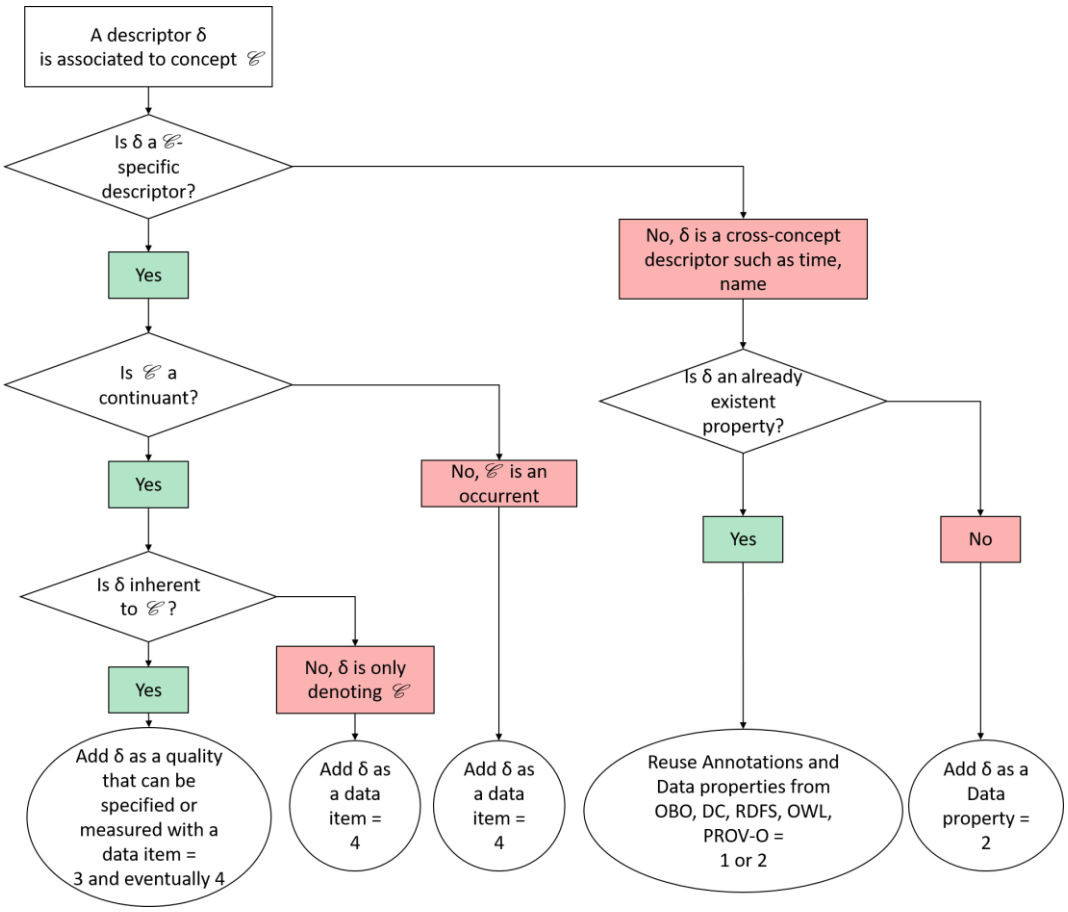


Figure 97 Les règles d'encodage de l'ontologie BMS-LM : Arbre de décision pour le choix de l'encodage OWL parmi les options dans ADQIV

Les feuilles de l'arbre Figure 97 montrent les options (et leurs identifiants numériques), résultats des décisions prises précédemment. Dans trois cas, un élément OWL:data item est utilisé. BFO:Quality est utilisé pour les descripteurs inhérents à un BFO:continuant. OWL:Data property ainsi que OWL:Annotation property sont utilisés pour les descripteurs non spécifiques.

V.2.5. Les règles de correspondances de l'ontologie noyau BMS-LM

Comme indiqué auparavant, l'ontologie BMS-LM est construite avec un maximum de réutilisation de concepts et d'ontologies existantes. Cette réutilisation est effectuée via des mises en correspondance entre les concepts de BMS-LM et les concepts réutilisés en se basant sur les principes MIREOT (Minimum Information to Reference an External Ontology Term) (Courtot et al., 2009) comme suit :

- L'URI de « l'ontologie source » et l'URI du « terme source » sont déclarés en utilisant la propriété IAO:imported from.
- La « classe mère » est référencée en utilisant la propriété (expérimentale) IAO:in branch.

Par exemple, le concept NCIT:acquisition est un concept externe importé du thesaurus NCIT selon les principes MIREOT. L'extrait OWL de l'ontologie BMS-LM de la Figure 98 suivante le décrit.

```
# 0. Prérequis
#NCIT:C43384 est le concept NCIT:acquisition
#obo:IAO_0000113 est l'annotation IAO:in branch
#obo:IAO_0000412 est l'annotation IAO:imported from

# 1. Assertions selon le syntaxe OWL functional-style et leurs explications en langage
naturel

AnnotationAssertion (obo:IAO_0000113 NCIT:C43384 "
http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C25404")
#c-à-d : NCIT:acquisition se trouve dans la branche suivante du thesaurus NCIT :
http://ncicb.nci.nih.gov/.../Thesaurus.owl#C25404

AnnotationAssertion (obo:IAO_0000412 NCIT:C43384 "
http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl")
# c-à-d : NCIT:acquisition est importé de cette adresse :
http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl
```

Figure 98 Exemple d'annotations en OWL pour l'import de concepts externes selon les principes MIREOT

Ensuite, le concept importé `NCIT:acquisition` a été mis en correspondance avec le concept `bmslm:data acquisition` selon l'axiome dans la Figure 99 ci-après.

```
EquivalentClasses  (
    bmslm:BMSLM_0000014
#bmslm :data acquisition
    ObjectIntersectionOf  (
        obo:OBI_0000011
#obo :planned process
        ObjectSomeValuesFrom  (skos2:broadMatch NCIT:C43384)
#NCIT :acquisition
    )
)
```

Figure 99 Exemple d'utilisation d'un concept externe dans BMS-LM

Ces concepts externes importés sont référencés par les concepts de l'ontologie BMS-LM afin d'augmenter l'interopérabilité de l'ontologie BMS-LM. Ils sont décrits en utilisant les `OWL:AnnotationProperty` en guise de « description informative » et ils sont aussi référencés par les concepts de l'ontologie BMS-LM via des « assertions exprimés plus formellement » (et donc compréhensibles par les raisonneurs et moteurs d'inférences) en utilisant des relations d'équivalence, des sous-classes et des propriétés telles que `skos:narrowMatch` et `skos:broadMatch`. Cette dernière signifie que le deuxième élément de la relation est plus large conceptuellement que son premier élément, elle est l'inverse de `skos:narrowMatch`.

Les entités externes importées peuvent être des concepts (classes) ou des propriétés (attributs). Ils peuvent appartenir au niveau noyau ou être des concepts spécifiques à des domaines précis. `NCIT:acquisition` par exemple est un concept du niveau noyau qui a été lié à l'ontologie BMS-LM via la relation d'équivalence `OWL:equivalentClass`. Pour les propriétés, il y a eu recours à la relation d'équivalence `OWL:equivalentProperty`. La Figure 100 ci-après, donne une synthèse graphique des différents liens possibles entre une entité de l'ontologie BMS-LM et une entité externe. La relation `rdfs:subClassOf` est utilisée pour formaliser un lien de parenté. Il est à noter à ce stade qu'un concept issu d'un domaine spécifique ne peut être lié à un concept de l'ontologie noyau BMS-LM que via une série de relations `rdfs:subClassOf`. De la même façon, les concepts de BMS-LM sont liés aux concepts de l'ontologie de haut niveau BFO via la relation `rdfs:subClassOf`. Les entités externes peuvent aussi être des définitions réutilisées depuis d'autres KOS. Elles sont référencées en utilisant les annotations `IAO:definition source` et `RDFS:isDefinedby` pour le reporting de la source de l'annotation `IAO:textual definition` d'une entité BMS-LM.

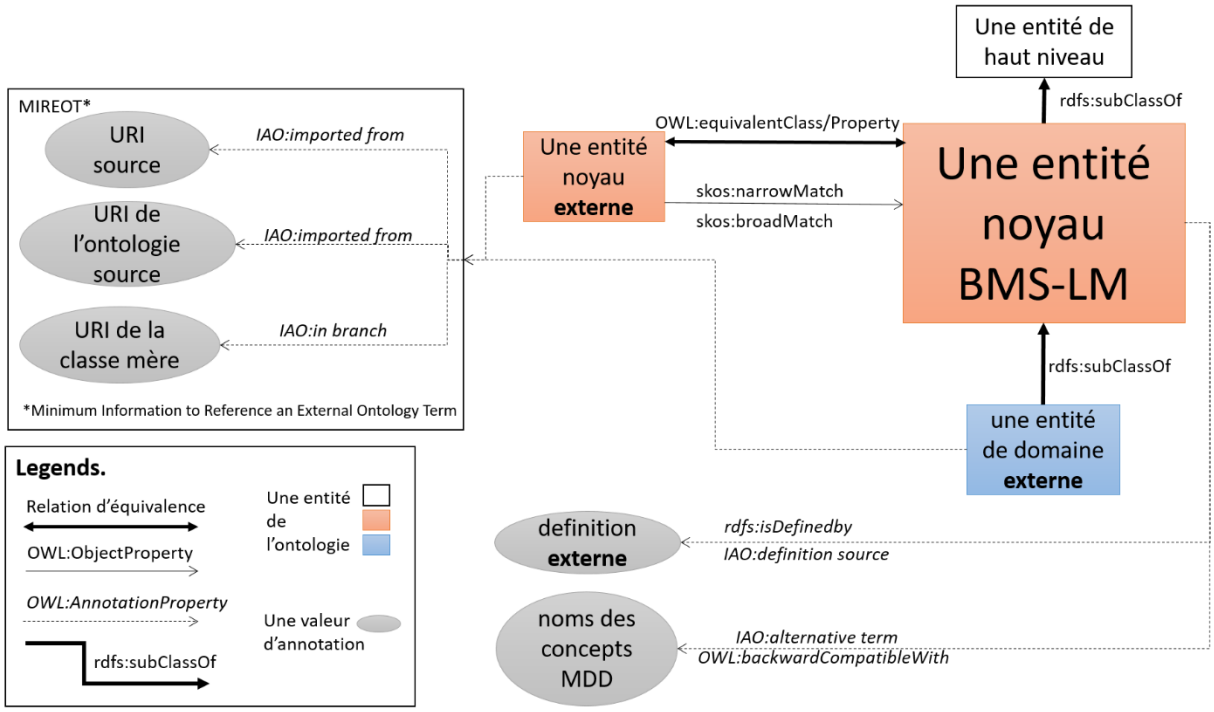


Figure 100 Les différentes méthodes de mise en correspondance utilisées lors de la construction de l'ontologie BMS-LM

V.2.6. CONSTRUCTION DES NIVEAUX DOMAINE ET LOCALE

À l'entrée de cette étape, les niveaux haut et noyau sont bien définis ainsi que les règles de construction en ingénierie ontologique de l'ontologie BMS-LM. Les niveaux domaine et local dépendent du domaine d'application comme expliqué auparavant. Dans ce paragraphe, nous explicitons comment les deux niveaux domaine et local sont construits et nous illustrons nos propos avec des exemples issus du laboratoire LRI.

V.2.6.1. Exemple de réutilisation de concepts depuis les KOS de domaine

Il faut d'abord identifier les ontologies, et plus largement les KOS, du domaine concerné. Par exemple, pour l'imagerie, nous avons identifié principalement le standard DICOM, l'ontologie QIBO et l'ontologie OntoVIP du projet OntoNeuroLog. Si tous les concepts d'une ontologie donnée sont à intégrer à la couche domaine, l'ontologie en question doit être importée en totalité depuis sa « release » officielle, avec identification de la version réutilisée. Ceci peut être effectué via l'outil d'édition d'ontologie « Protégé » à l'aide de la directive `owl:import`. Dans le cas où un sous-ensemble de l'ontologie est réutilisable seulement, il faut préparer un fichier personnalisé afin de permettre un import du sous-ensemble identifié uniquement.

Ensuite, il faut étudier et identifier les liens entre la couche noyau de l'ontologie BMS-LM et les concepts réutilisés des différents KOS de la couche domaine. Il s'agit d'une étape importante dans la construction de la couche domaine qui doit être réalisé par un/une « ontologiste ». Il/elle saura bien identifier les liens sémantiques afin de garder la cohérence de l'ensemble « haut/noyau/domaine/local ».

Par exemple, dans le cadre de la réalisation du module pour l'imagerie, le concept `QIBO:Xenograft` a été identifié comme un concept intéressant pour le laboratoire. Pour l'intégrer, il faut d'abord vérifier s'il est défini, dans son ontologie source, en cohérence avec les règles de construction BMS-LM. `QIBO:Xenograft` se présente comme dans la capture d'écran Figure 101, ci-après : `QIBO:Xenograft` est une sous-classe de `QIBO:Surgical Procedure` qui est une sous-classe de `QIBO:Biological Intervention`.

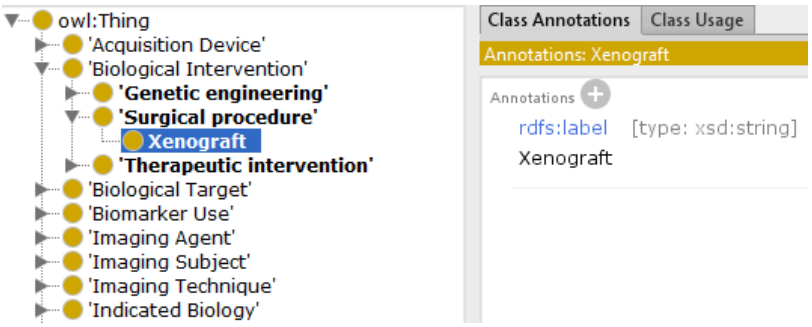


Figure 101 Le concept « Xenograft » dans l'ontologie QIBO

L'exploration de la branche d'au-dessus, `QIBO:Biological Intervention`, semble pertinente pour le domaine imagerie. Elle représente un sous-ensemble du concept noyau `BMS-LM:Biomedical Intervention`. Deux cas de figure peuvent être envisagés :

- Établir un lien `skos:narrowMatch` avec le `QIBO:Biological Intervention` et importer la branche entière.
- Importer uniquement le concept `QIBO:Xenograft` via des annotations MIREOT. Cependant, il n'a pas de définition textuelle dans QIBO, il faut alors soit le remplacer par un autre concept d'une autre ontologie, soit chercher à l'annoter via des définitions issues d'autres ontologies. Ces définitions peuvent être identifiées via BioPortal.
    - La définition de `QIBO:Xenograft` peut être importée de l'ontologie CRISP qui le définit dans http://purl.bioontology.org/ontology/CSP/2935-6859 comme une procédure:
        - « *procedures in which live cells, tissues, or organs derived from one species are transplanted, implanted, or infused into another species; includes those procedures from an animal into a human or from one animal into another; also includes procedures in which human body fluids, cells, tissues, or organs are removed from the body, come into contact with live animal cells, tissues, or organs, and are then placed back into a human patient; this definition is based upon the DHHS definition; where appropriate, index donor and recipient species.* »
    - Elle ne peut cependant pas être importée du Thesaurus NCIT qui le définit dans http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C12932 comme étant un échantillon biologique.
        - « *cells, tissues, or organs from a donor that are transplanted into a recipient of another species.* »

La recherche du terme « Xenograft » dans BioPortal a permis de mieux comprendre ce terme et ses différentes utilisations et définitions dans la communauté biomédicale. En effet, certaines définitions des ontologies et KOS trouvés le définissent en étant une entité matérielle (OBI, NCIT, MESH) et d'autres le définissent comme un processus (EFO, CRISP, QIBO). Ceci a permis aussi d'identifier des synonymes à « Xenograft » comme « Xenotransplantation » qui se présente comme un terme moins ambigu, toujours selon les données de BioPortal.

Si l'utilisation de QIBO dans ce contexte n'est pas une obligation, il est plus judicieux de remplacer le concept d'intérêt `QIBO:Xenograft` par « `CRISP:xenotransplantation` » de l'ontologie CRISP à cette adresse : http://purl.bioontology.org/ontology/CSP/2935-6859

Notre analyse de l'exemple de `QIBO:Xenograft` peut être généralisée comme une démarche de réflexion, à prendre en compte quand il faut intégrer un « nouveau concept » d'un KOS publié à l'ontologie BMS-LM:

1. Vérifier et compléter sa définition textuelle (chercher sur BioPortal)
2. Chercher des équivalents dans d'autres ontologies et choisir le mieux explicité sémantiquement et ontologiquement (le cas de `CRISP:xenotransplantation`)
3. Étudier la possibilité d'élargir vers d'autres concepts qui sont potentiellement intéressants pour le domaine d'application (le cas de la branche `QIBO:Biological Intervention`)

En appliquant ces principes pour chaque nouveau concept, nous garantissons une cohérence avec les principes de construction de l'ontologie BMS-LM entre le niveau noyau et le niveau domaine. De proche en proche, pour chaque concept ou terme du « lot initial des termes », il faut effectuer les mêmes réflexions afin de construire la couche domaine.

V.2.6.2. Exemples d'application pour le laboratoire LRI et intérêt pour la couche locale

En appliquant les principes et méthodes présentés jusque-là, et en application de notre approche à des données réelles, nous avons construit les niveaux local et domaine, relatifs aux données images des figures ci-après (Figure 102, Figure 103, Figure 104, voir la version en ligne pour les couleurs). Nous les avons également liés aux niveaux noyau et haut. Tout commence avec l'identification de l'image d'intérêt qui permettra de guider l'entretien avec le chercheur. Cette image est le support de la discussion. Elle et permet d'identifier un « lot initial de termes » qui lui correspond. Pour la Figure 102, l'image TEP-TDM montre une « souris » à laquelle du « FDG » a été injecté et où l'on peut tracer des zones d'intérêt des « VOI/ROI ». D'autres termes peuvent aussi être collectés à partir des données et métadonnées de l'image, le « KVP » par exemple dans le fichier « DICOM TDM ». La Figure 103 fournit un autre exemple où des termes comme « lectine » « FOV » « immunofluorescence » ont pu être collectés. Pour quelqu'un de l'extérieur, ce lot de termes (à côté de l'image dans les figures Figure 102, Figure 103) n'est pas forcément signifiant. Nous avons procédé à l'explicitation des termes initiaux en cherchant sur BioPortal des concepts équivalent dans des KOS de domaines et ayant une définition textuelle. Le `NCIT:Volume of Interest` a été trouvé comme équivalent au « VOI », le `QIBO :18-Fluorodeoxyglucose` pour remplacer « FDG », et le `NCIT : Field of View` pour remplacer « FOV », voir rectangles en bleu plein Figure 102 et Figure 103. Tous les concepts en bleu ont été ajoutés à la couche de domaine de l'ontologie pour le laboratoire LRI. La recherche dans des KOS publiés permet aussi de comprendre un terme donné, surtout quand le domaine d'application n'est pas maîtrisé par l'ontologiste. Elle permet aussi d'identifier le parent au niveau noyau du concept de domaine identifié. Les rectangles en orange (Figure 102 et Figure 103) représentent les concepts du niveau noyau. À titre d'exemple, les deux concepts de domaine `NCIT:Volume of Interest` (Figure 102) et `NCIT:Field of View` (Figure 103) que nous avons identifié préalablement ont comme parent le concept noyau `BMSLM:acquisition spatial region`. Les concepts noyau lient « les concepts de domaine » et « les concepts du niveau haut » via des relations de parenté (`rdfs:subClassOf`). Tous le long de ce processus de liens entre termes de proche en proche, le lot de termes initiaux est remplacé par des concepts plus standards, puisque publiés dans des ontologies. Au fur et à mesure, les termes initiaux seront remplacés par des concepts de domaine de la couche domaine. Cette « traduction » des termes locaux vernaculaires, collectés dans le « lot initial de termes », en concepts publiés et reconnus par la communauté a une valeur ajoutée pour la compréhension par un plus grand nombre de personnes ; qu'il s'agisse de personnes extérieures au centre de production lors d'un dépôt public, ou qu'il s'agisse d'une personne du centre de production n'ayant pas contribué elle-même à la production des données. Lors du partage ultérieur, par exemple, de l'image de la Figure 103, les termes de domaine bien définis sont partagés en accompagnement de l'image, renforçant ainsi la compréhension des données grâce à la mise en œuvre de l'interopérabilité sémantique décrite dans ce chapitre.

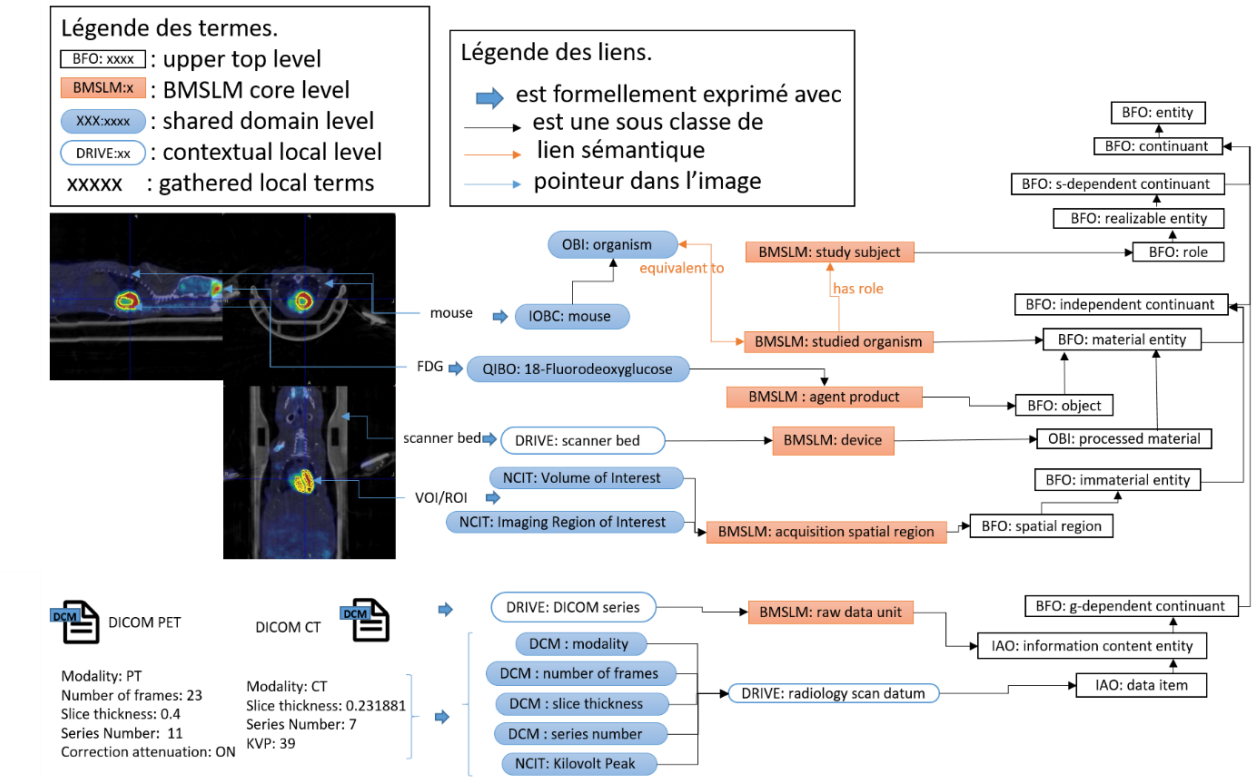


Figure 102 Exemple d'annotation d'une image TEP-TDM en utilisant les niveaux de l'ontologie BMS-LM

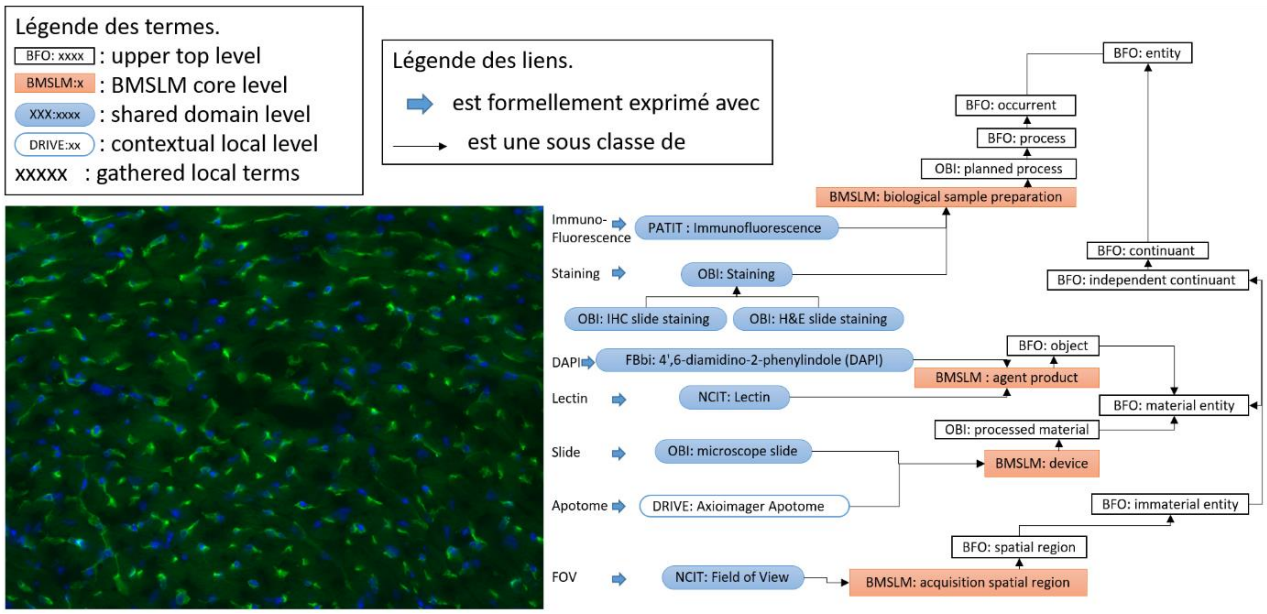


Figure 103 Exemple d'annotation d'une image d'histologie en utilisant les niveaux de l'ontologie BMS-LM

Cependant, des termes peuvent rester sans équivalents de domaine, il s'agit des termes locaux comme définis dans la section V.1.4. le concept `DRIVE:scanner bed` est un exemple de concept ajouté au niveau local de l'ontologie puisqu'aucune équivalence dans les KOS publiés n'a été trouvée. Le `scanner bed` est le compartiment d'installation des animaux dans l'appareil d'acquisition. Sa description est pertinente pour le contrôle qualité de l'expérimentation, le suivi de maintenance, etc. Elle a donc un impact sur le bon fonctionnement de l'acquisition des données, mais pas forcément sur la valeur scientifique du résultat. Il s'agit d'une prise en compte des spécificités du terrain qui s'avère très utile pour la standardisation des données scientifiques.

Nous avons effectué un test avec trois utilisateurs au laboratoire LRI afin d'évaluer la pertinence des exemples donnés Figure 102 et Figure 103 et 2 d'entre eux ont progressé en matière de compréhension des concepts du niveau noyau et de l'intérêt de la nouvelle méthode de modélisation (les documents utilisés pour le test sont présents en Annexe C).

La Figure 104 présente une collection de termes initiaux de plusieurs domaines identifiés au laboratoire LRI : **protéomique**, *histologie*, imagerie in vivo et leur intégration de proche en proche dans l'ontologie multi-niveaux BMS-LM. La figure se lie de bas en haut et suit la même charte graphique que les figures Figure 102 et Figure 103 : le niveau local en rectangle bleu vide, le niveau domaine en rectangle bleu plein, le niveau noyau en rectangle orange, et le niveau haut en rectangle noir vide.

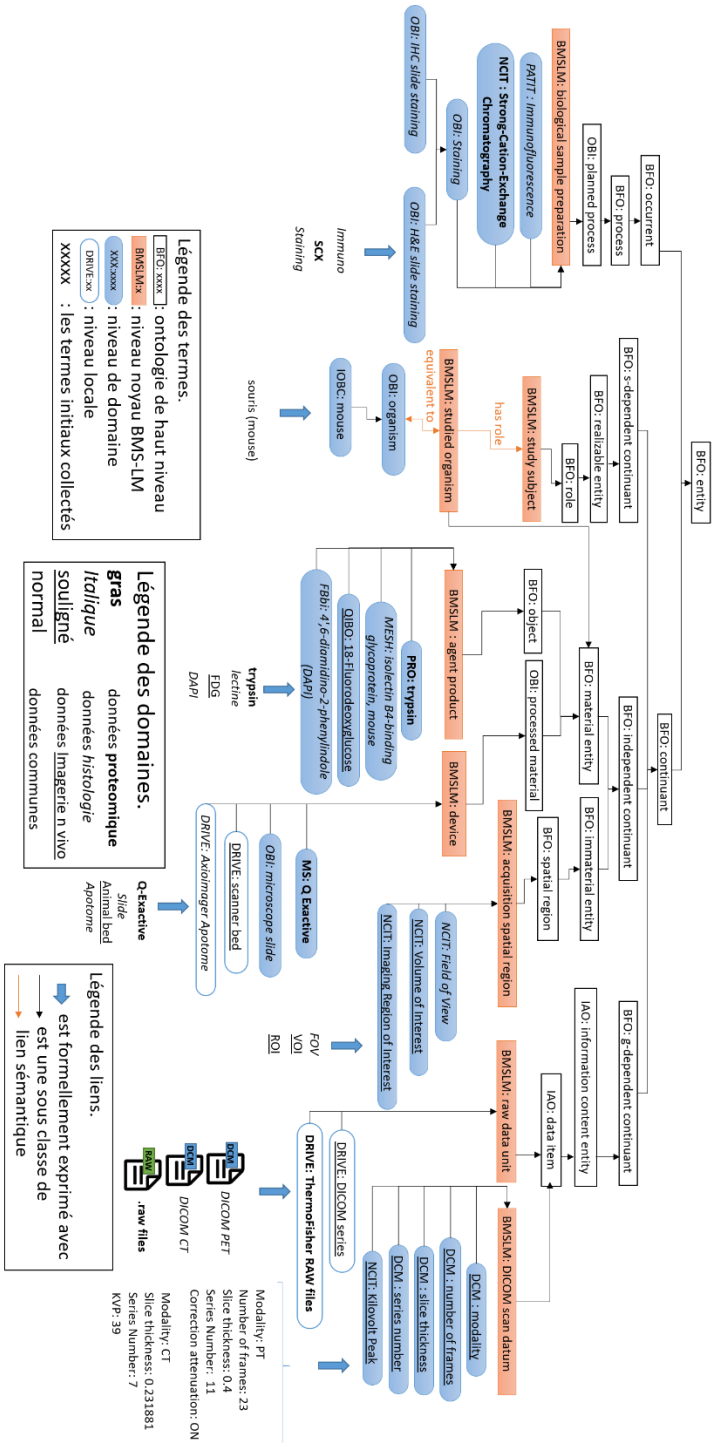


Figure 104 Exemple de l'ontologie multi-niveaux BMS-LM appliqué aux données du laboratoire LRI

V.2.6.3. Déploiement pour le laboratoire LRI et alignement avec la « Classification »

Une instance de Webprotégé[82], la version web et allégée de « Protégé » ; a été installé au LRI. Ceci permet de collaborer pour l'édition de l'ontologie spécifique au laboratoire LRI avec les *data manager* de l'entreprise Fealinx et les personnes du laboratoire LRI averties de l'importance de la standardisation. Des fonctionnalités d'éditions collaboratives comme la possibilité de commenter et suivre les modifications sont d'un grand intérêt pour le maintien et l'évolution de la version de l'ontologie « locale » et « de domaine » au laboratoire. L'instance de « Webprotégé » a été déployée sur le cloud universitaire CUMULUS et accessible à l'adresse suivante : http://pf-01.lab.parisdescartes.fr:1376/webprotege/ . L'utilisation de « Webprotégé » est proposée dans le cadre de collecte et standardisation des terminologies locales qui s'intègrent, plus tard, dans la couche domaine de l'ontologie BMS-LM, et dans la « Classification » du système BMS-LM. En effet, la « Classification » est la projection de la couche domaine et le MDD est la projection de la couche noyau de l'ontologie multi-niveaux BMS-LM.

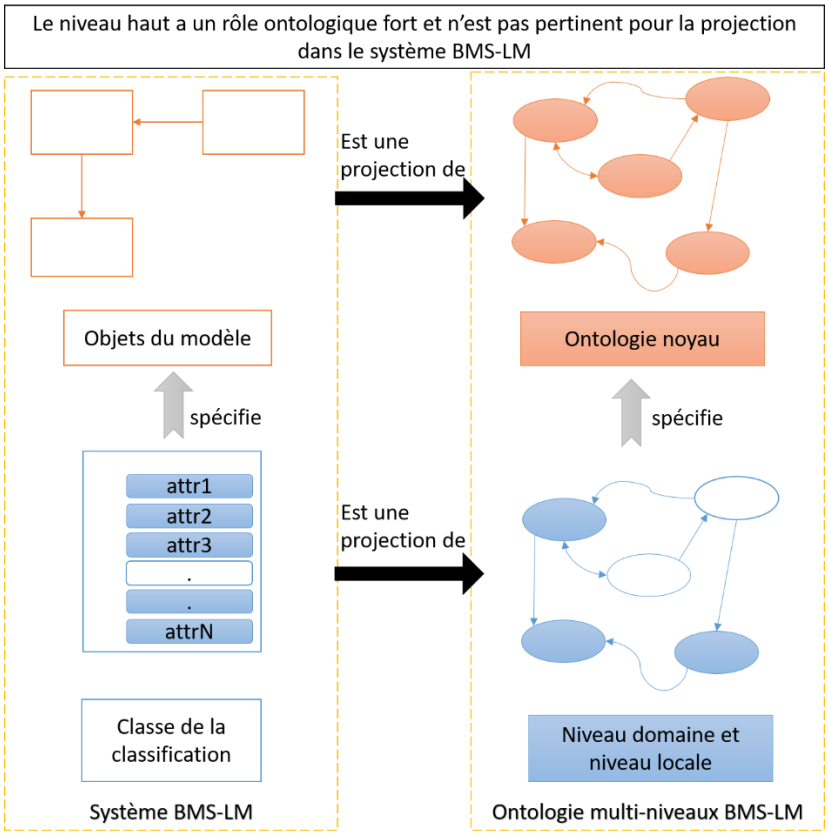


Figure 105 Relation entre le système BMS-LM et l'ontologie multi-niveaux BMS-LM

Une fois la couche domaine construite, la traduction au niveau « Classification » s'effectue en simplifiant les niveaux de l'ontologie. L'ontologie de haut niveau n'est pas représentée, les concepts intermédiaires non utilisés pour annoter les données ne sont gardés que dans le niveau ontologique et sont retirés du niveau « Classification ». Une conversion des `IAO : data item` en attributs de classification est effectuée ainsi que des autres propriétés de la liste OWL ADQIV décrites en Figure 96 §V.2.5. Un filtre ne gardant que les classes de domaine pertinentes pour un projet de recherche donné est utilisé. En effet, dans le système BMS-LM, il y a la possibilité de masquer des classes et des attributs s'ils ne sont pas pertinents pour le projet en cours. Nous avons réutilisé cette fonctionnalité pour simplifier au maximum l'arbre de « Classification » de manière à en augmenter sa pertinence. Les feuilles de l'arbre de classification présenté dans la Figure 74 §IV.1.3, sont alors directement liés avec

[82] https://github.com/protegeproject/webprotege

les classes de domaine ou locales qui sont pertinentes pour le projet de recherche en cours (voir Figure 106). Les attributs de ces classes sont aussi filtrés. Des mises en correspondance doivent ensuite être réalisées afin de pouvoir effectuer les traductions d'un KOS à un autre (Classification – ontologie multi-niveaux) et ainsi garantir l'interopérabilité sémantique des données manipulées via le système BMS-LM avec l'extérieur.

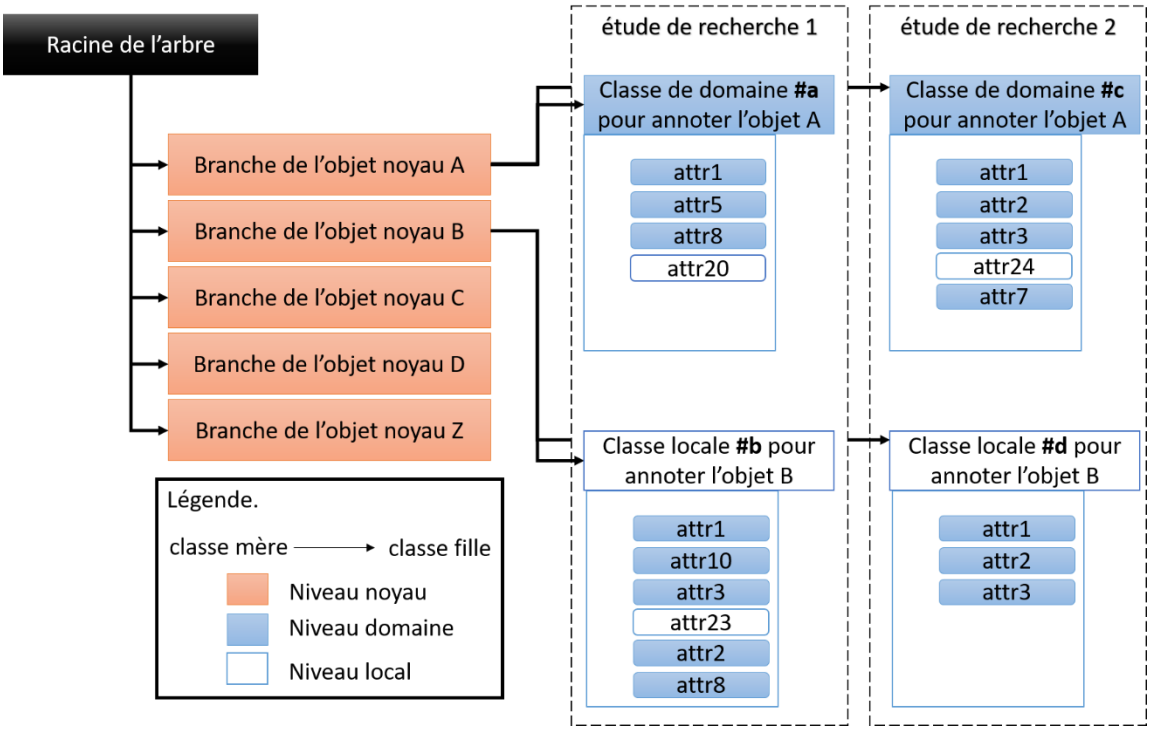


Figure 106 Projection de l'ontologie multi-niveaux dans la « Classification » du système BMS-LM

CONCLUSION DU CHAPITRE V

Dans ce chapitre, nous avons présenté notre méthode de construction d'ontologies pour la mise en place de l'interopérabilité sémantique. La méthode de construction d'ontologies « haut / noyau / domaine » de (Patel et al., 2005) a été étendue par le niveau « local ». Ce niveau, une fois ajouté, permet d'intégrer les termes locaux d'un groupe d'opérateurs et de chercheurs et de standardiser leur terminologie en préparation des partages et utilisations ultérieures des données de leurs projets.

Nous avons appliqué la méthode « haut / noyau / domaine / local » proposée, au KOS du système BMS-LM composé d'un modèle de donnée MDD et une « Classification ». L'ontologie qui en résulte est construite en utilisant l'ontologie BFO et les règles d'ingénierie ontologique qui y sont liées. La liste d'annotations et l'arbre de décision ADQIV ont été proposés pour expliciter la méthode de description et enrichissement sémantique de l'ontologie multi-niveaux BMS-LM en cohérence avec BFO. Le niveau noyau correspond à une évolution du MDD BMS-LM. Les concepts et les relations entre eux ont été renommés et mieux définis au niveau de l'ontologie noyau BMS-LM. La construction multi-niveaux a été appliquée aux données du laboratoire LRI avec réutilisation de KOS de domaine. Une « liste de termes initiaux », composée de termes d'histologie, d'imagerie et de protéomique, a été élaborée. À partir de cette liste, nous avons fourni de proche en proche les concepts de domaines, les concepts noyaux, et les concepts de haut niveau correspondants, en assurant un maximum de réutilisation des KOS publiés. L'examen du schéma Figure 104 permet de comprendre, même sans être un expert du domaine, le périmètre des termes initiaux. L'ontologie multi-niveaux qui en résulte est une ontologie interopérable. Elle permet au système BMS-LM de gérer et d'annoter les données en assurant l'interopérabilité sémantique avec les standards du domaine. En effet, les alignements entre la

« Classification » et le niveau domaine, et entre le MDD et le niveau noyau BMS-LM ont été expliqués. Le schéma de la figure ci-après résume les différents KOS utilisés pour assurer l'interopérabilité sémantique des données gérées par le système BMS-LM ainsi que leur agencement avec le SI présenté au chapitre IV précédent.

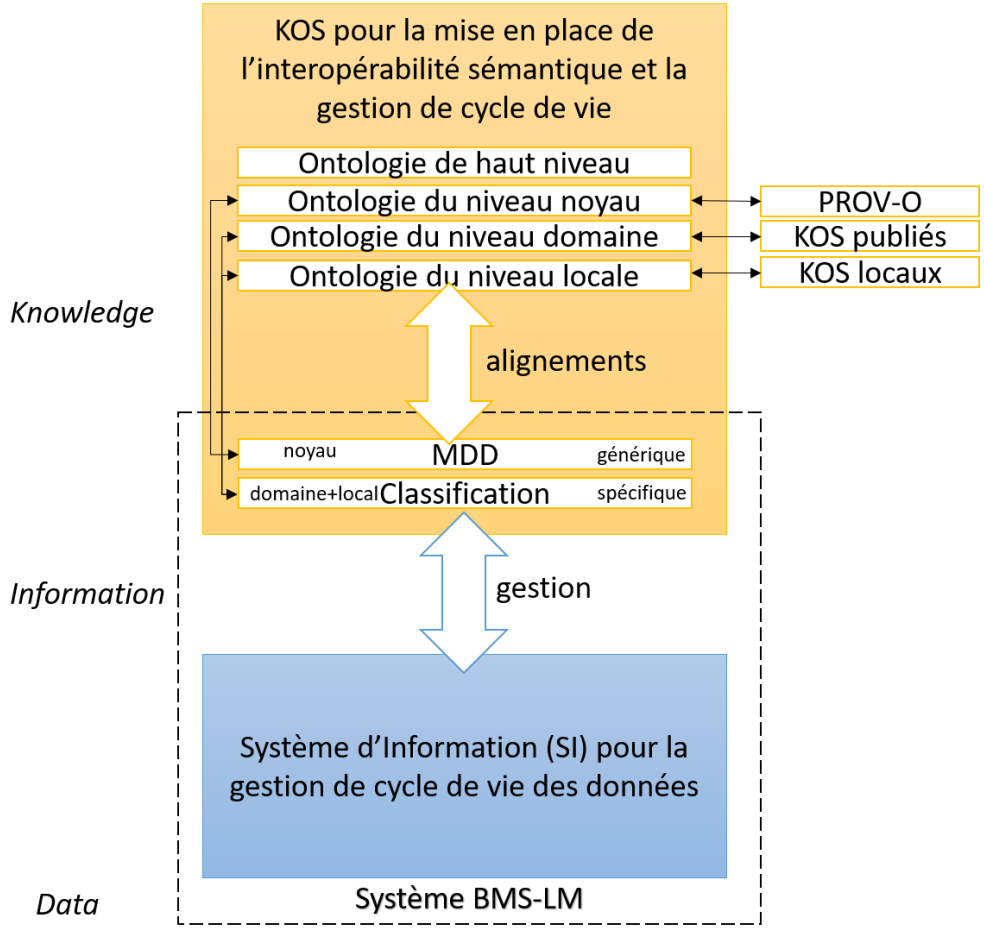


Figure 107 Résumé des KOS pour la gestion de cycle de vie et la mise en œuvre de l'interopérabilité sémantique

Chapitre VI. Application de la gestion de cycle de vie et de la provenance des études de recherche au laboratoire LRI de recherche préclinique

Les deux chapitres précédents ont été consacrés à la présentation détaillée de nos propositions, afin de répondre à la problématique de gestion de données hétérogènes de la recherche biomédicale, en assurant leur provenance et leur compréhension pour une réutilisation ultérieure. Nous avons défini le paradigme BMS-LM (BioMedical Study – Lifecycle Management) et expliqué la conception et la mise en œuvre d'un système BMS-LM pour la gestion des données et de leur provenance tout au long du cycle de vie. Nous avons également montré comment nous avons construit une ontologie multi-niveaux BMS-LM assurant l'interopérabilité sémantique avec les KOS de domaines.

Dans ce chapitre, nous appliquons nos propositions à la gestion des données du laboratoire d'accueil en recherche préclinique. Nous expliquons notre démarche pour la mise en œuvre d'un système BMS-LM pour le laboratoire et nous décrivons les différentes réalisations que nous avons effectuées dans ce cadre. L'application dans le laboratoire a commencé par un audit effectué en 2018 qui a permis de comprendre ses besoins et d'adapter nos propositions à la réalité terrain. Nous décrivons dans ce chapitre les résultats de l'audit et les différents plans d'expérimentation que nous avons conduits pour répondre aux besoins des chercheurs dans le laboratoire LRI.

VI.1. MÉTHODE D'AUDIT DANS LE LABORATOIRE LRI

Dans le but de permettre une gestion intégrée des données hétérogènes de recherche biomédicale, il est nécessaire d'identifier les besoins spécifiques du laboratoire d'accueil. Nous avons donc appliqué une méthode d'audit que nous avons essayé d'adapter aux laboratoires à effectif restreint, où les données, les projets et le personnel changent fréquemment. Nous présentons cette méthode et les résultats obtenus dans cette section.

VI.2. PRINCIPE ET ÉTAPES

Nous avons mis en place la méthode d'audit en nous basant sur l'expertise de l'entreprise Fealinx en « audit de systèmes d'information » et de l'équipe SIPP (Système d'Information Produit Process) du laboratoire Roberval en « conduite du changement ». Nous avons également adapté l'audit au contexte du laboratoire LRI. L'audit avait comme vision générale les citations ci-après (Spada, 2013) (Autissier & Moutot, 2016)(Autissier et al., 2018):

« Pour qu'une technologie nouvelle génère un changement, il faut que les individus lui trouvent un sens puisqu'ils l'utilisent pour servir leur but. »

« Sans utilisateur régulier et assidu, le nouveau logiciel demeurera un outil inexploité dont les coûts plomberont les comptes de l'entreprise »

« Il est important de leur donner un rôle, de les impliquer, de les faire agir et, finalement, de les faire adhérer à la nécessité du changement qui, à leur échelle, s'illustrera par une modification de leurs pratiques quotidiennes »

Il était alors primordial dans cet audit d'arriver à sensibiliser les futurs utilisateurs du système BMS-LM à l'importance de la gestion de cycle de vie des études de recherche biomédicale, et à les convaincre de l'utiliser pour leur recherche. Ceci permettra une traçabilité renforcée des données et constituera la mémoire du laboratoire. L'intérêt de cet audit pour notre thèse est de :

- Identifier les besoins spécifiques des chercheurs dans le laboratoire LRI.
- Tester le système BMS-LM dans un cadre adapté et avec des données variées, via la collecte de plans d'expérimentation pertinents pour sa validation.

Avant de commencer, trois groupes de personnes pertinents ont été identifiés (voir Figure 108 suivante) : L'équipe de scientifiques, des ingénieurs et des techniciens en recherche biomédicale d'un côté, et l'équipe des consultants en système d'information (SI) de l'autre côté, et entre les deux, une équipe déterminante de personnes médiatrices d'interface qui, de préférence, appartiennent aux deux équipes et comprennent le jargon de celles-ci.

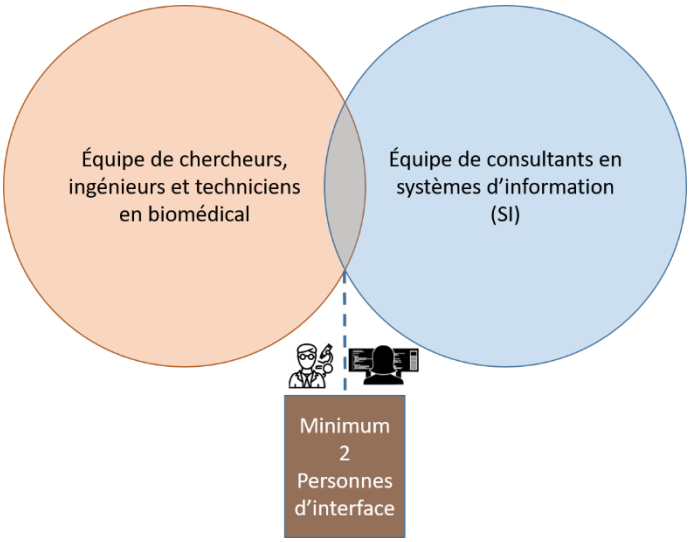


Figure 108 Les groupes de personnes identifiées au départ de l'audit

Dans le cas précis de notre audit, les personnes médiatrices étaient: une personne « Dan » du laboratoire, qui dispose d'une maîtrise de nombreux aspects fonctionnels et techniques de l'équipe de recherche biomédicale et qui a une ouverture sur les systèmes d'information (SI) et l'informatique en général ; et une personne « Ame » , moi-même, de l'équipe des consultants en SI, qui était, à plus de 50% du temps, accueillie au laboratoire pour ainsi se former sur les métiers de l'équipe de recherche biomédicale.

Ensuite, nous avons identifié les différents acteurs du système en nous référant à (Spada, 2013):

- Les « utilisateurs clés (*key users*) » qui seront impactés par l'introduction du nouveau système d'information basé sur la provenance et le cycle de vie.
- Les « top-managers » ou les « porteurs du projet » qui ont pour rôle de promouvoir le projet dans leur organisme et de parler de son issue finale (2 personnes : une personne du laboratoire « BTA » et une personne de l'équipe PLM « PBO »).
- Les « coordinateurs du projet », et les cadres intermédiaires qui doivent faire l'interface entre l'équipe PLM et l'équipe de recherche biomédicale (2 personnes : une personne de chaque groupe afin de traduire les jargons d'un contexte à un autre)
- Les « participants au projet » et « conseillers du projet » qui participent à la prise de décision concernant l'adoption (ou non) d'un processus métier et sa validation, ainsi que la promotion du changement auprès de leurs équipes respectives.

Nous avons organisé l'audit en cycles d'une semaine comme suit (voir Figure 109 suivante) :

- Entretiens de découverte du terrain avec un utilisateur clé
- Identification et validation d'un « processus pilote » avec cet utilisateur clé
- Étude technique de la réalisation du processus pilote avec les consultants SI
- Implémentation d'une « Preuve de Concept (POC) » en collaboration avec les consultants SI
- Démonstration du POC avec l'utilisateur clé concerné

Figure 109 Étapes de démarrage de l'audit sur la première semaine pour chaque utilisateur clé

Tous les lundis matin (temps libre en général dans le laboratoire), un entretien avec un utilisateur clé (Key user) a été programmé avec une grille de questions sous format carte mentale (MindMap), voir Figure 110 ci-après. Cela m'a permis de découvrir le quotidien de la personne interviewée (Activités, rôles, outils) et son entourage professionnel (inputs, outputs, objectifs, etc.). Il s'agissait d'une première étape pour la sélection des profils et personnes clés dans le processus d'audit et de conduite du changement.

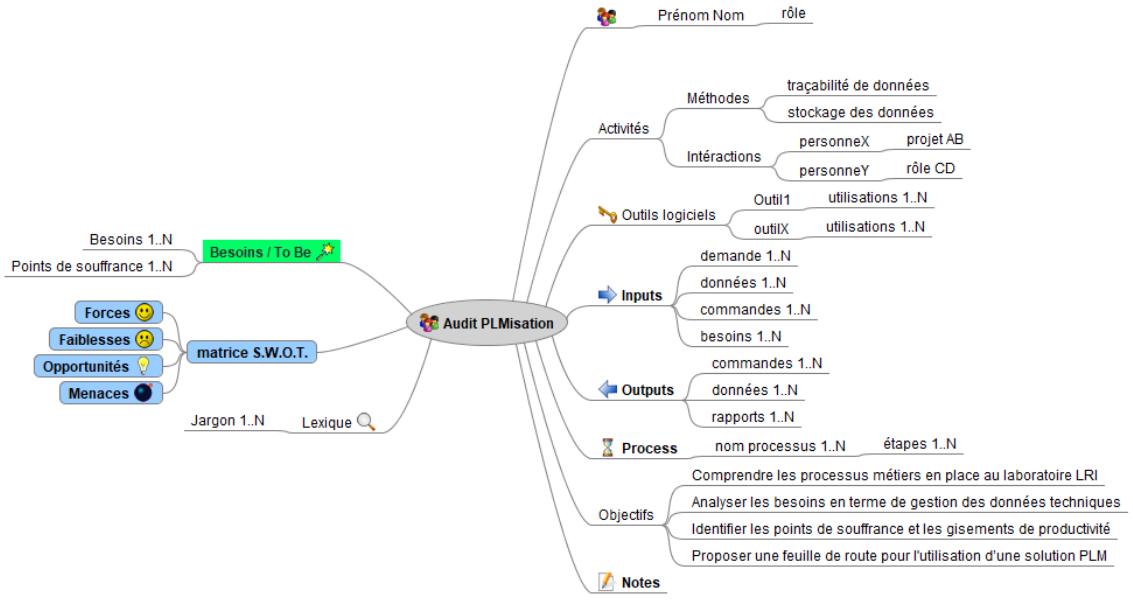


Figure 110 Points clé pour conduire les entretiens lors de l'audit au laboratoire LRI

La carte mentale n'est qu'un outil d'initiation des échanges et d'explicitation de la notion du PLM au cours de l'entretien. Au-delà du remplissage des cases, l'auditeur doit lire entre les lignes afin d'orienter la discussion vers les besoins exprimés par l'utilisateur clé.

À l'issue de ce premier entretien, il s'agissait pour moi (l'auditrice) d'identifier un processus en lien avec les besoins et réalisable via le système BMS-LM. Après l'entretien de découverte, d'autres entretiens ont suivi afin de mieux cerner ce processus : ses entrées, ses sorties, son utilité et les besoins auxquels il répond. Nous l'appelons : Processus pilote.

Avant de l'implémenter, ce processus pilote a été modélisé et validé avec chaque utilisateur clé. Nous avons envisagé deux moyens de modélisation : en utilisant BPMN « *Business Process Model and Notation* » (OMG, 2013) ou SADT « *Structured Analysis and Design Technique* » (Colquhoun et al., 1993). Finalement, nous avons choisi une version simplifiée de SADT, compréhensible plus aisément par les utilisateurs clés. Ci-après, un des diagrammes SADT que nous avons tracés et validés avec un utilisateur clé, le reste des processus pilotes en schémas SADT est en Annexe D.

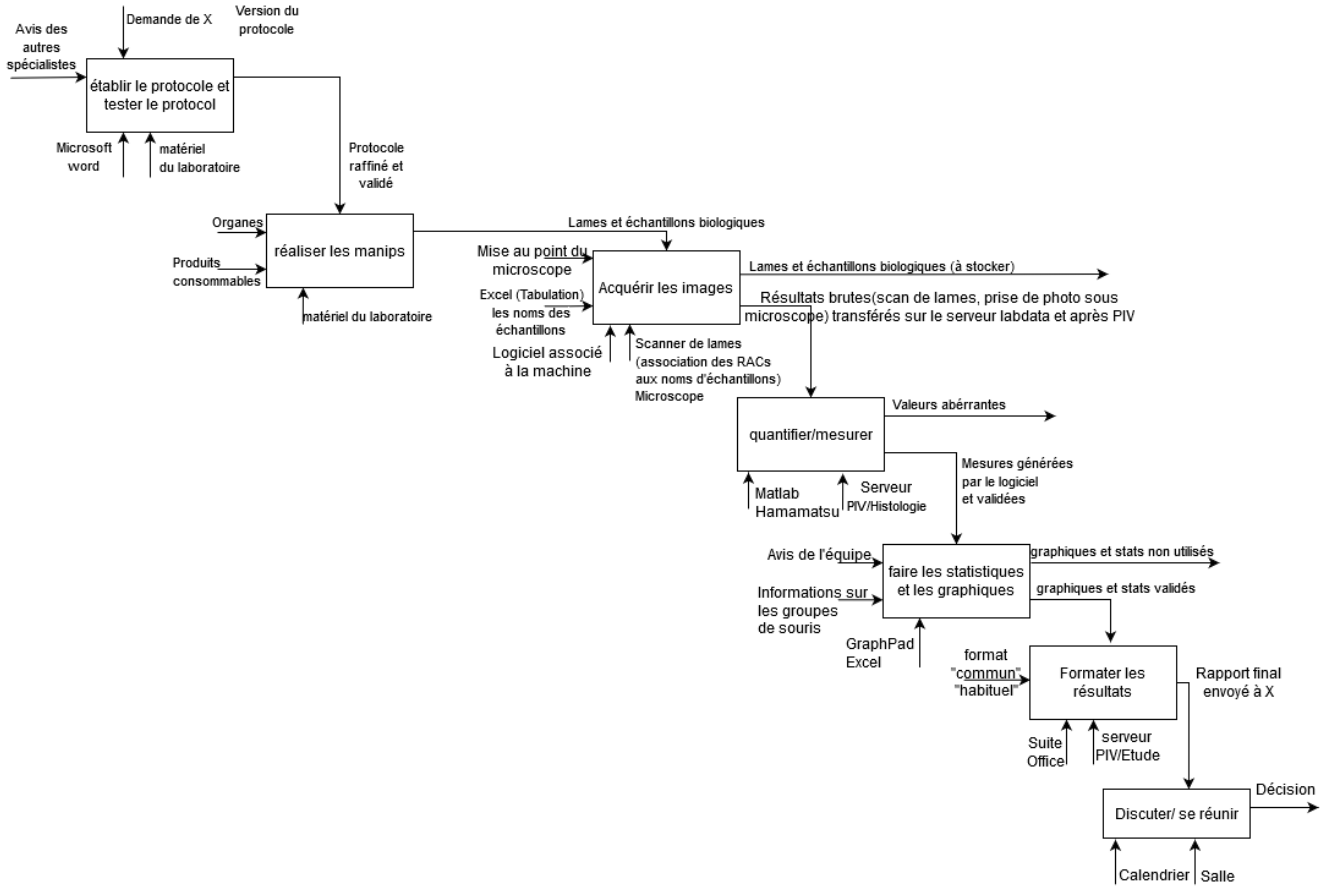


Figure 111 Diagramme SADT modélisé et validé avec l'ingénieur en histologie « ACE » au laboratoire LRI

La validation du diagramme était généralement effectuée dans la journée et au plus tard le lendemain. Pour chaque diagramme validé, une étude technique était réalisée afin d'identifier le processus qui pouvait être réalisable via le système BMS-LM. Pour le diagramme SADT de la figure 82 précédente, l'utilité du système BMS-LM commence après l'acquisition des images. Celles-ci seront (1) intégrées dans le système BMS-LM après leur acquisition, (2) rattachées à leur protocole source via le système BMS-LM pour la traçabilité, et (3) analysées via le système BMS-LM. Nous présenterons l'implémentation de ce processus pilote dans la section VI.4.

Tous les lundis après-midi, une réunion de pilotage du projet a eu lieu afin de présenter les nouveautés et formaliser des décisions. Étaient présents dans cette réunion, les top-managers du projet, les coordinateurs du projet, et les conseillers du projet.

Au cours de la semaine, une réunion technique avec les consultants SI concernés par l'implémentation du processus pilote était effectuée pour préparer une preuve de concept (POC) pour un utilisateur. Le projet étant en mode agile et itératif, le POC était présenté en fin de semaine à l'utilisateur clé pour pouvoir recommencer un nouveau cycle la semaine suivante avec un autre utilisateur clé. Ce nouveau cycle pouvait aussi concerner le même utilisateur, afin de tester le POC avec différents jeux de données et ainsi l'adapter à ses nouveaux besoins.

Le deuxième cycle avec le même utilisateur clé était assez spécifique à chaque difficulté rencontrée dans chaque processus pilote. Il était caractérisé par des réunions de tests avec l'utilisateur et des réunions techniques avec les consultants afin de faire converger les deux efforts vers une implémentation du processus fidèle aux besoins de l'utilisateur clé. Lors de cette phase, nous avons exploité les leviers identifiés précédemment (Tableau 8 §II.1.1), à savoir, le L1-collabroation bio-info, Le L2-changement, le L3-adaptation, le L4-mirroring, et le L5-intérêt.

VI.3. PLMBOOST : DÉROULEMENT DE L'AUDIT ET RÉSULTATS

Nous avons appelé l'audit PLMBoost, il s'est déroulé du 27 février 2018 au 28 juin 2018 et a donné lieu à un total de 28 réunions entre entretiens, pilotage, mise en œuvre, et formations (voir Tableau 22 ci-après).

Tableau 22 Résumé des réunions effectuées au laboratoire LRI

Réunions de pilotage	**Entretiens des utilisateurs clés**	**Réunions Techniques**	**Réunions utilisateur**
8 réunions de 1h **Une 9e réunion de synthèse d'une demi-journée**	5 entretiens Tous les lundis (ACE, DBA, TYO, TVI, CFA)	8 réunions avec les consultants SI Fealinx	4 formations « DBA » 2 formations « ACE » 1 réunion de spécification « ACE »

Nous avons identifié au départ 5 utilisateurs clé « ACE », « DBA », « TYO », « TVI », et « CFA ». Au cours de l'audit, nous avons défini « CFA » comme un utilisateur occasionnel du système et donc nous avons décidé de nous focaliser sur les autres. Le Tableau 23 ci-après liste les scénarios d'utilisation et les processus pilotes que nous avons identifiés à l'issue des entretiens avec chaque utilisateur clé.

Tableau 23 Liste des scénarios d'utilisation identifiés pour chaque utilisateur clé

Identifiant	***Besoins***	***Personne concernée***	***Objectif***
Intégration_1	B1-Archivage, B16-contrôle qualité, B11-annotation des données	ACE– utilisateur clé	Mettre en place une intégration des données d'Histologie vers le PLM par « ACE », toutes les semaines.
CUBE_1	B1-Archivage, B16-contrôle qualité, B11-annotation des données	TYO– utilisateur clé	Mettre en place un export configuré pour le projet COS-TEP de « TYO » depuis le système BMS-LM
Client_1	B1-Archivage, B16-contrôle qualité, B11-annotation des données	TYO— utilisateur clé	Vérifier la qualité des données
Intégration_2	B1-Archivage, B16-contrôle qualité, B11-annotation des données	TYO— utilisateur clé	Mettre en place une intégration par l'utilisateur des données TEP-TDM vers le système BMS-LM
Trait_1	B6-Traitements, B8-Traçabilité et B11-bonnes pratiques	DBA– utilisateur clé	Mettre en place une procédure d'intégration (totale/partielle) des traitements Matlab dans le système BMS-LM
Client_2	Exploration des données en vue d'analyse B3, B4, B17, B6	CFA– utilisateur occasionnel	Exploiter et exporter des données depuis le système BMS-LM

Client_3	Exploration des données en vue d'analyse B3, B4, B17, B6	TVI – utilisateur clé	Explorer la base de données TEP-TDM dans le PLM : recherche, modification, export, commentaires
CUBE_2	Exploration des données en vue d'analyse B3, B4, B17, B6	BTA – porteur du projet	Exporter et personnaliser la base de données du système BMS-LM via Excel

Au cours de l'audit, nous avons été confrontés à la réalité terrain de la faible disponibilité des chercheurs pour ce rythme de collaboration continue et aussi, de l'indisponibilité des consultants SI pour rendre les POCs à temps. Pour cette raison, les deux seuls utilisateurs avec lesquels nous avons réalisé les implémentations pendant la période de l'audit ont été : « ACE », l'ingénieur en histologie et « DBA », l'ingénieur en calcul scientifique dans le laboratoire. Concernant « TYO », nous avons initié le processus « CUBE1 », mais il s'est avéré plus demandeur en temps que le temps de l'audit. Pour « TVI » nous avons fait avancer les implémentations pour le plan d'expérimentation « intégration_2 ». Nous présentons dans les sections suivantes des exemples issus de ce que nous avons réalisé en collaboration avec eux et plus spécifiquement, les processus « intégration_1 », « intégration_2 », « trait_1 » que nous avons explorés.

VI.4. INTÉGRATION ET ANALYSE DES DONNÉES EN HISTOLOGIE

En lien avec le processus pilote « intégration_1 », nous avons lancé, avec l'utilisateur clé « ACE », un plan d'expérimentation qui se focalise sur l'intégration de données d'histologie. Nous avons ensuite utilisé les données d'histologie pour le processus pilote « trait_1 » en étroite collaboration avec l'utilisateur clé « DBA ». Nous avons ainsi mis en œuvre une intégration partielle d'un traitement d'analyse d'images d'histologie au laboratoire LRI. Dans cette section, nous présentons les deux réalisations pour l'intégration de données et de calcul scientifique en Histologie.

VI.4.1. INTÉGRATION DE DONNÉES D'HISTOLOGIE

À la suite de la série d'entretiens réalisée dans le cadre de PLMBoost, les contraintes ci-dessous ont été identifiées pour l'intégration de données :

- Les données sont acquises à l'aide de microscopes et scanners de lames qui sont externes à l'équipe (hôpital, plateforme d'histologie). Elles sont transférées via disques durs et ne sont donc pas connectables directement au système BMS-LM.
    - Les noms des fichiers produits à la suite de l'acquisition des données sont saisis à la console de la machine d'acquisition.
- Un serveur de stockage est utilisé pour la collecte centralisée des données issues de différentes sources. Ce serveur est géré principalement par l'utilisateur clé « ACE », mais son utilisation est partagée avec tous les chercheurs concernés.
    - Le stockage est semblable à une arborescence de dossiers/fichiers gérée par plusieurs personnes.
- Les formats de données brutes (qptiff, ndpi, im3, …etc.) ne sont utilisés que pour un export en format TIFF utilisable plus tard dans les analyses. Elles sont gardées sur le serveur pour leur archivage principalement.

VI.4.1.1. Préparation des données d'histologie

Nous avons identifié un manque de traçabilité et de standardisation. Pour cela, nous avons défini les paramètres/variables/annotations clés qui doivent être décrits lors d'une acquisition de données en histologie. La carte mentale dans la Figure 113 suivante a été réalisée en étroite collaboration avec « ACE ». Pour chaque concept noyau (EXA, SAR, etc.) impliqué, nous avons listé les annotations

(Opérateur, Date, etc.) qui doivent l'accompagner ainsi que leurs valeurs permises si possible (x20, x40 pour le Grossissement).

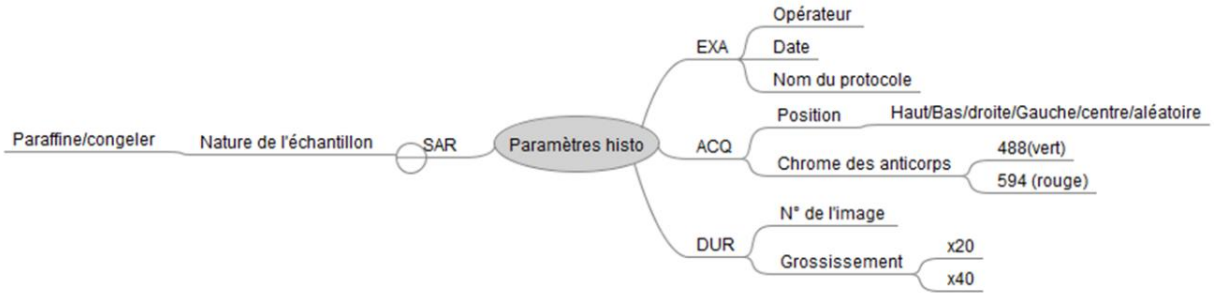


Figure 112 Présentation des différents paramètres pour le reporting des acquisitions d'histologie

Ensuite, nous avons étudié la meilleure façon de les annoter dans le contexte décrit précédemment. Le but étant d'appliquer les leviers L3-adaptation et L4-mirroring, qui consistent à éviter de changer trop radicalement le processus actuel des chercheurs, mais plutôt à s'adapter et se greffer à ses habitudes, pour garantir une durabilité du changement. Pour le processus pilote « intégration_1 », nous avons choisi de ne pas passer par un tableau Excel (comme proposé en IV.2.1), mais d'intégrer la liste des variables dans les noms de fichiers et de dossiers sur le serveur de stockage qui a été structuré par niveaux. L'arborescence de dossiers, décrite dans la capture d'écran ci-après (voir Figure 113), représente un résultat de ce travail de réorganisation pour l'intégration de données dans le système BMS-LM.

```
Z:\SDHB\Serie2>tree
Structure du dossier pour le volume HISTO
Le numéro de série du volume est DC37-1D44
Z:.
├───Rein
│   ├───Microscope-Apotome
│   ├───AnalyseFibrose-Matlab
│   └───LameScan-Hamamatsu
│       └───Colorations
│           ├───RougeSirius
│           └───HemalunEosine
├───Tumeur
│   ├───Microscope-Apotome
│   ├───AnalyseQuantificationCD31-IF-Matlab
│   ├───LameScan-Hamamatsu
│   │   ├───Colorations
│   │   │   ├───RougeSirius
│   │   │   └───TrichromeDeMasson
│   │   └───Immunohistochimie
│   └───LameScan-Polaris
│       └───Immunofluorescence
│           └───CD31-594
└───Coeur
    ├───LameScan-Hamamatsu
    │   └───Colorations
    │       ├───RougeSirius
    │       └───HemalunEosine
    ├───Microscope-Apotome
    └───AnalyseFibrose-Matlab
```

Figure 113 Structure de dossiers sur le serveur de stockage d'histologie

Le premier niveau de l'arborescence correspond au niveau « projet » (SDHB dans la Figure 113). Ensuite, est décrit le niveau « organe » étudié (Rein, Tumeur, Cœur). Les données sont ensuite regroupées en fonction de leur « protocole » (LameScan, AnalyseFibrose, Visualisation au Microscope, AnalyseQuantificationCD31-IF), « logiciel » (Matlab) et/ou « machine » (Apotome, Hamamatsu) qui sont utilisés. La différenciation entre données dérivées et données brutes s'effectue avec l'étiquette « Analyse ». Lorsque l'étiquette de protocole « LameScan » est utilisée, il faut spécifier dans un niveau supplémentaire le protocole (« Colorations » ou « Immunofluorescence »). Il faut ensuite spécifier, dans le niveau au-dessous, le type de la coloration (RougeSirius, TrichromeDeMasson, HemalunEosine) et le type de l'‘immunofluorescence (CD31-594). Des niveaux plus spécifiques peuvent être ajoutés selon

le besoin. Ces modifications doivent passer par des validations mutuelles entre les chercheurs et les data-managers.

La nomenclature se poursuit avec des informations supplémentaires délivrées par les noms de fichiers. Notamment, l'identifiant de souris, l'identifiant projet, l'identifiant organe, le nom de protocole, le type d'image, le grossissement, et l'identifiant de lame d'origine comme dans l'exemple suivant :

IDMouse_Projet(SDHB)_organe(coeur)_nameProtocol(coloration)_ImageType(RS)_Magnification(X20)_IDLame(L1)

Dans la capture d'écran, Figure 114 , un exemple de données brutes de « LameScan » est présenté. Ce scan est effectué via le scanner Polaris. Il correspond à l'Immunofluorescence GLUT1 et CD31. Il représente un scan de tumeur du projet « SDHB » avec un grossissement X20. Le prélèvement tissulaire appartient à la souris S000938, et la lame n°1. Toutes ces informations ont être collectées à partir des chemins et des noms de fichiers.

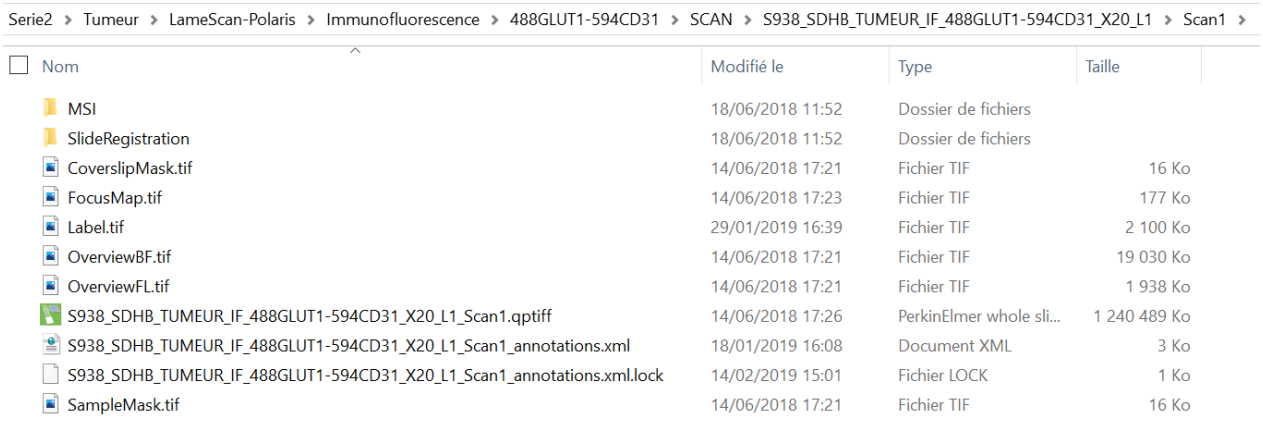


Figure 114 Exemple du serveur d'histologie, capture présentant des données brutes

Les images correspondantes aux données brutes exportées en format TIFF depuis le scanner Polaris ont été importées dans le système BMS-LM via la méthode d'intégration « générique » (§IV.2.1) adapté au contexte des données d'histologie. Pour adapter la méthode d'intégration de données, nous avons ajouté un script de prétraitement « tree-2-csv » à la procédure initiale d'intégration « générique » de données. Ce script permet de traduire l'arborescence de dossiers/fichiers en un tableau. Il fournit ainsi l'entrée à la procédure d'intégration « génétique » standard. Ce tableau peut être enrichi si besoin avec d'autres informations. À ce stade, il a été utilisé tel quel comme point d'entrée pour l'ETL d'intégration « csv-2-xml ».

VI.4.1.2. Préparation du système BMS-LM

Après la première phase de préparation des données d'histologie, nous avons procédé à la phase de préparation du système BMS-LM. Nous avons ainsi instancié les objets de « Provenance » du MDD et ajouté les classes spécifiques du domaine d'histologie à la « Classification » comme expliqué dans la section IV.2.1.1.2. Le résultat est explicité dans les captures d'écran Figure 115 et Figure 116 ci-après.

Elles sont issues d'interfaces du système BMS-LM (décrites en détail en Annexe A) : l'explorateur de l'arborescence d'objets MDD Figure 115 et l'explorateur des classes de la « Classification » Figure 116. Les captures d'écran de ces deux explorateurs seront réutilisées à chaque fois que nous souhaitons expliciter les ajouts effectués dans le système BMS-LM, au niveau MDD et au niveau « Classification ».

Le panneau de gauche de la Figure 115 montre sous forme d' arborescence deux protocoles d'histologies modélisés avec les objets de « Provenance » EXD, ACD, DUD du MDD. Dans l'espace de l'ontologie BMS-LM, ces objets sont une projection des concepts `BMS-LM : examination protocol`,

`BMS-LM : data acquisition protocol` et `BMS-LM : expected data unit`. La partie de droite montre les informations de l'objet sélectionné à gauche. Dans la capture d'écran Figure 115, l'objet « EXD_LameScam_Hamamatsu_RS » est sélectionné. Sa « Description » montre qu'il s'agit d'un protocole de scan de lames, acquis avec le scanner Hamamatsu, d'images marquées avec la coloration « Rouge Sirius ».

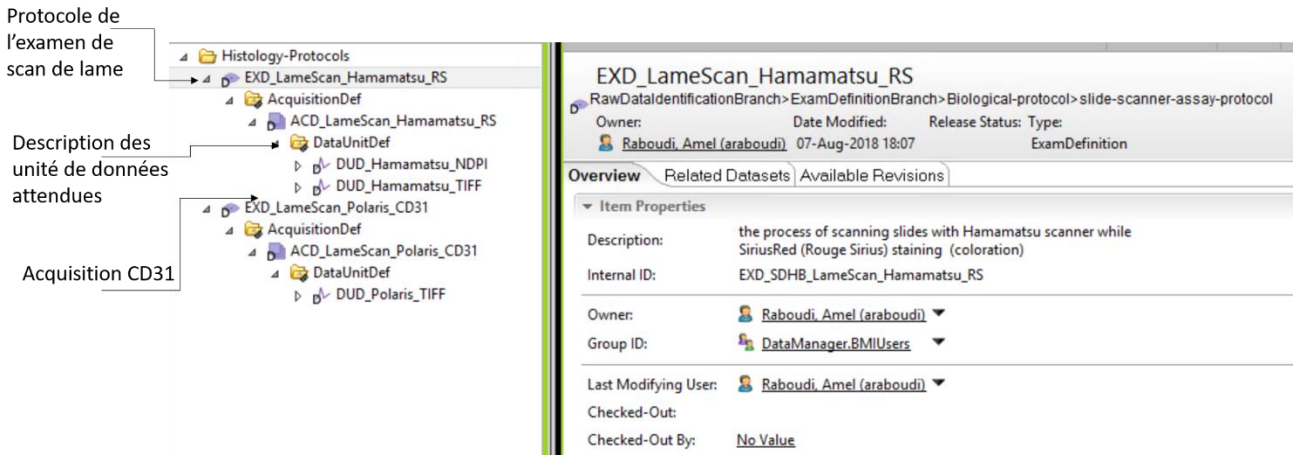


Figure 115 Objets ajoutés au MDD pour spécifier les protocoles d'histologie

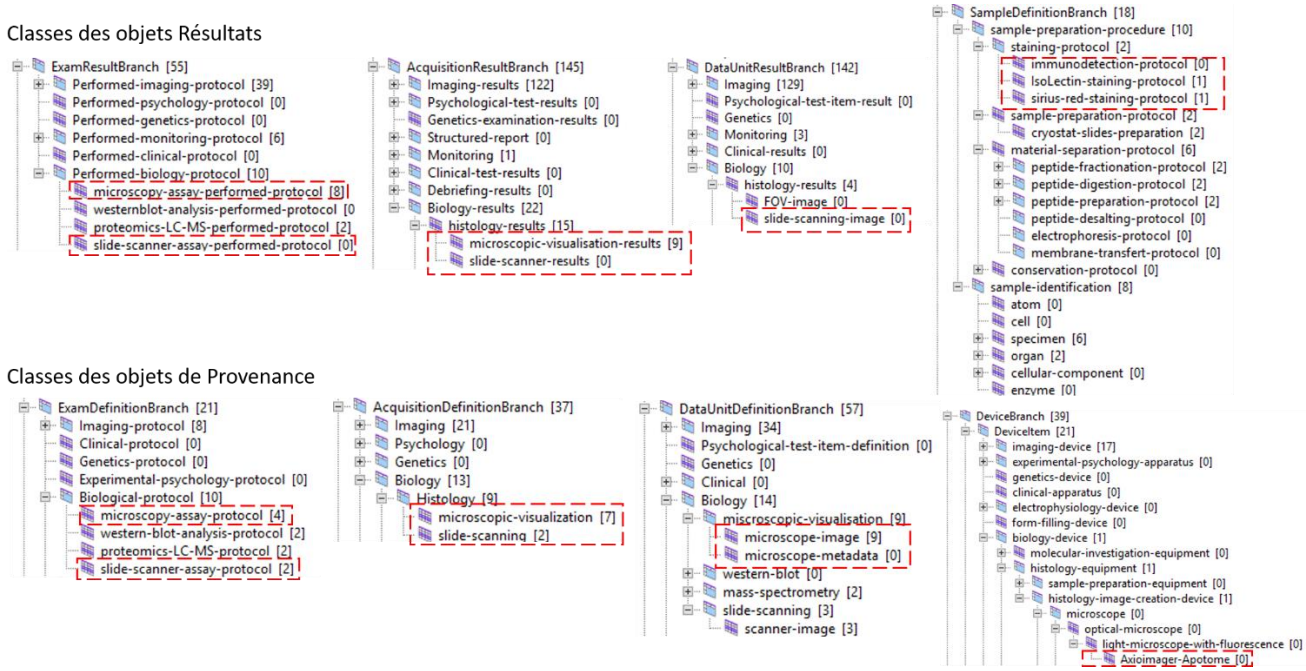


Figure 116 Classes de « Classification » ajoutées pour la prise en compte des données d'histologie

La Figure 116 montre des parties de la « Classification » du système BMS-LM. Les ajouts de classes pour les données d'histologie sont encadrés en rouge. Les classes feuilles encadrées sont celles utilisées directement pour l'annotation des données d'histologie. Une classe feuille est une classe qui constitue une extrémité de l'arborescence. Elle caractérise un objet sans avoir de descendances. Pour connaître l'objet du MDD à qui a été attribuée la classe ajoutée, il faut regarder le nom de la branche de « Classification » correspondante. Par exemple, la classe « Axioimager-Apotome » de la branche « DeviceBranch » spécifie l'objet « Device (DVC) », et « microscopic-visualization » de la branche « AcquisitionDefinitionBranch » est une spécification de l'objet « Acquisition Definition (ACD) ».

VI.4.1.3. Déroulement de l'intégration des données

L'intégration des données d'histologie a été effectuée pour des données de l'étude « Renotox » consacrée à l'évaluation de la toxicité d'un agent anti-angiogénique (Sunitinib) sur les reins. L'outil « tree-2-csv » a été exécuté par l'utilisateur clé « ACE » afin de générer le tableau contenant les descripteurs des données. Ce tableau a servi d'entrée à la procédure de transformation de données (ETL) « csv-2-xml » après avoir été enrichi par des informations nécessaires au bon fonctionnement de l'ETL par le data manager « Ame ». Les actions entreprises par « ACE » et par « Ame » sont décrites dans le diagramme BPMN[83] dans la Figure 117 suivante.

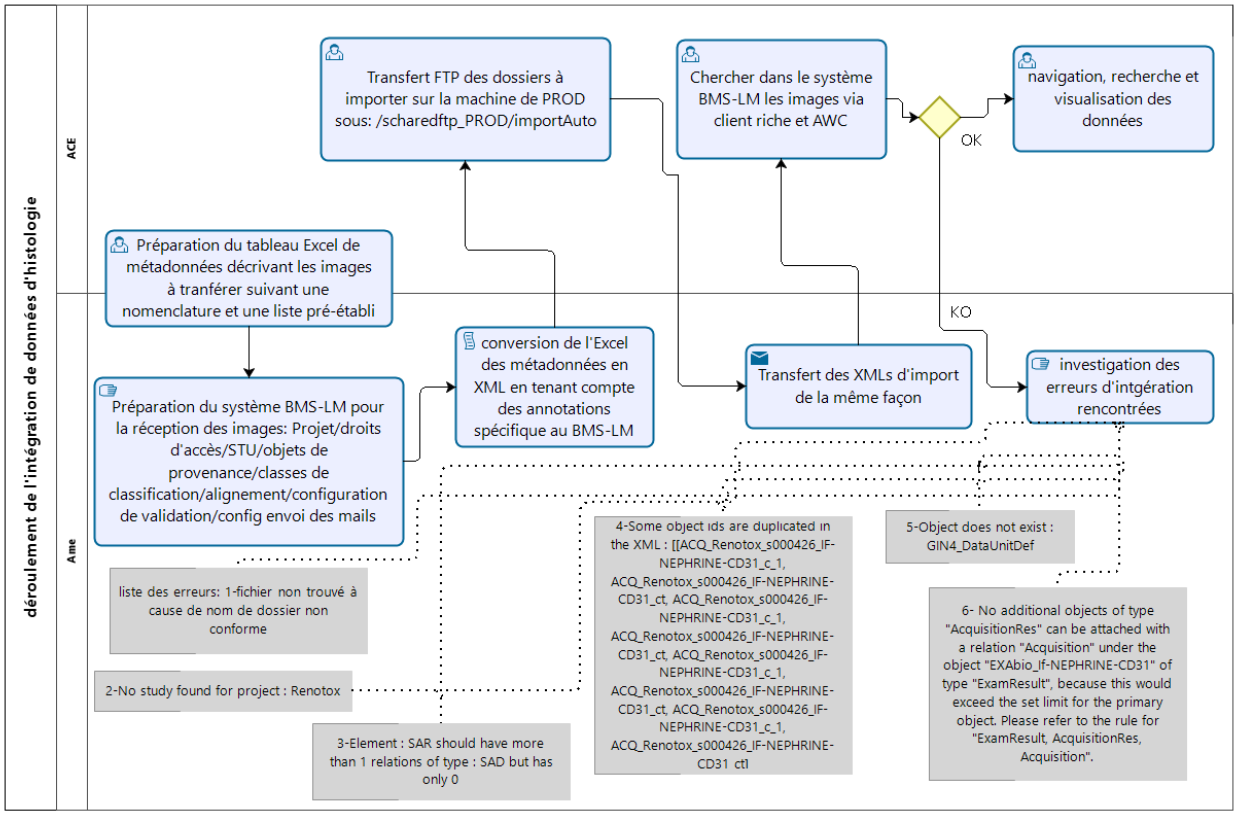


Figure 117 Étapes du déroulement de l'import des données d'histologie pour le plan d'expérimentation « intégration_1 »

Le transfert des fichiers depuis le serveur local d'histologie au laboratoire LRI vers l'espace temporaire de stockage du serveur BMS-LM « importAuto » a été effectué par l'utilisateur clé « ACE » via FTP, le 17/04/2018 (voir Figure 118) qui présente le dossier « DATA CD31 NEPHRIM 20171117 » transféré

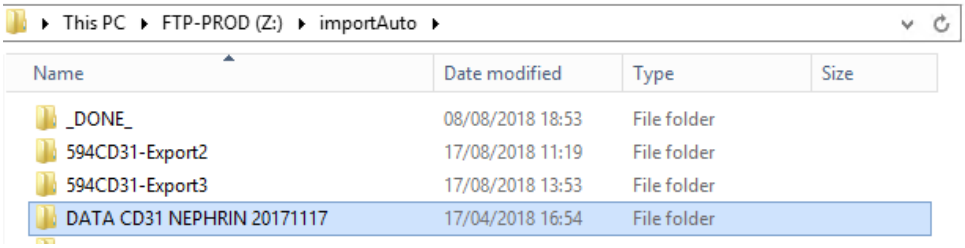


Figure 118 Dossiers transférés dans le cadre du plan d'expérimentation « intégration_1 »

[83] Pour cette fois-ci, nous avons choisi le BPMN (et pas le SADT) à cause de sa richesse de modélisation et parce que ce diagramme n'est pas à destination des utilisateurs, mais des data managers.

Ensuite, les fichiers XML d'import ont été transférés dans le dossier « importAuto » par « Ame » (voir Figure 119). Ce dossier est un *watch folder* qui est consulté périodiquement par le service web d'intégration de données. Quand les fichiers XML (« sample-SAR-Nephrine.xml » et « import-EXA.xml » dans ce plan d'expérimentation) sont détectés, l'intégration de données est démarrée. Une fois achevée, les fichiers XML sont déplacés à un dossier « success » pour notifier du succès de l'opération, ou « failed » accompagné d'un message d'erreur sinon. Parallèlement, une notification de la fin de l'import est envoyée par courriel.

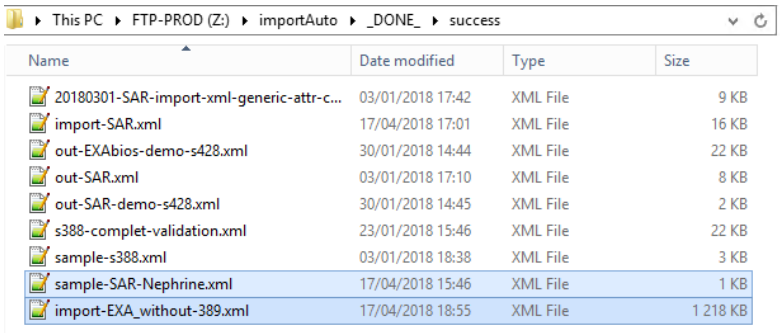


Figure 119 Fichiers XML correspondants au plan d'expérimentation « intégration_1 »

L'intégration de données a pu être réalisée (comme dans la capture d'écran Figure 120) après plusieurs essais infructueux à la suite de la détection et correction d'erreurs syntaxiques dans les fichiers XML d'entrée : des règles non respectées, des identifiants non reconnus, etc. (voir les notes du diagramme BPMN de la Figure 117). L'image sélectionnée dans la partie gauche de la Figure 120 est visualisée dans la partie droite de celle-ci. Elle présente des glomérules de reins marqués à la « Néphrine » (vert), les vaisseaux au « CD31 » (rouge) et des noyaux cellulaires marqués au « DAPI » (bleu). Elle est une image composite. Les objets du MDD (EXA, ACQ, DUR) permettent de structurer le lot d'images d'entrée afin de mieux comprendre les liens de Provenance des images.

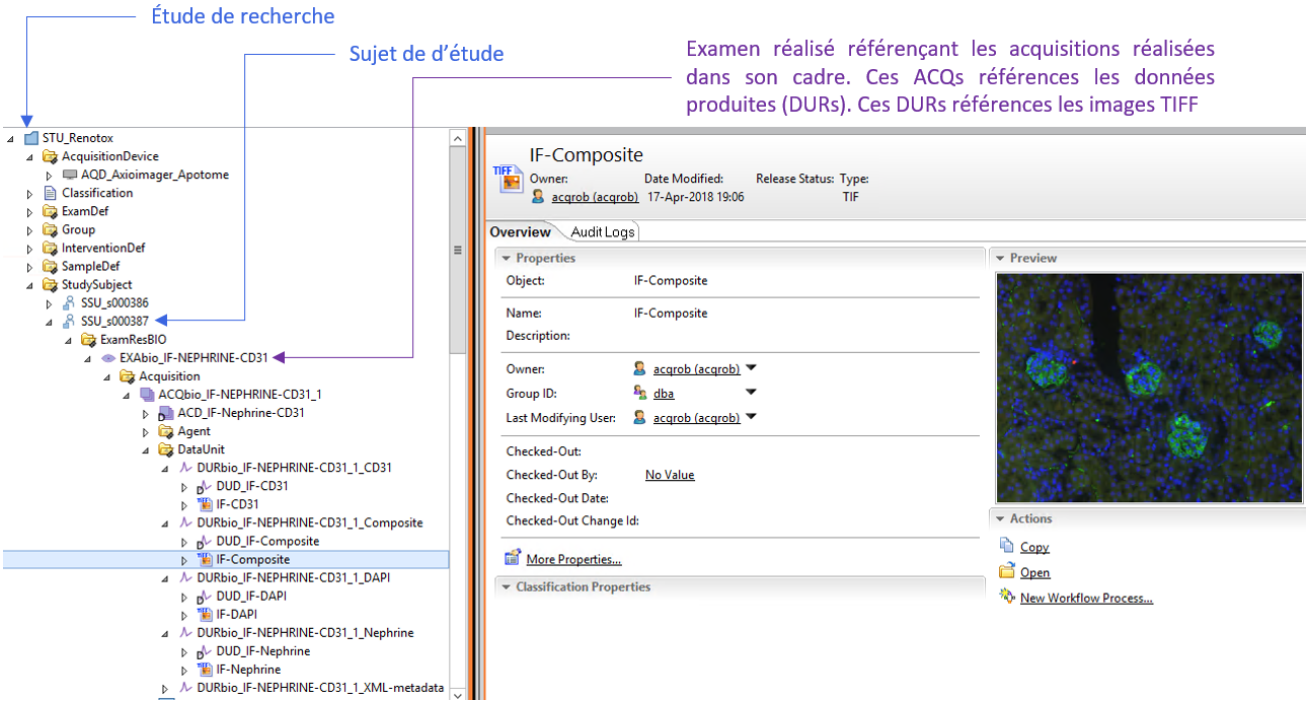
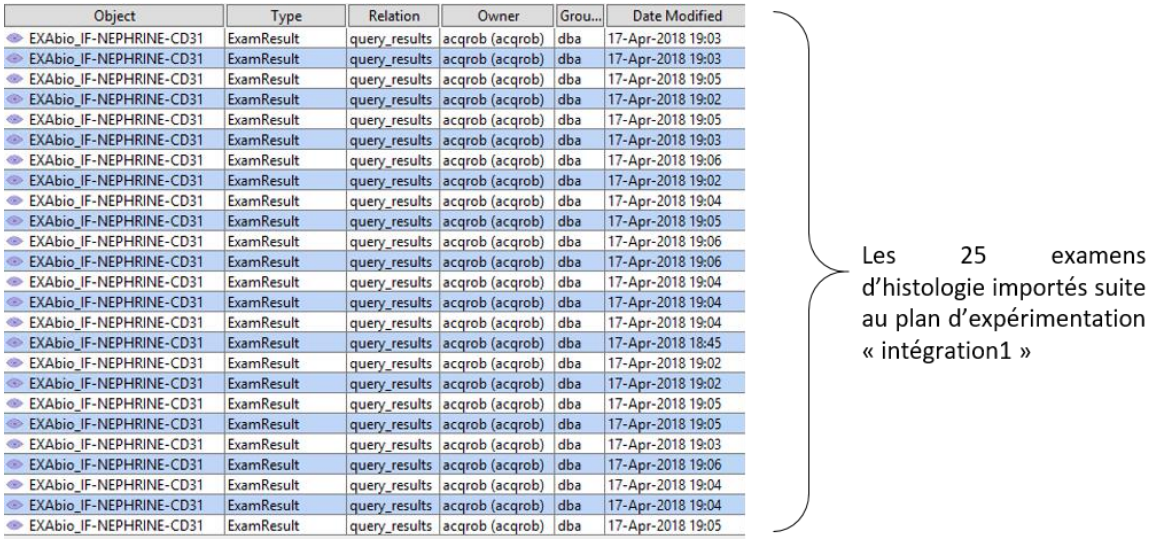


Figure 120 Exemple d'image d'histologie consultable depuis le système BMS-LM

Finalement 25 examens (voir liste de la Figure 121) ont été importés qui regroupent 1683 images TIFF de glomérules de reins.



Object	Type	Relation	Owner	Grou...	Date Modified
EXAbio_IF-NEPHRINE-CD31	ExamResult	query_results	acqrob (acqrob)	dba	17-Apr-2018 19:03
EXAbio_IF-NEPHRINE-CD31	ExamResult	query_results	acqrob (acqrob)	dba	17-Apr-2018 19:03
EXAbio_IF-NEPHRINE-CD31	ExamResult	query_results	acqrob (acqrob)	dba	17-Apr-2018 19:05
EXAbio_IF-NEPHRINE-CD31	ExamResult	query_results	acqrob (acqrob)	dba	17-Apr-2018 19:02
EXAbio_IF-NEPHRINE-CD31	ExamResult	query_results	acqrob (acqrob)	dba	17-Apr-2018 19:05
EXAbio_IF-NEPHRINE-CD31	ExamResult	query_results	acqrob (acqrob)	dba	17-Apr-2018 19:03
EXAbio_IF-NEPHRINE-CD31	ExamResult	query_results	acqrob (acqrob)	dba	17-Apr-2018 19:06
EXAbio_IF-NEPHRINE-CD31	ExamResult	query_results	acqrob (acqrob)	dba	17-Apr-2018 19:02
EXAbio_IF-NEPHRINE-CD31	ExamResult	query_results	acqrob (acqrob)	dba	17-Apr-2018 19:04
EXAbio_IF-NEPHRINE-CD31	ExamResult	query_results	acqrob (acqrob)	dba	17-Apr-2018 19:05
EXAbio_IF-NEPHRINE-CD31	ExamResult	query_results	acqrob (acqrob)	dba	17-Apr-2018 19:06
EXAbio_IF-NEPHRINE-CD31	ExamResult	query_results	acqrob (acqrob)	dba	17-Apr-2018 19:06
EXAbio_IF-NEPHRINE-CD31	ExamResult	query_results	acqrob (acqrob)	dba	17-Apr-2018 19:04
EXAbio_IF-NEPHRINE-CD31	ExamResult	query_results	acqrob (acqrob)	dba	17-Apr-2018 19:04
EXAbio_IF-NEPHRINE-CD31	ExamResult	query_results	acqrob (acqrob)	dba	17-Apr-2018 19:04
EXAbio_IF-NEPHRINE-CD31	ExamResult	query_results	acqrob (acqrob)	dba	17-Apr-2018 18:45
EXAbio_IF-NEPHRINE-CD31	ExamResult	query_results	acqrob (acqrob)	dba	17-Apr-2018 19:02
EXAbio_IF-NEPHRINE-CD31	ExamResult	query_results	acqrob (acqrob)	dba	17-Apr-2018 19:02
EXAbio_IF-NEPHRINE-CD31	ExamResult	query_results	acqrob (acqrob)	dba	17-Apr-2018 19:05
EXAbio_IF-NEPHRINE-CD31	ExamResult	query_results	acqrob (acqrob)	dba	17-Apr-2018 19:05
EXAbio_IF-NEPHRINE-CD31	ExamResult	query_results	acqrob (acqrob)	dba	17-Apr-2018 19:03
EXAbio_IF-NEPHRINE-CD31	ExamResult	query_results	acqrob (acqrob)	dba	17-Apr-2018 19:06
EXAbio_IF-NEPHRINE-CD31	ExamResult	query_results	acqrob (acqrob)	dba	17-Apr-2018 19:04
EXAbio_IF-NEPHRINE-CD31	ExamResult	query_results	acqrob (acqrob)	dba	17-Apr-2018 19:04
EXAbio_IF-NEPHRINE-CD31	ExamResult	query_results	acqrob (acqrob)	dba	17-Apr-2018 19:05

Figure 121 Liste des examens créés dans le système BMS-LM à la suite de l'exécution du plan d'expérimentation « intégration_1 »

Le lendemain de l'intégration (18/04/2018), un guide pour l'exploration des données dans le système BMS-LM a été envoyé à l'utilisateur clé « ACE ». Son contenu peut être consulté en annexe D. Il décrit la manière de se connecter à l'outil et indique comment naviguer pour retrouver les images. De plus, nous avons défini un exercice simple d'exploration de données afin d'aider et d'encourager « ACE » à se servir de l'outil.

Pour résumer, afin d'exécuter le plan d'expérimentation « intégration_1 », nous avons suivi les étapes suivantes :

1. La définition d'une nomenclature de nommage des fichiers d'histologie et d'une arborescence de dossier pour la standardisation du stockage sur le serveur. Cette nomenclature est entretenue par l'utilisateur clé « ACE ». Des vérifications périodiques des données sur le serveur doivent être réalisées pour garantir son maintien.
2. L'extraction des métadonnées qui sont présentes dans les noms et l'arborescence de dossier fichier via un l'ETL « tree-2-csv ». Cet ETL a généré le tableau d'entrée à la procédure d'intégration « générique » de donnée.
3. La vérification du tableau et son enrichissement éventuel par d'autres informations issues d'autres sources.
4. La préparation du système BMS-LM pour accueillir les données via l'ajout d'objets de Provenance dans le MDD et de classes dans la « Classification ».
5. La conversion du tableau en XML via l'ETL d'intégration « csv-2-xml »
6. Le transfert des images TIFF et des fichiers XML vers le *watch folder* « importAuto » via FTP et le lancement de l'intégration.
7. La définition d'exercices et scénarios pour l'exploration des données importées.

Les besoins qui ont été considérés dans ce plan d'expérimentation sont : B1-archivage, et B2-import des données d'histologie, B8-traçabilité, B11-standardisation via l'utilisation de nomenclature, B15-flexibilité via l'adaptation de la méthode « générique » d'intégration de données.

VI.4.2. INTÉGRATION « PARTIELLE » D'UN TRAITEMENT D'ANALYSE DE DONNÉES EN HISTOLOGIE

En étroite collaboration avec l'utilisateur clé « DBA », nous avons mis en œuvre une intégration « partielle » d'une application logicielle « maison » pour l'analyse d'images d'histologie au laboratoire LRI dans le cadre d'un encadrement de stage de Master 1, réalisé par Roberto Duarte. Ce traitement

était utilisé par l'utilisateur clé « ACE » au laboratoire. Nous appliquons les principes de l'intégration « partielle » expliquée dans la section IV.2.2.2 , afin de l'intégrer dans le système BMS-LM.

VI.4.2.1. Le calcul scientifique de quantification histologique

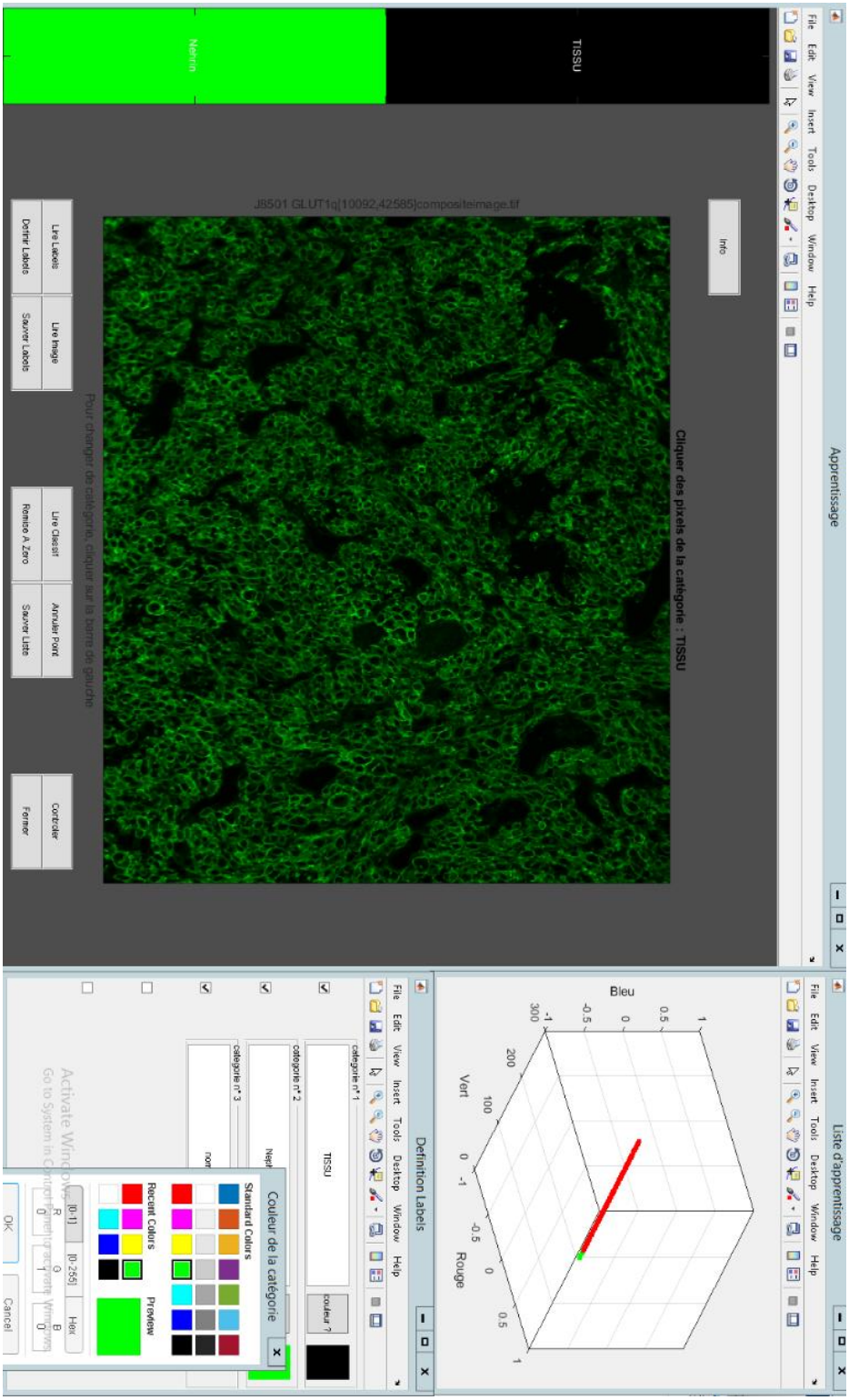


Figure 122 Interface graphique de l'application logicielle d'analyse d'images d'histologie utilisée pour l'apprentissage

L'application logicielle étudiée est un outil de quantification d'images histologiques. Elle est développée dans l'environnement Matlab et se base sur un algorithme de classification des pixels de l'image par

apprentissage supervisé. Il est composé de deux modules : le module de « préparation à l'apprentissage » où l'utilisateur définit ce qu'il souhaite détecter et alimente ainsi un « modèle de prédiction », et le module de « production » qui exécute le modèle de prédiction sur l'ensemble des images de l'étude.

Pendant la préparation de l'apprentissage, lors d'une utilisation typique du logiciel, l'utilisateur définit, charge ou sauvegarde une liste de « Labels » (voir boutons de la Figure 122). Les Labels sont définis par des paires « nom/couleur », par exemple, « Néphrine/vert », « CD31/rouge », « DAPI/bleu ». Ces labels correspondent à des "classes" pour l'algorithme de classification. Ensuite, sur une image sélectionnée, l'opérateur active un Label et clique sur l'image pour associer un pixel au Label activé (par exemple, associer un pixel bleu au Label "DAPI"). Après plusieurs « clics », il en résulte une liste de pixels labélisés, appelée « liste de classification », qui servira à la génération du modèle de classification pour la production. Un bouton de test ou de préproduction permet de contrôler la qualité de l'apprentissage au fil de l'eau.

Le module de production utilise les connaissances recueillies lors de l'apprentissage pour paramétrer un modèle de prévision (ou classifieur) qui sera exécuté sur l'ensemble des pixels d'un corpus d'images comparables à celles utilisées pour l'apprentissage. L'utilisateur choisit comme entrée du processus, une liste de points labélisés et un dossier dans lequel sont rangées les images. L'application logicielle applique exécute alors automatiquement l'algorithme de classification. Quelques techniques de traitement d'image sont utilisées optionnellement en post-traitement pour régulariser le résultat final. À l'issue de cette procédure, le comptage du nombre de pixels de chaque catégorie (i.e. Label) est effectué et compilé pour toute l'étude sous forme d'un tableur Excel.

Dans le cas d'un workflow très visuel et graphique comme celui du module de « préparation à l'apprentissage » (définition des Labels, choix des images une à une, association des points aux Labels), l'intégration « partielle » est la seule qui peut être envisagée. Cette intégration permettra de :

- Assurer la traçabilité des listes de Labels et listes des associations Label-pixels (liste de classification) produites par les utilisateurs.
- Réutiliser une liste de points créée par quelqu'un d'autre en ayant connaissance de sa provenance (confiance).
- Familiariser les utilisateurs avec les notions de gestion de cycle de vie sans impacter leur quotidien.

VI.4.2.2. La modélisation et structuration du calcul scientifique

Le stagiaire que j'ai encadré a restructuré la partie logicielle de « préparation de l'apprentissage » pour qu'elle soit modulaire et mieux intégrable au système BMS-LM. Le diagramme BPMN de la Figure 123 ci-après met en avant les différentes étapes identifiées ainsi que les fichiers échangés tout au long du processus. La liste des fichiers et données échangées est la suivante :

- Des fichiers de Labels : couples de valeurs (nom/couleur)
- Des répertoires définis par convention pour les données des images sources et les données dérivées
- Des fichiers de « liste de classification » : pour lister les points sélectionnés (en RGB), les Labels associés, et les noms d'images sources.

Cette structuration nous a aidés à identifier les unités et la chaîne de traitement de la « préparation de l'apprentissage » comme expliquées dans le diagramme BPMN. Une première unité (PUD_DefinitionSubstratClassif) consiste à définir les sources images et les Labels qui seront utilisés pour la « liste de classification ». Une deuxième unité (PUD_SelectionClassifExistante) permet de reprendre une procédure existante et sauvegarder la « liste de classification » et une troisième unité permet de la modifier (PUD_ModifierClassification).

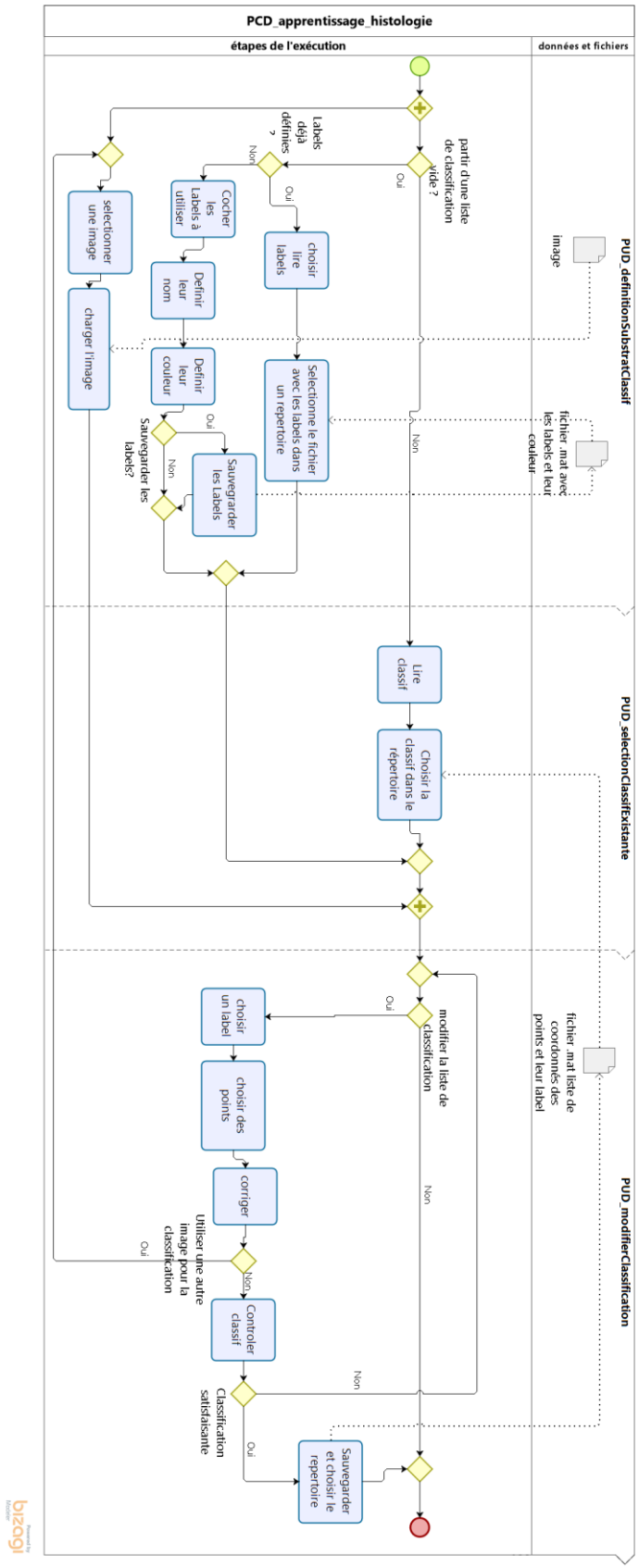


Figure 123 Étapes de la préparation de l'apprentissage pour la quantification d'images histologiques

VI.4.2.3. Précâblage des éléments au niveau du système BMS-LM et de la machine locale

Nous avons ajouté les objets dans les captures de la figure 117 suivante dans le système BMS-LM afin de permettre le suivi de l'exécution locale du traitement scientifique : l'objet spécifiant les entrées du workflow (WFI_Prep_Apprentissage_Histo_Fibrose), les types des données d'entrée (ACD_LameScan_Hamamatsu_RS, DUD, EXD), la machine sur laquelle le workflow est exécuté (AQD_machine_CEDRE), ainsi que la chaîne de traitement (PCD_Matlab_Prep_Apprentissage_Histo, PUD_Prep_Labels, etc.) et ses paramètres (PCP_Prep_Labels) utilisés dans ce workflow. Nous avons veillé à stocker en objets « résultats d'unité du traitement (PUR) », les principales données réutilisables : « la liste de classification », les labels, et des éventuels paramétrages.

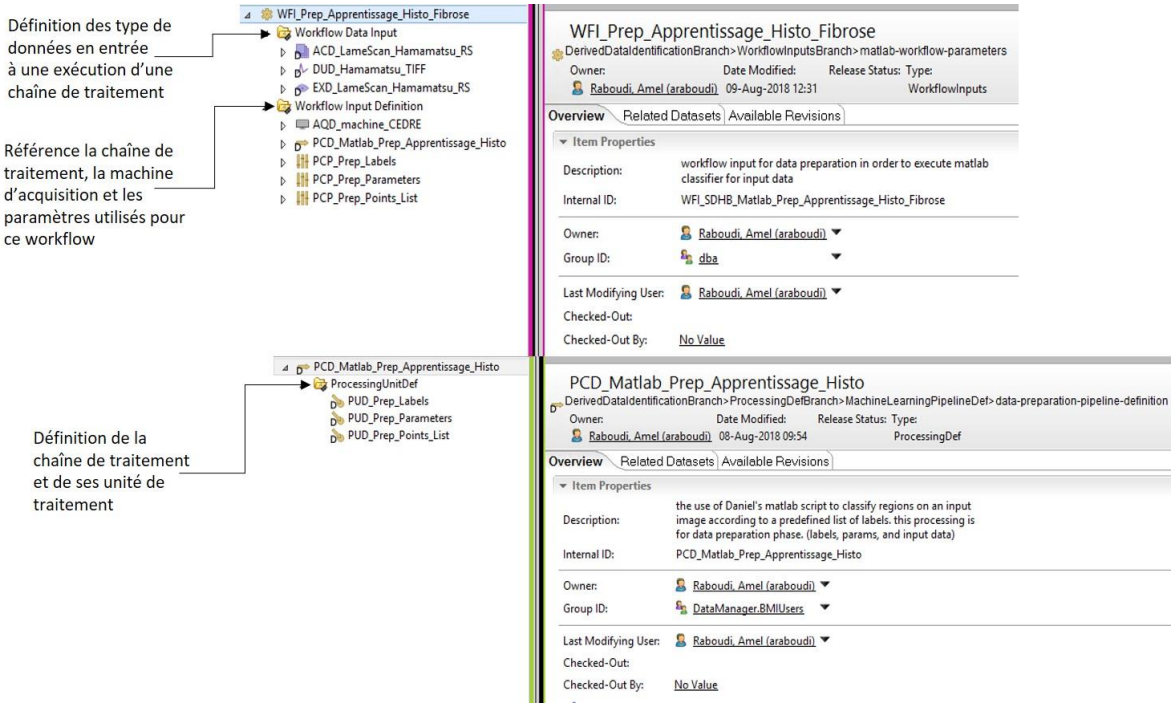


Figure 124 Objets du MDD ajoutés au système BMS-LM dans le cadre du plan d'expérimentation « trait1 »

Lors des échanges entre le système BMS-LM distant et l'application Matlab locale, des upload/download de fichiers sont exécutés. Un emplacement local temporaire est utilisé pour assurer l'exécution en local. Nous avons défini l'arborescence de dossiers/fichiers de la Figure 125 qui permet de stocker les fichiers de « listes de classification » dans le dossier « ClassRef » et les paires « nom/couleur » dans le dossier « Labels ». Les images téléchargées du système BMS-LM sont regroupées par groupe de sujet (SGP) et par sujet dans l'étude (SSU) sous le dossier correspondant à l'exécution en cours de la chaîne de traitement (cf. Figure 125 Execution_SGP_rduarte_201807...).

Figure 125 Arborescence prédéfinie du stockage local des fichiers échangés entre l'application Matlab et le système BMS-LM

VI.4.2.4. Ajout des modules de connexion au système BMS-LM

Nous nous sommes concentrés sur la traçabilité des échanges entre le système BMS-LM et l'outil logiciel de quantification histologique via l'API REST du système BMS-LM (voir section 0IV.2.2.2). Le diagramme de séquence de la Figure 126 présente les différents échanges que nous avons mis en place entre le système BMS-LM et l'application Matlab.

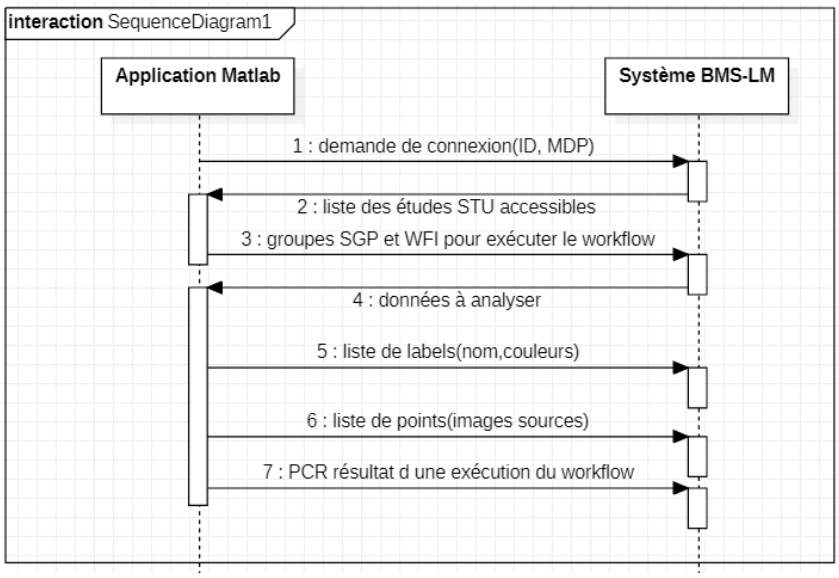


Figure 126 Diagramme de séquence des échanges entre le système BMS-LM et l'application Matlab d'apprentissage histologique

Nous avons été amenés à ajouter un module de connexion au système BMS-LM qui permet de télécharger les données d'intérêt avant de commencer l'apprentissage. Ci-après, nous décrivons la série des interfaces graphiques que nous avons mises en place pour le téléchargement (download) des données depuis le système BMS-LM. Pour télécharger les données, l'utilisateur doit, soit sélectionner un groupe de sujets (SGP) déjà présent dans le système, soit, créer un groupe à partir des sujets dans l'étude (SSU) affichés dans l'interface graphique. Aussi, il doit également sélectionner le WFI adapté à ces données d'entrées (il y a un WFI pour chaque ensemble différent de données d'entrée). L'interface interactive de sélection de données images à télécharger est présentée dans la Figure 127. Son utilisation est décrite dans le diagramme de séquence de la Figure 128. Au lancement de l'application Matlab, la liste des études de recherche à laquelle l'utilisateur est autorisé s'affiche et il doit sélectionner l'étude qui

l'intéresse. Une fois l'étude sélectionnée, la liste des « groupes de sujets (SGP) » présente dans le système BMS-LM s'affiche à gauche et celle des « sujets dans l'étude (SSU) » qui les composent à droite. L'utilisateur peut réutiliser un SGP existant ou créer un nouveau. Ce sont les images en lien avec les sujets de ce groupe qui seront téléchargés plus tard. La dernière étape de sélection est la sélection du « workflow input WFI » qui référence les types de données d'entrée et les configurations à utiliser pour l'exécution en cours (déjà préconfiguré à l'étape de préparation du système BMS-LM). Le téléchargement des images depuis le système BMS-LM dans l'espace temporaire (défini lors de la préparation de la machine locale) s'effectue en faisant la jointure entre les types de données référencées par le WFI (EXD, ACD, DUD) et ceux liés aux SSUs des SGPs (EXA, ACQ, DUR).

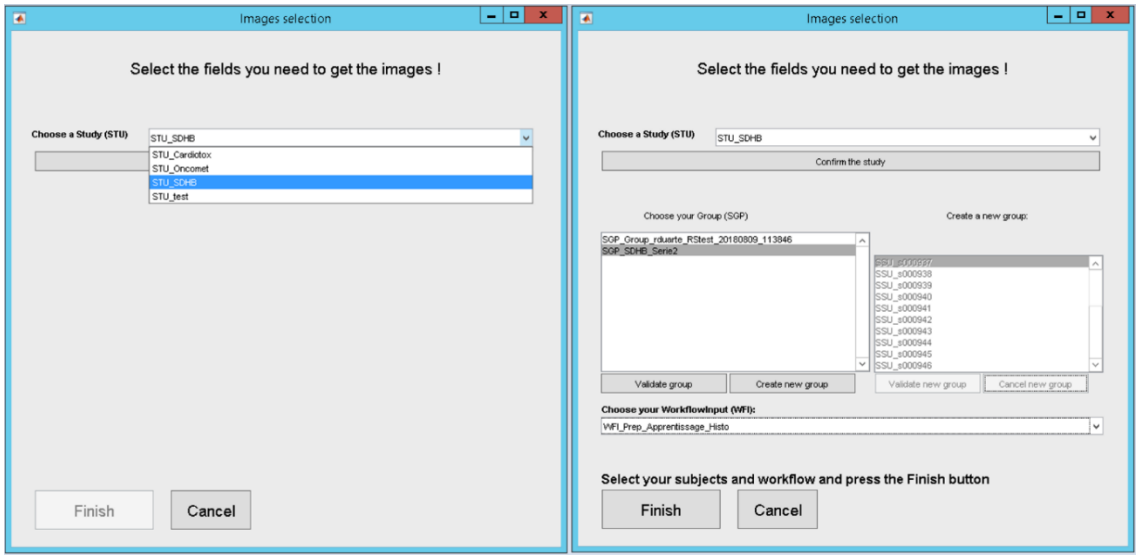


Figure 127 Interface interactive de téléchargement des données depuis le système BMS-LM

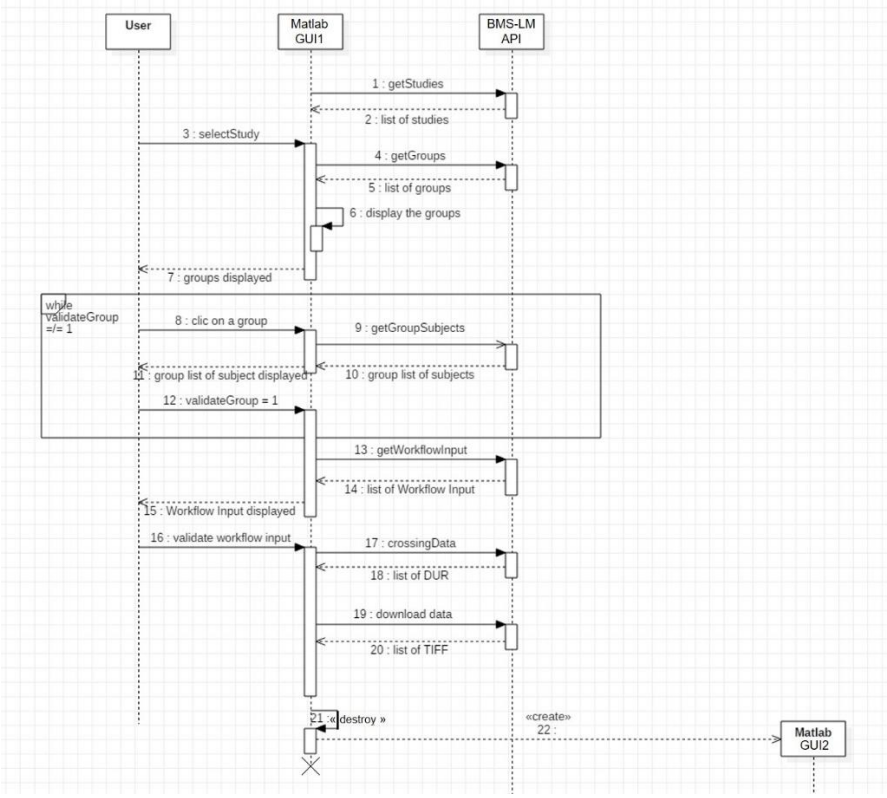


Figure 128 Diagramme de séquence décrivant les différents échanges pour télécharger les images depuis le système BMS-LM (diagramme réalisé par Roberto Duarte lors de son stage)

Lors du développement des interfaces graphiques, nous avons veillé à les rendre les plus ergonomiques possibles. L'interface de téléchargement des données de la Figure 127 est interactive pour répondre à ce besoin : à chaque sélection, elle se met à jour et met en évidence la section à remplir par l'utilisateur (l'autre étant grisé) afin de le guider. Nous avons aussi rendu l'interface principale du logiciel plus modulaire en regroupant les boutons concernant la liste des points ensemble, les Labels ensemble, et nous avons ajouté un bouton d'envoi au système BMS-LM des données résultat « Finir et envoyer au BMS-LM ».

Pour la récupération des fichiers de Labels et de « listes de classification » depuis le système BMS-LM, nous avons défini des requêtes se lançant respectivement via les boutons de l'interface principale « Lire labels » et « Lire liste ». Les listes des « PUR_Labels » et des « PUR_listeClassification » existant dans le système BMS-LM s'affichent et l'utilisateur sélectionne et télécharge celles qui l'intéressent. Pour l'envoi au système BMS-LM, nous avons proposé une interface graphique qui se lance après le clic sur « sauver Labels » et « sauver Liste », elle demande à l'utilisateur de choisir un nom au « résultat d'unité de traitement (PUR) » qu'il allait envoyer au système BMS-LM avant de l'envoyer. (Les captures d'écran associées aux interfaces graphiques non présentes ici peuvent être consultées en Annexe D)

Les PCRs et PURs envoyés au système BMS-LM doivent référencer leur WFI d'origine comme dans la capture d'écran de la Figure 129 ci-après.

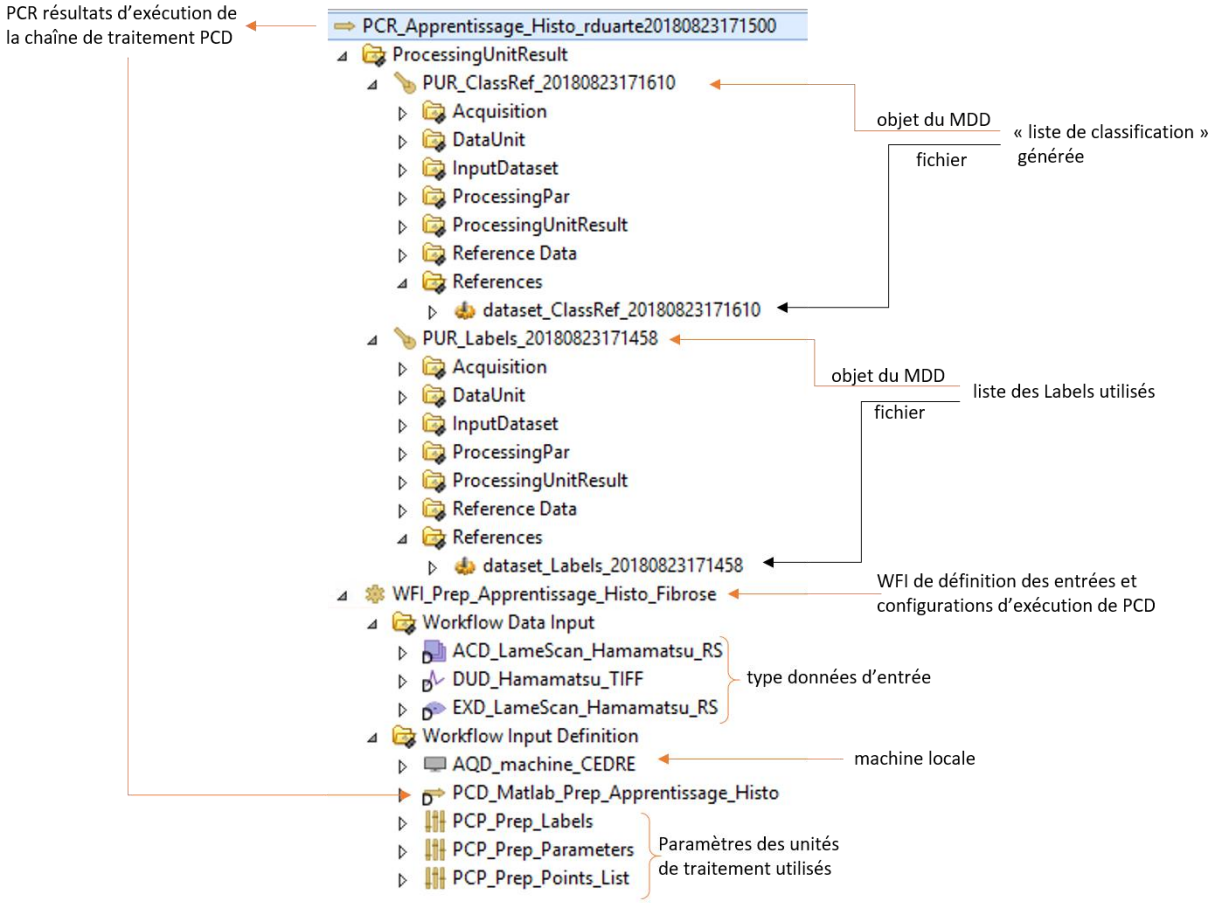


Figure 129 Des fichiers de « liste de classification » et de Labels envoyés au système BMS-LM et gérés via les objets du MDD

VI.4.2.5. Test avec des données réelles et validations utilisateur

Le logiciel d'histologie était utilisé dans sa version antérieure par l'utilisatrice « ACE ». Il était donc important de lui faire tester la nouvelle version qui intégrait la traçabilité tout au long de l'exécution du traitement de calcul scientifique. Nous avons réalisé deux questionnaires : un premier avant la formation,

et un autre après la formation. Après un bref tutoriel sur les objets du MDD BMS-LM, nous avons présenté un document qui faisait office de mode d'emploi de la nouvelle version de l'application Matlab. Chaque étape y était expliquée tout en faisant référence aux notions de BMS-LM impliquées. Ainsi elle était prête pour l'utilisation du logiciel. Ci-après, Figure 130, une capture d'écran montrant un objet SGP généré lors de l'un de ces tests utilisateur via l'interface de téléchargement des données Figure 127. Les grilles d'évaluation distribuées en début puis en fin de démonstration pour évaluer la progression de « ACE » sont présentées en Annexe D. Le bilan de cette évaluation est très intéressant, il nous permet de voir que les notions associées au système BMS-LM ont été assimilées et que la nouvelle version logicielle a été acceptée sans réticence contre par rapport au changement effectué.

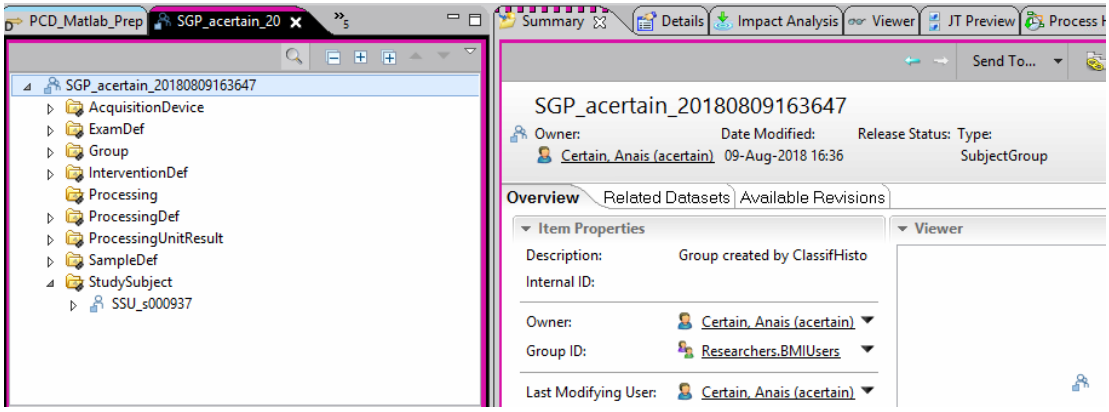


Figure 130 SGP créé par l'utilisateur clé « ACE » lors du test de l'application Matlab dans sa version BMS-LM

En résumé, afin d'exécuter le plan d'expérimentation « trait1 », nous avons suivi les étapes suivantes :

1. La modélisation du traitement pilote sous forme d'une chaîne de traitement (PCD) et la collecte d'informations et de données importantes à tracer : restructuration du traitement d'un côté et ajout d'objets de Provenance dans le MDD BMS-LM de l'autre (PCD, WFI, PUD, etc.).
2. L'utilisation de la bibliothèque Matlab et de l'API REST pour formuler les requêtes et définir les messages échangés entre l'application locale et le système BMS-LM.
3. L'ajout d'interfaces graphiques adaptées à l'application locale. L'utilisateur ne communique avec le système BMS-LM que via ces interfaces.

Les besoins qui ont été considérés dans ce plan d'expérimentation sont : B1-archivage des données dérivées, B6-Analyse avec l'ajout de ce workflow dans le catalogue géré par le système BMS-LM, B8-traçabilité, B11-standardisation avec la restructuration du code et standardisation des entrées sorties, B15-flexibilité avec une intégration qui ne change pas le quotidien de l'utilisateur. B19- suivi avec suivi des activités de recherche liées à l'utilisation d'applications logicielles maison. Les leviers sont L3-adaptation et L4-mirroring.

VI.5. INTÉGRATION ET ANALYSE DES DONNÉES EN IMAGERIE TEP-TDM

Les données les plus volumineuses et les plus centrales du laboratoire sont les données de TEP-TDM. Elles sont en lien avec un des axes principaux de recherche du laboratoire – l'imagerie nucléaire. Dans cette section sont d'abord présentées, les réalisations en lien avec le plan d'expérimentation « intégration_2 » et l'utilisateur clé « TVI » pour la gestion des données d'imagerie TEP-TDM sont présentées en premier lieu. Ensuite, un test d'intégration « totale » d'un calcul scientifique est commenté. Il analyse la réponse impulsionnelle (RI) du cœur en réutilisant des données TEP.

VI.5.1. INTÉGRATION DES DONNÉES TEP-TDM DANS LE SYSTÈME BMS-LM

Nous explicitons dans ce paragraphe comment nous avons appliqué et adapté la méthode « générique » d'intégration de données hétérogènes aux données TEP-TDM générées par le scanner Mediso au laboratoire LRI.

VI.5.1.1. La gestion de données TEP-TDM au laboratoire LRI

Avant de décrire les différentes implémentations réalisées, nous présentons dans ce paragraphe l'état de lieux en gestion de données TEP-TDM au démarrage de l'intégration. Lors de l'acquisition des données, il y a plusieurs annotations et champs de formulaires à remplir dans l'interface logicielle de commande du scanner Mediso (voir Figure 131 suivante), le contenu de ces champs est ensuite stocké dans les fichiers DICOM résultants de l'acquisition.

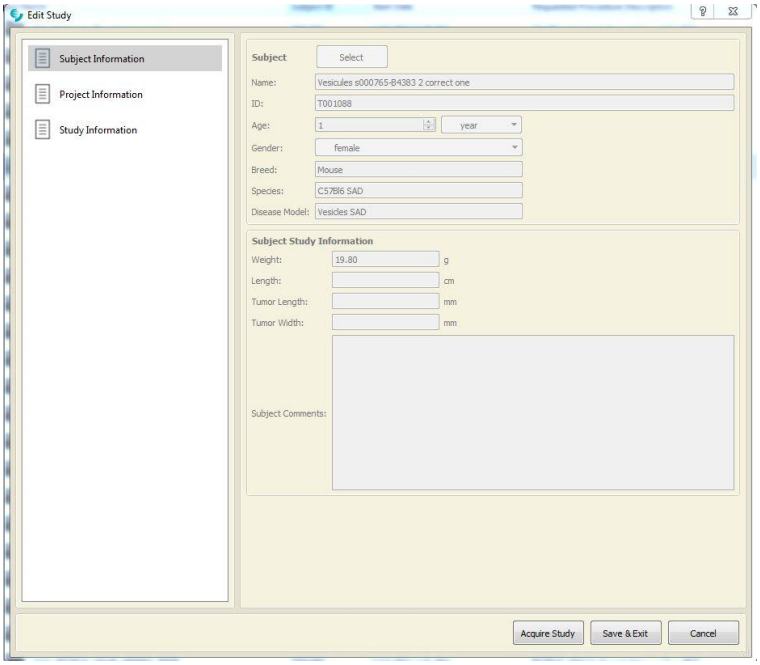


Figure 131 Une interface de saisie d'informations de scan TEP-TDM, informations sur le sujet de l'étude

En plus des annotations fournies par le logiciel d'acquisition, un cahier des expériences est maintenu pour décrire, au moment où l'opérateur est en train d'expérimenter, les paramètres et les informations concernant l'acquisition en cours. Plus tard, ces informations, notées sur le cahier des expériences, sont retranscrites sur un tableau Excel partagé avec tous les membres du laboratoire concernés (voir Figure 132). Ce tableau est une source d'information importante. Il est versionné sur le serveur tous les jours afin d'éviter la perte ou le changement d'information.

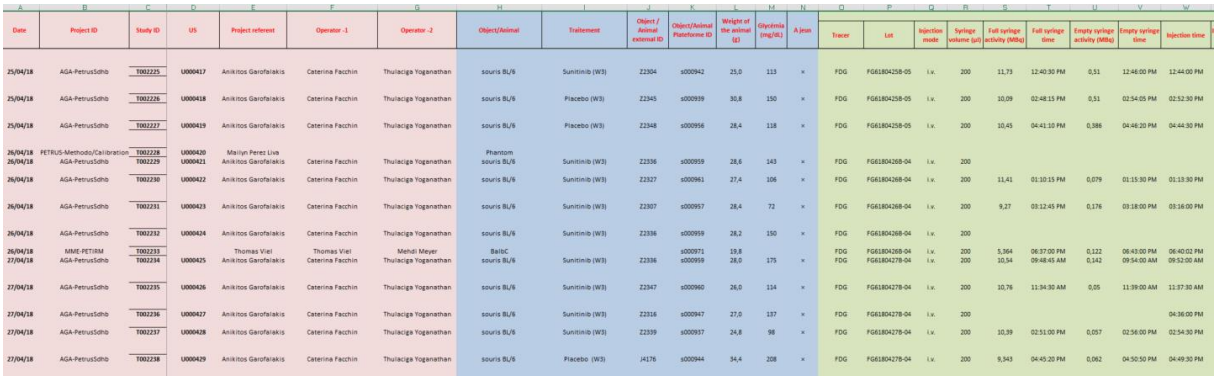


Figure 132 Tableau Excel listant les expériences TEP-TDM et leurs descriptions

Les données acquises sont stockées sur la machine d'acquisition, le temps de les transférer sur le serveur de stockage local du laboratoire, dimensionné à 54To. Elles sont gérées par le logiciel du scanner Mediso et ne sont pas consultables facilement. Pour pouvoir les utiliser localement en vue d'un traitement quelconque, un script de copie de données, du serveur sur la machine locale de travail « copieDicomTep2017 » a été développé par l'ingénieur en calcul scientifique.

VI.5.1.2. Intégration des données TEP-TDM dans le système BMS-LM

Les données issues du scanner TEP suivent un standard reconnu par la communauté des imageurs : le standard DICOM. Les annotations saisies via la console se trouvent dans ce qui est appelé « DICOM Header » où il existe une liste d'annotations avec une identification unique (groupe, élément) comme dans les exemples suivants :

- PatientName (0010,0010)
- PatientID (0010,0020)
- PatientWeight (0010,1030)
- PatientSize (0010,1020)
- StudyDescription (0008,1030)
- PerformingPhysicianName (0008,1050)
- ReferringPhysicianName (0008,0090)
- RequestedProcedureDescription (0032,1060)

Le « DICOM Header » contient également de nombreuses informations fournies automatiquement pas la machine d'acquisition, telles que les dates et les paramètres d'acquisition. Les en-têtes DICOM des images, ainsi que le fichier Excel de suivi des expérimentations couvrent les différents descripteurs nécessaires à l'intégration de données automatiquement dans le système BMS-LM. Cette automatisation a fait l'objet de la première expérimentation d'intégration des données TEP-TDM dans le système BMS-LM. Nous l'avons appelée Mediso2PLM v1. Elle a été effectuée avant le début de la thèse par les consultants Fealinx. Lors de cette expérience d'intégration, plusieurs contraintes et freins ont été soulignés :

- Une incohérence entre les données DICOMs et les données Excel. Après investigation, les données Excels sont les plus fiables puisque le logiciel d'acquisition d'images ne permet pas de corriger les erreurs dans les fichiers DICOM aisément ce qui a poussé les chercheurs à les corriger au niveau du fichier Excel seulement.
- Une abondance de descripteurs DICOM qui, parfois, sont redondants, privés et incompréhensibles. Il est donc difficile de savoir si ces descripteurs sont pertinents pour un examen et de décider comment les ranger dans le MDD et la « Classification » BMS-LM.
- Une utilisation détournée du standard DICOM par le constructeur de l'appareil d'acquisition, à laquelle se sont adaptés les membres du laboratoire LRI. Cela justifie l'impossibilité de la réutilisation de l'import DICOM développé antérieurement (Allanic et al., 2017) (voir section I.4.3.2)
- Une rigidité du système BMS-LM qui par nécessité de traçabilité et de couverture de la provenance ne permet que l'intégration des données déjà prédéclarées dans le système via les informations concernant : le projet, les droits d'accès, les protocoles, les versions et configurations du logiciel utilisé, etc.
- Un rejet de la méthode d'intégration de données via le protocole DICOM dans le laboratoire LRI. La méthode courante était d'effectuer des copies via Matlab.

Lors d'une deuxième version et un essai d'intégration (Mediso2PLM v2), nous avons mis en œuvre la méthode « générique » d'intégration de données sans utiliser les attributs DICOM, mais en n'exploitant que le tableau Excel. Ceci non plus n'a pas fonctionné de point de vue utilisation puisque les en-têtes

DICOM ont été négligés complètement, ce qui a diminué la valeur informationnelle des données importées dans le système BMS-LM.

En somme, l'intégration des données TEP-TDM s'est avérée plus compliquée que prévu. Nous avons donc défini une feuille de route dans le cadre du projet DRIVE. Elle nous a permis de travailler par itération et de prioriser pour chaque itération la résolution d'un problème précis parmi la liste des problèmes à traiter.

VI.5.1.3. Mediso2PLM v3

Nous avons alors défini une version 3 avec des objectifs bien clairs et en étroite collaboration avec l'ingénieur en calcul scientifique du laboratoire LRI « DBA ». Le principal objectif de cette itération était « d'importer des données TEP-TDM à la demande dans le système BMS-LM à partir d'un poste de travail, par le chercheur lui-même en utilisant une méthode simple mettant en œuvre les outils informatiques du laboratoire ». Nous avons alors mis en place une architecture qui se base sur Matlab (voir Figure 133) et nous avons fourni un guide d'utilisation qui l'accompagne (voir Annexe D).

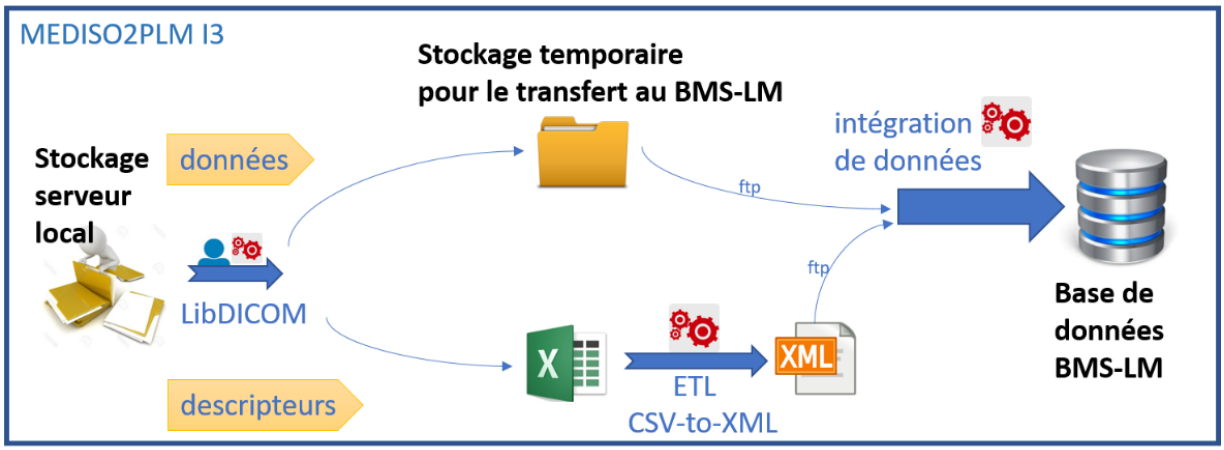


Figure 133 Les éléments et étapes de l'intégration de données TEP-TDM

Le premier élément de l'architecture est l'outil logiciel Matlab LibDICOM qui est une évolution et une adaptation du script « copieDicomTep2017 ». Cet outil permet de copier l'ensemble des fichiers DICOM dont le « Header » contient des valeurs d'annotation prédéterminées. Il a été enrichi d'interfaces graphiques (voir la capture de la Figure 134 ci-après et annexe D) pour la sélection des examens à importer. La sélection utilise des attributs DICOMs validés avec l'ingénieur TEP « TVI » comme suit : une première sélection à partir des champs « Private_0009_10DA », « Private_0009_10D5 », « Study Description », « Study Date », « Patient ID » permettant de réduire la recherche aux examens d'intérêt ; puis une seconde sélection permettant de sélectionner les acquisitions parmi ces examens, à l'aide des champs : « Series Description », « Series Date », « Modality », « Image Comments ». L'outil Matlab LibDICOM effectue ensuite la copie des examens sélectionnés à un emplacement temporaire pour les importer plus tard via FTP dans le système BMS-LM. LibDICOM génère également une liste d'attributs DICOM sous forme de tableau décrivant les données sélectionnées. Cette liste a été identifiée comme liste pertinente par l'ingénieur TEP. Ce tableau servira comme entrée à notre méthode « générique » d'intégration de données hétérogènes expliquée en IV.2.1.

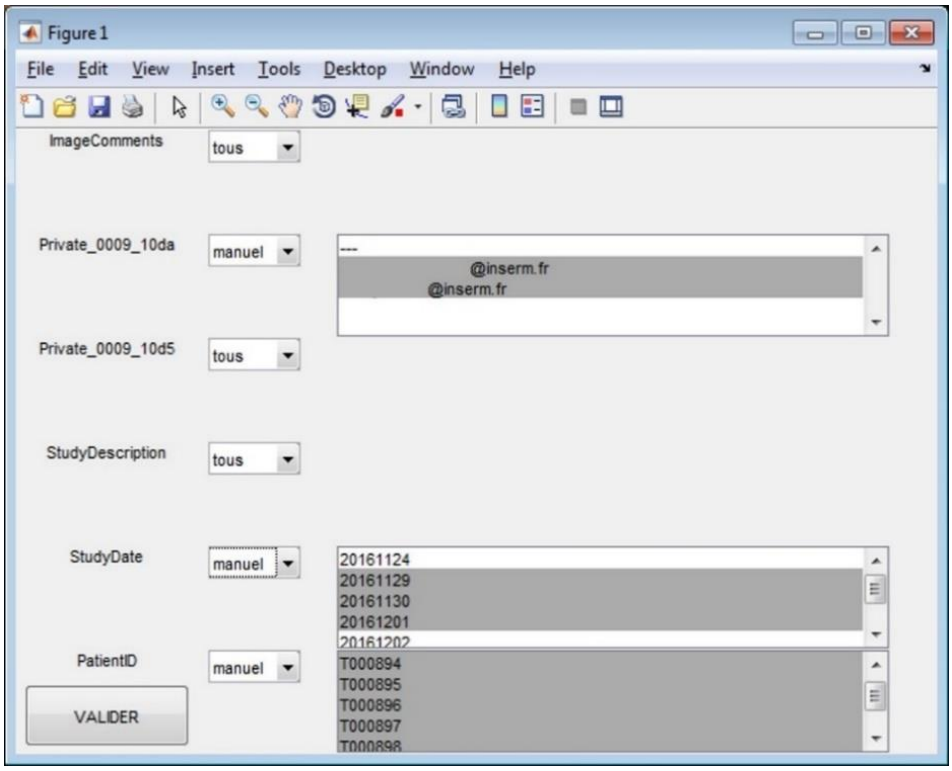


Figure 134 L'interface n°1 de l'outil LibDICOM pour la sélection des images d'intérêt à partir des headers DICOMs

Le tableau Excel, après vérification par l'utilisateur est transformé via l'ETL « csv-to-xml » à un fichier XML contenant les mêmes informations, mais enrichies par la structuration de BMS-LM. En effet, des objets de MDD et des classes de « Classification » ont été ajoutés au système BMS-LM pour le préparer à la réception des données TEP-TDM comme dans les captures d'écran Figure 135 et Figure 136 ci-après.

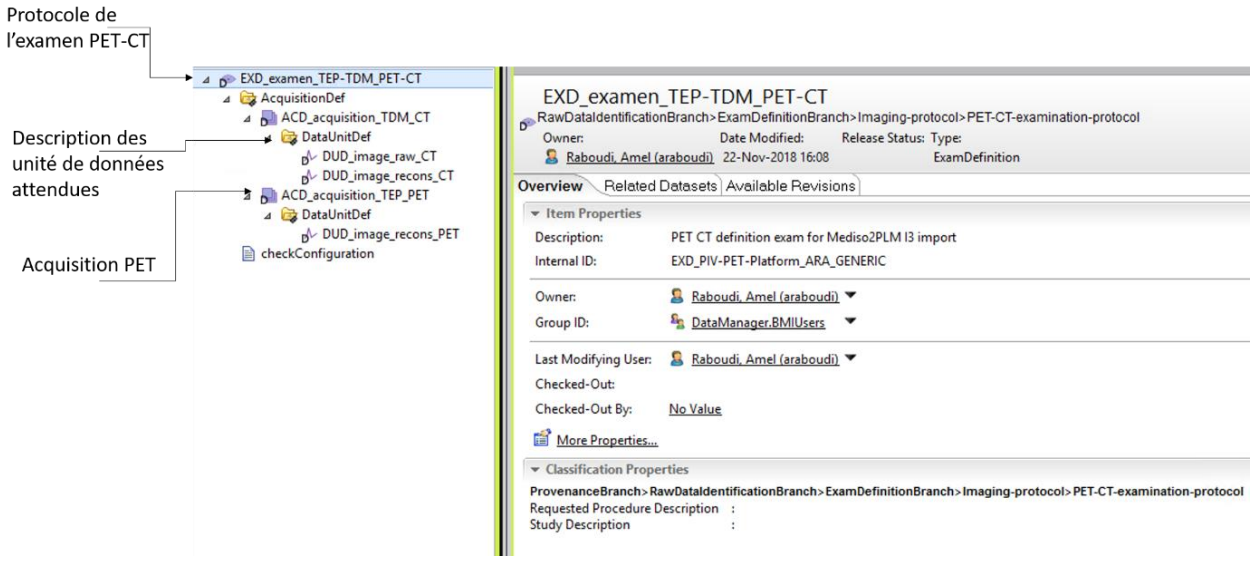


Figure 135 Les objets de provenance ajoutés au système BMS-LM

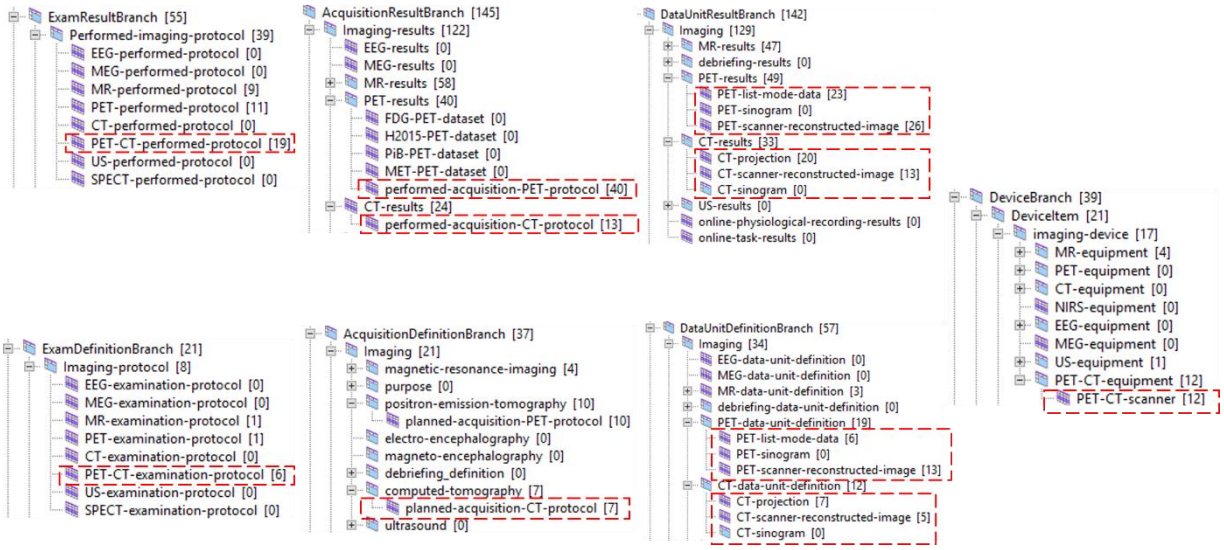


Figure 136 Les classes de la « Classification » ajoutées pour annoter les données TEP-TDM

Le lancement de l'import s'effectue via le transfert des dossiers et des fichiers XMLs via FTP. Le guide d'utilisation livré aux utilisateurs est présent en Annexe D. À la suite à la livraison de cette étape, deux utilisateurs clés ont été formés et ont utilisé l'outil. À la fin de l'utilisation, un questionnaire a été restitué par les deux utilisateurs : par « TVI » le 11/01/2019 et par « TYO » le 14/01/2019. La copie des réponses est présente en Annexe D et la liste des questions est reprise dans le tableau ci-après :

Tableau 24 Liste des questions d'évaluation de l'utilisation de « Mediso2PLM v3 »

1-Quelle est ton appréciation générale de la formation ?
A - Très satisfaisant B – Satisfaisant C – Moyen D – Insuffisant E - Échec
2-Quelle est ton appréciation concernant la clarté des instructions données lors de la formation ?
A - Très clair B – clair C – Moyennement clair D – clarté Insuffisante E – pas clair
3-Quelle est ton appréciation concernant la clarté du support de formation ?
A - Très clair B – clair C – Moyennement clair D – clarté Insuffisante E – pas clair
4-Quelle est ton appréciation concernant l'accessibilité de l'information de la formation ?
A - Très accessible B – Accessible C - Moyennement accessible D - Accessibilité insuffisante E – pas accessible
5-Quelle est ton appréciation générale concernant le temps de préparation des données pour l'envoi ?
A – très acceptable B – acceptable C – moyennement acceptable D – difficilement acceptable E – inacceptable
6-Quelle est ton appréciation générale concernant la maîtrise de toute la chaîne d'envoi de données vers le PLM ?
A – facilement maîtrisable B – maîtrisable C – moyennement maîtrisable D – difficilement maîtrisable E – non maîtrisable
7-Quelle est ton appréciation générale concernant la simplicité de l'utilisation de l'outil Mediso2PLM ?
A – Très simple B – simple C – moyennement simple D – simplicité insuffisante E – pas simple
8-Quelle est ton appréciation générale de l'outil Mediso2PLM ?
A - Très satisfaisant B – Satisfaisant C – Moyen D – Insuffisant E - Échec
9-Quelles sont les attentes satisfaites par l'outil ? .....
10-Quelles sont les attentes non satisfaites par l'outil ? .....
11-Attribuer une note de A à E pour ton niveau de familiarité avec :
– L'utilisation du client PLM SWOMed : _
– L'utilisation de la LibDICOM : _
– L'utilisation de Filezilla : _
– L'utilisation du CSV2XML : _
12-Quand l'envoi vers le PLM se déclenche-t-il ?
– Après le Matlab
– Après le CSV2XML
– Après le ftp
13-Quelle est la fréquence avec laquelle, tu penses utilisez l'outil Mediso2PLM ? et comment ? .....
14-Autres choses à nous dire ? .....

En référence aux réponses des utilisateurs, nous retenons que la formation s'est bien déroulée et était satisfaisante. Elle a été évaluée comme claire ainsi que son support. L'outil Mediso2PLM a été bien

reçu ainsi que le temps investi pour l'utiliser. La question 12 (voir Tableau 24) a été posée aux deux utilisateurs clés pour tester la maîtrise de la solution et ils ont bien répondu. Enfin, nous avons reçu une remarque concernant la vérification du fichier Excel, demandée systématiquement aux utilisateurs avant l'envoi de données via FTP (voir guide en Annexe D). Un des utilisateurs interviewés demande que le fichier Excel soit réorganisé pour mieux expliciter les informations importantes. Comme proposition finale (réponse question 14), cet utilisateur a demandé un complément de formation sur l'utilisation du système BMS-LM, ce qui confirme que cette expérimentation a augmenté l'intérêt que peut porter un utilisateur pour l'outil. La capture d'écran Figure 137 ci-après montre un des examens TEP-TDM importés à l'issue de ces formations.

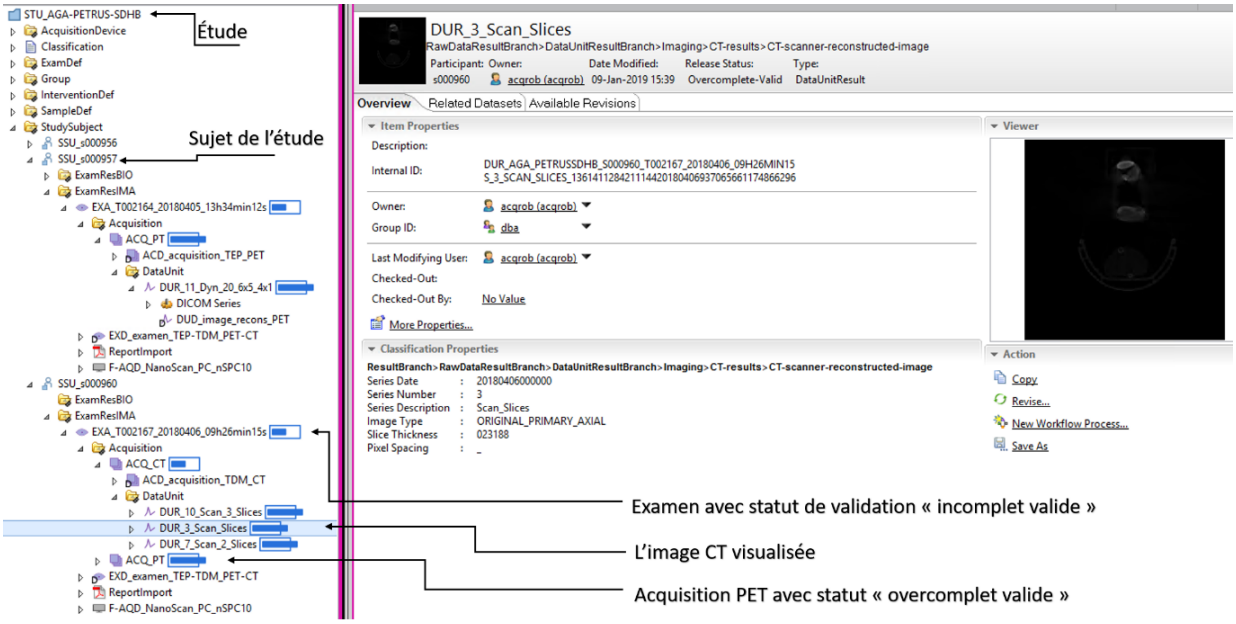


Figure 137 Données TEP-TDM importées via la méthode d'import « générique » dans sa version Mediso2PLM v3

VI.5.1.4. Mediso2PLM v4

La version 4 du sous-projet Mediso2PLM a fait l'objet du stage de Master2 de Ismael Bakayoko et constituait une version exploratoire afin de tester si le passage à une solution 100% web et libre (sans Matlab) améliorerait l'utilisabilité de l'outil de gestion de données ou non. Lors de cette version, nous avons aussi exploré la possibilité d'éditer les informations des fichiers DICOM avant l'envoi au système BMS-LM et de préconfigurer automatiquement le système BMS-LM. L'architecture qui a été mise en place lors de ce stage, ainsi que les différentes interfaces graphiques sont détaillées en Annexe E.

En bref, les travaux se sont déroulés en trois étapes comme suit:

1. La conversion de la bibliothèque LibDICOM Matlab (fournie à l'issue de Mediso2PLM version 3) en Python. Une formation de l'utilisateur clé « DBA » à son utilisation et maintenance a été effectuée.
2. La mise en place d'architecture web pour regrouper toutes les étapes de l'intégration « générique » dans un même outil web : du côté du laboratoire LRI, la bibliothèque LibDICOM a été encapsulée dans un serveur Django accessible via des messages URI GET depuis une interface web, et du côté du système BMS-LM, un client web personnalisé a été mis en place pour faciliter la consultation de données via l'API REST du système BMS-LM.
3. La mise en place d'interfaces graphiques adaptées et les plus ergonomiques possibles pour la sélection de données, leur correction, et leur envoi ainsi que leur consultation après envoi au système BMS-LM.

VI.5.2. INTÉGRATION « TOTALE » D'UN TRAITEMENT D'ANALYSE DE DONNÉES EN IMAGERIE

Comme pour les données d'histologie, nous avons identifié un calcul scientifique au laboratoire qui réutilise des données TEP-TDM dans le cadre du plan d'expérimentation « trait1 » défini lors de l'audit au laboratoire LRI. Il s'agit d'un calcul d'estimation de la Réponse Impulsionnelle (RI) tissulaire du muscle cardiaque à partir de mesures prises sur des images TEP-TDM de l'aorte et du myocarde. Il a été développé par l'ingénieur en calcul scientifique du laboratoire « DBA » et ne nécessite pas d'interaction utilisateur via des interfaces graphiques. Nous l'avons alors intégré en utilisant la méthode d'intégration « totale » des calculs scientifiques. Nous l'avons appelé « RI-Hermite ».

VI.5.3. Le calcul scientifique de la réponse impulsionnelle « RI-Hermite »

Le calcul « RI-Hermite » a été développé en Matlab. Il a pour objectif d'analyser la transition entre le signal vasculaire et le signal tissulaire de manière neutre, afin de comparer deux populations d'individus sans a priori physiologique. Il estime la Réponse Impulsionnelle (RI) tissulaire du cœur, à partir d'une mesure vasculaire nommée « AIF » et d'une mesure prise sur le muscle cardiaque, nommée « TAC ». Ces mesures ont été prises sur l'ensemble des individus de trois groupes de souris d'une étude au laboratoire LRI appelé « Cardiotox ». Elles sont calculées au niveau de l'artère myocardique à partir d'images TEP-TDM, après une injection du glucose radioactif FDG. Le traitement Matlab modélise la perfusion et le flux métabolique via un modèle de déconvolution appelé modèle FTPH (Free Time Point Hermit) et élaboré au laboratoire. En sortie, il fournit des paramètres du modèle d'Hermite permettant de « fitter » les données, afin de permettre leur analyse statistique par les chercheurs. Il fournit aussi les graphiques des réponses impulsionnelles (RI) associées, avec un code couleur différent pour chaque groupe d'appartenance des souris étudiées. La Figure 147 suivante schématise les différentes étapes et nœuds du workflow « RI-Hermite ».

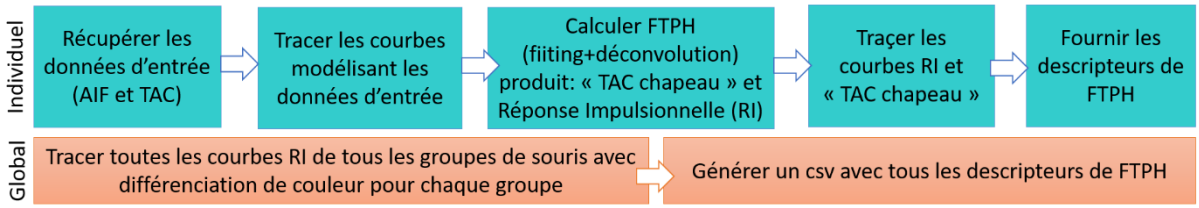


Figure 138 Étapes du calcul Matlab utilisant le modèle Hermite (FTPH) et se basant sur les mesures TAC et AIF

VI.5.4. La modélisation et la structuration du calcul scientifique « RI-Hermite »

Avant de pouvoir automatiser un calcul scientifique, il faut pouvoir le modéliser en unités de traitement indépendantes, avec des entrées et des sorties. Pour ceci, nous avons utilisé le standard de modélisation BPMN. Pour chaque unité de traitement, il faut en outre connaître ses paramètres de configuration.

La modélisation du workflow, avec explicitation des nœuds de la chaîne, ses entrées/sorties, ses paramètres est l'objet de la Figure 139 suivante. Cette représentation a permis, entre autres, à l'ingénieur en calcul scientifique de restructurer le code informatique, le rendre plus modulaire et plus lisible et donc plus facilement modifiable. Dans le diagramme BPMN présenté Figure 139, le panneau du haut décrit le workflow initial et le panneau du bas le workflow après rationalisation/structuration du code. Cette étape de rationalisation et de structuration est une étape primordiale pour faciliter l'intégration du workflow dans le système BMS-LM. Il est à noter que dans la nouvelle version du workflow, plusieurs factorisations ont été réalisées. Le traçage de courbes à l'aide de fenêtres pop-up pendant l'exécution, a été remplacé par un rapport PDF contenant les graphiques, pour permettre une exécution en mode serveur (voir Figure 139).

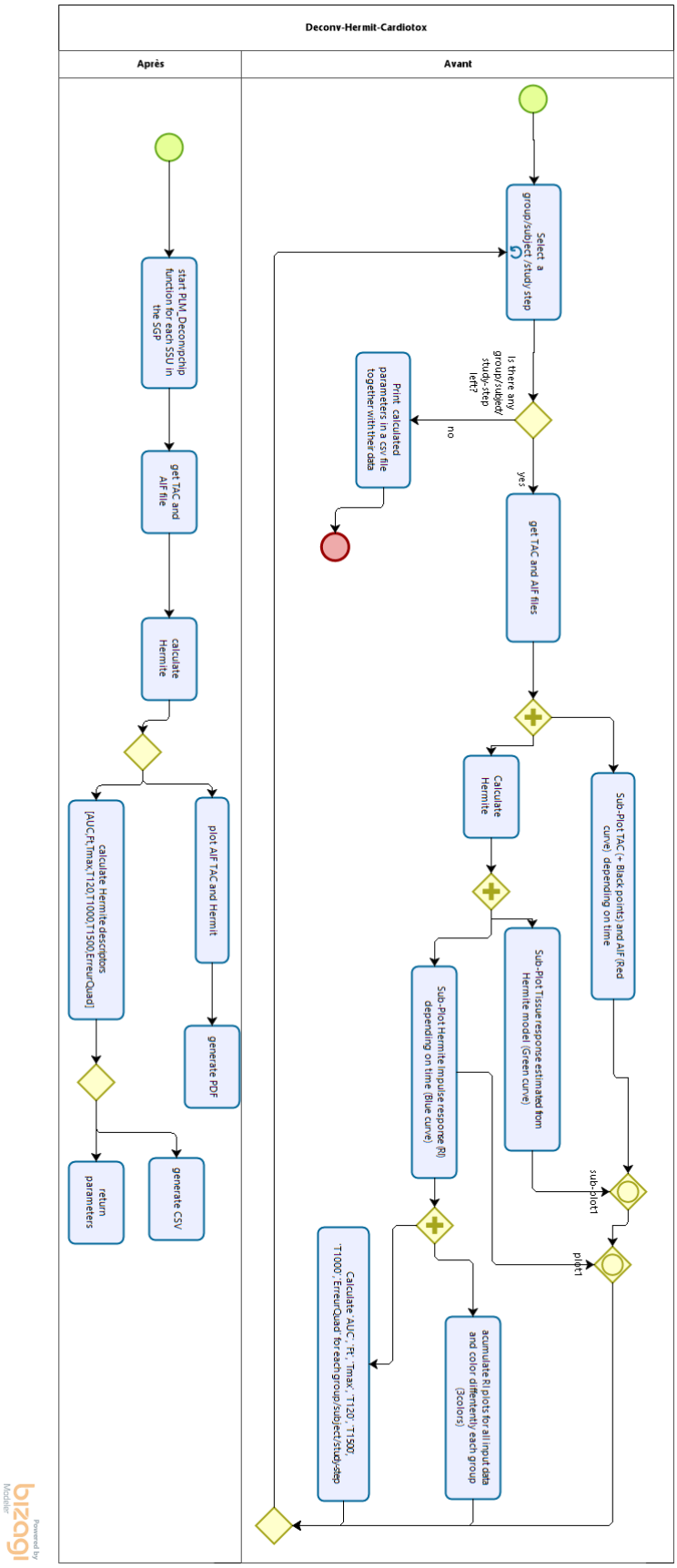


Figure 139 Modélisation BPMN du calcul Matlab « RI-Hermite » appliqué aux données TEP-TDM

VI.5.5. La préparation du système BMS-LM et de la machine de calcul

Pour pouvoir intégrer le calcul « RI-Hermite » dans le système BMS-LM, il faut préparer ce dernier à son exécution en définissant objets du MDD BMS-LM qui s'y rapportent. Les objets du MDD relatif à la machine d'exécution du traitement (ACD) et la chaîne de traitement (PCD, PUD) ont été ajoutés respectivement comme dans les captures d'écran Figure 140 et Figure 141 ci-après.

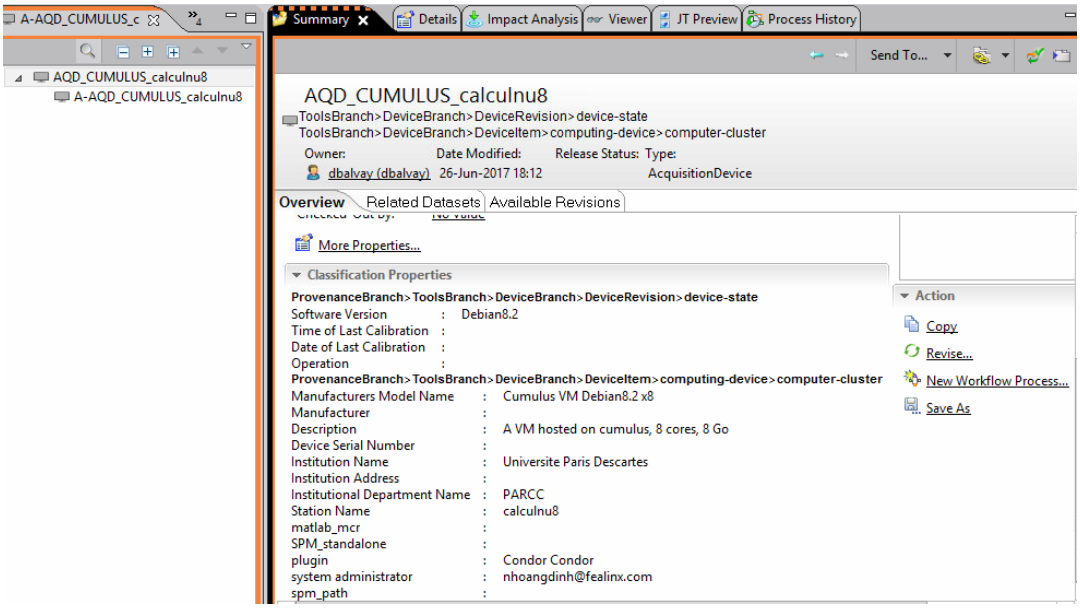


Figure 140 Ajout des informations sur la machine d'exécution des analyses dans le système BMS-LM

Les objets modélisant la chaîne de traitement dans la Figure 141 (version utilisateur, partie haut) ont été ajoutés par l'utilisateur clé « DBA » le 27 avril 2018 à la suite d'une formation. Pour l'exécution réelle du workflow, plus tard, nous avons utilisé une chaîne de traitement générée par les outils Fealinx « PCD_ima_deconv_hermit » (partie du bas Figure 141).

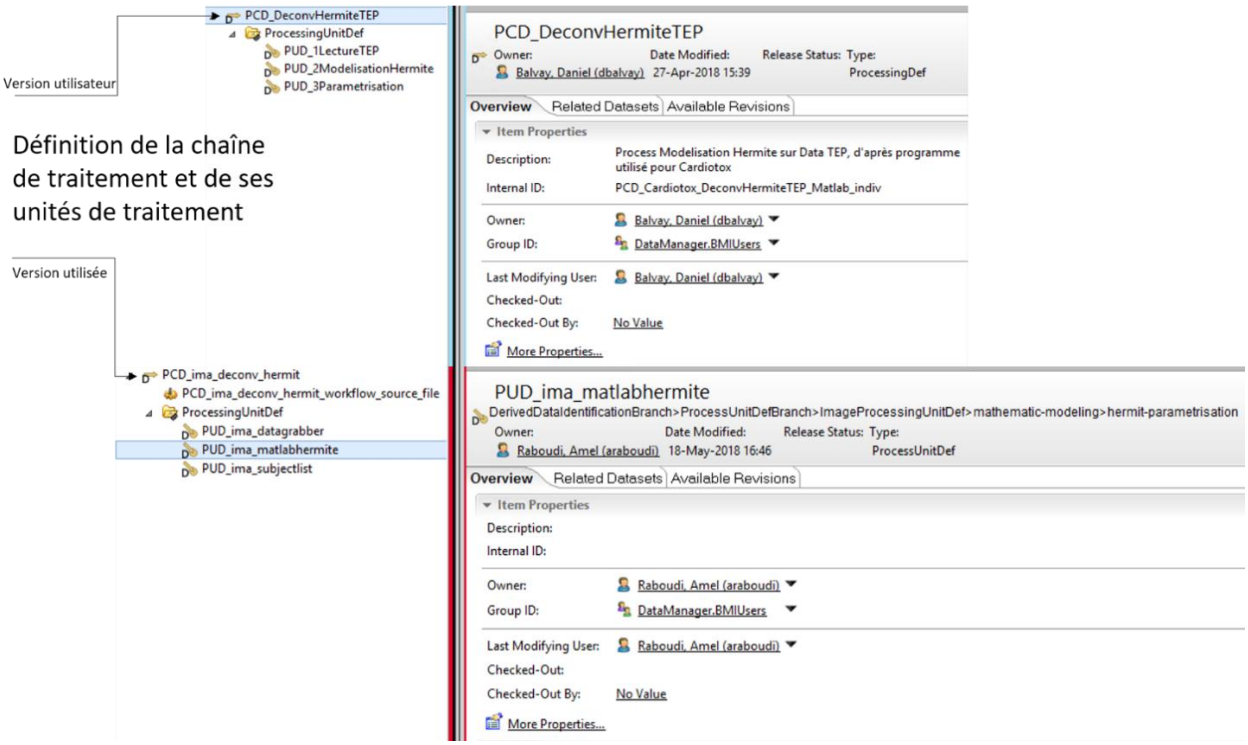


Figure 141 Les objets représentants le workflow Hermite (FTPH) dans le système BMS-LM

Ensuite, des données d'entrée (AIF et TAC) ont été préparées dans un groupe de sujets (SGP) pour tester l'intégration « totale » (voir Figure 142). Comme « AIF » et « TAC » ne sont pas des données brutes, mais des données issues de mesures réalisées par un expert sur des images PET, elles ont été modélisées via des classes de données dérivées (PUR, PCR, WFI, etc.). Les objets résultants dans le système BMS-LM sont présentés dans la capture d'écran Figure 142 ci-après. Celle-ci décrit les mesures caractérisant les paramètres AIF et TAC sur de souris issues de l'étude « Cardiotox ». Les objets de la Figure 142 sont des résultats d'exécution d'un autre workflow, « WFI_Heart_Melabolic_Flux_Indiv_Quantification », qui sont exploités en entrée du workflow « RI-Hermite ». La méthode d'intégration « générique » (décrite en §IV.2.1) a été appliquée pour importer ces données dérivées.

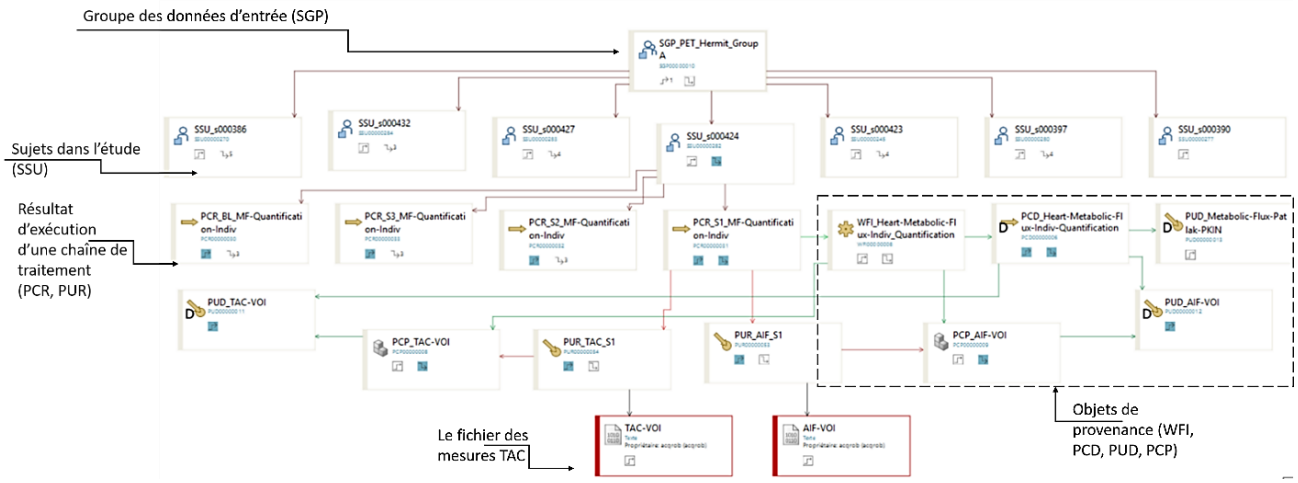


Figure 142 Données d'entrées au workflow « RI-Hermite » importées et tracées dans le système BMS-LM en préparation de son intégration « totale »

Ces données d'entrée sont référencées par le WFI correspondant au traitement Matlab « RI-Hermite » comme dans la Figure 143.

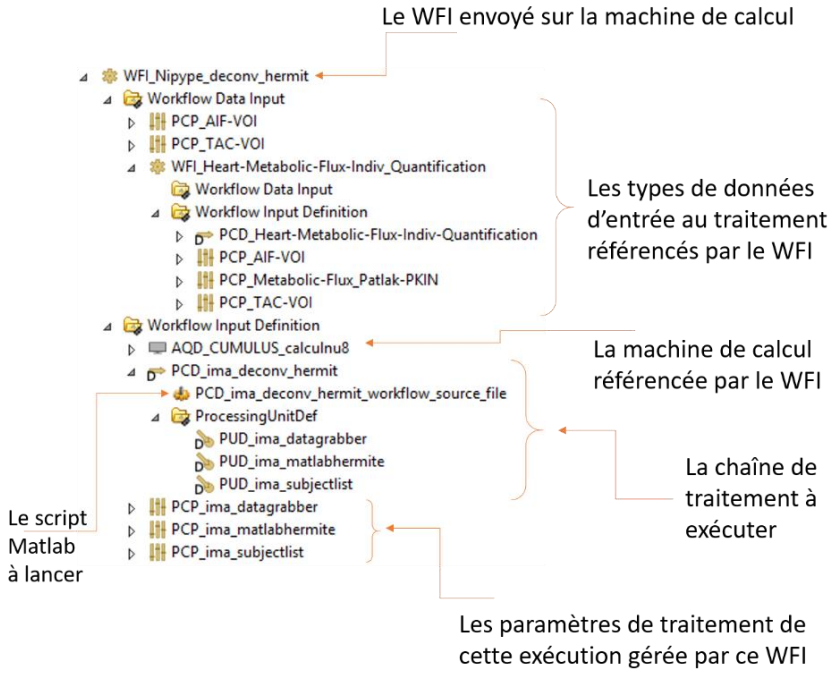


Figure 143 Le WFI de lancement du traitement Matlab « RI-Hermite » préconfiguré dans le système BS-LM

Ensuite, nous avons implémenté le nœud Nipype (voir section IV.2.2.1, Étapes de l'intégration « totale ») décrivant le workflow « RI-Hermite » pour préparer la machine de calcul distante (Tableau 25).

Tableau 25 Script de l'interface Nipype ajoutée pour configurer le workflow Matlab « RI-Hermite » sur la machine de calcul

```
"""
Created on 26 avr. 2018
@author: Amel Raboudi
"""
from nipype.pipeline.engine import Workflow, Node, MapNode
from nipype.interfaces.utility import IdentityInterface
from nipype.interfaces.io import DataGrabber
from pyplm.interfaces.PARCC import MatlabHermite
import os

# dummy args are a set of dummy parameters. The real kwargs come from BMSLM system.
# SUBJECT_LIST and BASE_DIR come from the dynamic parameters
dummy_args = {'SUBJECT_LIST': ['BMSLM::SUBJECT_LIST'],
              'SCRIPT' : ['BMSLM::SCRIPT'],
              'BASE_DIR' : os.path.normpath(os.path.expanduser('~'))}

def genWorkflow(**kwargs):
    """Launch matlab script for deconv hermit in Nipype."""
    mathw = Workflow('Deconv_Hermit')
    mathw. base_dir = kwargs['BASE_DIR']

    # get a list of subjects to iterate on
    subjectList = Node(IdentityInterface(fields=['subject_id'],
                       mandatory_inputs=True),
                       name="subjectList")
    subjectList.iterables = ('subject_id', kwargs['SUBJECT_LIST'])

    # file selection
    dataGrabber = Node(DataGrabber(infields=['subject_id'],
                                   outfields=['TACfile', 'AIFfile']),
                       name='dataGrabber')
    dataGrabber.inputs.base_directory = kwargs['BASE_DIR']
    dataGrabber.inputs.raise_on_empty = True
    dataGrabber.inputs.sort_filelist = True
    dataGrabber.inputs.template = '%s/%s/*_BL.%s'
    dataGrabber.inputs.template_args = {'TACfile': [['subject_id', 'TAC', 'tac']],
```

```
                                    'AIFfile': [['subject_id', 'AIF', 'crv']]}


    mathw.connect(subjectList, 'subject_id', dataGrabber, 'subject_id')
    matlabHermite = Node(MatlabHermite(),
                         name='matlabHermite')
    matlabHermite.inputs.script = kwargs['SCRIPT']
    mathw.connect(subjectList, "subject_id", matlabHermite, "souris")
    mathw.connect(dataGrabber, "TACfile", matlabHermite, "TACfile")
    mathw.connect(dataGrabber, "AIFfile", matlabHermite, "AIFfile")
    return mathw
```

Dans ce script, la fonction `genWorkflow` est appelé par Nipype pour l'exécution du script Matlab « RI-Hermite ». Les arguments en entrée sont `kwargs['SCRIPT']` qui référencent le script Matlab et `kwargs['BASE_DIR']` qui référence les dossiers des données d'entrée au script. Le script permet de sélectionner les données d'entrée de chaque SSU envoyé par le système BMS-LM via les unités de traitement (`subjectList` et `dataGrabber`). Les données sont transférées à l'unité suivante via la commande `mathw.connect(dataGrabber,"TACfile", matlabHermite, "TACfile")`, et ainsi de suite jusqu'à ce que toutes les unités de la chaîne de traitement soient déclarées ainsi que leurs entrées/sorties.

VI.5.6. Test avec des données réelles et retour utilisateur

Les tests d'intégrations ont été réalisés entre le 21 et le 25 mai 2018. Les objets présents dans la capture d'écran (Figure 144) ci-après, correspondent au résultat de l'intégration « totale » du calcul scientifique Matlab « RI-Hermite » dans le système BMS-LM. La Figure 144 montre la liste des PCRs (« Processing Results ») envoyées automatiquement depuis la machine « AQD_CUMULUS_calculnu8 » dans le système BMS-LM à la suite de l'exécution de la chaîne de traitement « RI-Hermite ».

Object	Type	Date Modified	Owner	Group ID	Relation	...	Relea
PCR_Deconv_Hermit	Processing	25-May-2018 15:30	Raboudi, Amel (araboudi)	DataManager.BMIUsers	query_results		Failed
PCR_Deconv_Hermit	Processing	25-May-2018 15:05	Raboudi, Amel (araboudi)	DataManager.BMIUsers	query_results		Failed
PCR_Deconv_Hermit	Processing	25-May-2018 15:05	Raboudi, Amel (araboudi)	DataManager.BMIUsers	query_results		Failed
PCR_Deconv_Hermit	Processing	25-May-2018 15:05	Raboudi, Amel (araboudi)	DataManager.BMIUsers	query_results		Failed
PCR_Deconv_Hermit	Processing	25-May-2018 15:05	Raboudi, Amel (araboudi)	DataManager.BMIUsers	query_results		Failed
PCR_Deconv_Hermit	Processing	22-May-2018 17:50	Raboudi, Amel (araboudi)	DataManager.BMIUsers	query_results		Failed
PCR_Deconv_Hermit	Processing	22-May-2018 17:49	Raboudi, Amel (araboudi)	DataManager.BMIUsers	query_results		Failed
PCR_Deconv_Hermit	Processing	22-May-2018 17:49	Raboudi, Amel (araboudi)	DataManager.BMIUsers	query_results		Failed
PCR_Deconv_Hermit	Processing	22-May-2018 17:49	Raboudi, Amel (araboudi)	DataManager.BMIUsers	query_results		Failed
PCR_Deconv_Hermit	Processing	21-May-2018 15:30	Raboudi, Amel (araboudi)	DataManager.BMIUsers	query_results		Failed
PCR_Deconv_Hermit	Processing	21-May-2018 15:30	Raboudi, Amel (araboudi)	DataManager.BMIUsers	query_results		Failed
PCR_Deconv_Hermit	Processing	21-May-2018 15:30	Raboudi, Amel (araboudi)	DataManager.BMIUsers	query_results		Failed
PCR_Deconv_Hermit	Processing	21-May-2018 12:29	Raboudi, Amel (araboudi)	DataManager.BMIUsers	query_results		Computed
PCR_Deconv_Hermit	Processing	21-May-2018 12:29	Raboudi, Amel (araboudi)	DataManager.BMIUsers	query_results		Computed
PCR_Deconv_Hermit	Processing	21-May-2018 12:29	Raboudi, Amel (araboudi)	DataManager.BMIUsers	query_results		Computed
PCR_Deconv_Hermit	Processing	21-May-2018 12:29	Raboudi, Amel (araboudi)	DataManager.BMIUsers	query_results		Computed
PCR_Deconv_Hermit	Processing	21-May-2018 11:59	Raboudi, Amel (araboudi)	DataManager.BMIUsers	query_results		Computed
PCR_Deconv_Hermit	Processing	21-May-2018 11:59	Raboudi, Amel (araboudi)	DataManager.BMIUsers	query_results		Computed
PCR_Deconv_Hermit	Processing	21-May-2018 11:59	Raboudi, Amel (araboudi)	DataManager.BMIUsers	query_results		Computed
PCR_Deconv_Hermit	Processing	21-May-2018 11:59	Raboudi, Amel (araboudi)	DataManager.BMIUsers	query_results		Computed
PCR_Deconv_Hermit	Processing	21-May-2018 11:59	Raboudi, Amel (araboudi)	DataManager.BMIUsers	query_results		Computed
PCR_Deconv_Hermit	Processing	21-May-2018 10:29	Raboudi, Amel (araboudi)	DataManager.BMIUsers	query_results		Computed
PCR_Deconv_Hermit	Processing	21-May-2018 10:29	Raboudi, Amel (araboudi)	DataManager.BMIUsers	query_results		Computed
PCR_Deconv_Hermit	Processing	21-May-2018 10:29	Raboudi, Amel (araboudi)	DataManager.BMIUsers	query_results		Computed
PCR_Deconv_Hermit	Processing	21-May-2018 10:29	Raboudi, Amel (araboudi)	DataManager.BMIUsers	query_results		Computed
PCR_Deconv_Hermit	Processing	18-May-2018 17:14	Raboudi, Amel (araboudi)	DataManager.BMIUsers	query_results		Computed
PCR_Deconv_Hermit	Processing	18-May-2018 17:14	Raboudi, Amel (araboudi)	DataManager.BMIUsers	query_results		Computed
PCR_Deconv_Hermit	Processing	18-May-2018 17:14	Raboudi, Amel (araboudi)	DataManager.BMIUsers	query_results		Computed
PCR_Deconv_Hermit	Processing	18-May-2018 17:14	Raboudi, Amel (araboudi)	DataManager.BMIUsers	query_results		Computed
PCR_Deconv_Hermit	Processing	18-May-2018 17:14	Raboudi, Amel (araboudi)	DataManager.BMIUsers	query_results		Computed

Les PCRs résultats de processing générés lors du de l'exécution de la chaîne de traitement « RI-Hermit » dans un workflow Nipype

Figure 144 Liste des PCRs générés lors du test d'intégration du workflow « RI-Hermite » dans le système BMS-LM

Deux de ces PCRs : un qui a été bien exécuté sur la machine de calcul (statut « *Computed* ») et un qui s'est arrêté à cause d'une erreur dans l'exécution (statut « *Failed* ») sont présents dans la capture d'écran de la Figure 145. Ces PCRs référencent le WFI utilisé pour les générer. Ils référencent aussi les fichiers utilisés pour leurs exécutions (« AIF-VOI », « TAC-VOI » sur la figure). Le rapport d'erreur du nœud *Failed* (montré en bas à gauche de la capture d'écran) est envoyé par la machine de calcul pour assurer la traçabilité et pour permettre d'étudier le problème.

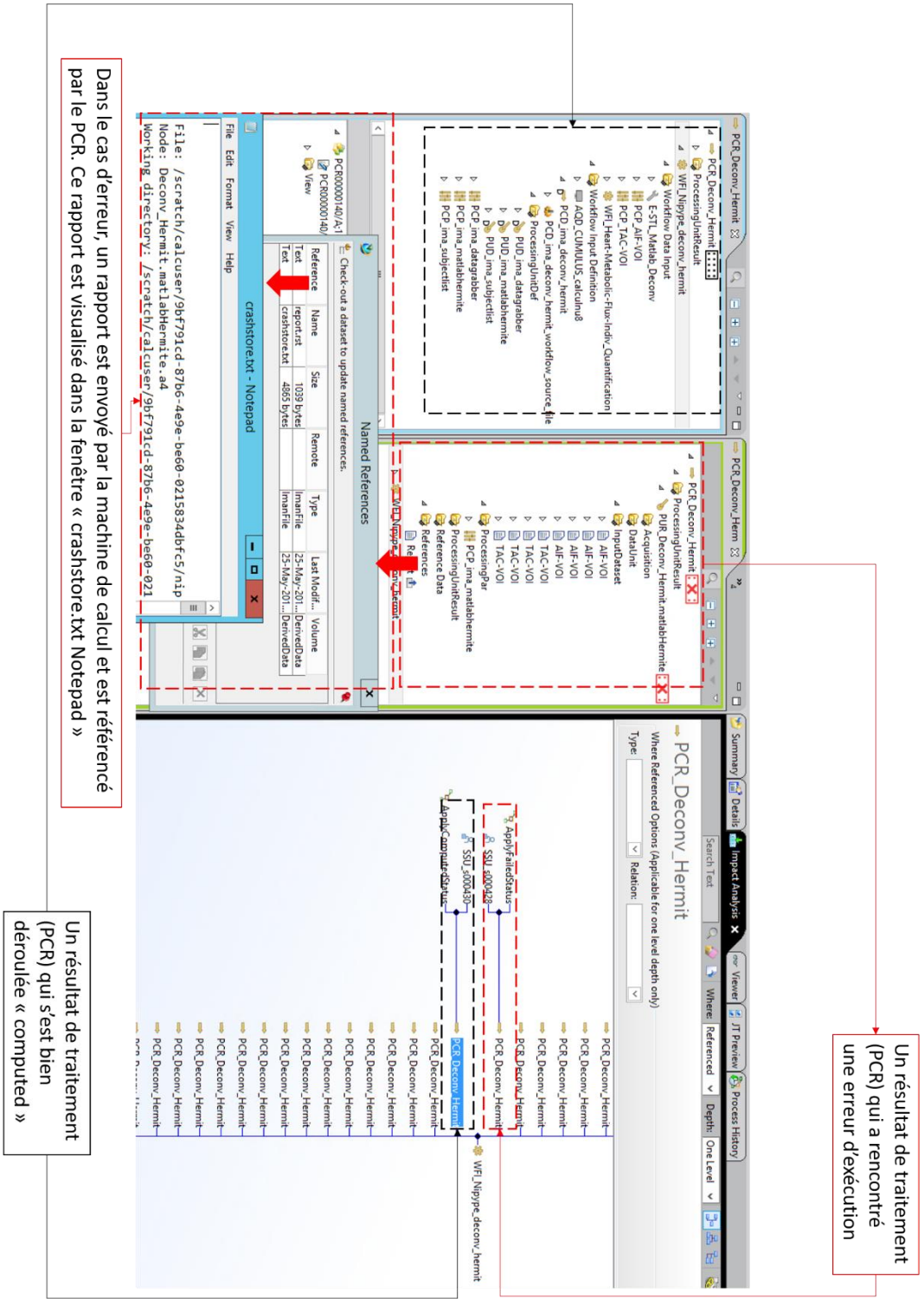


Figure 145 Deux PCRs (un réussi et un avec rapport d'erreur) résultant de l'intégration du workflow Matlab « RI-Hermite » dans le système BMS-LM

L'intégration « totale » du workflow « RI-Hermite » a permis de satisfaire les besoins en B1-archivage, B2-import des données dérivées, B6-Analyse, B8-Traçabilité, B10- automatisation, B11 standardisation et rationalisation des chaînes de traitement, B16- vérification avec les rapports d'erreur et les statuts (Failed et Computed), B19- suivi de l'activité d'exécution de chaînes de traitement.

Pour résumer quant au plan d'expérimentation « trait1 », nous avons testé l'intégration « totale » avec le workflow « RI-Hermite » et l'intégration « partielle » avec le workflow de quantification de données d'histologie (§VI.4.2), et nous pouvons conclure que l'intégration est réalisable dans les deux cas de figure avec un investissement de temps de la part de l'ingénieur de calcul scientifique. Il faut alors bien choisir le workflow que l'on souhaite intégrer afin de rentabiliser le temps investi. Dans le cas du workflow « RI-Hermite », l'intérêt à moyen terme n'était pas clair. Le workflow a été créé essentiellement à titre exploratoire pour les données TEP. Il n'a pas été utilisé depuis. Le choix de ce workflow n'était finalement pas pertinent pour le laboratoire et nous avons donc abandonné la suite de son intégration. Inversement, dans le cas du workflow d'histologie (§VI.4.2), l'intérêt était fort et clair pour le laboratoire. L'outil d'apprentissage en histologie qui a été mieux structuré lors de l'intégration avec le système BMS-LM est maintenant utilisé comme un outil de plateforme de recherche, au-delà du laboratoire, et va être publié prochainement en accès libre.

CONCLUSION DU CHAPITRE VI

Dans ce chapitre, nous avons présenté deux cas d'application de l'approche de gestion de cycle de vie des études biomédicales BMS-LM que nous avons introduit dans cette thèse. Un troisième cas d'application, pour les données protéomique est présenté en Annexe E. Nous avons appliqué nos méthodes d'intégration proposées aux données et calculs scientifiques en imagerie, en histologie et en protéomique et nous avons pu valider la pertinence de nos implémentations via des questionnaires utilisateurs. À chaque nouvelle intégration, nous avons veillé à exploiter les quatre leviers identifiés dans la littérature via des adaptations de la méthode de collecte de descripteurs de données ou la méthode d'intégration du calcul scientifique au quotidien de l'utilisateur clé. Cette approche nous a beaucoup aidés pour améliorer la standardisation et la traçabilité de la gestion des données au laboratoire LRI.

Pour identifier les scénarios d'utilisation pertinents du système BMS-LM, nous avons effectué un audit adapté à un laboratoire de recherche. Il nous a permis d'identifier les processus pilotes, qui une fois implémentés, auront une forte valeur ajoutée pour le laboratoire LRI. Faute de temps, nous n'avons pas pu les implémenter tous. Ils constituent la carte de route de la suite de la mise en œuvre du système BMS-LM pour le laboratoire LRI.

Chapitre VII. Discussion, Perspectives et Conclusion

Dans ce chapitre, nous concluons notre manuscrit avec une synthèse rappelant nos objectifs et les résultats obtenus au cours de nos travaux de recherche effectués, et avec une discussion de ces différents résultats en passant en revue leurs apports, les points d'amélioration et les perspectives scientifiques envisageables.

VII.1. SYNTHÈSE

Nous avons abordé dans cette thèse la problématique de la « gestion de données hétérogènes de recherche biomédicale tout au long d'une étude de recherche en vue de partage et de réutilisation ». Notre thèse est inscrite dans les initiatives nationales et internationales pour la science et les données ouvertes, notamment les principes FAIR et le DMP qui façonnent la gestion de données de recherche (RDM). Nous sommes partis du constat que pour pouvoir réutiliser des données scientifiques provenant de dépôts publics ou des archives d'une équipe de recherche, il faut (1) leur accorder un certain degré de confiance et (2) les comprendre. Nous avons donc suivi un raisonnement hypothético-déductif stipulant les deux hypothèses suivantes :

1. L'application du paradigme de la gestion de cycle de vie en industrie (PLM) à la recherche biomédicale permet de renforcer la confiance dans les données lors de leur partage et leur réutilisation ultérieurs. Cela s'inscrit en continuité des travaux de thèse de (Allanic, 2015).
2. La mise en œuvre d'une interopérabilité sémantique entre KOS des données hétérogènes permet de renforcer leur compréhension facilitant leur partage et leur réutilisation.

Pour tester nos hypothèses de recherches, nous avons réalisé nos expérimentations sur des données en recherche préclinique issues du laboratoire LRI. La recherche préclinique a été définie dans cette thèse comme « la recherche biomédicale effectuée sur le petit animal avec ou sans débouchés thérapeutiques chez l'être humain ». Elle est une recherche multimodale qui se pratique in vivo et in vitro et qui exploite différentes techniques d'imagerie, de bio-imagerie et d'analyse protéomique. Elle est représentative du large domaine de la recherche biomédicale avec des données d'hétérogénéité « multiple » et des outils d'analyse spécifiques de données expliquées en chapitre I.

Dans le chapitre II, nous avons donné un état de l'art des besoins et des systèmes en gestion de données dans le domaine de la recherche biomédicale en général, et préclinique en particulier. Une liste de 19 besoins a été définie (voir Tableau 7) : archivage, import, export, requête, partage, analyse, réutilisation, traçabilité, simplicité, automatisation, standardisation, ergonomie, efficacité, évolutivité, flexibilité, vérification, reporting, sécurité, suivi. Quant aux systèmes, nous avons identifié des systèmes en silos, en fonction de la phase de la recherche translationnelle (préclinique ou clinique), et des types de données : images, analyses biologiques et rapports textuels. Nous avons montré qu'aucun des systèmes identifiés ne permet de répondre à tous les besoins de la communauté biomédicale en gestion de données de recherche. Afin de formuler une proposition de gestion de données hétérogènes, nous avons identifié 4 leviers à utiliser pour réussir la mise en œuvre d'un système de gestion de données dans le domaine biomédical, repris dans le Tableau 26 ci-après.

Le système BMS-LM (BioMedical Study – Lifecycle Management), détaillé au chapitre IV, a été proposé afin d'améliorer la confiance dans les données de recherche lors d'une réutilisation ultérieure. Ce système est le fruit de l'application des paradigmes du PLM au domaine de la recherche biomédicale. Son ancêtre est la plateforme BIOMIST où le PLM a été appliqué précédemment à des données en neuroimagerie (Allanic, 2017). Le BMS-LM est défini dans cette thèse comme « une démarche intégrée

de gestion de données, informations, connaissances et de leurs provenances avec traçabilité de toutes les étapes d'une étude de recherche biomédicale (i.e. cycle de vie) : (1) la spécification, (2) l'acquisition de données brutes, (3) la production de données dérivées, et (4) la valorisation scientifique des résultats qui serviront à (1) la spécification d'autres études. »

Tableau 26 Liste des leviers pour la gestion de données de recherche

	LEVIER	DESCRIPTION
L1	collaboration bio-info	La collaboration entre biologistes, imageurs et autres disciplines de recherche et les ingénieurs informaticiens, doit être étroite dès le début de la conception de la solution logicielle de gestion de données
L2	changement	La prise en compte des changements au cours d'une étude est primordiale dès le début de la mise en place de la solution logicielle
L3	adaptation	Même avec la prise en compte des changements lors de la conception d'une solution logicielle, son adaptation à la réalité terrain est nécessaire. En recherche, il y a beaucoup de « bricolage » puisque les standards des domaines n'évoluent pas rapidement. Le système doit pouvoir s'adapter aux nouveautés tout en respectant les standards
L4	mirroring	Les saisies d'informations dans le système doivent suivre le workflow de recherche (protocole d'acquisition ou d'analyse) pour permettre une meilleure utilisation du système
L5	intérêt	Toute utilisation du système par un chercheur doit être liée à un intérêt majeur pour son étude de recherche

Le système BMS-LM propose un ensemble de 17 modules de gestion de données répartis tout au long du cycle de vie d'une étude de recherche biomédicale (voir Figure 70). Parmi eux des modules de base des plateformes PLM sont utilisés : les modules de « Traçabilité automatique des différentes utilisations » et de « Gestion documentaire (GED) et Gestion de Données Techniques (SGDT) ». Une reprise des modules de la plateforme BIOMIST a été effectuée, notamment les modules de « Requêtes personnalisées » ou de « Communication avec les clusters de calcul ». Les modules qui ont été construits ou qui ont fait l'objet d'évolution dans cette thèse sont : le module d'« intégration de données hétérogènes », le module d'« intégration de calcul scientifique », le module de « modélisation de données via MDD et « Classification » », et le module d' « intégration de terminologies publiées du domaine ». Ces quatre modules ajoutés font l'objet des chapitres 4 et 5 et 6. Au chapitre IV et en Annexe B, nous avons expliqué comment l'ensemble des 17 modules répond aux 19 besoins identifiés de la communauté. Cependant, le niveau de réponse varie selon le besoin. Par exemple, les besoins en ergonomie et en simplicité d'utilisation ne sont satisfaits que lors d'utilisation de « Clients personnalisés », tandis que le besoin en traçabilité est satisfait pour tout type d'utilisation du système.

Tout d'abord, nous avons fait évoluer le MDD et la « Classification » du système BMS-LM pour l'adapter au contexte de la recherche biomédicale. Trois concepts ont été ajoutés : l'Agent, l'Échantillon et l'Intervention. Les objets du MDD ont été répartis en deux groupes : de Provenance et de Résultats et 6 objets ont été ajoutés (SAR, SAD, AGT, AGD, ITR, ITD). Nous avons ensuite proposé la méthode « générique » d'intégration de données hétérogènes dans le système BMS-LM. Celle-ci consiste à contextualiser et annoter les données via la collecte de descripteurs de Provenance en réponse aux questions QQOQCCP, à les structurer en utilisant les objets de MDD et les classes et attributs de « Classification ». Pour la mise en œuvre, une procédure de transformation de données a été mise en place et adaptée à chaque type de données. Nous avons appliqué cette méthode à l'import de données d'histologie et de TEP-TDM en recherche préclinique dans le chapitre VI et nous avons veillé à son adaptabilité aux habitudes des utilisateurs clés.

Concernant l'intégration des calculs scientifiques, nous avons distingué deux méthodes : l'une « totale » et l'autre « partielle ». La méthode « totale » consiste à lancer le calcul scientifique depuis le système BMS-LM encapsulé dans un workflow Nipype avec une traçabilité de l'exécution. Elle a été proposée dans le cadre du projet BIOMIST et nous l'avons appliquée à un calcul scientifique de la réponse impulsionnelle du cœur dans des images TEP-TDM « RI-Hermite ». La méthode d'intégration « partielle » d'un calcul scientifique vise à assurer la traçabilité dans le système BMS-LM des données scientifiques fournies en entrées/sorties des applications logicielles locales. Les échanges entre l'application locale et le système BMS-LM sont transparents pour l'utilisateur qui ne doit pas s'adapter au système BMS-LM. L'inverse est vrai, ce dernier s'adapte aux pratiques locales. En même temps, l'application logicielle locale doit être ajustée pour pouvoir communiquer de manière cohérente avec le BMS-LM. L'intégration « partielle » a été conçue en exploitant les leviers L5-mirroring et L3-adaptation. Elle a été mise en œuvre pour une application logicielle Matlab pour la quantification d'images histologiques. Elle a été testée avec un utilisateur régulier et a donné lieu à une acceptabilité d'utilisation satisfaisante. Dans cette configuration, l'application logicielle s'intercale entre l'utilisateur et le système BMS-LM afin de combiner les avantages d'un environnement familier et d'une traçabilité renforcée.

Le système BMS-LM a été proposé afin de traiter la question de confiance dans les données de recherche lors d'une réutilisation ultérieure. La gestion de cycle de vie, la traçabilité et la gestion de la Provenance des données via le MDD BMS-LM apportent des éléments de réponse à cette question (voir Figure 165 Données en protéomique tracées depuis la spécification de l'étude jusqu'à la publication dans le système BMS-LM, Annexe B). Les liens entre objets (composition, contribution, qualification, ingestion, production, réalisation, transformation et provenance), et les annotations correspondantes permettent de conserver l'information telle qu'elle a été produite par l'utilisateur au cours de son étude et de fournir son contexte. La réutilisation ultérieure de données pourra être effectuée avec une meilleure confiance.

Par ailleurs, le dernier module ajouté au système BMS-LM « intégration de terminologies publiées du domaine » vise à traiter la question de la compréhension de données, par une personne externe, lors d'une réutilisation par un tiers ou plusieurs années après la clôture d'une étude de recherche. Cette compréhension est freinée par les termes vernaculaires locaux non conventionnels utilisés pour l'annotation des données. Pour formaliser notre proposition, nous avons effectué un état de l'art en gestion, organisation, et ingénierie des connaissances présenté en chapitre III.

Nous avons noté que le couple « MDD, Classification » du système BMS-LM représente son KOS (Systèmes d'Organisation de Connaissances). Nous avons retenu également que l'utilisation de la représentation ontologique des KOS permet une meilleure compréhension des données, et que la méthode de construction d'ontologies « haut/noyau/domaine » facilite l'interopérabilité sémantique de l'ontologie. Nous avons proposé dans le chapitre V, une méthode de construction d'ontologies interopérables en 4 niveaux « haut/noyau/domaine/local » et nous l'avons appliquée au KOS du système BMS-LM.

L'ontologie multi-niveaux BMS-LM a été ainsi construite en cherchant à avoir un maximum de réutilisation des KOS publiés pour une meilleure interopérabilité sémantique. Nous avons choisi pour le niveau haut, l'ontologie BFO afin d'avoir un terrain d'interopérabilité commun avec les 220 ontologies de l'OBO Foundry. Le domaine de la gestion de cycle de vie étant un domaine large, l'ontologie correspondante au MDD BMS-LM se situe au niveau noyau. Le niveau domaine a été construit à partir de la collecte d'une « liste initiale de termes » où chaque terme a été repris et remplacé par un concept d'un KOS publié. Pour le dernier niveau, nous avons étendu la méthode « haut/noyau/domaine » à un niveau local pour lequel les termes locaux (les termes restants du lot initial des termes) ont été collectés. Ce niveau permet de renforcer la standardisation de données à l'échelle locale de l'étude ou de l'équipe de recherche.

Nous avons appliqué notre méthode à l'ontologie du projet DRIVE (laboratoire LRI) et nous avons fourni des exemples de chaque élément décrit précédemment (voir§V.2.6) : le lot initial des termes, le niveau local, le niveau domaine, le niveau noyau et le niveau haut. La construction multi-niveaux permet d'améliorer la compréhension du lot initial des termes de proche en proche jusqu'au niveau haut. Nous avons effectué un test avec trois utilisateurs au laboratoire LRI afin d'évaluer la pertinence de l'ontologie par rapport au MDD BMS-LM et deux d'entre eux ont progressé en matière de compréhension des concepts du niveau noyau (les documents utilisés pour le test sont présents en Annexe C).

VII.2. **DISCUSSION ET PERSPECTIVES**

Nous avons proposé de faire évoluer le KOS du système BMS-LM (MDD+Classification) en une ontologie multi-niveaux en plusieurs étapes. Dans les paragraphes suivants, nous discutons chaque étape et ses perspectives d'évolution.

Pour le choix de l'ontologie BFO au niveau haut, nous avons identifié, dans notre thèse, des critères de « popularité » afin de maximiser l'interopérabilité sémantique de l'ontologie BMS-LM. Néanmoins, d'autres critères peuvent aussi être considérés surtout que le choix d'une ontologie de haut niveau est toujours accompagné d'une adhésion à son engagement ontologique. Il faut alors bien étudier la « vision du monde » qu'une ontologie de niveau haut véhicule. L'ontologie BFO est connue par sa vision basée sur le réalisme qui ne fait pas de distinction entre le « particulier » et l'« universel », tandis que, l'ontologie DOLCE est connue pour sa vision cognitive qui distingue entre eux. Le concept « abstract » de DOLCE n'existe pas dans BFO. Cette dernière présume que toute réflexion humaine est un terrain de subjectivité et ne doit pas être modélisée au niveau haut (Mascardi et al., 2007) (Seyed, 2009). Dans notre thèse, nous avons adhéré à la vision basée sur le réalisme de BFO. Néanmoins, pour atteindre plus d'interopérabilité sémantique, l'alignement de l'ontologie noyau BMS-LM avec d'autres ontologies de haut niveau semble incontournable : notamment, l'ontologie DOLCE, l'ontologie SUMO et l'ontologie GFO. Il faut noter que des efforts de rapprochement entre les deux ontologies BFO et DOLCE sont en train d'être déployés (Guarino, 2017).

Le niveau noyau n'est pas toujours présent, dans la plupart des ontologies publiées réutilisant une ontologie de niveau haut. Des concepts noyaux peuvent alors être retrouvés dans des ontologies de domaine. Par exemple, l'ontologie de domaine OBI, qui se base sur l'ontologie BFO, définit des concepts noyau pour les investigations biomédicales. Des recouvrements entre les concepts de l'ontologie BMS-LM et OBI existent notamment pour les concepts `OBI:protocol` et `OBI:planned process`. Néanmoins, la portée de chaque ontologie est différente. Tout d'abord, l'ontologie OBI ne couvre pas la provenance et tout le cycle de vie de l'étude. Deuxièmement, il ne s'agit pas d'une ontologie noyau, même si elle contient des concepts génériques. Une discussion avec la communauté BFO et OBI pourrait aider à éliminer les recouvrements identifiés dans les prochaines versions des deux ontologies. Soulignons que la soumission de l'ontologie noyau BMS-LM au consortium OBO Foundry et sa publication dans BioPortal sont prévues dans un futur proche.

Au niveau domaine, nous avons effectué une agrégation de KOS publiés dans différents domaines spécifiques (ici la recherche préclinique) et au niveau local, nous avons regroupé les termes locaux du contexte d'application (ici le laboratoire LRI). Nous avons testé notre proposition en montrant les trois niveaux : local, domaine, et noyau, des exemples d'applications §V.2.6 à trois utilisateurs clés. Deux ont bien passé le test et étaient préalablement avertis quant à l'intérêt de la modélisation de données. Une troisième personne, malgré sa légère évolution après avoir consulté les exemples, a exprimé le fait que le test était trop technique et hors de son domaine de compétence (ingénierie d'histologie). En effet, le domaine des ontologies est un domaine généralement étranger au monde de la recherche préclinique et à la réalité des laboratoires. Cela peut faire peur, générer une incompréhension et induire une forme de non-adhésion voire de rejet, d'où la pertinence de l'utilisation d'une version allégée de l'ontologie via la « Classification ».

Alors que la « Classification » est utilisée pour l'annotation de données au sein d'un groupe de chercheurs, l'ontologie multi-niveaux est utilisée en arrière-plan comme un chef d'orchestre qui effectue les correspondances entre les annotations utilisateurs et les KOS publiés. Dans cette thèse nous avons étudié les différentes « mises en correspondance (*matching*) » possibles de l'ontologie multi-niveaux BMS-LM afin de préparer la mise en place d'« alignement automatique (*mapping)* » entre KOS. Le domaine de « *matching* » et « *mapping* » entre ontologies est un domaine de recherche actif (Shvaiko & Euzenat, 2013) (Diallo, 2014) (Groß et al., 2016) et la mise en place d'un « *mapping* » efficace demande d'aller encore plus loin dans nos recherches. Les correspondances de l'ontologie multi-niveaux étant hétérogènes, évolutives, variables selon le contexte d'application, leur alignement automatique est un problème de recherche difficile.

Nous proposons l'alignement de manière assistée (par un expert en ontologies) comme première perspective de recherche. Il permettrait de « traduire » les données manipulées via le système BMS-LM et annotées via la « Classification », en des données annotées utilisant les standards de domaine et prêtes à être publiées en « open access ». Typiquement, nous pourrions avoir le scénario suivant : des objets du MDD sont sélectionnés et envoyés à un outil d'export dédié pour l'« open access ». Les termes locaux sont remplacés par des termes MESH, HL7, LOINC, NCIT, SNOMED-CT etc. et sont exportés, après retrait d'éventuelles informations confidentielles, sous format d'un tableau de descripteurs pour accompagner les fichiers des données. Le tout est envoyé sur un repository public tel que zenodo[84.]

Lors de ce scénario, des correspondances sont mises en œuvre entre la « Classification », le MDD, l'ontologie multi-niveaux BMS-LM, et les KOS publiés de domaine. Nous avons présenté comment nous envisageons d'effectuer ces correspondances, mais nous n'avons pas encore résolu le problème de leur maintenance dans le contexte actuel de l'évolution rapide en recherche biomédicale. L'évolution continue de ces alignements est inévitable, principalement entre les niveaux domaine et local de l'ontologie et la « Classification ». Le domaine de maintenance d'ontologie et des alignements entre elles est très actif en recherche. Nous nous inspirons des travaux de (Oliveira & Pesquita, 2018) (Zablith et al., 2015) (Dos Reis et al., 2015) pour continuer dans cette voie de recherche.

Le système et l'ontologie BMS-LM seront mis en œuvre dans un futur proche dans les projets PsyCare (RHU), Shiva (RHU) et PACIFIC (PSPC) dont l'entreprise Fealinx est partenaire. Les deux niveaux haut et noyau de l'ontologie BMS-LM seront réutilisés pour chaque projet. Les niveaux domaine (pour les domaines concernés) et local (pour les laboratoires partenaires) seront construits selon les principes expliqués dans cette thèse. Cette séparation des niveaux domaine et local de l'ontologie BMS-LM pour chaque contexte d'application risque de créer des silos de concepts, autrement dit, des lots de concepts dissociés qu'il faudra systématiquement reconstruire pour chaque contexte. Afin de créer des passerelles entre les différentes instances de l'ontologie et ainsi enrichir l'une par les avancées de l'autre, une stratégie de fusion des différentes ontologies de domaine doit être mise en place. Nous souhaitons explorer la possibilité de construire des modules réutilisant les KOS publiés de domaine (imagerie, biologie, psychologie, etc.) et prêts à être réutilisés pour les contextes d'application (projets Psycare, PACIFIC, Shiva, etc.). La Figure 146 résume la méthode de construction avec ajout de modules de domaines et met en avant les différents alignements mis en jeu (interopérabilité, projection, spécification, réutilisation).

[84] https://zenodo.org

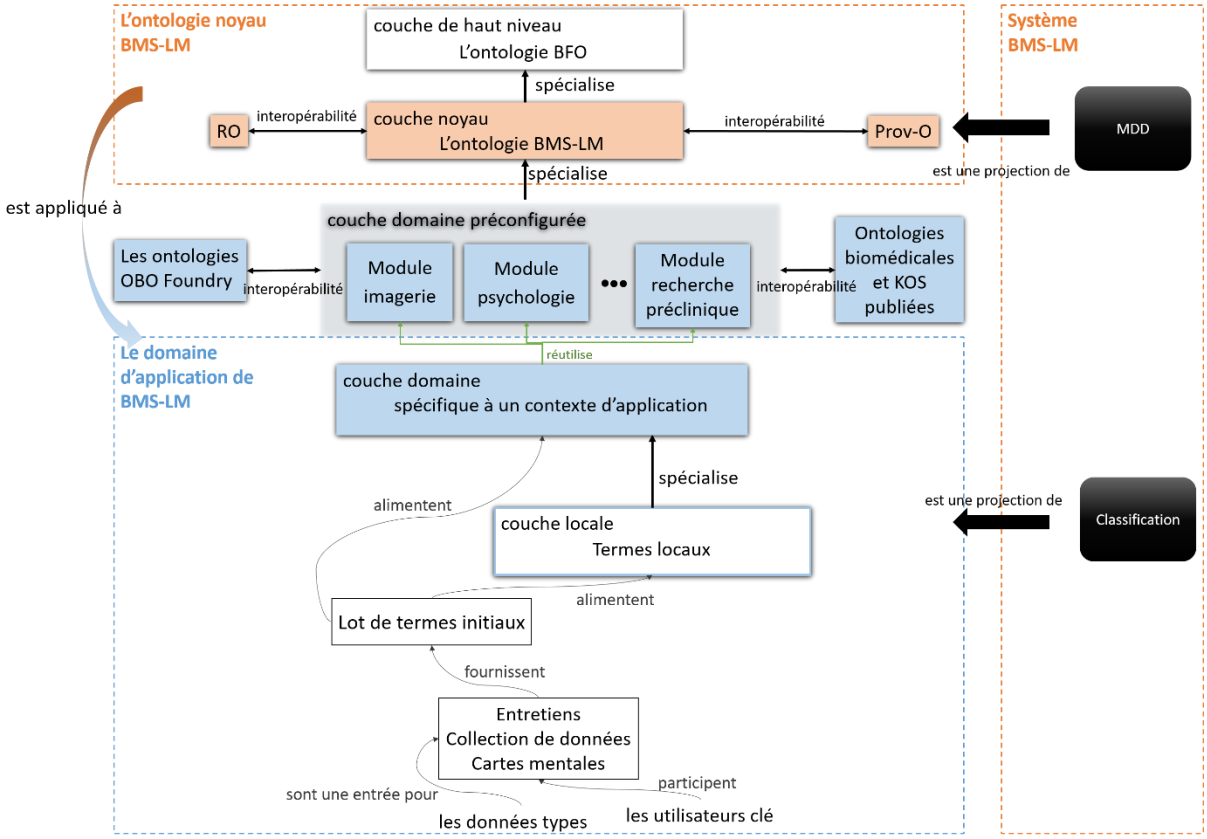


Figure 146 Ajout de modules préconfigurés de la couche domaine pour faciliter l'application de l'ontologie BMS-LM à un nouveau domaine

Le système BMS-LM, considéré ici comme un Système d'Information (SI) pour la gestion de données hétérogènes (tout le long du cycle de vie d'une étude biomédicale régi par un KOS, est une plateforme PLM « Teamcenter » qui a été adaptée à nos besoins. Il n'est ni un système à base de connaissances (KBS) ni un système de gestion de connaissances (KMS). Plusieurs architectures de KBS ont été proposés par le domaine de recherche, très actif, sur les connaissances et le PLM (§III.3.3). Nous envisageons dans les prochaines versions de faire évoluer le système BMS-LM en un KBS afin d'intégrer les fonctionnalités des moteurs de règles, moteurs d'inférences et du raisonnement logique. Une telle évolution serait un véritable atout pour faciliter la gestion de données hétérogènes et les différents « mapping » et alignements du système BMS-LM.

Le système BMS-LM proposé dans cette thèse est constitué de briques fonctionnelles issues de différentes sources : PLM, résultats du projet BIOMIST, et nouveaux résultats issus des propositions qui ont été faites et testées dans la thèse, mais qui ne sont pas encore déployés en « production » pour les utilisateurs. Le Tableau 27 présente en détail l'avancement réalisé au cours de notre thèse.

En outre, le système BMS-LM est basé sur une solution logicielle particulière : la plateforme Teamcenter. Il serait pertinent de l'implémenter à partir d'une autre plateforme PLM afin de mettre l'accent sur la généricité des différentes briques logicielles proposés et leur indépendance de la plateforme choisie.

Tableau 27 Liste des modules fonctionnels du système BMS-LM et leurs niveaux de maturité
De gauche à droite : axe du temps. X : élément développé pour cette étape d'avancement. Vert clair : élément existant pouvant être réutilisé. Vert sombre : fonctionnalité assurée

Module\Système BMS-LM	**PLM**	**BIO MIST**	**BMS-LM**	**Futur**
(1) (GED) et (SGDT)				
(2) Gestion des utilisateurs, des droits d'accès et des sites distants		X		
(3) Traçabilité des utilisations des données				
(4) Annotation des données par les informations de provenance		X	X	
(5) Intégration de données		X	X	
(6) Intégration de calcul scientifique		X	X	
(7) Communication avec les applications tierces			X	X
(8) Communication avec les clusters de calcul		X	X	
(9) Modèle De Données (MDD) et Classification		X	X	
(10) Contrôle qualité			X	X
(11) Sécurité, Réplication et Archivage		X		
(12) Requêtes personnalisées		X		
(13) Exploration des données via applications clientes		X	X	X
(14) Intégration des terminologies publiées du domaine		X	X	
(15) Reporting de données avec des formats standards				X
(16) Partage et réutilisation avec contextualisation et traçabilité				
(17) Export personnalisé et sécurisé de la base de données				X

Tout au long du développement de la version actuelle du système BMS-LM, nous avons pris en compte les quatre leviers qui sont : collaboration bio-info (L1), changement (L2), adaptation (L3), mirroring (L4) et intérêt (L5). Lors de notre plan d'« intégration_1 » (§VI.4.1) par exemple, la méthode « générique » d'import de données a été adaptée aux usages du terrain en étroite collaboration avec l'utilisateur clé « ACE » (L1-L3-L5). Le plan d'« intégration_2 » (§VI.5.1) « Mediso2PLM » des données TEP-TDM a été modifié et enrichi par une étape de filtrage des données d'intérêt via Matlab pour permettre aux utilisateurs d'intégrer le BMS-LM dans leur pratique quotidienne (L1-L3-L4-L5). Les méthodes d'intégration de données et de calcul scientifique proposées dans cette thèse visent à être le plus inclusive possible et compatibles avec les changements inhérents à la recherche scientifique (L2). L'ajout de la couche locale à l'ontologie BMS-LM avait également pour but d'intégrer la pratique réelle de l'utilisateur et de réutiliser ses termes afin de l'impliquer dans la gestion et l'annotation de données tout au long du cycle de vie de son étude (L1-L2-L3-L5). Nous avons identifié des scénarios d'utilisation du système BMS-LM lors de l'audit et nous avons veillé à ce qu'ils présentent un maximum d'intérêt pour le chercheur. L'un d'eux, l'intégration du calcul « RI-Hermite » a pu aboutir sur le plan technique, cependant, son intégration s'est révélée sans intérêt pratique pour le laboratoire LRI (Levier L5 - Intérêt faible). Néanmoins, il représente un cas d'application pratique de notre travail de thèse.

Les propositions et résultats présentés dans cette thèse constituent des contributions originales de recherche à plusieurs niveaux : l'application du PLM à d'autres domaines, la gestion de données de recherche (RDM) via un outil logiciel intégré aux pratiques quotidiennes d'un laboratoire, la gestion des données de la recherche préclinique et l'annotation interopérable des données hétérogènes. Les méthodes de mise en œuvre du système et de l'ontologie BMS-LM que nous avons proposées dans ce manuscrit ont été testées et validées via des POCs à l'échelle de nos travaux de recherche, en collaboration avec des ingénieurs, chercheurs, et experts en imagerie, en calcul scientifique, en SI, en

PLM et en data management. Nous avons, en effet, mené nos recherches dans un cadre pluridisciplinaire avec une grande variété de profils et d'interlocuteurs.

Pour présenter ce paragraphe se rapportant aux activités complémentaires au projet de thèse, je me permets de rédiger à la première personne du singulier ses tâches : au cours de cette thèse, j'ai été amenée à proposer deux stages et encadrer deux étudiants en master I et II, pour la mise en œuvre d'outils logiciels adaptés aux spécificités du laboratoire LRI. J'ai aussi assuré la formation et la vulgarisation scientifique du PLM et du BMS-LM au sein du laboratoire LRI, une mission d'audit et de pilotage du projet DRIVE, et j'ai complété mes compétences par cette expérimentation dans le domaine de la gestion des données de recherche (RDM).

La pérennisation des POCs que nous avons proposés, demande un travail de collaboration bio-info entre les acteurs pour la mise en place d'une stratégie de conduite de changement adapté au fort turn-over d'un laboratoire de recherche. Désormais, le passage en « production » du système BMS-LM est un sujet d'ingénierie avec une vocation utilitaire qui dépasse le cadre de cette thèse.

VII.3. CONCLUSION

La gestion de données de recherche est un domaine complexe où plusieurs facteurs sont à prendre en compte, en plus des questions administratives et réglementaires liées à l'« open science ». Une prise de conscience générale est en train de se développer partout en France et dans le monde quant à l'urgence d'agir et de pérenniser les données scientifiques. Dans cette thèse, nous avons exploré le sujet d'un point de vue bottom-up : en partant du laboratoire pour inclure les recommandations des instituts et tutelles scientifiques. Nous avons mené nos recherches en étroite collaboration avec les acteurs de la recherche préclinique du laboratoire LRI, afin de pouvoir répondre à leurs besoins de gestion de données scientifiques au quotidien. Cette collaboration étroite a influencé nos propositions et a structuré nos plans d'expérimentations. À chaque fois que nous avons proposé une méthode ou implémenté un outil logiciel, nous avons cherché à masquer sa complexité et l'hétérogénéité « multiple » des données et à la rendre la plus simple et adaptée aux producteurs et utilisateurs de ces données. À chaque fois, les explications techniques ont été accompagnées d'une explication fonctionnelle et d'un guide d'utilisation pour la personne concernée.

Dans cette thèse, le paradigme BMS-LM (BioMedical Study-Lifecycle Management) a été introduit pour la gestion et la structuration des données et des workflows scientifiques tout au long du cycle de vie d'une étude de recherche biomédicale. Nous sommes conscients de la nature préliminaire de nos recherches, mais aussi de leur valeur fondatrice. Nous avons fait appel aux avancées dans les domaines de la gestion, l'ingénierie et l'organisation des connaissances afin de résoudre la problématique de la compréhension dans le long terme des données hétérogènes gérées par le système BMS-LM. L'adoption du BMS-LM comme paradigme de gestion de données en recherche biomédicale devrait participer à sa maturation. Une fois adopté, il permet aux pratiques et données des laboratoires d'être interopérables avec les standards des différents domaines et les principes FAIR et donc maîtriser la RDM. Nous avons aussi participé à la mise en place de bonnes pratiques de gestion de données au laboratoire LRI. Nous espérons avec ce manuscrit et ces travaux de recherche avoir apporté un éclairage nouveau à la gestion de données scientifiques dans le domaine biomédical en général, et dans le domaine préclinique en particulier.

Liste des classes du MDD BMS-LM

Classe générique	Trigramme	Rôle : modéliser les informations sur …	Type	Icône
Acquisition Result	ACQ	…les différentes modalités d'acquisition de données utilisées dans un examen et leurs paramètres.	2 : R	
Acquisition Definition	ACD	…le protocole prédéfini de l'acquisition.	1 : D	
Agent Result	AGR	…les produits et leurs doses et leur utilisation réelle lors d'une expérimentation.	2 : R	
Agent Definition	AGD	…le protocole prédéfini pour l'administration et l'utilisation de l'agent en question.	1 : D	
Device	AQD/DVC	…l'appareil d'acquisition ou la machine d'analyse de données, ses configurations et versions.	1 : D	
Bibliographical reference	BBR	…tout document référencé par une étude de recherche : article de journal, de conférence, etc.	3 : A	
Data Unit Result	DUR	…tout jeu de données produit lors d'une acquisition de données.	2 : R	
Data Unit Definition	DUD	…les jeux de données attendus d'une acquisition de données.	1 : D	
Exam Result	EXA	…les paramètres et configurations d'un examen réalisé sur un sujet d'étude (humain, animal, etc.)	2 : R	
Exam Definition	EXD	…le protocole prédéfini de l'examen.	1 : D	
Intervention Result	ITR	…le déroulement d'une intervention sur un sujet d'étude.	2 : R	
Intervention Definition	ITD	…le protocole de l'intervention.	1 : D	
Processing Result	PCR	…le résultat d'exécution d'une chaîne de traitement, composé de plusieurs PURs.	2 : R	
Processing Definition	PCD	…le protocole d'exécution d'une chaîne de traitement, composé de plusieurs PUDs.	1 : D	
Processing Unit Result	PUR	…le résultat de l'exécution avec les paramètres (PCP) d'une unité de traitement (PUD)	2 : R	
Processing Unit Definition	PUD	…la spécification d'une unité de traitement : son algorithme, ses types d'entrées et ses types de sorties	1 : D	
Processing Parameters	PCP	…le (ou les) paramétrage(s) possible(s) d'une unité de traitement (PUD) utilisé(s) pour produire ses résultats (PUR) lors d'une exécution.	1 : D	
Reference Data	RFD	…les données utilisées comme référence dans une étude ou publiées comme référence pour d'autres études : atlas du cerveau, liste de protéines, etc.	3 : A	
Sample Result	SAR	…les échantillons biologiques prélevés sur des sujets d'étude et le déroulement du prélèvement.	2 : R	
Sample Definition	SAD	…le protocole du prélèvement d'échantillons biologiques.	1 : D	
Software Tool	STL	…l'outil logiciel utilisé dans le cadre de l'étude pour l'analyse de données.	1 : D	
Study	STU	…l'étude de recherche.	2 : R	
Study Subject	SSU	…le sujet dans le cadre de l'étude : souris, lapin, être humain (les informations sont collectées après consentement de la personne).	2 : R	
Subject	SUB	…le sujet unique dans la base de données, ses liens de parenté avec les autres sujets et les SSU qui lui sont liés.	1 : D	
Subject Group	SGP	…le groupe de sujets dans l'étude et son rôle (contrôle, traité, placebo, etc.).	3 : A	
Workflow Input	WFI	…les données, l'algorithme, le logiciel et la machine nécessaires pour l'exécution d'une chaîne de traitement (PCD).	1 : D	

Références

Abidi, L., Azzag, H., Benbernou, S., Bentounsi, M., Cérin, C., Duong, T., Garteiser, P., Lebbah, M., Ouziri, M., Sahri, S., & Smadja, M. (2019). A Big Data Platform for Enhancing Life Imaging Activities. In J. Darmont & S. Loudcher (Eds.), *Utilizing Big Data Paradigms for Business Intelligence* (pp. 39–71). IGI Global. https://doi.org/10.4018/978-1-5225-4963-5.ch002

Abi-Zeid, I., & Lamontagne, L. (2000). *Le modèle de connaissances CommonKADS pour la recherche et sauvetage* [Rapport Technique]. Defence R&D Canada – Valcartier.

Ackoff, R. L. (1989). From data to wisdom. *Journal of Applied Systems Analysis*, *16*(1), 3–9.

Adiba, M., & Portal, D. (1978). A cooperation system for heterogeneous data base management systems. *Information Systems*, *3*(3), 209–215. https://doi.org/10.1016/0306-4379(78)90004-2

Adloff, J. P. (1999). The laboratory notebooks of Pierre and Marie Curie and the discovery of polonium and radium. *Czechoslovak Journal of Physics*, *49*(1), 15–28. https://doi.org/10.1007/s10582-999-0002-y

Afoutni, Z., Le-Duigou, J., Abel, M.-H., & Eynard, B. (2017). Towards a Proactive Interoperability Solution in Systems of Information Systems: A PLM Perspective. In J. Ríos, A. Bernard, A. Bouras, & S. Foufou (Eds.), *Product Lifecycle Management and the Industry of the Future* (pp. 580–589). Springer, Cham. https://doi.org/10.1007/978-3-319-72905-3_51

Agarwal, T. K., & Sanjeev. (2012). Vendor neutral archive in PACS. *The Indian Journal of Radiology & Imaging*, *22*(4), 242–245. https://doi.org/10.4103/0971-3026.111468

Alavi, M., & Leidner, D. E. (2001). Review: Knowledge Management and Knowledge Management Systems: Conceptual Foundations and Research Issues. *MIS Quarterly*, *25*(1), 107–136. https://doi.org/10.2307/3250961

Allan, C., Burel, J.-M., Moore, J., Blackburn, C., Linkert, M., Loynton, S., MacDonald, D., Moore, W. J., Neves, C., Patterson, A., Porter, M., Tarkowska, A., Loranger, B., Avondo, J., Lagerstedt, I., Lianas, L., Leo, S., Hands, K., Hay, R. T., … Swedlow, J. R. (2012). OMERO: Flexible, model-driven data management for experimental biology. *Nature Methods*, *9*(3), 245–253. https://doi.org/10.1038/nmeth.1896

Allanic, M. (2015). *Gestion et visualisation de données hétérogènes multidimensionnelles: Application PLM à la neuroimagerie* [Thèse de Doctorat, Université de Technologie de Compiègne]. https://www.theses.fr/199231729

Allanic, M., Hervé, P.-Y., Durupt, A., Joliot, M., Boutinaud, P., & Eynard, B. (2016). Processing and Visual Analyze of Heterogeneous and Multidimensional Data in Biomedical PLM Context. In R. Harik, L. Rivest, A. Bernard, B. Eynard, & A. Bouras (Eds.), *Product Lifecycle Management for Digital Transformation of Industries* (pp. 385–398). Springer, Cham. https://doi.org/10.1007/978-3-319-54660-5_35

Allanic, M., Hervé, P.-Y., Pham, C.-C., Lekkal, M., Durupt, A., Brial, T., Grioche, A., Matta, N., Boutinaud, P., Eynard, B., & Joliot, M. (2017). BIOMIST: A Platform for Biomedical Data Lifecycle Management of Neuroimaging Cohorts. *Frontiers in ICT*, *3*. https://doi.org/10.3389/fict.2016.00035

Almuhaideb, A., Papathanasiou, N., & Bomanji, J. (2011). 18F-FDG PET/CT imaging in oncology. *Annals of Saudi Medicine*, *31*(1), 3–13. https://doi.org/10.4103/0256-4947.75771

Aloulen, Z., Belhajjame, K., Grigori, D., & Acker, R. (2019). A Domain-Independent Ontology for Capturing Scientific Experiments. In D. Kotzinos, D. Laurent, N. Spyratos, Y. Tanaka, & R. Taniguchi (Eds.), *Information Search, Integration, and Personalization* (Vol. 1040, pp. 53–68). Springer, Cham. https://doi.org/10.1007/978-3-030-30284-9_4

Anderson, N. R., Lee, S., Brockenbrough, J. S., Minie, M. E., Fuller, S., Brinkley, J., & Tarczy-Hornoch, P. (2007). Issues in biomedical research data management and analysis: Needs and barriers. *Journal of the American Medical Informatics Association*, *14*(4), 478–488. https://doi.org/10.1197/jamia.M2114

Ardila-Mejia, C. C., López-Gualdrón, C. I., & Martínez-Gómez, J. M. (2018). Product Lifecycle Management Strategy for the Definition and Design Process of Face Implants Oriented to Specific Patients. In P. Chiabert, A. Bouras, F. Noël, & J. Ríos (Eds.), *Product Lifecycle Management to Support Industry 4.0* (pp. 181–190). Springer, Cham. https://doi.org/10.1007/978-3-030-01614-2_17

Article L1121-1—Code de la santé publique—Légifrance, (2008). https://www.legifrance.gouv.fr/codes/article_lc/LEGIARTI000006685827/2008-01-29

Assouroko, I. (2012). *Gestion de données et dynamiques des connaissances en ingénierie numérique: Contribution à l'intégration de l'ingénierie des exigences, de la conception mécanique et de la simulation numérique* [Thèse de Doctorat, Université de Technologie de Compiègne]. http://www.theses.fr/2012COMP2030

Autissier, D., & Moutot, J.-M. (2016). *Méthode de conduite du changement - 4e éd.: Diagnostic, Accompagnement, Performance*. Dunod, Malakoff.

Autissier, D., Vandangeon, I., & Vas, A. (2018). *Conduite du changement: Concepts-clés - 3e éd.: 60 ans de pratiques héritées des auteurs fondateurs*. Dunod, Malakoff.

Awad, E. M., & Ghaziri, H. (2004). *Knowledge management* (1st ed). Prentice Hall. Upper Saddle River, N.J. https://trove.nla.gov.au/work/16685926

Azhari, H. (2010). *Basics of Biomedical Ultrasound for Engineers*. John Wiley & Sons, Hoboken, NJ.

Barbau, R., Krima, S., Rachuri, S., Narayanan, A., Fiorentini, X., Foufou, S., & Sriram, R. D. (2012). OntoSTEP: Enriching product model data using ontologies. *Computer-Aided Design*, *44*(6), 575–590. https://doi.org/10.1016/j.cad.2012.01.008

Barillot, C., Bannier, E., Commowick, O., Corouge, I., Baire, A., Fakhfakh, I., Guillaumont, J., Yao, Y., & Kain, M. (2016). Shanoir: Applying the Software as a Service Distribution Model to Manage Brain Imaging Research Repositories. *Frontiers in ICT*, *3*. https://doi.org/10.3389/fict.2016.00025

Barillot, C., Bannier, E., Commowick, O., Corouge, I., Guillaumont, J., Yao, Y., & Kain, M. (2015). Shanoir: Software as a Service Environment to Manage Population Imaging Research Repositories. *MICCAI Workshop on MAnagement and Processing of Images for Population ImagiNG, Oct 2015*, 22–30. https://www.hal.inserm.fr/inserm-01244551

Baskarada, S., & Koronios, A. (2013). Data, information, knowledge, wisdom (DIKW): A semiotic theoretical and empirical exploration of the hierarchy and its quality dimension. *Australasian Journal of Information Systems*, *18*(1). https://doi.org/10.3127/ajis.v18i1.748

Becerra-Fernandez, I., & Sabherwal, R. (2010). *Knowledge management: Systems and processes*. M.E. Sharpe, Armonk, NY.

Begley, C. G., & Ioannidis, J. P. A. (2015). Reproducibility in science: Improving the standard for basic and preclinical research. *Circulation Research*, *116*(1), 116–126. https://doi.org/10.1161/CIRCRESAHA.114.303819

Belhajjame, K., B'Far, R., Cheney, J., Coppens, S., Cresswell, S., Gil, Y., Groth, P., Klyne, G., Lebo, T., McCusker, J., & others. (2013). *Prov-dm: The prov data model*. W3C Recommendation; World Wide Web Consortium (W3C). http://www.w3.org/TR/2013/REC-prov-dm-20130430/

Belkadi, F., Troussier, N., Eynard, B., & Bonjour, E. (2010). Collaboration based on product lifecycles interoperability for extended enterprise. *International Journal on Interactive Design and Manufacturing*, *4*(3), 169–179. https://doi.org/10.1007/s12008-010-0099-z

Bellgard, M., Beroud, C., Parkinson, K., Harris, T., Ayme, S., Baynam, G., Weeramanthri, T., Dawkins, H., & Hunter, A. (2013). Dispelling myths about rare disease registry system development. *Source Code for Biology and Medicine*, *8*(1), 21. https://doi.org/10.1186/1751-0473-8-21

Bergman, M. (2007). An Intrepid Guide to Ontologies [Professional Blog]. *Adaptive Information Adaptive Innovation Adaptive Infrastructure*. https://www.mkbergman.com/374/an-intrepid-guide-to-ontologies

Bonifati, A., & Velegrakis, Y. (2011). Schema Matching and Mapping: From Usage to Evaluation (Tutorial). *14th International Conference on Extending Database Technology, EDBT'2011 March 21-25*. http://staff.icar.cnr.it/staff/bonifati/public_html/pubs/BonifatiV11.pdf

Borst, W. N. (1997). *Construction of Engineering Ontologies for Knowledge Sharing and Reuse* [PhD Thesis, University of Twente, Enschede, Netherlands]. https://research.utwente.nl/en/publications/construction-of-engineering-ontologies-for-knowledge-sharing-and-

Brahaj, A., Razum, M., & Schwichtenberg, F. (2012). Ontological Formalization of Scientific Experiments Based on Core Scientific Metadata Model. In P. Zaphiris, G. Buchanan, E.

Rasmussen, & F. Loizides (Eds.), *Theory and Practice of Digital Libraries* (Vol. 7489, pp. 273–279). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-33290-6_29

Branke, J., Farid, S. S., & Shah, N. (2016). Industry 4.0: A vision for personalized medicine supply chains? *Cell and Gene Therapy Insights*, *2*(2), 263–270. https://doi.org/10.18609/cgti.2016.027

Briguet, A., Fanet, H., Sappey-Marinier, D., & Vray, D. (2014). *Imagerie médicale à base de champs magnétiques et d'ultrasons*. Hermes-Lavoisier, Cachan. https://hal.archives-ouvertes.fr/hal-00976650

Bruns, E., Metz, T., & Asfaw, M. (1993). LIMS. Data management and information system for animal production. *Berichte Der Gesellschaft Fuer Informatik in Der Land-, Forst-Und Ernaehrungswirtschaft (Germany)*, *5*, 69–72. https://agris.fao.org/agris-search/search.do?recordID=DE94R0432

Buckler, A. J., Ouellette, M., Danagoulian, J., Wernsing, G., Liu, T. T., Savig, E., Suzek, B. E., Rubin, D. L., & Paik, D. (2013). Quantitative Imaging Biomarker Ontology (QIBO) for Knowledge Representation of Biomedical Imaging Biomarkers. *Journal of Digital Imaging*, *26*(4), 630–641. https://doi.org/10.1007/s10278-013-9599-2

Butler, D. (2005). Electronic notebooks: A new leaf. *Nature*, *436*(7047), 20–21. https://doi.org/10.1038/436020a

Buvat, I. (2007). Les limites du SUV. *Médecine Nucléaire*, *31*(4), 165–172.

Camarasu-Pop, S., Cervenansky, F., Cardenas, Y., Nief, J.-Y., & Benoit-Cattin, H. (2010, May). *Overview of Medical Data Management Solutions for Research Communities*. 2010 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing, 17-20 May, Melbourne, VIC, Australia. https://doi.org/10.1109/CCGRID.2010.55

Cassina, J., Tomasella, M., Taisch, M., & Matta, A. (2009). A new closed-loop PLM Standard for mass products. *International Journal of Product Development*, *8*(2), 141--161. https://doi.org/10.1504/IJPD.2009.024185

Cavelaars, M., Rousseau, J., Parlayan, C., de Ridder, S., Verburg, A., Ross, R., Visser, G. R., Rotte, A., Azevedo, R., Boiten, J.-W., Meijer, G. A., Belien, J. A. M., & Verheul, H. (2015).

OpenClinica. *Journal of Clinical Bioinformatics*, *5*(1), S2. https://doi.org/10.1186/2043-9113-5-S1-S2

CHAI. (2014). *Guide d'audit des systèmes d'information* (p. 118). Comité d'Harmonisation de l'Audit interne Interministériel, France. https://www.economie.gouv.fr/files/guide_d_audit_des_si_v1-2.pdf

Chakravarthy, A., Beales, R., Matskanis, N., & Yang, X. (2009). OntoFilm: A Core Ontology for Film Production. In T. S. Chua, Y. Kompatsiaris, B. Merialdo, W. Haas, G. Thallinger, & W. Bailer (Eds.), *International Conference on Semantic and Digital Media Technologies* (Vol. 5887, pp. 177–181). Springer Berlin Heidelberg.

Charlet, J. (2002). *L'ingénierie des connaissances: Développements, résultats et perspectives pour la gestion des connaissances médicales* [Habilitation à Diriger des Recherches]. Université Pierre et Marie Curie.

Chen, D., Doumeingts, G., & Vernadat, F. (2008). Architectures for enterprise integration and interoperability: Past, present and future. *Computers in Industry*, *59*(7), 647–659. https://doi.org/10.1016/j.compind.2007.12.016

Chen, X., & Wu, M. (2017). Survey on the Needs for Chemistry Research Data Management and Sharing. *The Journal of Academic Librarianship*, *43*(4), 346–353. https://doi.org/10.1016/j.acalib.2017.06.006

Chen, Y.-J., Chen, Y.-M., & Chu, H.-C. (2009). Development of a mechanism for ontology-based product lifecycle knowledge integration. *Expert Systems with Applications*, *36*(2), 2759–2779. https://doi.org/10.1016/j.eswa.2008.01.049

Ciccarese, P., Soiland-Reyes, S., Belhajjame, K., Gray, A. J., Goble, C., & Clark, T. (2013). PAV ontology: Provenance, authoring and versioning. *Journal of Biomedical Semantics*, *4*(1), 37. https://doi.org/10.1186/2041-1480-4-37

CIGREF. (2019). *Guide d'audit de la gouvernance du système d'information de l'entreprise numérique* (2ème Édition) [Rapport Technique]. AFAI, CIGREF et IFACI. https://www.cigref.fr/wp/wp-content/uploads/2019/03/2019-Guide-Audit-Gouvernance-Systeme-Information-Entreprise-Numerique-2eme-edition-Cigref-Afai-Ifaci.pdf

Clunie, D. A. (2000). *DICOM Structured Reporting*. PixelMed Publishing. https://books.google.fr/books?id=EVjOolUJNGUC

Cohrs, R., Martin, T., Ghahramani, P., Bidaut, L., Higgins, P., & Shahzad, A. (2015). Translational Medicine definition by the European Society for Translational Medicine. *New Horizons in Translational Medicine*, *2*(3), 86–88. https://doi.org/10.1016/j.nhtm.2014.12.002

Colquhoun, G. J., Baines, R. W., & Crossley, R. (1993). A state of the art review of IDEFO. *International Journal of Computer Integrated Manufacturing*, *6*(4), 252–264. https://doi.org/10.1080/09511929308944576

Corcho, O., Fernández-lópez, M., & Gómez-pérez, A. (2006). Ontological Engineering: Principles, Methods, Tools and Languages. In *Ontologies for software engineering and software technology* (pp. 1--48). Springer, Berlin, Heidelberg.

Courtot, M., Gibson, F., Lister, A. L., Malone, J., Schober, D., Brinkman, R. R., & Ruttenberg, A. (2009). MIREOT: The Minimum Information to Reference an External Ontology Term. *Nature Precedings*. International Conference on Biomedical Ontology, July 24-26, 2009, Buffalo, NY. https://doi.org/10.1038/npre.2009.3574.1

Danciu, I., Cowan, J. D., Basford, M., Wang, X., Saip, A., Osgood, S., Shirey-Rice, J., Kirby, J., & Harris, P. A. (2014). Secondary use of clinical data: The Vanderbilt approach. *Journal of Biomedical Informatics*, *52*, 28–35. https://doi.org/10.1016/j.jbi.2014.02.003

Daniel-Le Bozec, C., Steichen, O., Dart, T., & Jaulent, M.-C. (2007). The role of local terminologies in electronic health records. The HEGP experience. *Studies in Health Technology and Informatics*, *129*(Pt 1), 780–784.

Danjou, C. (2015). *Ingénierie de la chaîne numérique d'industrialisation: Proposition d'un modèle d'interopérabilité pour la conception-fabrication intégrées* [Thèse de Doctorat, Université de Technologie de Compiègne]. http://www.theses.fr/2015COMP2234

Danjou, C., Duigou, J. L., & Eynard, B. (2017). Manufacturing knowledge management based on STEP-NC standard: A Closed-Loop Manufacturing approach. *International Journal of Computer Integrated Manufacturing*, *30*(9), 995–1009. https://doi.org/10.1080/0951192X.2016.1268718

Danjou, C., Le Duigou, J., & Eynard, B. (2013, July). Interopérabilité des systèmes PLM: Un état de l'art sur la chaîne numérique conception/industrialisation. *JD/JN MACS 2013*. https://hal.archives-ouvertes.fr/hal-01058826

Deserno, T. M., Haak, D., Brandenburg, V., Deserno, V., Classen, C., & Specht, P. (2014). Integrated image data and medical record management for rare disease registries. A general framework and its instantiation to theGerman Calciphylaxis Registry. *Journal of Digital Imaging*, *27*(6), 702–713. https://doi.org/10.1007/s10278-014-9698-8

Deutsch, E. W. (2010). Mass Spectrometer Output File Format mzML. *Methods in Molecular Biology*, *604*, 319–331. https://doi.org/10.1007/978-1-60761-444-9_22

DG RTD. (2017). *Guidelines to the rules on open access to scientific publications and open access to research data in Horizon 2020* (Version 3.2; p. 11). European Commission Directorate-General for Research & Innovation. https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf

Diallo, G. (2014). An effective method of large scale ontology matching. *Journal of Biomedical Semantics*, *5*, 44. https://doi.org/10.1186/2041-1480-5-44

Dianat, O., Paris, C., & Wan, S. (2013). A Study: From Electronic Laboratory Notebooks to Generated Queries for Literature Recommendation. *Proceedings of the Australasian Language Technology Association*. ALTA Workshop, Dec 4-6, 2013, Brisbane, Australia. https://www.aclweb.org/anthology/U13-1009

Dillenseger, J.-P. (2017). *Imagerie préclinique multimodale chez le petit animal: Qualification des instruments et des méthodes (IRM, µTDM et µTEMP)* [Thèse de Doctorat, Université De Strasbourg]. http://www.theses.fr/2017STRAD026

Ding, L., Bao, J., Michaelis, J. R., Zhao, J., & McGuinness, D. L. (2010). Reflections on Provenance Ontology Encodings. In D. L. McGuinness, J. R. Michaelis, & L. Moreau (Eds.), *Provenance and Annotation of Data and Processes* (Vol. 6378, pp. 198–205). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-17819-1_22

Dos Reis, J. C., Pruski, C., Da Silveira, M., & Reynaud-Delaître, C. (2015). DyKOSMap: A framework for mapping adaptation between biomedical knowledge organization systems. *Journal of Biomedical Informatics, 55*(Supplement C), 153–173. https://doi.org/10.1016/j.jbi.2015.04.001

Ducellier, G. (2008). *Gestion de règles expertes en ingénierie collaborative: Applications aux plateformes PLM* [Thèse de Doctorat, Université de Technologie de Troyes]. http://www.theses.fr/2008TROY0008

Eliot, T. S. (1934). *The Rock* (1st edition). Faber & Faber London.

EMA. (2019). *Du laboratoire au patient: Le voyage d'un médicament évalué par l'EMA* (p. 27) [European Report]. European Medicines Agency. https://www.ema.europa.eu/en/documents/other/laboratory-patient-journey-centrally-authorised-medicine_fr.pdf

Eynard, B. (2005). *Gestion du cycle de vie des produits et dynamique des connaissances industrielles en conception intégrée* [Habilitation à Diriger des Recherches, Université de Technologie de Compiègne]. https://hal.archives-ouvertes.fr/tel-03128022

Eynard, B., Gallet, T., Nowak, P., & Roucoules, L. (2004). UML based specifications of PDM product structure and workflow. *Computers in Industry*, *55*(3), 301–316. https://doi.org/10.1016/j.compind.2004.08.006

Fenves, S. J., Foufou, S., Bock, C., & Sriram, R. D. (2008). CPM2: A Core Model for Product Data. *Journal of Computing and Information Science in Engineering*, *8*(1), 014501. https://doi.org/10.1115/1.2830842

Fielding, E. A., McCardle, J. R., Eynard, B., Hartman, N., & Fraser, A. (2014). Product lifecycle management in design and engineering education: International perspectives: *Concurrent Engineering*, *22*(2), 123–134. https://doi.org/10.1177/1063293X13520316

Fiorentini, X., Gambino, I., Liang, V.-C., Rachuri, S., Mani, M., & Bock, C. (2007). *An ontology for assembly representation* (NISTIR 7436). National Institute of Standards and Technology & U.S. Department of Commerce. https://doi.org/10.6028/NIST.IR.7436

Fishman, E. K., Ney, D. R., Hennessey, J. G., Magid, D., & Kuhlman, J. E. (1991). Data base management in radiology: A simplified approach. *Journal of Digital Imaging*, *4*(3), 185–187.

Forsberg, D., Rosipko, B., Sunshine, J. L., & Ros, P. R. (2016). State of Integration Between PACS and Other IT Systems: A National Survey of Academic Radiology Departments. *Journal of the American College of Radiology*, *13*(7), 812-818.e2. https://doi.org/10.1016/j.jacr.2016.01.018

Fortineau, V., Paviot, T., & Lamouri, S. (2013). 5 root concepts for a meta-ontology to model product along its whole lifecycle*. *IFAC Proceedings Volumes*, *46*(7), 47–52. https://doi.org/10.3182/20130522-3-BR-4036.00061

Frank, N., Riedesel, H., & Lenz, R. (1991). The laboratory animal management system–an animal housing management data-processing system. *Journal of Experimental Animal Science*, *34*(4), 140–146.

Frohmann, B. (1994). The Social Construction of Knowledge Organization: The Case of Melvil Dewey. *Advances in Knowledge Organization*, *4*, 109–117. https://www.ergon-verlag.de/isko_ko/downloads/aikovol04199415.pdf

Gagalova, K. K., Elizalde, M. A. L., Portales-Casamar, E., & Görges, M. (2020). What You Need to Know Before Implementing a Clinical Research Data Warehouse: Comparative Review of Integrated Data Repositories in Health Care Institutions. *JMIR Formative Research*, *4*(8), e17687. https://doi.org/10.2196/17687

Gangemi, A., Guarino, N., Masolo, C., Oltramari, A., & Schneider, L. (2002). Sweetening Ontologies with DOLCE. In A. Gómez-Pérez & V. R. Benjamins (Eds.), *Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web* (Vol. 2473, pp. 166–181). Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-45810-7_18

Garcelon, N., Neuraz, A., Salomon, R., Faour, H., Benoit, V., Delapalme, A., Munnich, A., Burgun, A., & Rance, B. (2018). A clinician friendly data warehouse oriented toward narrative reports: Dr. Warehouse. *Journal of Biomedical Informatics*, *80*, 52–63. https://doi.org/10.1016/j.jbi.2018.02.019

Garets, D., & Davis, M. (2006). Electronic medical records vs. electronic health records: Yes, there is a difference. *Policy White Paper. Chicago, HIMSS Analytics*, 1–14.

Gecevska, V., Chiabert, P., Anisic, Z., Lombardi, F., & Cus, F. (2010). Product lifecycle management through innovative and competitive business environment. *Journal of Industrial Engineering and Management*, *3*(2), 323–336. https://doi.org/10.3926/jiem.2010.v3n2.p323-336

Gennari, J. H., Musen, M. A., Fergerson, R. W., Grosso, W. E., Crubézy, M., Eriksson, H., Noy, N. F., & Tu, S. W. (2003). The evolution of Protégé: An environment for knowledge-based systems development. *International Journal of Human-Computer Studies*, *58*(1), 89–123. https://doi.org/10.1016/S1071-5819(02)00127-1

Gibbon, G. A. (1996). A brief history of LIMS. *Laboratory Automation & Information Management*, *32*(1), 1–5. https://doi.org/10.1016/1381-141X(95)00024-K

Giraldo, O., García, A., López, F., & Corcho, O. (2017). Using semantics for representing experimental protocols. *Journal of Biomedical Semantics*, *8*(1), 52. https://doi.org/10.1186/s13326-017-0160-y

Goldberg, I. G., Allan, C., Burel, J.-M., Creager, D., Falconi, A., Hochheiser, H., Johnston, J., Mellen, J., Sorger, P. K., & Swedlow, J. R. (2005). The Open Microscopy Environment (OME) Data Model and XML file: Open tools for informatics and quantitative analysis in biological imaging. *Genome Biology*, *6*(5), R47. https://doi.org/10.1186/gb-2005-6-5-r47

González-Beltrán, A., Maguire, E., Sansone, S.-A., & Rocca-Serra, P. (2014). linkedISA: Semantic representation of ISA-Tab experimental metadata. *BMC Bioinformatics*, *15*(14), S4. https://doi.org/10.1186/1471-2105-15-S14-S4

Gorgolewski, K., Burns, C. D., Madison, C., Clark, D., Halchenko, Y. O., Waskom, M. L., & Ghosh, S. S. (2011). Nipype: A Flexible, Lightweight and Extensible Neuroimaging Data Processing Framework in Python. *Frontiers in Neuroinformatics*, *5*. https://doi.org/10.3389/fninf.2011.00013

Graham, M. M. (2012). Clinical molecular imaging with radiotracers: Current status. *Medical Principles and Practice: International Journal of the Kuwait University, Health Science Centre*, *21*(3), 197–208. https://doi.org/10.1159/000333552

Gregorio, J.-L. (2020). *Contribution à la définition d'un jumeau numérique pour la maîtrise de la qualité géométrique des structures aéronautiques lors de leurs processus d'assemblage* [Thèse de Doctorat, Université Paris-Saclay]. https://hal.archives-ouvertes.fr/tel-02616138

Groß, A., Pruski, C., & Rahm, E. (2016). Evolution of biomedical ontologies and mappings: Overview of recent approaches. *Computational and Structural Biotechnology Journal*, *14*, 333–340. https://doi.org/10.1016/j.csbj.2016.08.002

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, *5*(2), 199–220. https://doi.org/10.1006/knac.1993.1008

Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing? *International Journal of Human-Computer Studies*, *43*(5–6), 907–928. https://doi.org/10.1006/ijhc.1995.1081

Guarino, N. (2017). BFO and DOLCE: So Far, So Close... *COSMOS + TAXIS: Studies In Emergent Order And Organization*, *4*(4), 9.

Guarino, N., Carrara, M., & Giaretta, P. (1994). Formalizing Ontological Commitment. In B. Hayes-Roth & R. E. Korf (Eds.), *Proceedings of the Twelfth National Conference on Artificial Intelligence* (Vol. 94, pp. 560–567). AAAI Press.

Harris, P. A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., & Conde, J. G. (2009). Research electronic data capture (REDCap)—A metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of Biomedical Informatics*, *42*(2), 377–381. https://doi.org/10.1016/j.jbi.2008.08.010

He, Y., Xiang, Z., Zheng, J., Lin, Y., Overton, J. A., & Ong, E. (2018). The eXtensible ontology development (XOD) principles and tool implementation to support ontology interoperability. *Journal of Biomedical Semantics*, *9*(1), 3. https://doi.org/10.1186/s13326-017-0169-2

Hecht, M. (2008). *PACS - Picture Archiving and Communication System* (p. 10) [Master Thesis]. Vienna University of Technology, University of Paderborn, Austria.

Helsens, K., Colaert, N., Barsnes, H., Muth, T., Flikka, K., Staes, A., Timmerman, E., Wortelkamp, S., Sickmann, A., Vandekerckhove, J., Gevaert, K., & Martens, L. (2010). Ms_lims, a simple yet

powerful open source laboratory information management system for MS-driven proteomics. *PROTEOMICS*, *10*(6), 1261–1264. https://doi.org/10.1002/pmic.200900409

Hervy, B., Laroche, F., Bernard, A., & Kerouanton, J.-L. (2017). Framework for historical knowledge management in museology. *International Journal of Product Lifecycle Management*, *10*(1), 44–68. https://doi.org/10.1504/IJPLM.2017.083001

Hidde, A. R. (1991). Management of information in a system of heterogeneous distributed data bases using the example of a PCB assemblage. *International Journal of Computer Integrated Manufacturing*, *4*(6), 323–330. https://doi.org/10.1080/09511929108944510

Hjørland, B. (2008). What is Knowledge Organization (KO)? *Knowledge Organization*, *35*(2–3), 86--101. http://arizona.openrepository.com/arizona/handle/10150/106183

Hodge, G. (2000). *Systems of Knowledge Organization for Digital Libraries: Beyond Traditional Authority Files*. Digital Library Federation, Council on Library and Information Resources, Washington, DC. https://eric.ed.gov/?id=ED440657

Houston, L., Yu, P., Martin, A., & Probst, Y. (2020). Heterogeneity in clinical research data quality monitoring: A national survey. *Journal of Biomedical Informatics*, *108*, 103491. https://doi.org/10.1016/j.jbi.2020.103491

Hunter, J. (2003). Enhancing the semantic interoperability of multimedia through a core ontology. *IEEE Transactions on Circuits and Systems for Video Technology*, *13*(1), 49–58. https://doi.org/10.1109/TCSVT.2002.808088

IEC 62264-1. (2003). *Enterprise-control system integration—Part 1: Models and terminology*. International Organization for Standardization. Geneva. https://www.iso.org/fr/standard/35480.html

IEEE. (1991). IEEE Standard Computer Dictionary: A Compilation of IEEE Standard Computer Glossaries. *IEEE Std 610*, 1–217. https://doi.org/10.1109/IEEESTD.1991.106963

ISACA. (2018). *COBIT 2019 Framework: Introduction and Methodology*. Information Systems Audit and Control Association. https://www.isaca.org/bookstore/bookstore-cobit_19-digital/wcb19fim

ISO 10303. (1994). *Industrial automation systems and integration—Product data representation and exchange—Part 1: Overview and fundamental principles*. International Organization for Standardization. Geneva. https://www.iso.org/fr/standard/20579.html

ISO 14258. (1998). *Industrial automation systems—Concepts and rules for enterprise models*. International Organization for Standardization. Geneva. https://www.iso.org/fr/standard/24020.html

ISO/IEC 27000. (2018). *Information technology—Security techniques—Information security management systems—Overview and vocabulary*. International Organization for Standardization. Geneva. https://www.iso.org/standard/73906.html

Jannot, A.-S., Zapletal, E., Avillach, P., Mamzer, M.-F., Burgun, A., & Degoulet, P. (2017). The Georges Pompidou University Hospital Clinical Data Warehouse: A 8-years follow-up experience. *International Journal of Medical Informatics*, *102*, 21–28. https://doi.org/10.1016/j.ijmedinf.2017.02.006

Jha, A. K., DesRoches, C. M., Campbell, E. G., Donelan, K., Rao, S. R., Ferris, T. G., Shields, A., Rosenbaum, S., & Blumenthal, D. (2009). Use of Electronic Health Records in U.S. Hospitals. *New England Journal of Medicine*, *360*(16), 1628–1638. https://doi.org/10.1056/NEJMsa0900592

Jonnalagadda, S. R., Adupa, A. K., Garg, R. P., Corona-Cox, J., & Shah, S. J. (2017). Text Mining of the Electronic Health Record: An Information Extraction Approach for Automated Identification and Subphenotyping of HFpEF Patients for Clinical Trials. *Journal of Cardiovascular Translational Research*, *10*(3), 313–321. https://doi.org/10.1007/s12265-017-9752-2

Jun, H. B., Kiritsis, D., & Xirouchakis, P. (2007). A primitive ontology model for product lifecycle meta data in the closed-loop PLM. In R. J. Gonçalves, J. P. Müller, K. Mertins, & M. Zelm (Eds.), *Enterprise Interoperability II* (pp. 729–740). Springer London. https://doi.org/10.1007/978-1-84628-858-6_80

Kadiri, S. E., & Kiritsis, D. (2015). Ontologies in the context of product lifecycle management: State of the art literature review. *International Journal of Production Research*, *53*(18), 5657–5668. https://doi.org/10.1080/00207543.2015.1052155

Kain, M. (2018, June 22). *France Life Imaging (FLI)-Information Analysis and Management (IAM) Provider of data storage and processing solutions for preclinical imaging studies*. Appning2018 - Workshop on Animal PoPulatioN ImagiNG, 22 June, Paris. https://hal.inria.fr/hal-01949362/document

Kain, M., Bodin, M., Loury, S., Chi, Y., Louis, J., Simon, M., Lamy, J., Barillot, C., & Dojat, M. (2020). Small Animal Shanoir (SAS) A Cloud-Based Solution for Managing Preclinical MR Brain Imaging Studies. *Frontiers in Neuroinformatics*, *14*. https://doi.org/10.3389/fninf.2020.00020

Khalil, M. (2017). Small Animal micro-PET imaging: An overview. *The Egyptian Journal Nuclear Medicine*, *14*(14), 8–27. https://doi.org/10.21608/egyjnm.2017.5435

Khundam, C., & Noël, F. (2017). Storytelling Platform for Virtual Museum Development: Lifecycle Management of an Exhibition. In J. Ríos, A. Bernard, A. Bouras, & S. Foufou (Eds.), *Product Lifecycle Management and the Industry of the Future* (pp. 416–426). Springer, Cham. https://doi.org/10.1007/978-3-319-72905-3_37

KPMG. (1998). *Knowledge management* [Research report]. KPMG Management Consulting. https://www.brint.com/papers/submit/knowmgmt.pdf

Krahe, M. A., Toohey, J., Wolski, M., Scuffham, P. A., & Reilly, S. (2020). Research data management in practice: Results from a cross-sectional survey of health and medical researchers from an academic institution in Australia. *Health Information Management Journal*, *49*(2–3), 108–116. https://doi.org/10.1177/1833358319831318

Lantada, A. D. (Ed.). (2013). *Handbook on Advanced Design and Manufacturing Technologies for Biomedical Devices*. Springer US. https://doi.org/10.1007/978-1-4614-6789-2

Lapinlampi, N., Melin, E., Aronica, E., Bankstahl, J. P., Becker, A., Bernard, C., Gorter, J. A., Gröhn, O., Lipsanen, A., Lukasiuk, K., Löscher, W., Paananen, J., Ravizza, T., Roncon, P., Simonato, M., Vezzani, A., Kokaia, M., & Pitkänen, A. (2017). Common data elements and data

management: Remedy to cure underpowered preclinical studies. *Epilepsy Research*, *129*, 87–90. https://doi.org/10.1016/j.eplepsyres.2016.11.010

Laudon, K. C., & Laudon, J. P. (2006). *Management information systems: Managing the digital firm*. Pearson/Prentice Hall, Upper Saddle River, NJ.

Le Duigou, J. (2010). *Cadre de modélisation pour les systèmes PLM en entreprise étendue. Application aux PME mécaniciennes.* [Thèse de Doctorat, Ecole Centrale de Nantes (ECN)]. https://tel.archives-ouvertes.fr/tel-00487196

Le Duigou, J. (2017). *Apports des ontologies et de l'apprentissage automatique à la conception de systèmes mécaniques* [Habilitation à Diriger des Recherches, Université de Technologie de Compiègne]. https://tel.archives-ouvertes.fr/tel-01662566/document

Leaders, F. E., van Hoose, M. C., & O'Kane, K. C. (1980). Computer-Based Systems for Acquisition, Management and Reporting of Animal Data—The Biomedical Scientist's Perspective. *Drug Information Journal*, *14*(1), 15–19.

Lebo, T., Sahoo, S., McGuinness, D., Belhajjame, K., Cheney, J., Corsar, D., Garijo, D., Soiland-Reyes, S., Zednik, S., & Zhao, J. (2013). *Prov-o: The prov ontology*. W3C Recommendation. http://www.w3.org/TR/2013/REC-prov-o-20130430/

Ledford, H., & Noorden, R. V. (2020). High-profile coronavirus retractions raise concerns about data oversight. *Nature*, *582*(7811), 160–160. https://doi.org/10.1038/d41586-020-01695-w

Lee, J.-H., & Suh, H.-W. (2007). OWL-Based Product Ontology Architecture and Representation for Sharing Product Knowledge on a Web. *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, *48035*, 853--861. https://doi.org/10.1115/DETC2007-35312

Lelong, R., Soualmia, L. F., Grosjean, J., Taalba, M., & Darmoni, S. J. (2019). Building a Semantic Health Data Warehouse in the Context of Clinical Trials: Development and Usability Study. *JMIR Medical Informatics*, *7*(4). https://doi.org/10.2196/13917

Leonelli, S. (2019). *La recherche scientifique à l'ère des Big Data*. Edition Mimésis, MI, Italie.

Lewkowicz, M. (2012). *Le rôle des modèles dans la conception de systèmes pour des pratiques collectives médiatisées: Contributions méthodologiques, conceptuelles et instrumentales à une*

recherche interdisciplinaire en CSCW [Habilitation à Diriger des Recherches]. Université de Technologie de Compiègne.

Li, Weiy., & Banks, K. (2006). Computers as data analysis and data management tools in preclinical development. *Computer Applications in Pharmaceutical Research and Development*, *2*, 51.

Loi Huriet-Sérusclat: Loi n° 88-1138 du 20 décembre 1988 relative à la protection des personnes qui se prêtent à des recherches biomédicales, 2000 (1988).

Loi Jardé: Loi n° 2012-300 du 5 mars 2012 relative aux recherches impliquant la personne humaine, 2012-300 (2012).

Madec, J., Bouzillé, G., Riou, C., Van Hille, P., Merour, C., Artigny, M.-L., Delamarre, D., Raimbert, V., Lemordant, P., & Cuggia, M. (2019). eHOP Clinical Data Warehouse: From a Prototype to the Creation of an Inter-Regional Clinical Data Centers Network. *Studies in Health Technology and Informatics*, *264*, 1536–1537. https://doi.org/10.3233/SHTI190522

Maier, R. (2007). *Knowledge management systems: Information and communication technologies for knowledge management* (3rd ed). Springer Berlin Heidelberg.

Mannheim, J. G., Kara, F., Doorduin, J., Fuchs, K., Reischl, G., Liang, S., Verhoye, M., Gremse, F., Mezzanotte, L., & Huisman, M. C. (2018). Standardization of Small Animal Imaging—Current Status and Future Prospects. *Molecular Imaging and Biology*, *20*(5), 716–731. https://doi.org/10.1007/s11307-017-1126-2

Mansoori, B., Erhard, K. K., & Sunshine, J. L. (2012). Picture archiving and communication system (PACS) implementation, integration & benefits in an integrated health system. *Academic Radiology*, *19*(2), 229–235.

Marée, R., Rollus, L., Stévens, B., Hoyoux, R., Louppe, G., Vandaele, R., Begon, J.-M., Kainz, P., Geurts, P., & Wehenkel, L. (2016). Collaborative analysis of multi-gigapixel imaging data using Cytomine. *Bioinformatics*, *32*(9), 1395–1401. https://doi.org/10.1093/bioinformatics/btw013

Martínez Gómez, J. M., López Gualdrón, C. I., Murillo Bohórquez, A. P., & Garnica Bohórquez, I. (2019). PLM Strategy for Developing Specific Medical Devices and Lower Limb Prosthesis at Healthcare Sector: Case Reports from the Academia. In John Stark (Ed.), *Product Lifecycle*

Management (Volume 4): The Case Studies (pp. 201–221). Springer, Cham. https://doi.org/10.1007/978-3-030-16134-7_16

Martínez-Romero, M., O'Connor, M. J., Dorf, M., Vendetti, J., Willrett, D., Egyedi, A. L., Graybeal, J., & Musen, M. A. (2017). Supporting Ontology-Based Standardization of Biomedical Metadata in the CEDAR Workbench. In J. D. Warrender & P. Lord (Eds.), *Proceedings of the 8th International Conference on Biomedical Ontology ICBO 2017: Vol. Vol-2137* (pp. 1–6). CEUR-WP, Newcastle-upon-Tyne, UK. http://ceur-ws.org/Vol-2137/

Mascardi, V., Cordì, V., & Rosso, P. (2007). A Comparison of Upper Ontologies. *Proceedings of WOA2007: Workshop Dagli Oggetti Agli Agenti (WOA) 24-25 Sept, Genova*, *vol 2007*, 55–64.

Masolo, C., Borgo, S., Gangemi, A., Guarino, N., & Oltramari, A. (2003). *WonderWeb deliverable D18 ontology library (final)* (IST Project 2001-33052, Del18 v1.0). Laboratory For Applied Ontology and ISTC-CNR.

Matsokis, A., & Kiritsis, D. (2010). An ontology-based approach for Product Lifecycle Management. *Computers in Industry*, *61*(8), 787–797. https://doi.org/10.1016/j.compind.2010.05.007

Maxhelaku, S., & Kika, A. (2020). Implementation of SPA in Radiology Information System. *2020 IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS), Bari, Italy, 2020*, 1–5. https://doi.org/10.1109/EAIS48028.2020.9122761

Mayer, G., Montecchi-Palazzi, L., Ovelleiro, D., Jones, A. R., Binz, P.-A., Deutsch, E. W., Chambers, M., Kallhardt, M., Levander, F., Shofstahl, J., Orchard, S., Antonio Vizcaíno, J., Hermjakob, H., Stephan, C., Meyer, H. E., & Eisenacher, M. (2013). The HUPO proteomics standards initiative- mass spectrometry controlled vocabulary. *Database*, *2013*. https://doi.org/10.1093/database/bat009

McIlwaine, I. C., & Mitchell, J. S. (2008). Preface to Special Issue "What is Knowledge Organization". *KNOWLEDGE ORGANIZATION*, *35*(2–3). https://nkos.slis.kent.edu/KO_35_2-3_ToC_Preface.pdf

Micard, E., Husson, D., Team, C.-I., & Felblinger, J. (2016). ArchiMed: A Data Management System for Clinical Research in Imaging. *Frontiers in ICT*, *3*. https://doi.org/10.3389/fict.2016.00031

Miksa, T., Simms, S., Mietchen, D., & Jones, S. (2019). Ten principles for machine-actionable data management plans. *PLOS Computational Biology*, *15*(3), e1006750. https://doi.org/10.1371/journal.pcbi.1006750

Mildenberger, P., Eichelberg, M., & Martin, E. (2002). Introduction to the DICOM standard. *European Radiology*, *12*(4), 920–927. https://doi.org/10.1007/s003300101100

Mongan, J., & Avrin, D. (2018). Impact of PACS-EMR integration on radiologist usage of the EMR. *Journal of Digital Imaging*, *31*(5), 611–614.

Moss, R. (1964). Categories and relations: Origins of two classification theories. *American Documentation*, *15*(4), 296–301. https://doi.org/10.1002/asi.5090150408

Murphy, S. N., Mendis, M., Hackett, K., Kuttan, R., Pan, W., Phillips, L. C., Gainer, V., Berkowicz, D., Glaser, J. P., Kohane, I., & Chueh, H. C. (2007). Architecture of the Open-source Clinical Research Chart from Informatics for Integrating Biology and the Bedside. *AMIA Annual Symposium Proceedings*, *2007*, 548–552. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2655844/

Myers, R., Hume, S., Bloomfield, P., & Jones, T. (1999). Radio-imaging in small animals. *Journal of Psychopharmacology*, *13*(4), 352–357. https://doi.org/10.1177/026988119901300404

Nelson, E. K., Piehler, B., Eckels, J., Rauch, A., Bellew, M., Hussey, P., Ramsay, S., Nathe, C., Lum, K., Krouse, K., Stearns, D., Connolly, B., Skillman, T., & Igra, M. (2011). LabKey Server: An open source platform for scientific data integration, analysis and collaboration. *BMC Bioinformatics*, *12*(1), 71. https://doi.org/10.1186/1471-2105-12-71

Newman, P. G., & Rozycki, G. S. (1998). The History of Ultrasound. *Surgical Clinics of North America*, *78*(2), 179–195. https://doi.org/10.1016/S0039-6109(05)70308-X

Ngo, T. N. (2018). *Une approche PLM pour supporter les collaborations et le partage des connaissances dans le secteur médical: Application aux processus de soins par implantation de prothèses* [Thèse de Doctorat, Ecole centrale de Nantes]. http://www.theses.fr/2018ECDN0013

Niessen, W. M. A. (2006). *Liquid Chromatography-Mass Spectrometry* (Third Edition). CRC Press - Taylor and Francis, Boca Raton, FL.

Ogden, C., & Richards, I. (1923). *The Meaning of Meaning: A study of the influence of thought and of the science of symbolism*. Harcourt, Brace & World, Inc. NY.

Ohly, H. P. (2012). Organization, Management and Engineering of Knowledge: Rivals or Complements? In C. Pérez Pais & M. de los Á. González Bonome (Eds.), *Minutes of the XX Congress ISKO-Spain. Ferrol, June 30—July 1, 2011* (pp. 541–551). Universidade da Coruña (España), Publications Service. http://ruc.udc.es/dspace/handle/2183/11639

Oliveira, D., & Pesquita, C. (2018). Improving the interoperability of biomedical ontologies with compound alignments. *Journal of Biomedical Semantics*, *9*(1), 1. https://doi.org/10.1186/s13326-017-0171-8

OMG. (2013). *Business Process Model and Notation (BPMN), Version 2.0* (p. 532) [Rapport Technique]. Object Management Group. https://www.omg.org/spec/BPMN/2.0.2/PDF/

Operto, G., Chupin, M., Batrancourt, B., Habert, M.-O., Colliot, O., Benali, H., Poupon, C., Champseix, C., Delmaire, C., Marie, S., Rivière, D., Pélégrini-Issac, M., Perlbarg, V., Trebossen, R., Bottlaender, M., Frouin, V., Grigis, A., Orfanos, D. P., Dary, H., … and the CATI Consortium. (2016). CATI: A Large Distributed Infrastructure for the Neuroimaging of Cohorts. *Neuroinformatics*, *14*(3), 253–264. https://doi.org/10.1007/s12021-016-9295-8

Panetto, H., Dassisti, M., & Tursi, A. (2012). ONTO-PDM: Product-driven ONTOlogy for Product Data Management interoperability within manufacturing process environment. *Advanced Engineering Informatics*, *26*, 334–348. https://doi.org/10.1016/j.aei.2011.12.002

Pasquetto, I. (2018). *From Open Data to Knowledge Production: Biomedical Data Sharing and Unpredictable Data Reuses* [PhD Thesis, University of California]. https://escholarship.org/uc/item/1sx7v77r

Patel, M., Koch, T., Doerr, M., Tsinaraki, C., Gioldasis, N., Golub, K., & Tudhope, D. (2005). *Semantic interoperability in digital library systems* (Research Report Project no.507618). DELOS Network of Excellence on Digital Libraries and UKOLN University of Bath, Bath, UK.

Pedrinaci, C., Domingue, J., & de Medeiros, A. K. A. (2008). A core ontology for Business Process Analysis. In S. Bechhofer, M. Hauswirth, J. Hoffmann, & M. Koubarakis (Eds.), *European Semantic Web Conference* (Vol. 5021, pp. 49–64). Springer Berlin Heidelberg.

Penciuc, D., Durupt, A., Belkadi, F., Eynard, B., & Rowson, H. (2014). Towards a PLM Interoperability for a Collaborative Design Support System. *Procedia CIRP*, *25*, 369–376. https://doi.org/10.1016/j.procir.2014.10.051

Peretti, C., Giami, A., Rolland, M., & Ott, M.-O. (2015). *Les évolutions de l'emploi scientifique: Constats et perspectives* (p. 215) [Rapport Public Français]. Inspection générale de l'Administration de l'Éducation nationale et de la Recherche. https://www.vie-publique.fr/rapport/35585-les-evolutions-de-lemploi-scientifique-constats-et-perspectives

Persoon, L., Hoof, S. van, van der Kruijssen, F., Granton, P., Sanchez Rivero, A., Beunk, H., Dubois, L., Doosje, J.-W., & Verhaegen, F. (2019). A novel data management platform to improve image-guided precision preclinical biological research. *The British Journal of Radiology*, *92*(1095), 20180455. https://doi.org/10.1259/bjr.20180455

Pham, C. C. (2017). *Multi-utilisation de données complexes et hétérogènes: Application au domaine du PLM pour l'imagerie biomédicale* [Thèse de Doctorat, Université de Technologie de Compiègne]. http://www.theses.fr/2017COMP2365

Poline, J.-B., Breeze, J. L., Ghosh, S. S., Gorgolewski, K., Halchenko, Y. O., Hanke, M., Helmer, K. G., Marcus, D. S., Poldrack, R. A., Schwartz, Y., Ashburner, J., & Kennedy, D. N. (2012). Data sharing in neuroimaging research. *Frontiers in Neuroinformatics*, *6*. https://doi.org/10.3389/fninf.2012.00009

Prajapati, V., & Dureja, H. (2012). Product lifecycle management in pharmaceuticals. *Journal of Medical Marketing*, *12*(3), 150–158. https://doi.org/10.1177/1745790412445292

Prather, J. C., Lobach, D. F., Goodwin, L. K., Hales, J. W., Hage, M. L., & Hammond, W. E. (1997). Medical data mining: Knowledge discovery in a clinical data warehouse. *Proceedings: A Conference of the American Medical Informatics Association. AMIA Fall Symposium*, 101–105.

Rachuri, S., Subrahmanian, E., Bouras, A., Fenves, S. J., Foufou, S., & Sriram, R. D. (2008). Information sharing and exchange in the context of product lifecycle management: Role of standards. *Computer-Aided Design*, *40*(7), 789–800. https://doi.org/10.1016/j.cad.2007.06.012

Rance, B., Canuel, V., Countouris, H., Laurent-Puig, P., & Burgun, A. (2016). Integrating Heterogeneous Biomedical Data for Cancer Research: The CARPEM infrastructure. *Applied Clinical Informatics*, *7*(2), 260–274. https://doi.org/10.4338/ACI-2015-09-RA-0125

Raynaud, M., Zhang, H., Louis, K., Goutaudier, V., Wang, J., Dubourg, Q., Wei, Y., Demir, Z., Debiais, C., Aubert, O., Bouatou, Y., Lefaucheur, C., Jabre, P., Liu, L., Wang, C., Jouven, X., Reese, P., Empana, J.-P., & Loupy, A. (2021). COVID-19-related medical research: A meta-research and critical appraisal. *BMC Medical Research Methodology*, *21*(1), 1. https://doi.org/10.1186/s12874-020-01190-w

Rich, D. A. (1997). A brief history of positron emission tomography. *Journal of Nuclear Medicine Technology*, *25*(1), 4–11.

Rocca-Serra, P., Brandizi, M., Maguire, E., Sklyar, N., Taylor, C., Begley, K., Field, D., Harris, S., Hide, W., Hofmann, O., Neumann, S., Sterk, P., Tong, W., & Sansone, S.-A. (2010). ISA software suite: Supporting standards-compliant experimental annotation and enabling curation at the community level. *Bioinformatics*, *26*(18), 2354–2356. https://doi.org/10.1093/bioinformatics/btq415

Romero-Reverón, R. (2011). Marcello Malpighi (1628-1694), Founder of Microanatomy. *International Journal of Morphology*, *29*(2), 399–402. https://doi.org/10.4067/S0717-95022011000200015

Rosa, L., Gilberto, F., Daniela, F., & Mauro, G. (2014). ImmunoDB: A web based tool to analyze preclinical data. *Studies in Health Technology and Informatics*, 438–442. https://doi.org/10.3233/978-1-61499-432-9-438

Rosenbloom, S. T., Miller, R. A., Johnson, K. B., Elkin, P. L., & Brown, S. H. (2006). Interface Terminologies: Facilitating Direct Entry of Clinical Data into Electronic Health Record Systems. *Journal of the American Medical Informatics Association : JAMIA*, *13*(3), 277–288. https://doi.org/10.1197/jamia.M1957

Roussey, C., Pinet, F., Kang, M. A., & Corcho, O. (2011). An Introduction to Ontologies and Ontology Engineering. In G. Falquet, C. Métral, J. Teller, & C. Tweed (Eds.), *Ontologies in Urban Development Projects* (Vol. 1, pp. 9–38). Springer London. https://doi.org/10.1007/978-0-85729-724-2_2

Rowley, J. (2007). The wisdom hierarchy: Representations of the DIKW hierarchy. *Journal of Information Science*, *33*(2), 163–180. https://doi.org/10.1177/0165551506070706

Sahoo, S. S., Sheth, A., & Henson, C. (2008). Semantic provenance for eScience—Managing the deluge of scientific data. *IEEE Internet Computing*, *12*(4), 46–54. https://doi.org/10.1109/MIC.2008.86

Sampaio, M., Ferreira, A. L., Castro, J. A., & Ribeiro, C. (2019). Training Biomedical Researchers in Metadata with a MIBBI-Based Ontology. In E. Garoufallou, F. Fallucchi, & E. William De Luca (Eds.), *Metadata and Semantic Research* (pp. 28–39). Springer, Cham. https://doi.org/10.1007/978-3-030-36599-8_3

Sansone, S.-A., McQuilton, P., Rocca-Serra, P., Gonzalez-Beltran, A., Izzo, M., Lister, A. L., & Thurston, M. (2019). FAIRsharing as a community approach to standards, repositories and policies. *Nature Biotechnology*, *37*(4), 358–367. https://doi.org/10.1038/s41587-019-0080-8

Schnöckel, U., Hermann, S., Stegger, L., Law, M., Kuhlmann, M., Schober, O., Schäfers, K., & Schäfers, M. (2010). Small-animal PET: A promising, non-invasive tool in pre-clinical research. *European Journal of Pharmaceutics and Biopharmaceutics*, *74*(1), 50–54. https://doi.org/10.1016/j.ejpb.2009.05.012

Schreiber, A. T., Schreiber, G., Akkermans, H., Anjewierden, A., Shadbolt, N., Hoog, R. de, Velde, W. V. de, & Wielinga, B. (2000). *Knowledge Engineering and Management: The CommonKADS Methodology*. MIT Press, Cambridge, MA.

Schulz, S., Rodrigues, J. M., Rector, A., & Chute, C. G. (2017). Interface terminologies, reference terminologies and aggregation terminologies: A strategy for better integration. In Z. Dongsheng, A. V. Gundlapalli, & J. Marie-Christine (Eds.), *MEDINFO 2017* (pp. 940–944). IOS Press, Amsterdam. https://doi.org/10.3233/978-1-61499-830-3-940

Seyed, A. P. (2009). BFO/DOLCE Primitive Relation Comparison. *Nature Precedings*, 1–1. https://doi.org/10.1038/npre.2009.3481.1

Sharma, N. (2008). *The origin of the data information knowledge wisdom (DIKW) hierarchy* [Research Report]. Google Inc. Mountain View, United States. https://www.researchgate.net/publication/292335202_The_Origin_of_Data_Information_Knowledge_Wisdom_DIKW_Hierarchy

Shongwe, M. M. (2016). An Analysis of Knowledge Management Lifecycle Frameworks: Towards a Unified Framework. *Electronic Journal of Knowledge Management*, *14*(3), 15.

Shvaiko, P., & Euzenat, J. (2013). Ontology matching: State of the art and future challenges. *IEEE Transactions on Knowledge and Data Engineering*, *25*(1), 158–176. http://ieeexplore.ieee.org/abstract/document/6104044/

Simmhan, Y. L., Plale, B., & Gannon, D. (2005). A survey of data provenance in e-science. *Association for Computing (ACM) Machinery Special Interest Group On Management of Data (SIGMOD) Record*, *34*(3), 31. https://doi.org/10.1145/1084805.1084812

Singh, A., & Anand, P. (2013). State of art in ontology development tools. *International Journal of Advances in Computer Science and Technology*, *2*(7), 96--101. http://www.warse.org/pdfs/2013/ijacst01272013.pdf

Singh, R., Chubb, L., Pantanowitz, L., & Parwani, A. (2011). Standardization in digital pathology: Supplement 145 of the DICOM standards. *Journal of Pathology Informatics*, *2*. https://doi.org/10.4103/2153-3539.80719

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L. J., Eilbeck, K., Ireland, A., Mungall, C. J., Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S.-A., Scheuermann, R. H., Shah, N., Whetzel, P. L., & Lewis, S. (2007). The OBO Foundry: Coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology*, *25*(11), 1251–1255. https://doi.org/10.1038/nbt1346

Smith, B. P., Kumar, A., & Bittner, T. (2005). *Basic Formal Ontology for bioinformatics*. IFOMIS Reports. https://philarchive.org/rec/KUMIR

Soldatova, L. N., Aubrey, W., King, R. D., & Clare, A. (2008). The EXACT description of biomedical protocols. *Bioinformatics*, *24*(13), i295–i303. https://doi.org/10.1093/bioinformatics/btn156

Soldatova, L. N., Nadis, D., King, R. D., Basu, P. S., Haddi, E., Baumle, V., Saunders, N. J., Marwan, W., & Rudkin, B. B. (2014). EXACT2: The semantics of biomedical protocols. *Bmc Bioinformatics*, *15*, S5. https://doi.org/10.1186/1471-2105-15-S14-S5

Souza, R. R., Tudhope, D., & Almeida, M. B. (2012). Towards a taxonomy of KOS: Dimensions for classifying Knowledge Organization Systems. *Knowledge Organization*, *39*(3), 179–192.

Spada, F. (2013). *La conduite du changement lors du déploiement d'un système d'information* (p. 33) [Rapport de Master]. http://www.mf-services.ch/dossiers/chmgmt2.pdf

Sriti, M.-F. (2008). *Démarche et logiciel de gestion des connaissances pour le cycle de vie des produits* [Thèse de Doctorat, Université de Technologie de Troyes]. http://www.theses.fr/2008TROY0009

Stark, J. (2015). *Product Lifecycle Management (Volume 1): 21st Century Paradigm for Product Realisation.* Springer, Cham. https://doi.org/10.1007/978-3-319-17440-2

Steckler, T., Brose, K., Haas, M., Kas, M. J., Koustova, E., & Bespalov, A. (2015). The preclinical data forum network: A new ECNP initiative to improve data quality and robustness for (preclinical) neuroscience. *European Neuropsychopharmacology*, *25*(10), 1803–1807. https://doi.org/10.1016/j.euroneuro.2015.05.011

Strickland, N. H. (2000). PACS (picture archiving and communication systems): Filmless radiology. *Archives of Disease in Childhood*, *83*(1), 82–86. https://doi.org/10.1136/adc.83.1.82

Studer, R., Benjamins, V. R., & Fensel, D. (1998). Knowledge engineering: Principles and methods. *Data & Knowledge Engineering*, *25*(1), 161–197. https://doi.org/10.1016/S0169-023X(97)00056-6

Szymanski, J. (2008). *An Integrated Informatics Infrastructure For Pre-Clinical Research-It Support* [PhD Thesis, Case Western Reserve University]. https://etd.ohiolink.edu/apexprod/rws_etd/send_file/send?accession=case1196266590

Tabard, A., Mackay, W. E., & Eastmond, E. (2008). From individual to collaborative: The evolution of prism, a hybrid laboratory notebook. *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work*, 569–578. https://doi.org/10.1145/1460563.1460653

Taira, R. K., Breant, C. M., Chan, H. M., Huang, L., & Valentino, D. J. (1996). Architectural design and tools to support the transparent access to hospital information systems, radiology information systems, and picture archiving and communication systems. *Journal of Digital Imaging*, *9*(1), 1–10.

Tanter, M., & Fink, M. (2014). Ultrafast imaging in biomedical ultrasound. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, *61*(1), 102–119. https://doi.org/10.1109/TUFFC.2014.2882

Taylor, C. F., Field, D., Sansone, S.-A., Aerts, J., Apweiler, R., Ashburner, M., Ball, C. A., Binz, P.-A., Bogue, M., Booth, T., Brazma, A., Brinkman, R. R., Michael Clark, A., Deutsch, E. W., Fiehn, O., Fostel, J., Ghazal, P., Gibson, F., Gray, T., … Wiemann, S. (2008). Promoting coherent minimum reporting guidelines for biological and biomedical investigations: The MIBBI project. *Nature Biotechnology*, *26*(8), 889–896. https://doi.org/10.1038/nbt.1411

Terzi, S., Bouras, A., Dutta, D., Garetti, M., & Dimitris Kiritsis. (2010). Product lifecycle management – from its history to its new role. *International Journal of Product Lifecycle Management*, *4*(4), 360–389. https://doi.org/10.1504/IJPLM.2010.036489

Tony Liu, D., & William Xu, X. (2001). A review of web-based product data management systems. *Computers in Industry*, *44*(3), 251–262. https://doi.org/10.1016/S0166-3615(01)00072-0

Tranquart, F., Correas, J.-M., & Bouakaz, A. (2007). *Echographie de contraste*. Springer-Verlag France.

Turban, E., Rainer, R. K., & Potter, R. E. (2004). *Introduction to Information Technology* (3 edition). Wiley, New York.

Umar, M. M. (2015). A Survey on State-of-the-Art Knowledge-based System Development and Issues. *Smart Computing Review*, *5*(6), 498--509. https://doi.org/10.6029/smartcr.2015 .06.001

Varma, D. (2012). Managing DICOM images: Tips and tricks for the radiologist. *Indian Journal of Radiology and Imaging*, *22*(1), 4. https://doi.org/10.4103/0971-3026.95396

Vreeman, D. J., & McDonald, C. J. (2005). Automated Mapping of Local Radiology Terms to LOINC. *AMIA Annual Symposium Proceedings*, *2005*, 769–773. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1560555/

Whetzel, P. L., Noy, N. F., Shah, N. H., Alexander, P. R., Nyulas, C., Tudorache, T., & Musen, M. A. (2011). BioPortal: Enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. *Nucleic Acids Research*, *39*(suppl_2), W541–W545. https://doi.org/10.1093/nar/gkr469

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*, 160018. https://doi.org/10.1038/sdata.2016.18

Williams, E., Moore, J., Li, S. W., Rustici, G., Tarkowska, A., Chessel, A., Leo, S., Antal, B., Ferguson, R. K., Sarkans, U., Brazma, A., Carazo Salas, R. E., & Swedlow, J. R. (2017). Image Data Resource: A bioimage data integration and publication platform. *Nature Methods*, *14*(8), 775–781. https://doi.org/10.1038/nmeth.4326

Yamoah, G. G., Cao, L., Wu, C. W., Beekman, F. J., Vandeghinste, B., Mannheim, J. G., Rosenhain, S., Leonardic, K., Kiessling, F., & Gremse, F. (2019). Data Curation for Preclinical and Clinical Multimodal Imaging Studies. *Molecular Imaging and Biology*. https://doi.org/10.1007/s11307-019-01339-0

Zablith, F., Antoniou, G., d'Aquin, M., Flouris, G., Kondylakis, H., Motta, E., Plexousakis, D., & Sabou, M. (2015). Ontology evolution: A process-centric survey. *The Knowledge Engineering Review*, *30*(1), 45–75.

Zaidi, H. (2007). Optimisation of whole-body PET/CT scanning protocols. *Biomedical Imaging and Intervention Journal*, *3*, e36. https://doi.org/10.2349/biij.3.2.e36

Zapletal, E., Rodon, N., Grabar, N., & Degoulet, P. (2010). Methodology of integration of a clinical data warehouse with a clinical information system: The HEGP case. *Studies in Health Technology and Informatics*, *160*(Pt 1), 193–197.

Zhang, W. Y., & Yin, J. W. (2008). Exploring Semantic Web technologies for ontology-based modeling in collaborative engineering design. *The International Journal of Advanced Manufacturing Technology*, *36*(9), 833–843. https://doi.org/10.1007/s00170-006-0896-5

Zunner, C., Bürkle, T., Prokosch, H.-U., & Ganslandt, T. (2012). Mapping local laboratory interface terms to LOINC at a German university hospital using RELMA V.5: A semi-automated approach. *Journal of the American Medical Informatics Association*, *20*(2), 293–297. https://doi.org/10.1136/amiajnl-2012-001063

Notice bibliographique

Les communications scientifiques qui ont eu lieu suite aux travaux présentés dans cette thèse sont décrites ci-après.

1. Revue à comité de lecture

Raboudi, A., Allanic, M., Balvay, D., Hervé, P.-Y., Durupt, A., Viel, T., Yoganathan, T., Certain, A., Boutinaud, P., Tavitian, B., & Eynard, B. (2021). BMS-LM: a core ontology for semantic interoperability, data reporting and lifecycle management throughout a biomedical research study. *Journal of Biomedical Informatics-submitted*.

2. Conférences internationales

Raboudi, A., Allanic, M., Hervé, P.-Y., Balvay, D., Durupt, A., Boutinaud, P., Tavitian, B., & Eynard, B. (2021). *Implementation of a Product Lifecycle Management framework for Biomedical Research*. IFIP International Conference on Product Lifecycle Management 11-14 July 2021, Curitiba, Brazil- submitted.

Raboudi, A., Yoganathan, T., Viel, T., Allanic, M., Balvay, D., & Tavitian, B. (2019, March). *Integrating PET-CT imaging research workflow with end-to-end traceability for small animal research*. European Molecular Imaging Meeting – EMIM'2019, 19-22 March, Glasgow, UK.

3. Conférences nationales et séminaires doctoraux

Raboudi, A. (2018, November). *A study of Information System mutations (changes) using Knowledge Organization System (KOS) applied to biomedical research study*. Doctoral workshop of 21st International Conference on Knowlege Engineering and Knowlege Management – EKAW'2018, Nancy, France.

Raboudi, A., & Allanic, M. (2018). *Mise en place d'un SI de gestion de cycle de vie d'une étude au sein d'un laboratoire de recherche biomédicale: Retour d'expérimence du projet DRIVE-SPC*. Congrès INFormatique des ORganisations t Systèmes d'Information et de Décision – INFORSID'2018, 28-31 mai, Nantes, France.

Raboudi, A., Allanic, M., Hervé, P.-Y., Boutinaud, P., Durupt, A., Balvay, D., & Eynard, B. (2017). *Traçabilité de l'intégration de données biomédicales hétérogènes dans le système SWOMed de gestion du cycle de vie des études biomédicales*. Symposium Ingénierie de l'Information Médicale - SIIM2017, 23-24 Novembre, Toulouse, France.

Raboudi, A., Allanic, M., Hervé, P.-Y., Balvay, D., Sourdon, J., Boutinaud, P., & Tavitian, B. (2017). *Integration and provenance control of proteomics data using SWOMed, a Product Lifecycle Management framework for biomedical research*. Conférence de Spectrométrie de Masse, Métabolomique et Fluxomique & Electrophorèse et Analyse Protéomique – SMMAP'2017, 2-5 octobre, Marne la Vallée, France.

ANNEXES

ANNEXE A : INTERFACES GRAPHIQUES COMMENTÉES DES CLIENTS DE LA PLATEFORME TEAMCENTER

Les applications clientes, fournies par le logiciel Teamcenter, sont au nombre de trois :

- Un client bureautique appelé « Client Riche » qui est installé sur une machine spécifique et utilisable par plusieurs personnes sans risque de sécurité.
- Un client web appelé « Client Léger » qui est une transcription du client bureautique, mais via le navigateur (ceci a été progressivement délaissé par Siemens)
- Un client web « moderne » appelé AWC (Active Workspace Client Il est une version « nouvelles technologies » et qui se veut ergonomique et compatible avec n'importe quels navigateurs et appareils utilisés.

Les différentes réalisations explicitées dans ce manuscrit ont été réalisées principalement via le « Client Riche » et occasionnellement via le « Client Web AWC ».

L'INTERFACE DU « CLIENT RICHE »

La Figure 147 ci-après présente l'interface du Client Riche de la plateforme Teamcenter. Beaucoup d'informations et d'outils sont présents dans cette interface. Elle est répartie en trois zones verticalement du gauche à droite. Premièrement, dans la zone tout à gauche, il y a la possibilité d'effectuer une recherche rapide pour trouver les objets qui nous intéressent si nous connaissons leur nom ou une partie. Après, il y a une liste personnalisée pour la personne connectée : les projets, les tâches, les recherches sauvegardés, etc. Il y a aussi une liste d'éléments ouverts pendant la session en cours (Open Items) ainsi qu'un historique d'éléments ouverts (History). En bas de la zone, différents outils sont mis à la disposition de l'utilisateur pour effectuer des requêtes, explorer les projets, et administrer la plateforme s'il en a les droits (Query Builder, Classification et Classification Admin, Organisation, Process, Access Manager, etc.).

Dans la zone du centre s'affichent les éléments ouverts pendant la session en cours. Par défaut, le dossier « Home » de l'utilisateur en cours est affiché. Cette zone est régie par des onglets qui s'ajoutent à chaque ouverture d'un nouvel élément. Pour chaque élément de l'onglet, on peut explorer ses relations descendantes (voir l'exemple « EXD_proteomics_LC-MS_3P5_2015 »). À chaque fois où un élément est sélectionné, ses informations sont affichées dans la dernière zone tout à droite qui est la zone d'exploration, affichage et visualisation des données.

Dans la dernière zone à droite, il y a aussi des onglets qui permettent d'avoir des informations sur l'élément sélectionné. Dans l'onglet « Summary » par exemple on trouve la description de l'élément, sa date de modification, la personne qui a effectué la modification. Si l'élément est classifié, on trouve la classe de classification qui le décrit dans « Classification Properties ».

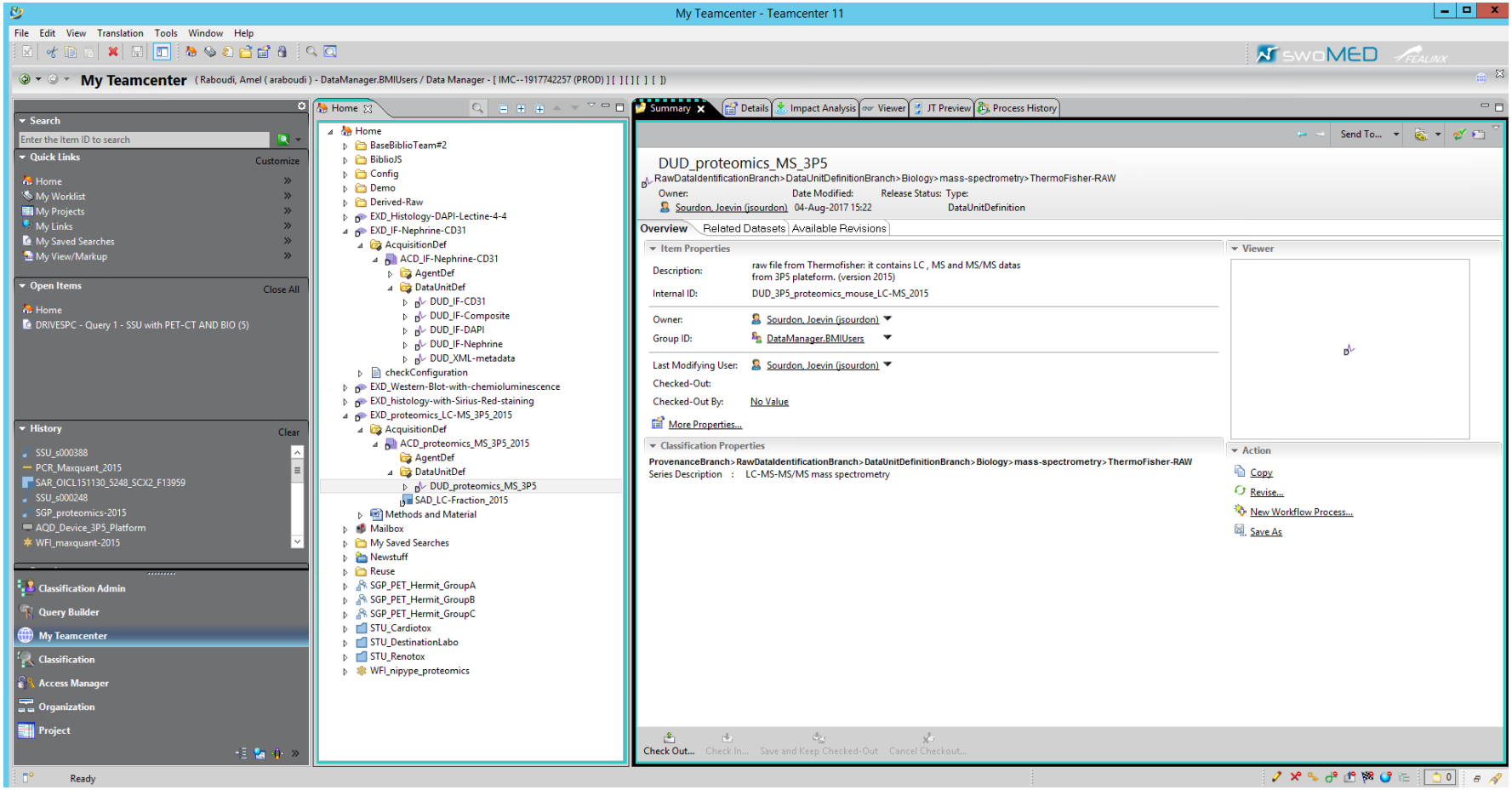


Figure 147 Interface d'accueil du client Riche de la plateforme Teamcenter

L'INTERFACE D'EXPLORATION DE LA « CLASSIFICATION »

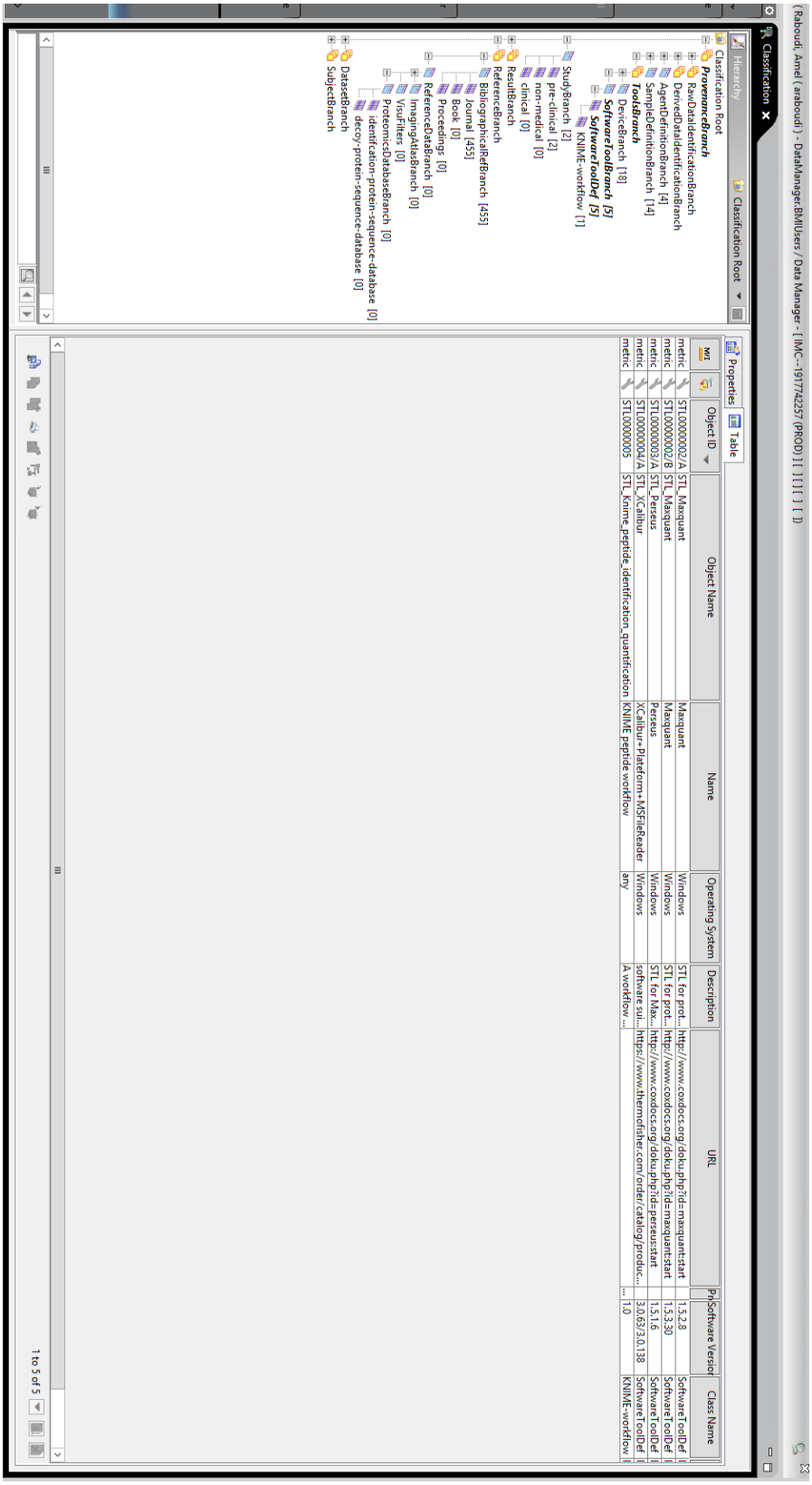


Figure 148 Interface d'exploration des classes et instances de la « Classification » du système BMS-LM

L'interface d'exploration de la classification est présentée Figure 148. À droite, il y a la possibilité d'explorer l'arbre de classification et de sélectionner des classes. Dans la capture d'écran, la classe

sélectionnée est « SoftwareToolDef », une classe associée à l'objet STL du MDD. À gauche, il y a la liste des instances de la base de données qui sont classifiés en utilisant la classe sélectionnée.

L'INTERFACE DU « CLIENT WEB AWC »

La Figure 149 représente la page d'accueil du client AWC de la plateforme Teamcenter. Nous pouvons y trouver un ensemble de cases sous forme de carré. Chaque case représente un service fourni par la plateforme. Par exemple, Il y a une boîte de réception où les tâches affectées par d'autres collaborateurs, ou en attente, sont comptabilisés (6/6/0). Nous pouvons avoir accès au dossier accueil personnalisé du client riche. Il y aussi une option d'exploration des recherches sauvegardées et des données dont la personne a accès.

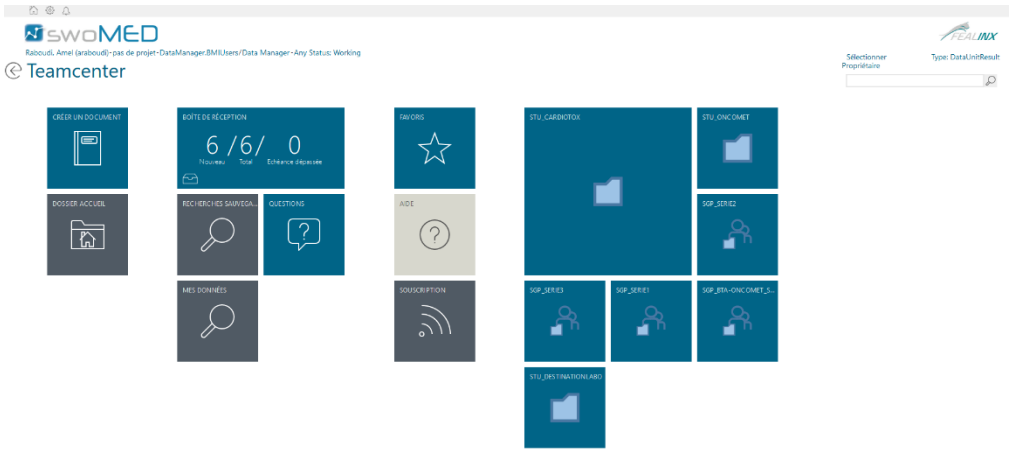


Figure 149 Page d'accueil du client web AWC

En plus des cases par défaut, il y a la possibilité d'épingler ceux qu'on utilise souvent, par exemple, l'objet « SGP_Serie2 » représente « un groupe de sujet dans une étude ». De plus, en haut à droite, il y a une case de recherche rapide qui est aussi accessible par toutes les pages web du « Client Web AWC ». La Figure 150 montre les relations que le SGP_Serie2 a avec d'autres objets dans la base de données. Il est entre autres lié aux sujets dans l'étude (SSU) qui le constituent.

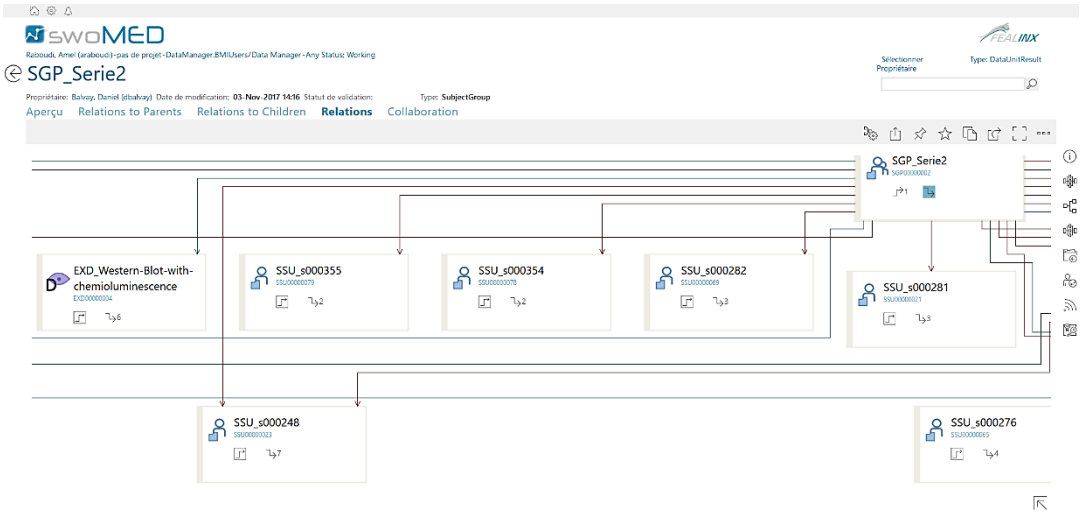


Figure 150 Interface d'exploration des relations du client web AWC

ANNEXE B : BESOINS DE LA COMMUNAUTÉ ET SYSTÈME BMS-LM

Dans cette annexe, nous expliquons en détail les réponses aux besoins de la communauté scientifique dans le domaine biomédical. Ces réponses sont données par la plateforme BIOMIST, version antérieure au système BMS-LM et par le système BMS-LM proposé dans cette thèse.

À L'ISSUE DU PROJET BIOMIST

Les réalisations effectuées par les travaux antérieurs lors de l'application de la gestion de cycle de vie à la neuroimagerie ont été réutilisées et adaptées pour le système BMS-LM. Les besoins déjà satisfaits sont les besoins d'archivage (B1), partage (B5), traçabilité (B8), sécurité (B18), automatisation (B10), vérification (B16), reporting (B17), fonctionnalités natives des plateformes PLM. Des configurations ont été ajoutées comme : les droits d'accès, et des noms de « volumes » ou « coffre-fort », les statuts, etc.

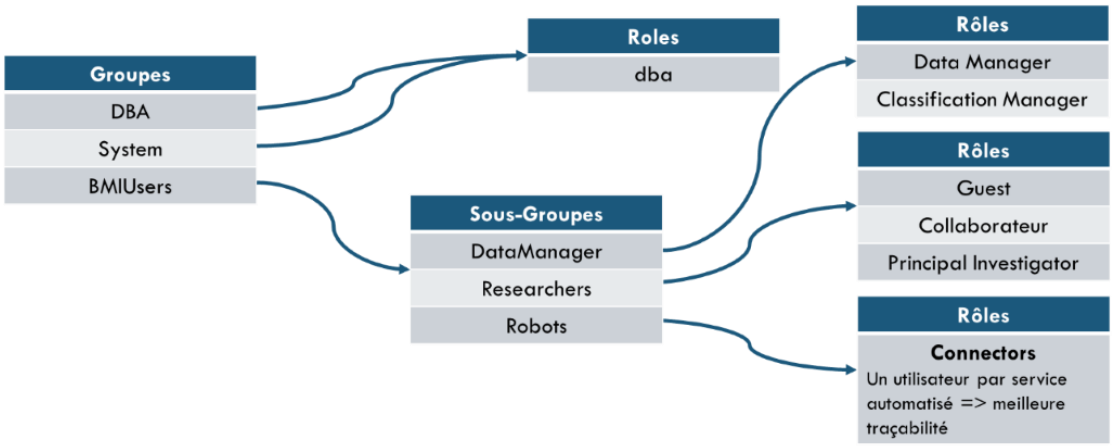


Figure 151 Les différents groupes, sous-groupes et rôles dans le système BMS-LM

La Figure 151 montre les 4 groupes, 3 sous-groupes et 7 rôles qui ont été ajoutés pour sécuriser l'accès aux données via le système BMS-LM. Les groupes « DBA » et « System » sont des groupes techniques utilisés par l'administrateur. Le sous-groupe « Researchers » est un groupe d'utilisateurs avec des rôles distincts en fonction de l'implication dans les travaux de recherche (« Guest », « Collaborateur », « Principal Investigator »). Un groupe spécifique d'administrateurs de projets pour la création et la modification des projets.

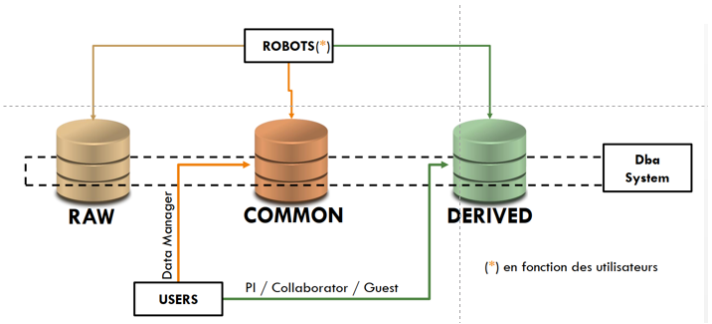


Figure 152 Les volumes (ou coffres-forts) définis pour stocker les données d'une étude de recherche

Les coffres de données sont importants, car ils sont aussi une variable des droits et accès et de la sécurité de la plateforme en général. 3 coffres de données ont été définis pour les données des études de recherche.

- Le coffre des données brutes ou RAW qui est destiné au stockage des données brutes inchangeables et primordiales pour les études de recherche. Ce stockage est archivé et répliqué plus que les deux autres.
- Le coffre des données dérivées ou DERIVED. Ce stockage gère les données dérivées d'analyse ou de traitement. Des données importantes pour la publication scientifique.
- Le coffre des données communes : qui sont des données partageables avec les autres personnes utilisant la plateforme sans risque de sécurité.

Le besoin en import (B2) a été satisfait avec la mise en place de l'import DICOM. L'utilisateur peut sélectionner ses données depuis un système PACS et les envoyer directement à la plateforme BMS-LM comme dans la Figure 153. Un exemple d'objets créé dans le système BMS-LM est donné en Figure 154. Les fichiers de la série DICOM sont stockés dans le dataset (en contour rouge) « DICOM_Series ».

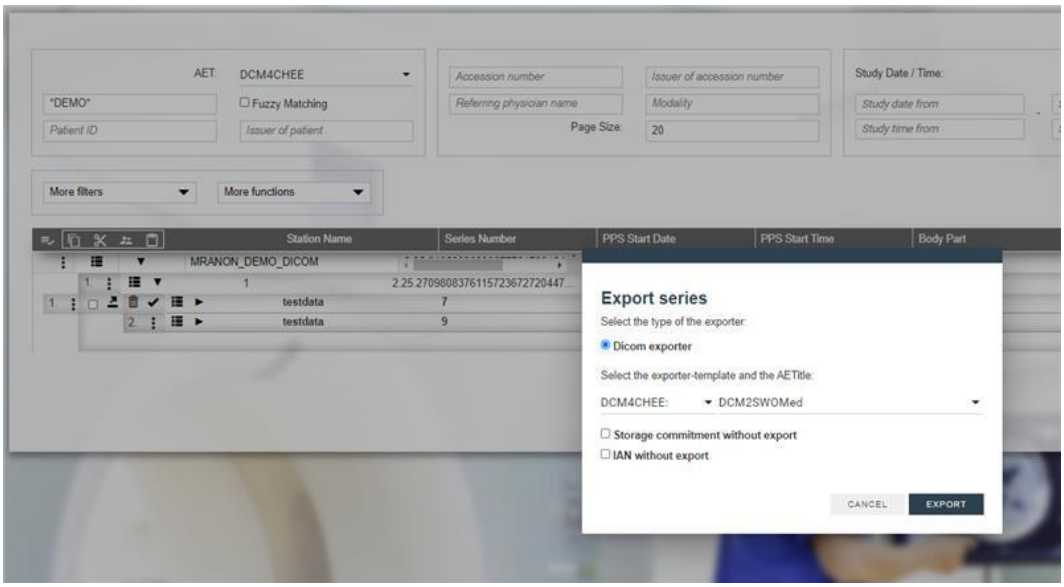


Figure 153 Interface du PACS DCM4CHEE depuis laquelle une série DICOM est envoyée au système BMS-LM

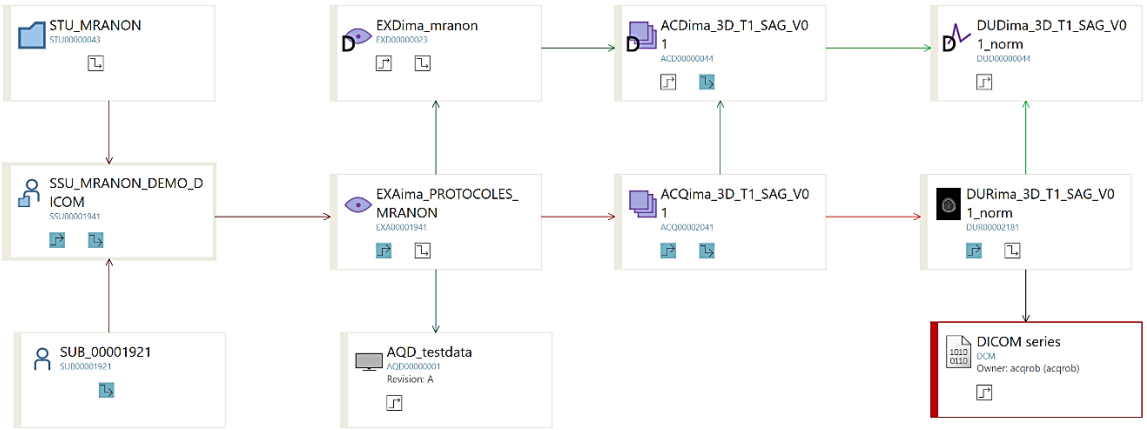


Figure 154 Objets du MDD créés lors de l'envoi d'une série DICOM au système BMS-LM

Les modules d'export (B3), et de requête (B4), quant à eux, nécessitent une configuration supplémentaire qui dépend du métier, du domaine, ce qui permet de répondre aux demandes formulées par les utilisateurs. Le besoin en export (B3) est réalisé via un export personnalisé des données dans une base de données SQL Server pour permettre la consultation et l'analyse dans des outils statistiques spécifiques (Tableau, Excel, GraphPad, etc.).

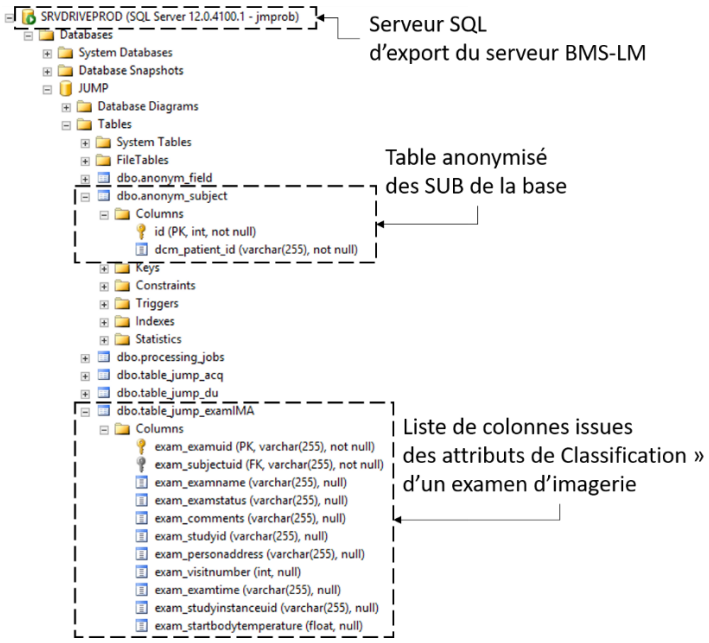


Figure 155 Serveur SQL d'export de données depuis le système BMS-LM

Le besoin en requête (B4) est satisfait en premier plan par les outils de recherche de la plateforme PLM native (Teamcenter). La capture d'écran Figure 156 ci-après montre l'interface de construction des requêtes personnalisées utilisée.

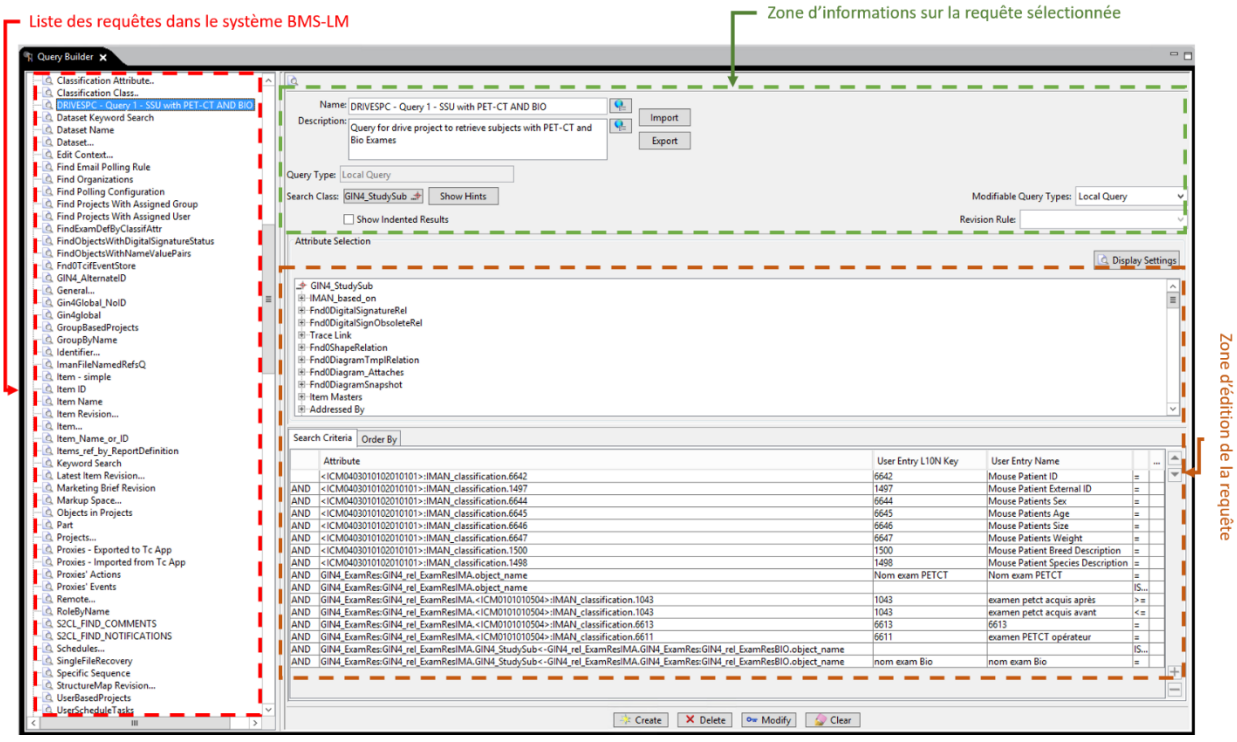


Figure 156 Interface de construction de requêtes personnalisées

La Figure 157 montre le résultat de l'exécution de la requête Figure 156 : la liste des SSU « male » qui ont à la fois des examens TEP-TDM et des examens biologiques (BIO) (la condition est formulée dans la zone d'édition des requêtes de la Figure 156).

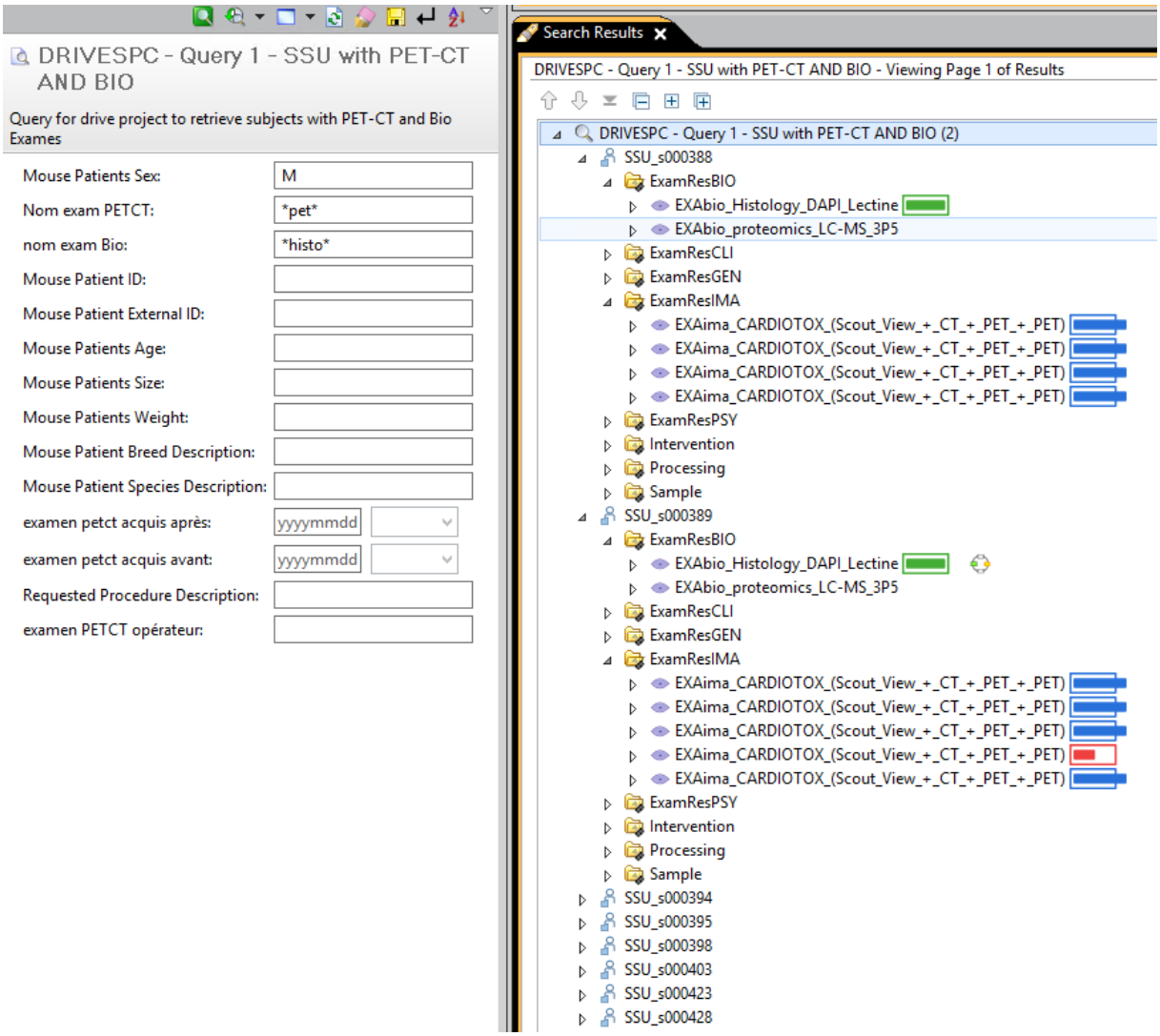


Figure 157 Résultat de l'exécution d'une requête personnalisée dans le système BMS-LM

L'export en une base de données SQL serveur, peut aussi servir pour l'exploration et la recherche des données dans le système BMS-LM via Excel en utilisant son module de connexion « PowerQuery ». Une requête depuis Excel a été exécutée pour récupérer les données de l'étude Cadiotox. Son résultat est dans la capture d'écran Figure 158 ci-après.

Figure 158 Requête PowerQuery dans Excel pour récupérer les informations sur les examens de l'étude Cardiotox

La dernière option de requête a été proposée lors de la thèse de (Pham, 2017). Il s'agit d'une méthode de requête de données hétérogènes et complexes en utilisant l'ontologie appelée « VAQUERO » pour « VisuaAlization and QUERy based Ontology ». L'interface de requête proposée lors de sa thèse est présente dans la Figure 159. Source : (Allanic et al., 2017).

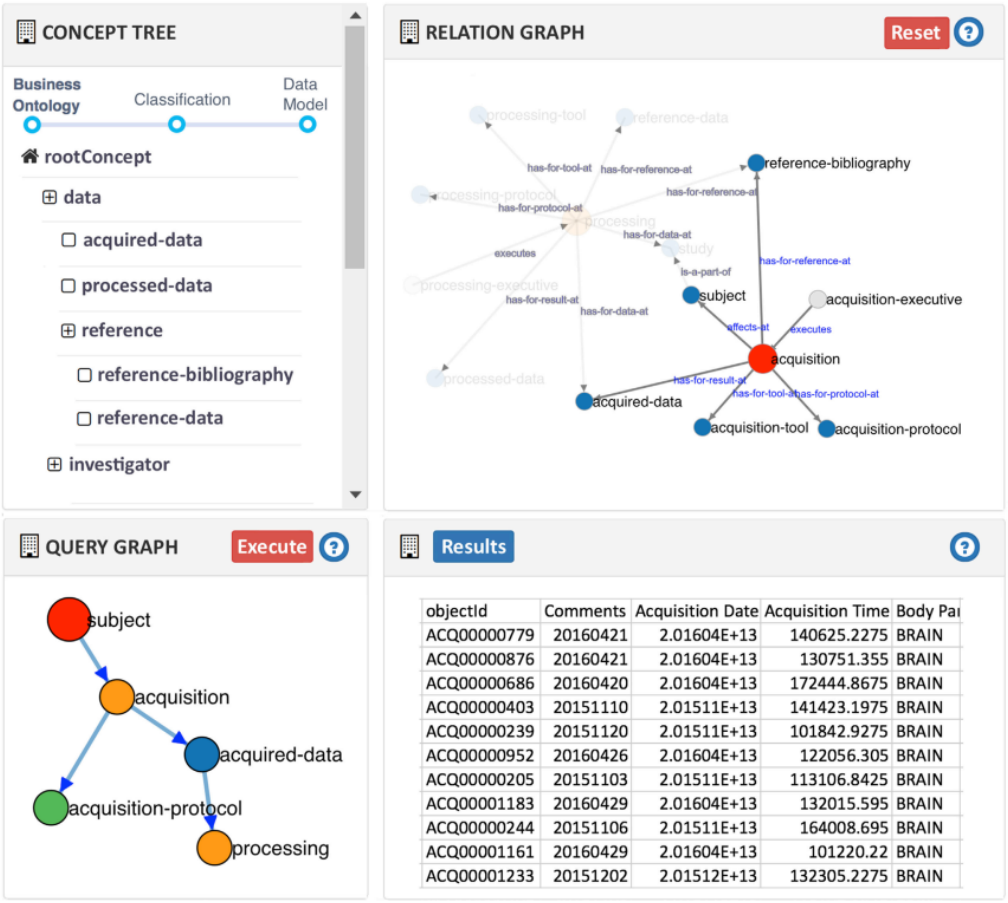


Figure 159 Interface de requête « VAQUERO » basé sur l'utilisation d'ontologies pour la formulation de requêtes (Allanic et al., 2017)

Les besoins en automatisation (B10) et analyse (B6) ont été satisfaits en partie via l'intégration du gestionnaire du workflow Nipype (Gorgolewski et al., 2011) pour l'automatisation des analyses des données en neuroimagerie. L'envoi de données à des clusters de calculs, et leur récupération se font depuis le système BMS-LM comme expliqué dans la Figure 160 ci-après. Reste à identifier les analyses pertinentes pour le contexte d'application.

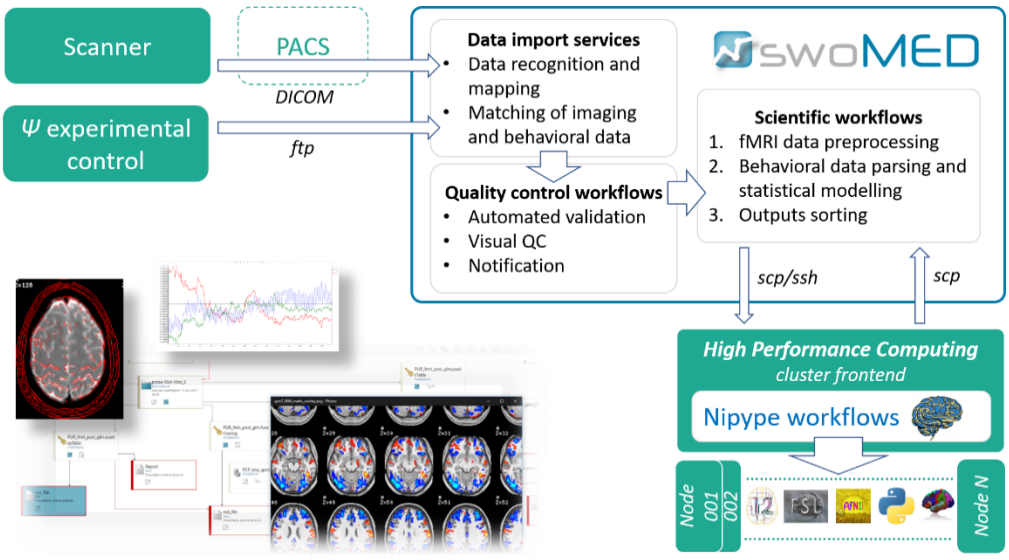


Figure 160 L'intégration des calculs scientifiques via Nipype en neuroimagerie dans le cadre du projet BIOMIST

Les besoins en standardisation (B11), évolutivité (B14), et flexibilité (B15) ont été en partie satisfaits lors du projet BIOMIST via la proposition du modèle de données BMI-LM et la réutilisation des ontologies de domaines pour la structuration de données en neuroimagerie dans l'arbre de « Classification ». Cet arbre est évolutif et flexible.

Pour le besoin en partage (B5), les plateformes PLM ont des fonctionnalités de base comme l'envoi d'un pseudo-email au sein du logiciel (voir Fenêtre « New Envelope » de la Figure 161) qui trace les différents partages. D'autres outils intéressants pour le travail collaboratif sont aussi disponibles comme l'attribution de taches aux autres collaborateurs de l'étude ou la souscription à un objet de la base de données afin de recevoir des notifications à chaque modification.

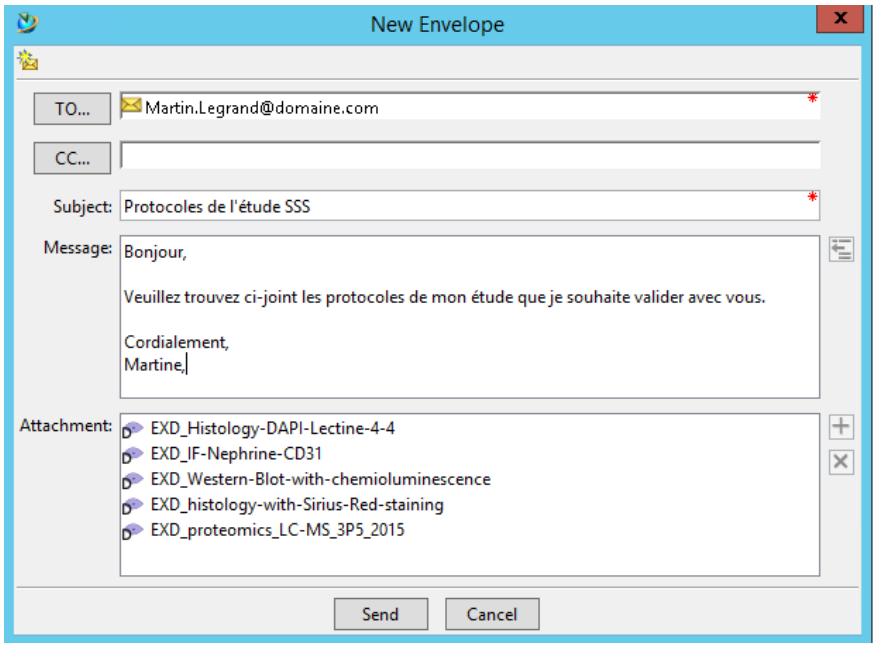


Figure 161 Partage d'objets au sein du système BMS-LM

LES NOUVEAUTÉS DANS LE SYSTÈME BMS-LM

En lien avec les besoins simplicité(B9), ergonomie(B12), efficacité(B13). Une exploration des données via l'utilisation d'un client personnalisé et de l'API REST a été mise en place pour le laboratoire LRI. Ce client a été proposé dans le cadre des développements pour Mediso2PLM v4 ; intégration de données TEP-TDM. Il permet à un utilisateur de naviguer dans les études auxquelles il a accès et d'afficher les images et les informations correspondantes à un SSU par exemple d'une manière simplifiée.

Figure 162 Interface d'accueil de l'outil SWOPARCC listant les études auxquelles la personne est autorisée

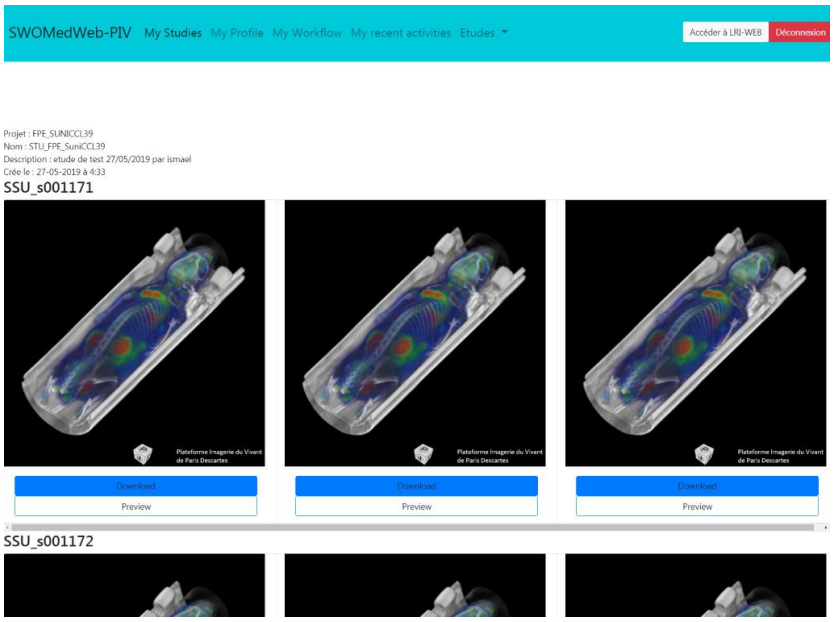


Figure 163 Interface web non finalisée pour le téléchargement d'images d'intérêt depuis le système BMS-LM

Le besoin en reporting (B17) personnalisé peut être satisfait via l'utilisation de l'outil de reporting native de la plateforme PLM Teamcenter comme dans la Figure 164.

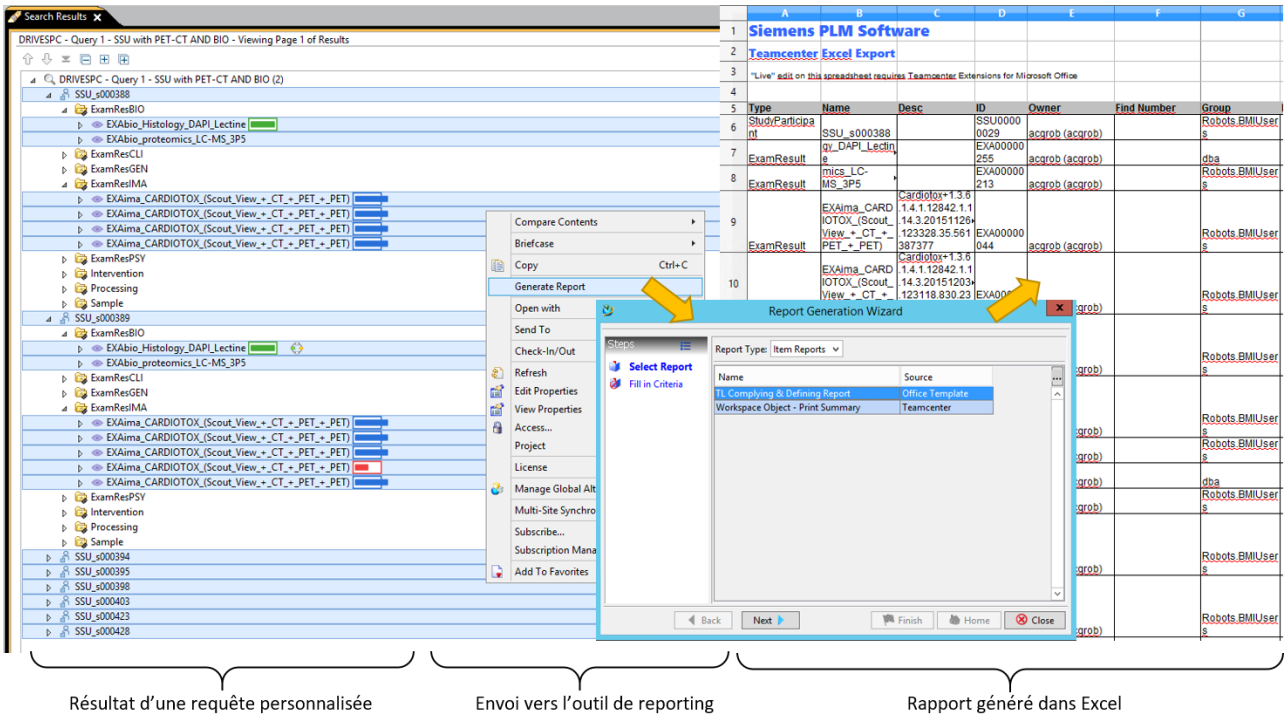


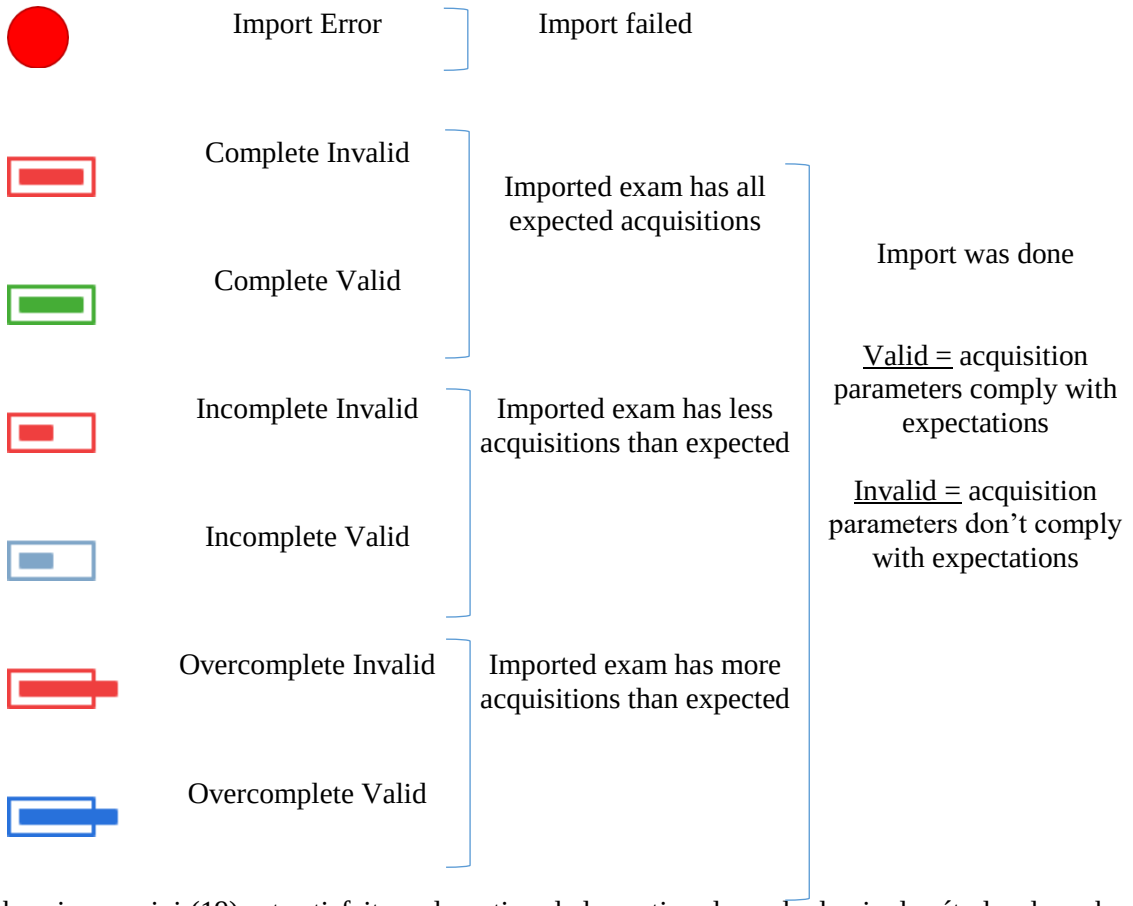
Figure 164 Génération de rapports depuis le système BMS-LM

Pour le besoin en vérification (B16), des statuts ont été définis dans le système BMS-LM afin de valider les données brutes et les données dérivées. Elles sont présentées dans les tableaux suivants (Tableau 28, Tableau 29). Les statuts « utilisable » ou non, sont ajoutés manuellement par l'utilisateur une fois le calcul exécuté pour vérifier s'il est bien réutilisable. Tous les autres statuts sont générés automatiquement lors d'un import de données ou d'une exécution d'un traitement.

Tableau 28 Liste des statuts attribués aux données dérivées dans le système BMS-LM

	Failed	Le calcul n'a pas abouti
	Computed	Le calcul a été exécuté comme prévu
	Unusable	Les résultats de calcul ne peuvent pas être exploités
	Usable	Les résultats de calcul peuvent être exploités

Tableau 29 liste des statuts attribués aux données brutes dans le système BMS-LM



Le besoin en suivi (19) est satisfait par la notion de la gestion de cycle de vie des études de recherche (BMS-LM). En effet, ce paradigme suit les activités du chercheur et permet de les tracer. Le collage de captures d'écran ci-après montre des données protéomiques importées dans le système BMS-LM avec liens de traçabilité de la spécification de l'étude jusqu'à sa publication, il s'agit des données de l'étude « Cardiotox ». Pareillement, le schéma de la montre une modélisation du flux d'acquisition et flux de traitement des données TEP-TDM en utilisant le MDD et la Classification (liste d'attributs encadrés en jaune) du système BMS-LM.

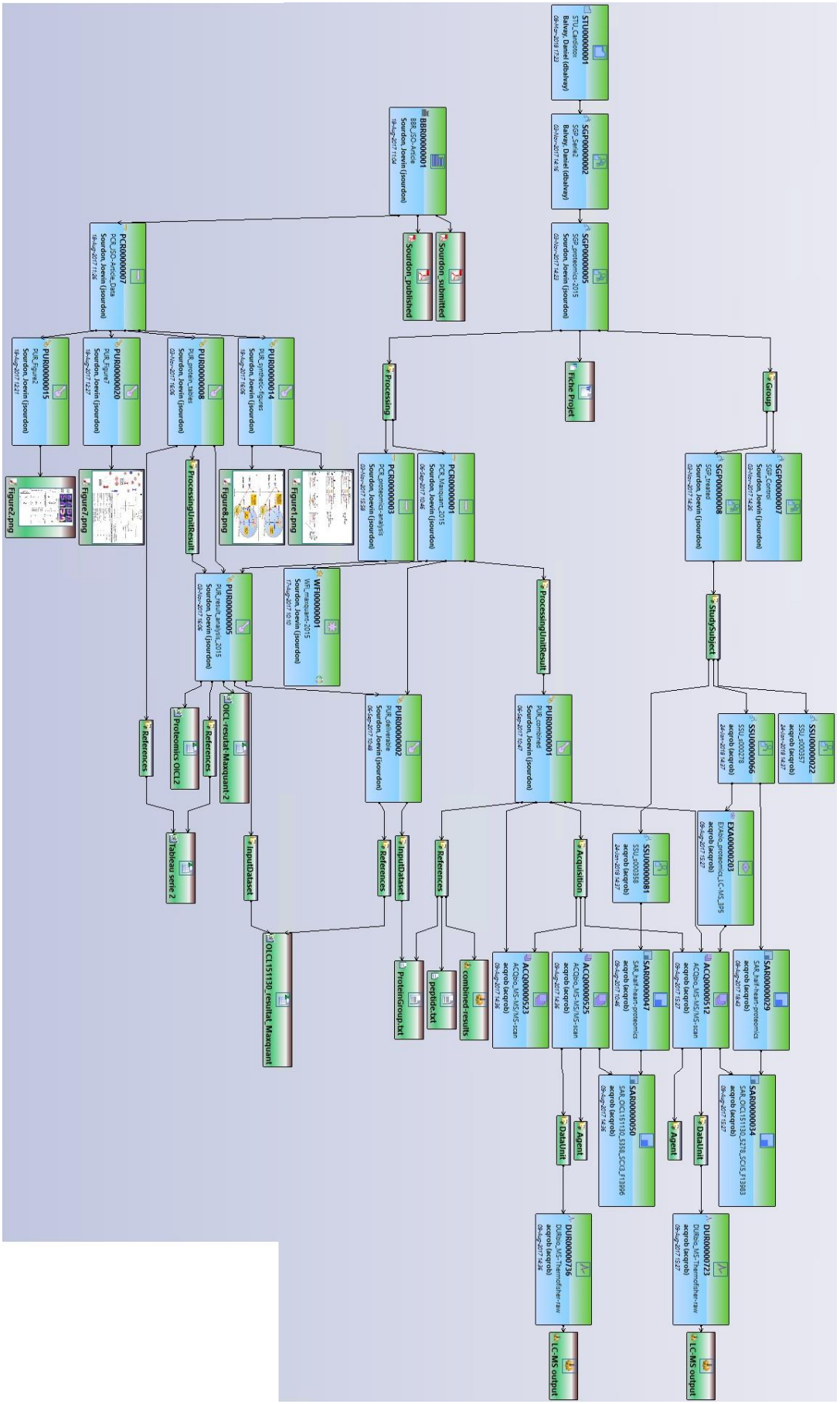


Figure 165 Données en protéomique tracées depuis la spécification de l'étude jusqu'à la publication dans le système BMS-LM

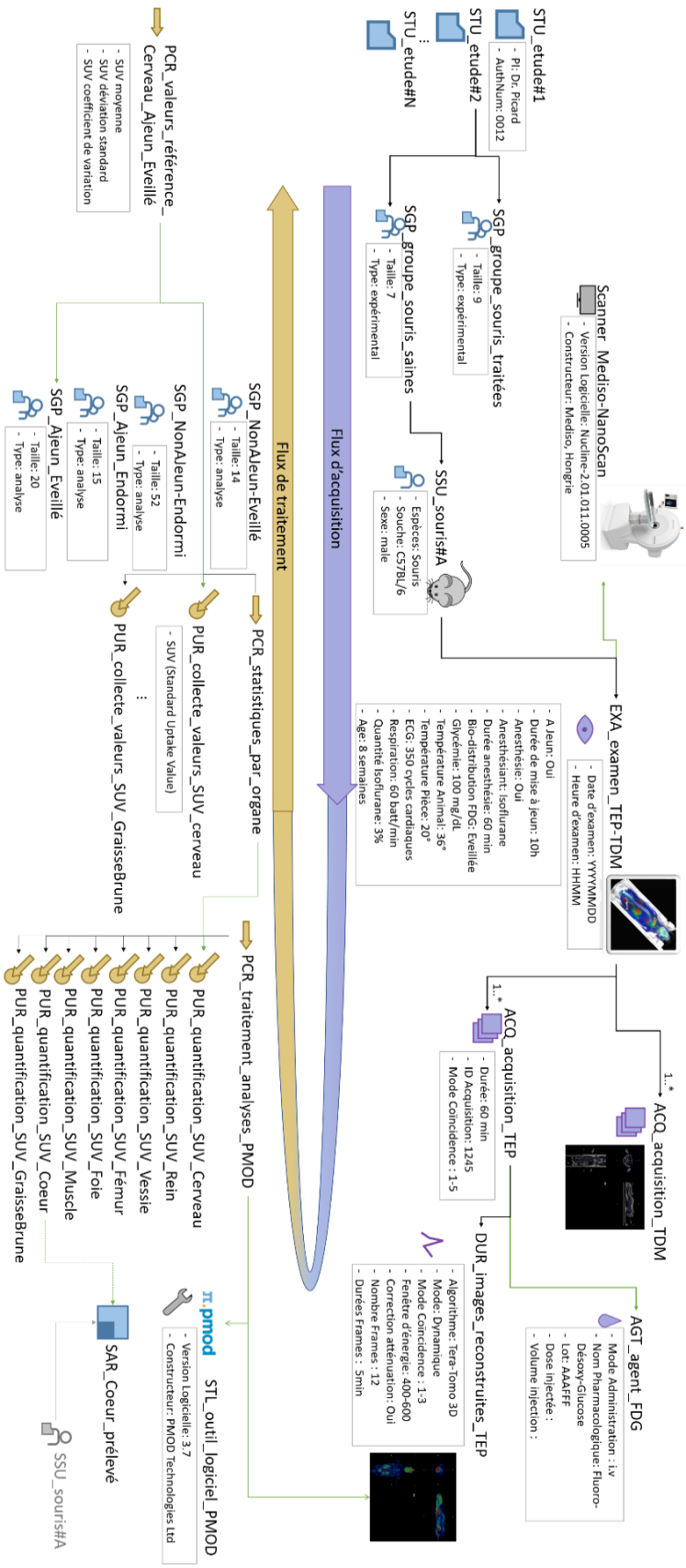


Figure 166 Flux d'acquisition et flux de traitement des données TEP-TDM

Pour le besoin en réutilisation (B7), nous avons documenté dans le cadre du projet DRIVE-SPC, une réutilisation de données entre deux projets de recherche en utilisant les fonctionnalités natives des plateformes PLM : des données d'examen TEP-TDM d'une souris dans le projet « Oncomet » réutilisées pour le projet « Cardiotox », comme le montre Figure 167.

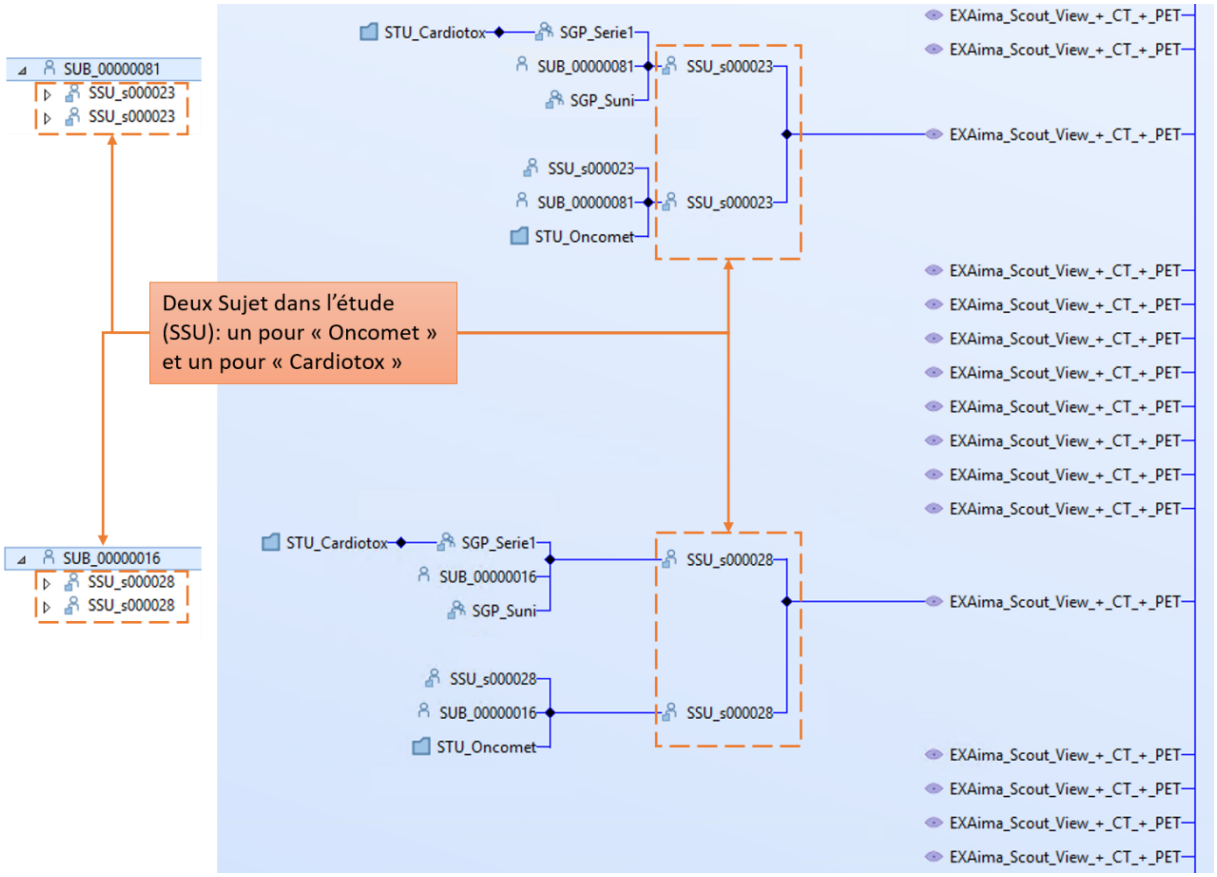


Figure 167 du système BMS-LM montrant deux examens de l'étude « Oncomet » réutilisés pour « Cardiotox »

ANNEXE C : ANALYSE DES KOS PUBLIÉS POUR LA CONSTRUCTION DE L'ONTOLOGIE BMS-LM

Dans cette annexe, nous présentons les différentes réflexions menées dans le cadre de la construction de l'ontologie BMS-LM. Nous présentons tout d'abord les scripts et algorithmes que nous avons utilisés pour l'identification des équivalences dans des KOS publiés (au niveau noyau et au niveau domaine) et nous expliquons après les exemples OBO sur lesquels est basée notre approche ADQIV. Pour finir, les supports des tests utilisateur sont donnés en troisième partie.

SCRIPT 1 : ANALYSE DES KOS DANS BIOPORTAL POUR CONSTRUIRE LES LIENS BMS-LM – BFO

Pour établir les alignements haut-noyau entre l'ontologie BFO et l'ontologie noyau BMS-LM, un script explorant les ontologies du BioPortal « script1 » a été utilisé (voir Figure 168). Il a été conçu via l'outil Knime[85] d'analyse scientifique, qui est une suite de fonctions (représentées via des boîtes) avec des entrées-sorties (représentées via des flèches). Dans la Figure 168, des noms explicatifs ont été donnés à chaque étape pour les expliquer. Le « script1 » a comme entrée les concepts génériques du MDD BMS-LM et donne comme résultat une liste de termes potentiellement équivalents dans les ontologies BioPortal ainsi que leurs parents et définitions respectives. La Figure 169 montre un exemple de résultat de réponse en utilisant le terme « study ».

Plusieurs choix de conception ont été faits en fonction de l'analyse des résultats du « script1 ». Tout d'abord, les choix de parents BFO des classes BMS-LM suivent autant que possible le consensus de la communauté sur les parents d'un concept. Deuxièmement, le changement de nom des concepts BMS-LM, et troisièmement la réutilisation des définitions déjà disponibles dans d'autres ontologies.

SCRIPT 2 : ANALYSE DES KOS DANS BIOPORTAL POUR IDENTIFIER LES CORRESPONDANCES AU NIVEAU DOMAINE

La première règle de conception de l'ontologie noyau BMS-LM est la réutilisation maximale des ontologies existantes et l'interopérabilité avec celles-ci. Un script explorant les ontologies du BioPortal « script2 » a été utilisé à cette fin (voir Figure 170). Comme « script 1 », il a été conçu via l'outil Knime d'analyse scientifique et des noms explicatifs ont été attribués à ces nœuds. Il aide à construire le niveau ontologique du domaine pour le contexte d'application. Pour l'utiliser, des termes ont été collectés à l'aide d'entretiens et de cartes mentales. Ils représentent le « lot de termes initiaux » et ont été fournis en tant qu'entrée au « Script2 » qui explore les ontologies du BioPortal et fournit comme sortie la liste des concepts de domaine identifiés, avec leur parent direct, leurs enfants, et les définitions. La Figure 171 montre un exemple de résultat de réponse en utilisant le terme « ».

[85] https://www.knime.com/

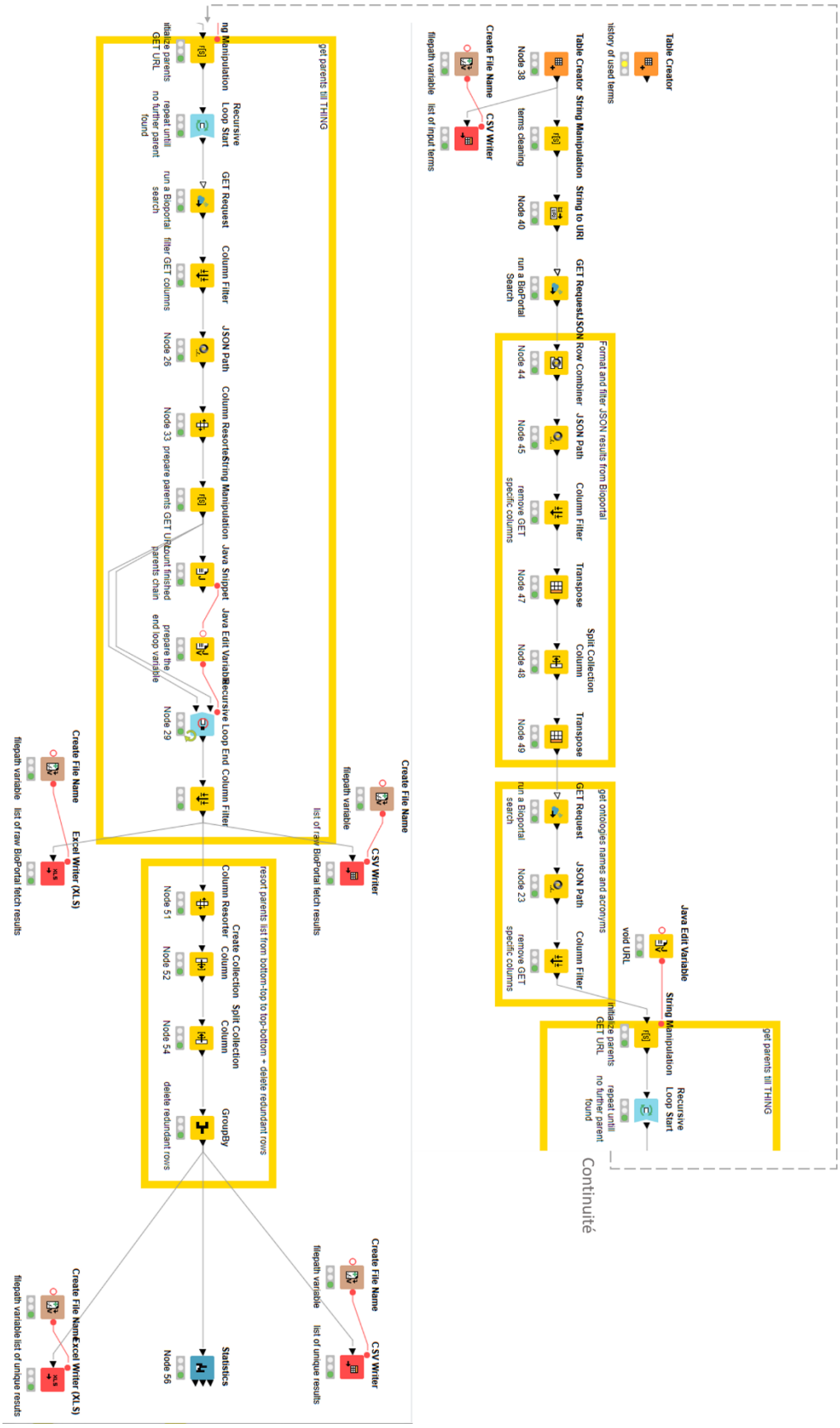


Figure 168 Capture d'écran du script1 codé en KNIME et utilisé pour l'exploration BioPortal des termes noyau

Liste de termes « noyau » d'entrée au script1 de recherche sur Bioportal

Liste des parents depuis la racine au parent direct

	prefLabels	definitions	name	acronym	parent (#24)	parent (#13)	parent (#12)	parent (#11)
1	study							
2	study subject		SNOMED CT	SNOMEDCT	Home	Location within home premises	Location inside building	Room of building
3	study participant		Logical Observation Identifier Names and Codes	LOINC				
4	research project		Read Codes, Clinical Terms Version 3 (CTV3)	RCD	Home environment	Location within home premises	Location inside building	Room of building
5	contrast agent	["A detailed examination, analysis, National Cancer Institute Thesaurus		NCIT	Activity			
6	biological sample		Children's Health Exposure Analysis Resource	CHEAR	entity	Clinical or Research Activity	procedure	investigation
7	chirurgical intervention		Children's Health Exposure Analysis Resource	CHEAR	Science Indicator	entity	process	Activity
8	exam	["A study is a process that realizes Semanticscience Integrated Ontology		SIO	entity	process	procedure	investigation
9	examination		Radiation Oncology Ontology	ROO	Idea or Concept	Functional Concept	DICOM Functional Concept	InformationEntity
10	acquisition	["A detailed examination, analysis, Radiation Oncology Ontology		ROO	Event	Activity	Occupational Activity	Research Activity
11	acquisition protocol		The Drug-Drug Interactions Ontology	DINTO	Information Resource Data_Resource		Clinical_Research_Data	
12	study acquisition	["<ncicp:ComplexDefinition> NanoParticle Ontology		NPO	entity	occurrent	processual entity	process
13	image acquisition	["a planned process that consists (Human Physiology Simulation Ontology		HUPSON	occurrent	processual_entity	process	planned process
14	data acquisition	["A study is a program of data coll Ontology of Clinical Research		OCRE	Entity			
15	dataset		Translational Medicine Ontology	TMO	processualEntity	planned process		
16	data unit	["A study is a collection of one or DICOM Controlled Terminology		DCM				
17	processing		KB_Bio_101	AURA	Event	Activity	Cognitive-Activity	
18	analysis		Pharmacogenomic Relationships Ontology	PHARE	Analysis			
19	scientific workflow		Sage Bionetworks Synapse Ontology	SYN	Activity	Clinical or Research Activity	Research Activity	
20	scientific calculus	["A detailed examination, analysis, Bilingual Ontology of Alzheimer's Disease and Rele ONTOAD		SEDI	process	Activities	Clinical or Research Activity	Research Activities
21	workflow		Semantic DICOM Ontology	ONS	informationEntity			
22	parameters		Ontology for Nutritional Studies	ONL-MSA	entity	occurrent	process	planned process
23	Input data	["[DEF] A STUDY is a research inve Mental State Assessment		ORTH	perdurant	action	non-physical action	conceptual action
24	output data		Orthology Ontology	JERM	entity	process	procedure	investigation
25	bibliographical reference	["a collection of experiments desi(Just Enough Results Model Ontology		BIOMO	process			
26	scientific article		Biological Observation Matrix Ontology	CANCO	entity	process	procedure	investigation
27	reference data	["A study is a collection of bioassa Cancer Chemoprevention Ontology		REPRODUCE-ME	Entity	continuant	dependent_continuant	generically_dependent_continuant
28	Device	["A process to examine and analyz REPRODUCE-ME		CTX		Plan	Plan	
29	Acquisition device		Cerebrotendinous Xanthomatosis Ontology	PREMEDONTO	Activity	Clinical or Research Activity	Research Activity	
30	Software	["A detailed examination, analysis, Precision Medicine Ontology		ONL-MR-DA	perdurant	action	non-physical action	conceptual action
31	tool	["[DEF] A STUDY is a research inve Magnetic Resonance Dataset Acquisition Ontolog		OntoVIP	perdurant	action	non-physical action	conceptual action
32	script	["[DEF] A STUDY is a research inve Medical Image simulation		HASCO	entity	process	procedure	investigation
33	group		Human-Aware Science Ontology		entity	process	Activity	
34	patient group		Human-Aware Science Ontology	HASCO	Entity			
35	lifecycle stage							
36	lifecycle phase							
37	life cycle							
38	biological							

Figure 169 Liste des termes noyaux d'entrée et exemple de résultat donné par le script1

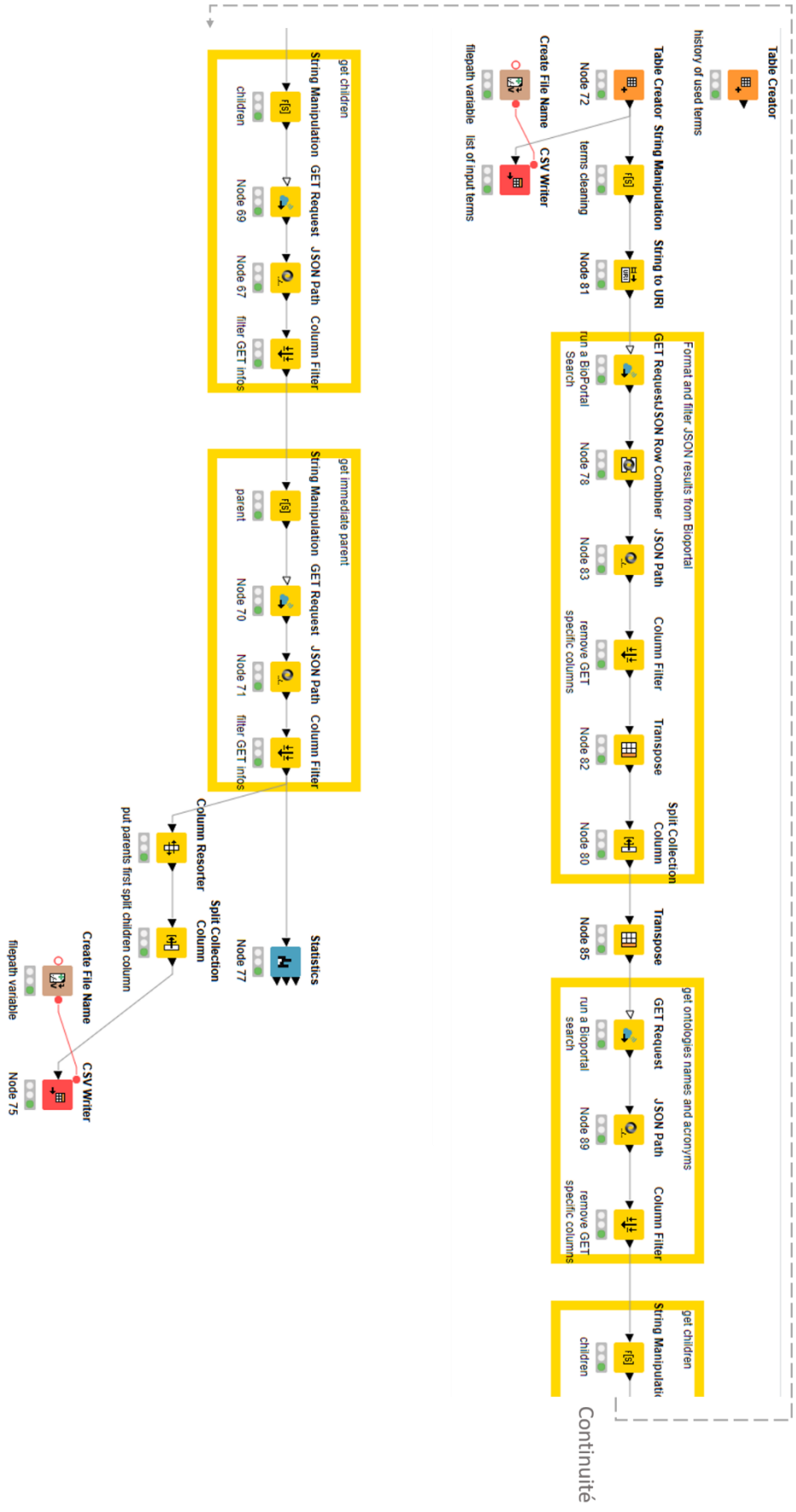


Figure 170 Capture d'écran du script2 codé en KNIME pour l'exploration Bioportal des termes de domaines

Liste de termes initiaux

fluorodésoxyglucose
animal position
FDG uptake
glycemia
isoflurane
anesthesia
gantry
semicircularmultifov
helical
full scan
half scan
projection number
zoom
voxel size
slice thickness
filter
filter cut-off
image reconstruction
reconstruction model
whole body image
dynamic reconstruction
gated reconstruction
coincidence mode
count rate mode
normal mode
high counts
time limited
count limited
radio tracer
syringe radioactivity
syringe volume
time of counting
resolution
image correction

Parent du concept trouvé

Liste des concepts frères au concept trouvé

term	ontology	parent	relevance	sibling1	sibling2	sibling3	sibling4
isoflurane	NDDF		ko				
Isoflurane	LOINC	General anesthetics	ok	Isoflurane gas delivered.total	Isoflurane setting	Isoflurane	Gas delive	Isoflurane liquid delivere
Isoflurane	LOINC		ko				
ISOFLURANE	VANDF		ko				
Isoflurane	IOBC	organic compound	ko				
Isoflurane	IOBC	NikkajiCompounds	ok				
isoflurane	NIFSTD	organofluorine compound	ko				
Isoflurane measurement	SNOMEDCT	Measurement of substance	ko				
Isoflurane poisoning	SNOMEDCT	Poisoning by gaseous anesthetic	ko	Intentional isoflurane poison	Isoflurane poisoning of unde	Accidental isoflurane poisoni	
Isoflurane overdose	SNOMEDCT	Ether overdose	ok	Accidental isoflurane overdo	Isoflurane overdose of undet	Intentional isoflurane overdo	
Isoflurane target	LOINC		ko				
Isoflurane setting	LOINC		ko				
Isoflurane target	LOINC	Isoflurane	ok	Target isoflurane	Airv			
Isoflurane setting	LOINC	Isoflurane	ok	Isoflurane setting	Ga			
Isoflurane^during expiration	LOINC		ko				
Isoflurane liquid delivered.total	LOINC		ko				
Anesthesia	MEDDRA		ko				
Anesthesia	LOINC		ko				
Anesthesia	LOINC		ko				
anesthesia	RADLEX	process	ok	regional anesthesia	local anesthesia	tumescent anesthesia	epidural anesthesia
Anesthesia	IOBC	Anesthesia and Analgesia	ok	Cardiac anesthesia	acupunctual anesthesia	ambulatory anesthesia	low flow anesthesia
anesthesia	HUPSON	medical procedure	ok				
anesthesia	GAMUTS		ko				
Absence of sensation	SNOMEDCT	Observation of sensation	ok	Thermal anesthesia	Absent body position sense	On examination - absence of	Absence of vibratory sen
Procedure anesthesia:Find:Pt:^Patient:	LOINC	CLINICAL.NOTYETCATEG	ko				
anaesthesia	ICPC2P		ko				
Loss of sensation	IOBC	perceptual disorder	ko				
Anesthesia procedures	ONTOAD	Medication Management	ko	General Anesthesia			
General anesthesia - under adequate a	SNOMEDCT	Under general anesthesia	ko				
Anesthesia duration	LOINC		ko				
W anesthesia	LOINC		ko				

Figure 171 Liste des termes noyaux d'entrée et exemple de résultat donné par le script2

Nous avons noté que pour décrire un « BFO : process » dans OBO, il faut ajouter une relation « object property » et un « data item ». Les qualités « quality » dans OBO sont réservées au « BFO :continuant ». Les exemples de la Figure 172 illustre ces propos.

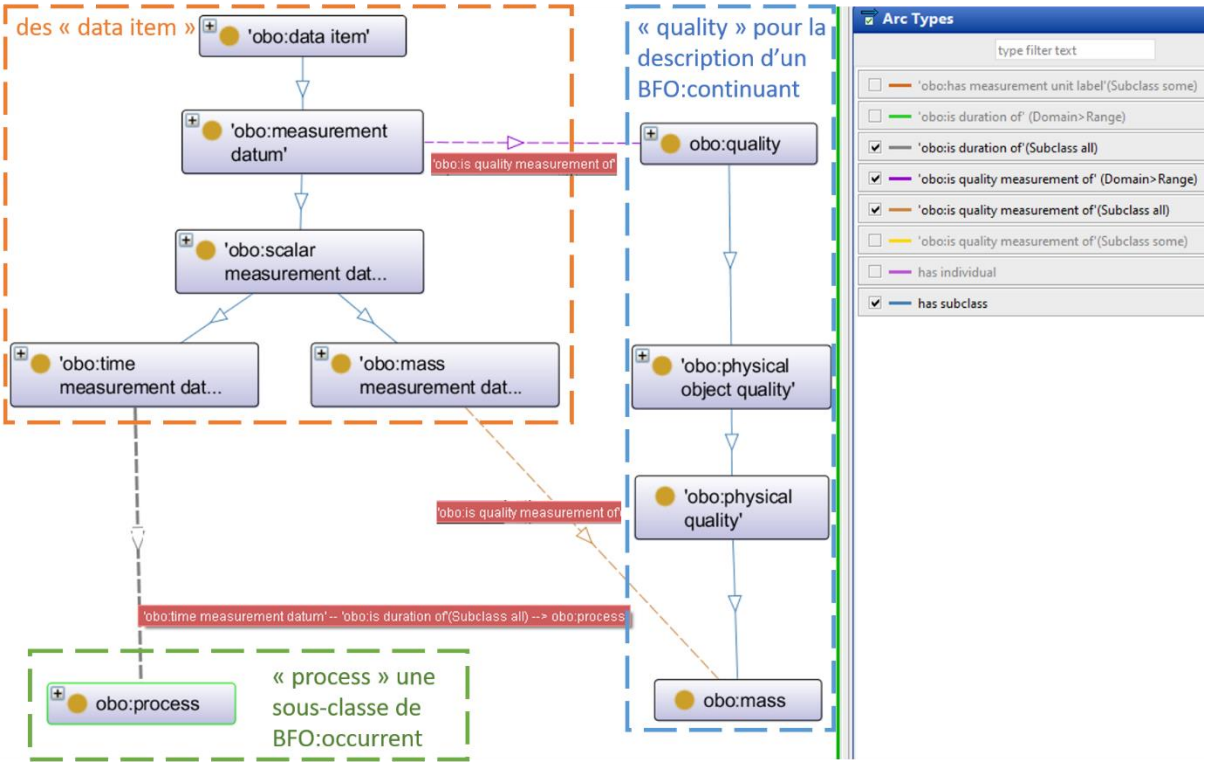


Figure 172 Exemple de description via des « data items » et des « qualities » des concepts « occurrent » et « continuant » dans OBO

Quant aux listes de valeurs, et afin de détecter les patterns utilisés pour leurs modélisations dans OBO, nous avons exploré les ontologies IAO et OBI. Nous avons repéré premièrement la classe « OBI : value specification » qui permet de les lister. Ensuite, nous avons exploré l'ontologie IAO et nous avons trouvé l'exemple de la capture d'écran Figure 173 ci-après. Elle montre que pour donner la liste des valeurs de la classe « OBO : curation status specification », les auteurs ont utilisé une liste d'individus OWL. Nous avons alors retenu cette option en plus de l'option de la classe « OBI : value specification ».

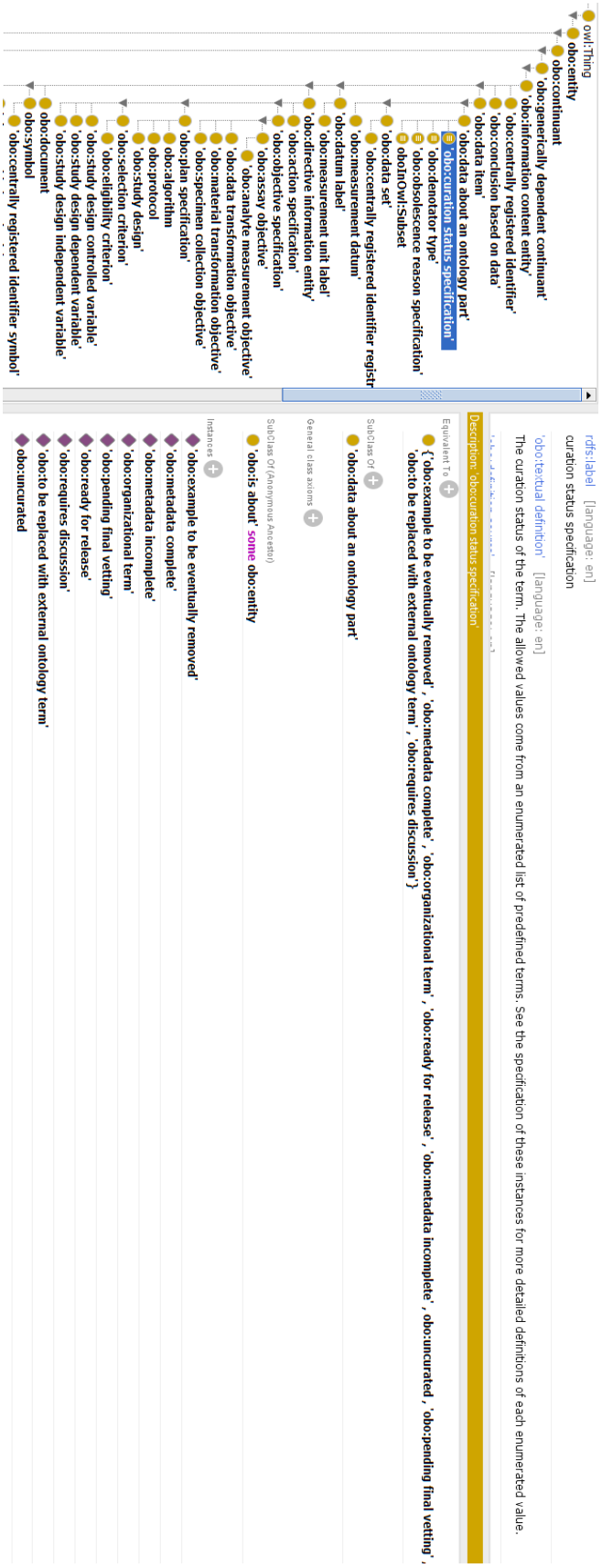
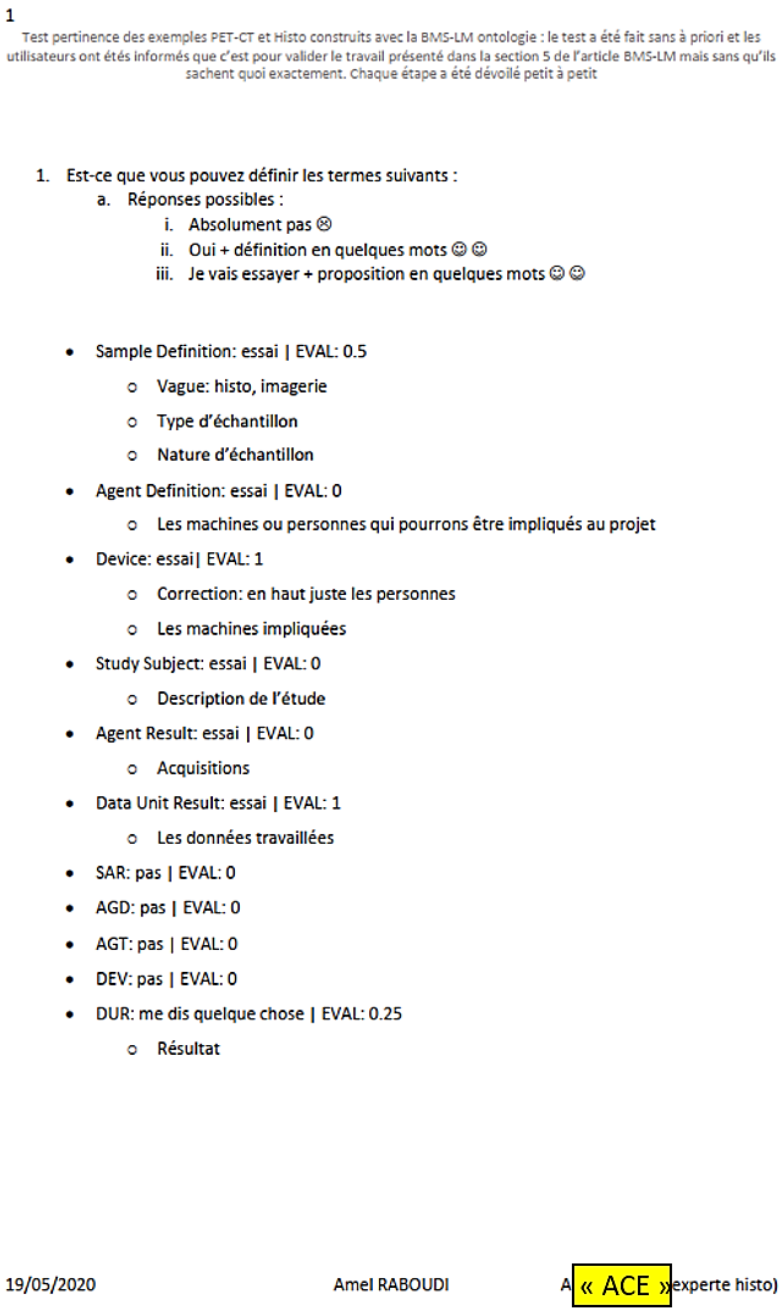


Figure 173 Exploration dans Protégé de l'ontologie IAO

RESTITUTIONS DES TESTS UTILISATEURS APRÈS L'APPLICATION DE L'ONTOLOGIE BMS-LM AUX DONNÉES DU LABORATOIRE LRI

Avec trois utilisateurs clés « ACE, « TVI, et « TYO », nous avons évalué les exemples TEP-TDM et Histologie présents en Figure 102et Figure 103. Nous n'avons pas dévoilé le but du test pour ne pas créer un biais dans les résultats. Les tests ont été effectués via zoom et pour chaque personne interviewée un questionnaire a été répondu avant et après l'explication des images montrées à l'écran. Les restitutions ont été envoyées et validées par e-mail. Pour les trois, une amélioration de la compréhension des concepts noyau a été notée.



1

Test pertinence des exemples PET-CT et Histo construits avec la BMS-LM ontologie : le test a été fait sans à priori et les utilisateurs ont étés informés que c'est pour valider le travail présenté dans la section 5 de l'article BMS-LM mais sans qu'ils sachent quoi exactement. Chaque étape a été dévoilé petit à petit

1. Est-ce que vous pouvez définir les termes suivants :
   a. Réponses possibles :
      i. Absolument pas ⊗
      ii. Oui + définition en quelques mots ☺ ☺
      iii. Je vais essayer + proposition en quelques mots ☺ ☺

- Sample Definition: essai | EVAL: 0.5
  - Vague: histo, imagerie
  - Type d'échantillon
  - Nature d'échantillon
- Agent Definition: essai | EVAL: 0
  - Les machines ou personnes qui pourrons être impliqués au projet
- Device: essai| EVAL: 1
  - Correction: en haut juste les personnes
  - Les machines impliquées
- Study Subject: essai | EVAL: 0
  - Description de l'étude
- Agent Result: essai | EVAL: 0
  - Acquisitions
- Data Unit Result: essai | EVAL: 1
  - Les données travaillées
- SAR: pas | EVAL: 0
- AGD: pas | EVAL: 0
- AGT: pas | EVAL: 0
- DEV: pas | EVAL: 0
- DUR: me dis quelque chose | EVAL: 0.25
  - Résultat

19/05/2020 Amel RABOUDI A « ACE » experte histo)

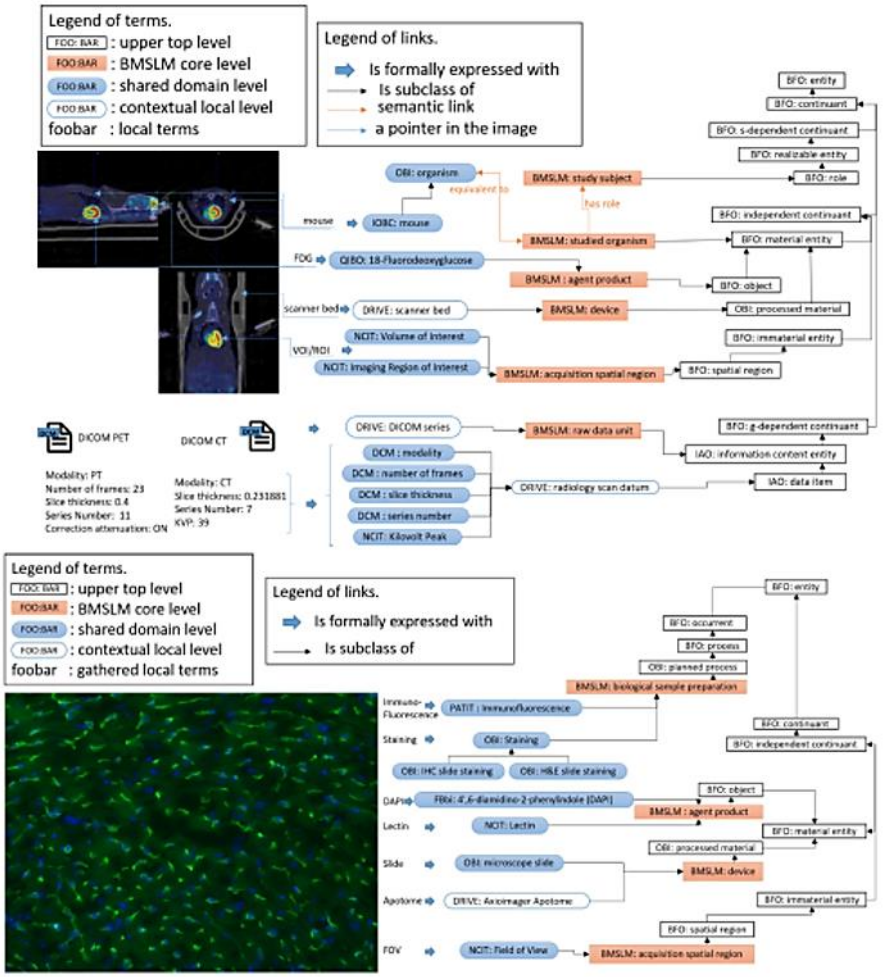
Figure 174 Questionnaire répondu par « ACE » avant de voir les exemples d'application de l'ontologie BMS-LM pour le préclinique

Somme EVAL : 1.75/11

2

Test pertinence des exemples PET-CT et Histo construits avec la BMS-LM ontologie : le test a été fait sans à priori et les utilisateurs ont étés informés que c'est pour valider le travail présenté dans la section 5 de l'article BMS-LM mais sans qu'ils sachent quoi exactement. Chaque étape a été dévoilé petit à petit

2. **Après avoir vu les exemples d'images annotés, il faut commencer par l'image, les termes à cotés qui sont directement liés à son contenu, et après leurs équivalents ou parents parmi les termes en bleu qui sont plus standard, et après leurs parents en orange qui sont eux génériques mais en lien avec le biomédical et enfin, les termes en noir et blanc qui relèvent de l'abstraction du monde.**



19/05/2020 Amel RABOUDI A « ACE » (experte histo)

Figure 175 Schémas montrés à « ACE », « TVI », « TYO » lors du test utilisateur

3

Test pertinence des exemples PET-CT et Histo construits avec la BMS-LM ontologie : le test a été fait sans à priori et les utilisateurs ont étés informés que c'est pour valider le travail présenté dans la section 5 de l'article BMS-LM mais sans qu'ils sachent quoi exactement. Chaque étape a été dévoilé petit à petit

3. Maintenant, en se basant sur les images, est-ce que vous pouvez définir les termes suivants:
    a. Réponses possibles :
        i. Absolument pas ⊗
        ii. Oui + définition en quelques mots ☺ ☺
        iii. Je vais essayer + proposition en quelques mots ☺ ☺

- biological sample preparation : essai | EVAL: 1
    - La technique utilisée pour marquer la lectine
- Agent product : oui | EVAL: 1
    - Les produits utilisés lors de la technique
- Device : oui | EVAL: 1
    - La machine d'acquisition
- Acquisition spatial region : oui | EVAL: 0
    - Type d'acquisition : *20, 40x, whole scan
- Study subject: oui | EVAL: 0
    - Étude
- Studied organism: pas | EVAL: 0
- Raw data unit : essai | EVAL: 1
    - Data brute

Question supplémentaire : Est-ce que tu as compris l'intérêt du test ?

Vérifier si les bons mots ont été utilisé => si c'était du chinois, je trouve que c'est très technique

19/05/2020 Amel RABOUDI A « ACE » n (experte histo)

Figure 176 Questionnaire répondu par « ACE » après explication des exemples d'application de l'ontologie BMS-LM

Somme EVAL : 4/6

1

Test pertinence des exemples PET-CT et Histo construits avec la BMS-LM ontologie : le test a été fait sans à priori et les utilisateurs ont étés informés que c'est pour valider le travail présenté dans la section 5 de l'article BMS-LM mais sans qu'ils sachent quoi exactement. Chaque étape a été dévoilé petit à petit

1. Est-ce que vous pouvez définir les termes suivants :
    a. Réponses possibles :
        i. Absolument pas ⊗
        ii. Oui + définition en quelques mots ☺ ☺
        iii. Je vais essayer + proposition en quelques mots ☺ ☺

- Sample Definition: essai | EVAL: 0.75
    - L'animal ou l'objet ou la lame d'histologie : nature, référence
- Agent Definition: essai | EVAL: 0.5
    - Agent (la personne)
    - Agent (traceur)
- Device: oui | EVAL: 1
    - Description de la machine sur laquelle on fait du TEP ou le microscope pour histo
- Study Subject: oui | EVAL: 1
    - Sample => mauvaise réponse
    - Souris, rat, objet passé dans la camera tep
- Agent Result: Absolument pas | EVAL: 0
- Data Unit Result: essai | EVAL: 1
    - Ce qui sort du processus: un fichier DICOM, une image JPEG
- SAR: pas | EVAL: 0
- AGD: essai | EVAL: 0.75
    - Agent G? Definition
- AGT: pas | EVAL: 0
- DEV: pas| EVAL: 0
- DUR: Data Unit Result | EVAL: 1

19/05/2020 Amel RABOUDI T « TVI » (expert PET-CT)

Figure 177 Questionnaire répondu par « TVI » avant de voir les exemples d'application de l'ontologie BMS-LM pour le préclinique.

Somme EVAL :6/11

3

Test pertinence des exemples PET-CT et Histo construits avec la BMS-LM ontologie : le test a été fait sans à priori et les utilisateurs ont étés informés que c'est pour valider le travail présenté dans la section 5 de l'article BMS-LM mais sans qu'ils sachent quoi exactement. Chaque étape a été dévoilé petit à petit

4. Maintenant, en se basant sur les images, est-ce que vous pouvez définir les termes suivants:
    a. Réponses possibles :
        i. Absolument pas ⊗
        ii. Oui + définition en quelques mots ☺ ☺
        iii. Je vais essayer + proposition en quelques mots ☺ ☺

- biological sample preparation : oui | EVAL: 1
    - Histology : deux grands types : IF et marquage d'immunochimie
    - Préciser le type de préparation de la lame d'histologie
- Agent product : oui | EVAL: 1
    - Réactif/molécules chimiques utilisés
- Device : oui | EVAL: 1
    - Description de la machine sur laquel on fait du TEP ou le microscope pour histo
- Acquisition spatial region : oui | EVAL: 1
    - Zone d'interet
- Study subject: oui | EVAL: 1
    - Souris, rat, objet passé dans la camera tep
- Studied organism: | EVAL: 1
    - La catégorie dans laquelle entre le Study Subject
- Raw data unit : oui | EVAL: 1
    - Le fichier numérique
    - Objet virtuel étudié :
        - transforme une lame en un fichier jpeg
        - transforme une souris en une image DICOM

Question supplémentaire : Est-ce que tu as compris l'intérêt du test ?

voir si l'ontologie est plus ou mois compréhensible, intuitive ou pas

19/05/2020 Amel RABOUDI T « TVI » (expert PET-CT)

Figure 178 Questionnaire répondu par « TVI » après explication des exemples d'application de l'ontologie BMS-LM.

Somme EVAL : 7/7

1

Test pertinence des exemples PET-CT et Histo construits avec la BMS-LM ontologie : le test a été fait sans à priori et les utilisateurs ont étés informés que c'est pour valider le travail présenté dans la section 5 de l'article BMS-LM mais sans qu'ils sachent quoi exactement. Chaque étape a été dévoilé petit à petit

1. Est-ce que vous pouvez définir les termes suivants :
    a. Réponses possibles :
        i. Absolument pas ⊗
        ii. Oui + définition en quelques mots ☺ ☺
        iii. Je vais essayer + proposition en quelques mots ☺ ☺

- Sample Definition: oui | EVAL: 0.5
    - Identités de l'échantillon, c'est des rats souris, femelle male,
    - Tous ce qui permet de bien définir l'échantillon
    - Echantillon sujet de l'étude (échantillon d'histo, animal)
- Agent Definition: oui | EVAL: 1
    - Les consommables, pour le TEP le traceur utilisé
    - Type quantité
- Device: oui | EVAL: 1
    - Appareil utilisé: TEP, US
- Study Subject: oui | EVAL: 0
    - Étude du sujet: objectif et problématique sur laquelle on se base pour tout choisir
- Agent Result: essai | EVAL: 0
    - Résultat en utilisant ces consommable
    - Pas le même selon la dose, les caractéristiques du consommable
- Data Unit Result: essai | EVAL: 1
    - Les données obtenues suite à une série de manip
- SAR: essai | EVAL: 0.75
    - Les caractéristiques de l'échantillon, sujet d'étude
- AGD: oui | EVAL: 1
    - Agent Definition : définition de tous ce qui est consommable
- AGT: pas | EVAL: 0.5
    - Agent
- DEV: oui | EVAL: 1
    - Device : tous les appareils utilisés
- DUR: oui | EVAL: 1
    - Data Unit Results
    - Données et résultats pour une étude donnée

19/05/2020 Amel RABOUDI « TYO »
(experte PET-CT, BDD souris, doctorante)

Figure 179 Questionnaire répondu par « TYO » avant de voir les exemples d'application de l'ontologie BMS-LM pour le préclinique

Somme Eval : 6.75/11

3

Test pertinence des exemples PET-CT et Histo construits avec la BMS-LM ontologie : le test a été fait sans à priori et les utilisateurs ont étés informés que c'est pour valider le travail présenté dans la section 5 de l'article BMS-LM mais sans qu'ils sachent quoi exactement. Chaque étape a été dévoilé petit à petit

3. Maintenant, en se basant sur les images, est-ce que vous pouvez définir les termes suivants:
    a. Réponses possibles :
        i. Absolument pas ☹
        ii. Oui + définition en quelques mots ☺ ☺
        iii. Je vais essayer + proposition en quelques mots ☺ ☺

- biological sample preparation : oui | EVAL: 1
    - Le protocole expérimental, par quel moyen tu arrives à générer le résultat que tu voulais
- Agent product : oui| EVAL: 1
    - Tous les consommables utilisés
    - Traceurs
- Device : oui| EVAL: 1
    - Tous les appareils utilisé pour suivre le protocole et générer les données
- Acquisition spatial region : oui| EVAL: 1
    - Région d'intérêt
- Study subject: oui| EVAL: 1
    - Caractéristiques du sujet d'étude
- Studied organism: oui| EVAL: 1
    - Le tissu étudié, l'animal étudié
    - Aussi sujet d'étude mais pas la même chose
- raw data unit : oui| EVAL: 1
    - Données brutes générées

Question supplémentaire : Est-ce que tu as compris l'intérêt du test ?

Tu vérifies si on connait ou pas le vocabulaire de base, les exemples vont mieux définir les choses qu'on définissait avant

19/05/2020 Amel RABOUDI « TYO »
(experte PET-CT, BDD souris, doctorante)

Figure 180 Questionnaire répondu par « TYO » après explication des exemples d'application de l'ontologie BMS-LM.

Somme EVAL :7/7

ANNEXE D : GUIDES ET ENQUÊTES ÉLABORÉS AU LABORATOIRE LRI

Nous avons dédié cette annexe aux résultats d'audit, enquêtes et guides réalisés dans le cadre de la mise en œuvre du système BMS-LM pour la recherche préclinique.

DIAGRAMMES SADT VALIDÉS LORS DE L'AUDIT PLMBOOST

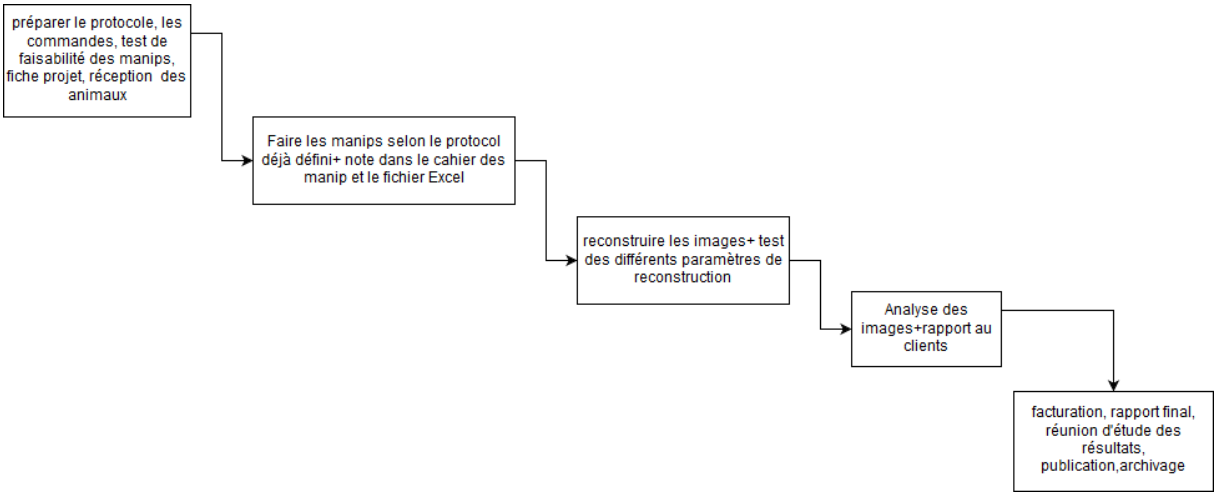


Figure 181 Diagramme SADT réalisé avec l'utilisateur clé « TVI »

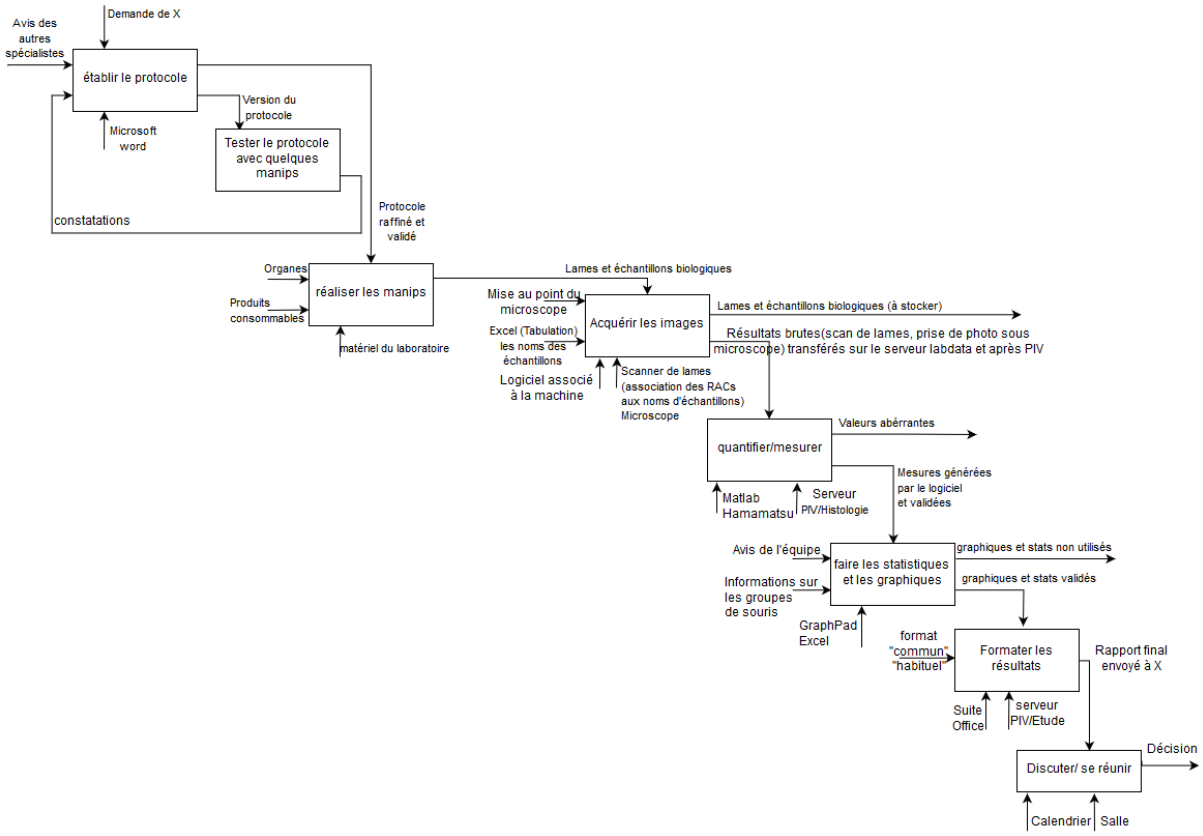


Figure 182 Diagramme SADT réalisé avec l'utilisateur clé « ACE »

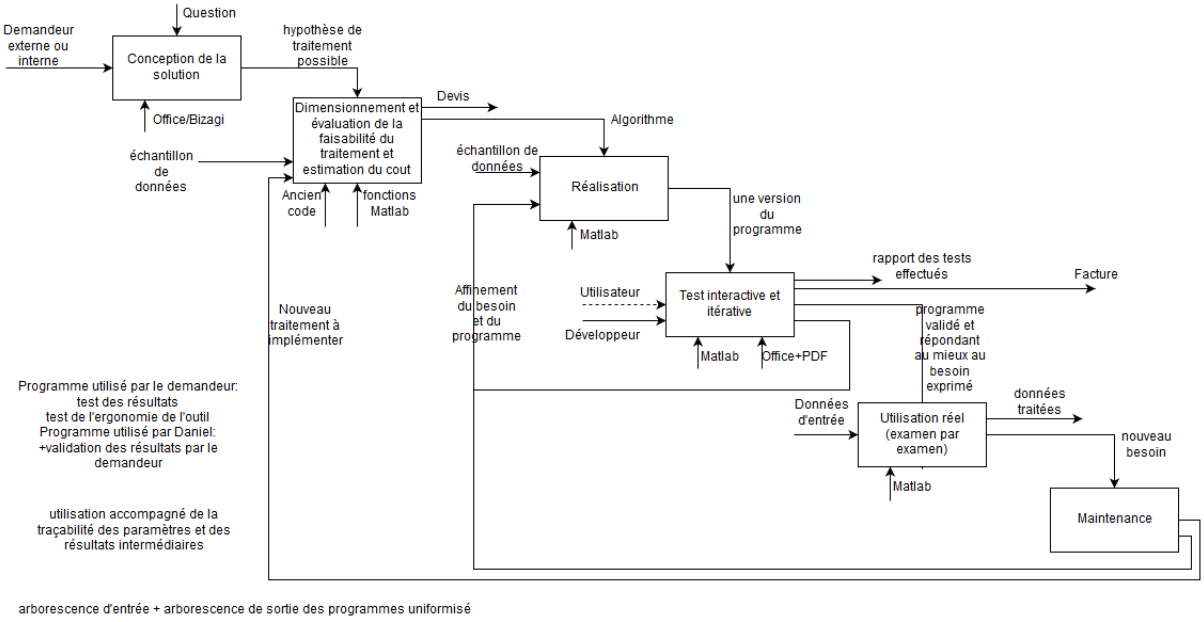


Figure 183 Diagramme SADT réalisé avec l'utilisateur clé « DBA »

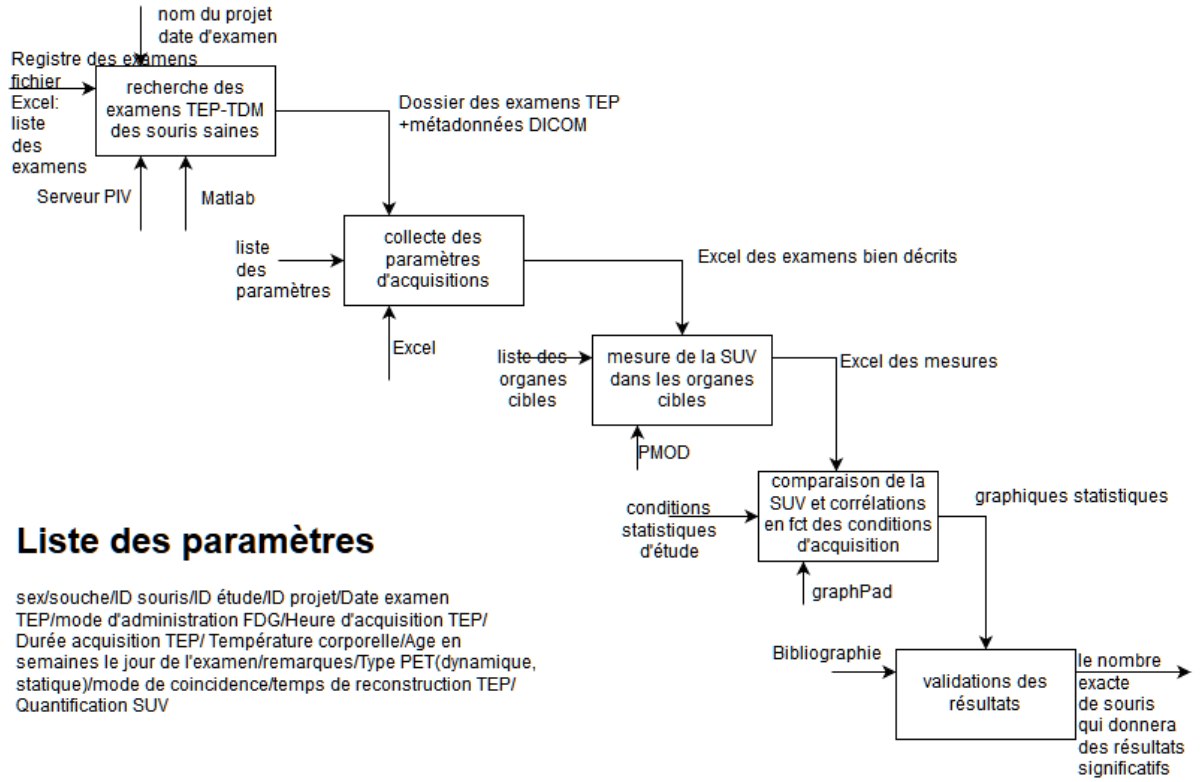
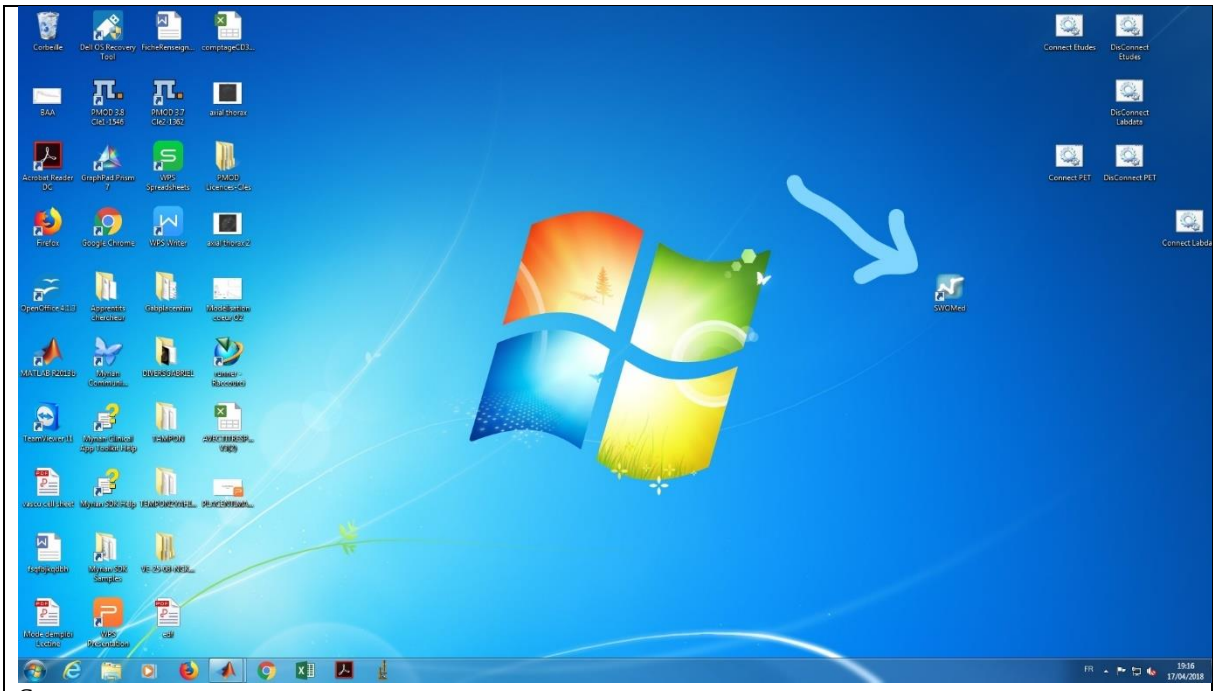


Figure 184 Diagramme SADT réalisé avec l'utilisateur clé « TYO »

GUIDE ET EXERCICE D'EXPLORATION DE DONNÉES APRÈS IMPORT DE DONNÉES D'HISTOLOGIE

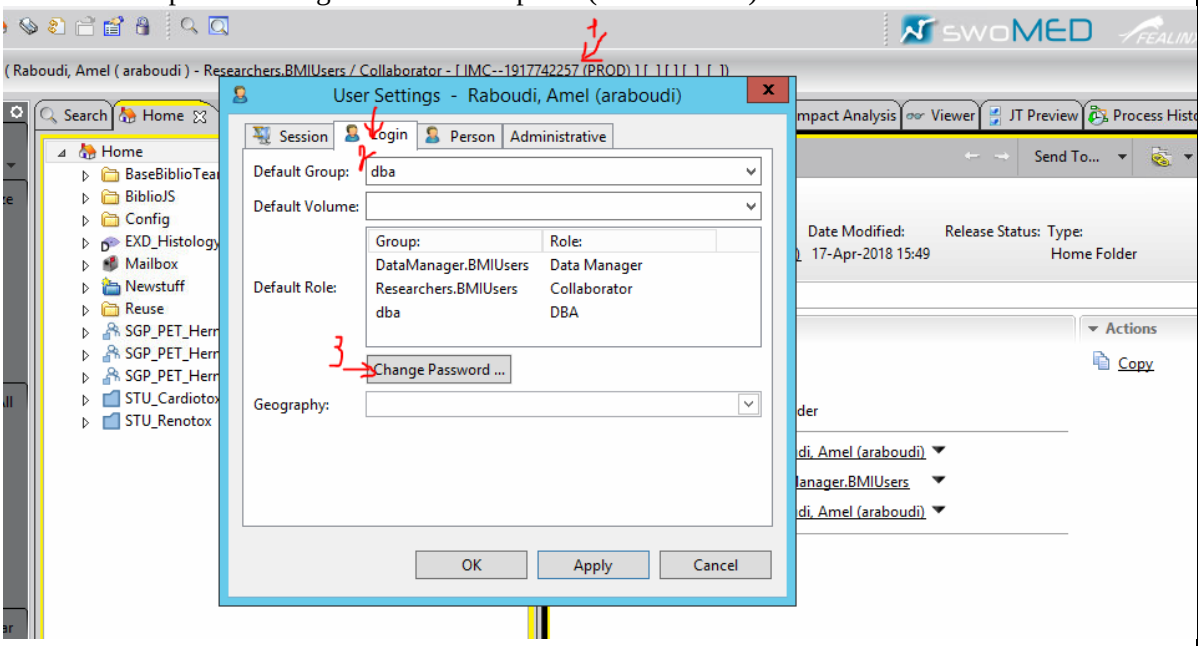
Lors du test d'import des données d'histologie avec l'utilisateur clé « ACE » le guide suivant a été fourni le 18/04/2018 afin de permettre à l'utilisateur d'explorer les données importées.
« Deux options de navigation possibles :

- Le navigateur via l'adresse : http://pf-01.lab.parisdescartes.fr:1327/awc/
- Le client sur la station de travail SAPIN

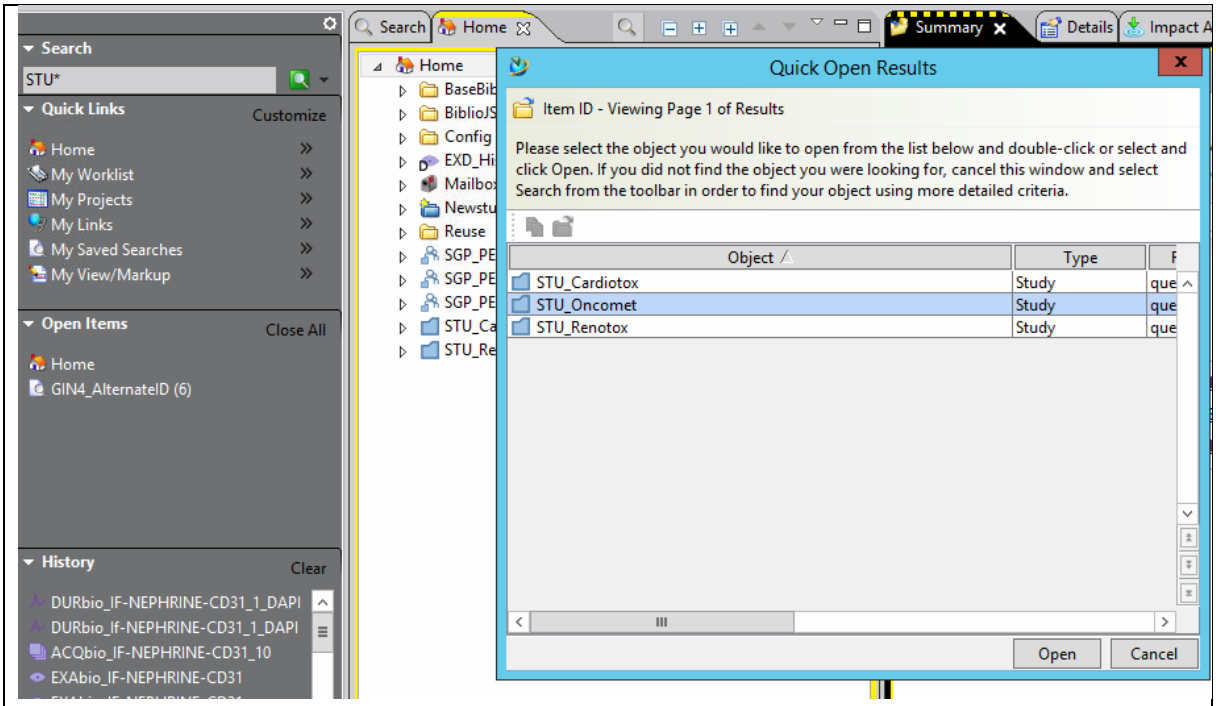


Se connecter avec votre compte :

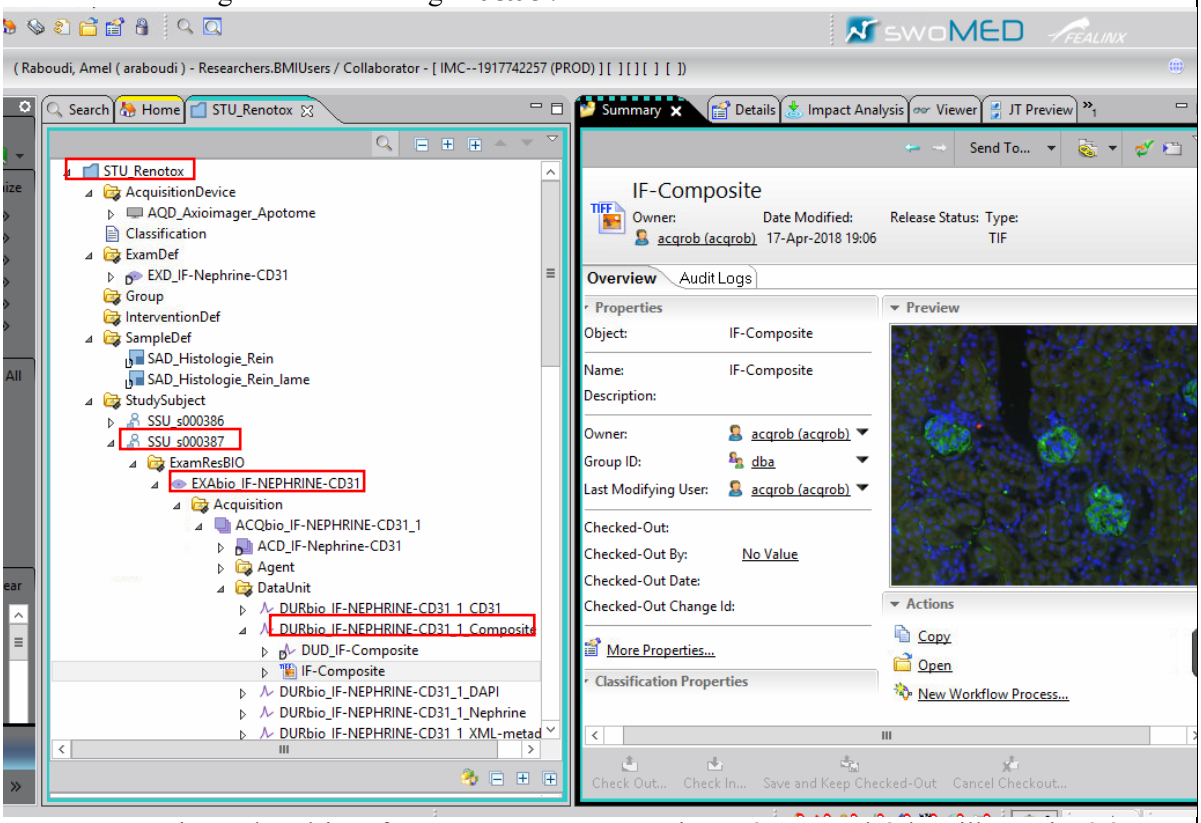
- PNOM/pnom
- Vous pouvez changer votre mot de passe (recommandé)



Dans la barre de recherche rapide, il faut écrire STU* puis faire entrée
Sélectionner l'étude concernée : Ici STU_Renotox



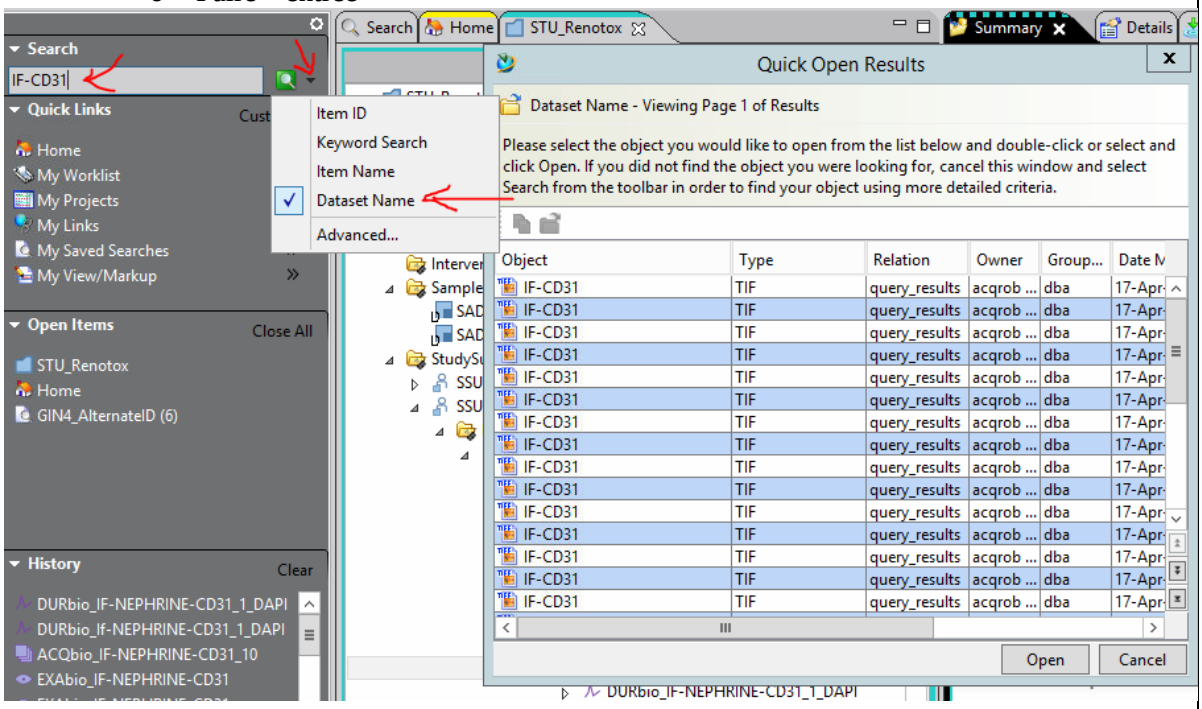
Double cliquer sur STU_Renotox et naviguer dans ce qu'il y a dedans via la petite flèche à gauche. Si vous voulez isoler un dossier pour une navigation plus clair, il faut double cliquer et ceci l'ouvre dans un nouvel onglet à côté de l'onglet actuel.



Exercice : Naviguer dans l'interface sous STU_Renotox, jusqu'à trouver l'échantillon qui a été acquis sous le microscope et qui correspond à l'examen Encadré dans la capture d'écran.
Autres propositions et possibilités d'exploration :

- Si on souhaite avoir accès à la liste des images TIFF qui représente le marquage CD31 de ces examens, il faut :
  - Revenir dans la barre de recherche rapide

- Écrire « IF-CD31 »
- Choisir dans la liste à côté « dataset »
- Faire « entrée »



INTERFACES ET ÉVALUATION DE L'ACCEPTABILITÉ DE LA MÉTHODE D'INTÉGRATION « PARTIELLE » DE CALCUL SCIENTIFIQUE

Dans cette partie de l'annexe, nous présentons les interfaces et les fiches d'évaluation de la méthode d'intégration « partielle » des calculs scientifiques qui a été appliqué aux quantifications histologiques

Interfaces ajoutées à l'outil Matlab de quantification de données d'histologie

Nous avons présenté en section les interfaces ajoutées pour assurer la connexion entre le système BMS-LM et l'application Matlab locale de quantification de données d'histologie. Les interfaces ajoutées sont présentes dans les figures ci-après.

Figure 185 Interface de connexion au système BMS-LM

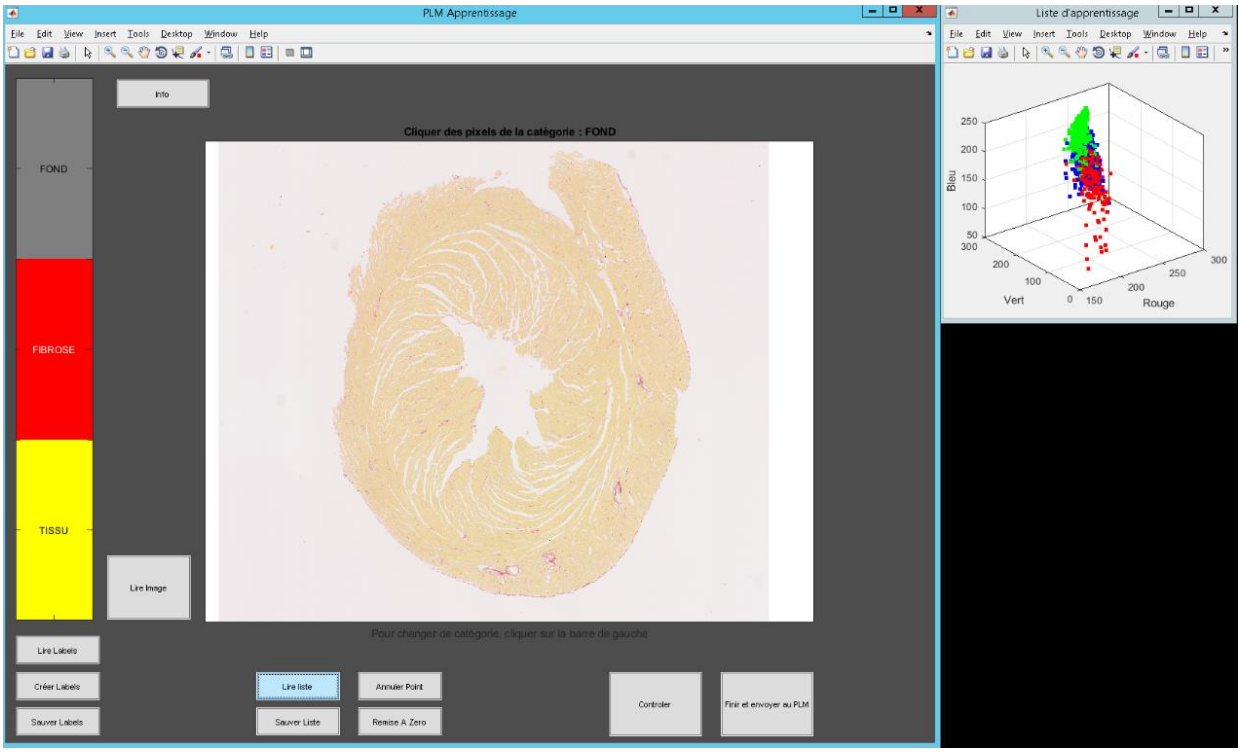


Figure 186 Réorganisation de l'interface graphique pour plus de modularité

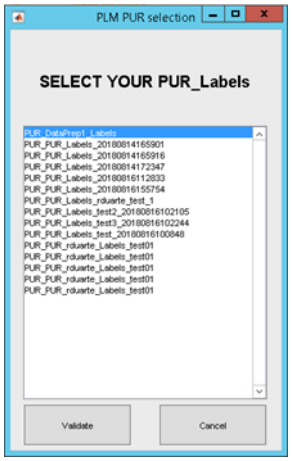


Figure 187 Récupération d'une liste de Labels déjà envoyés au système BMS-LM lors des exécutions précédentes

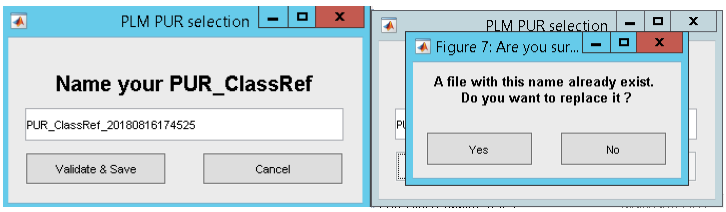


Figure 188 Sauvegarde de la liste des points dans le système BMS-LM

Fiches d'évaluation utilisateurs de l'intégration « partielle »

L'évaluation s'est déroulée le 09/08/2018 avec l'utilisateur clé « ACE ». Les figures ci-après montre les fiches restituées avant et après la formation.

A « ACE »

Un petit test avant de commencer ☺

1- Avis sur le temps d'exécution du téléchargement des données depuis le serveur Histo
   a. Rapide-Intéressant
   b. Normal- Rien de spécial
   c. Acceptable – logique
   d. Lent- Inexplicable

S937 - RS - Coeur (711 No) 1min 38

2- Avis sur l'organisation des données selon la logique PLM
(il faut expliquer en un mot le trigramme (XXX) suivant)
   a. STU : Aide
   b. SGP :
   c. SSU : sujet (x aide de Thula)
   d. WFI :
   e. DUR : Resultat
   f. PUR :
   g. PLM : product Life management

~16h30 ~15h30 : 1h pour faire la démo
~2:31 pour le téléchargement des TIFF

Amel Raboudi - PLMBoost - jeudi 9 août 2018

Signature de l'utilisateur :
User's signature:

Date : 09/08/2018

Témoin (prénom, nom) : Roberto DUARTE
Witness (first name, last name):

Date : 09/08/2018
Signature :

42

Figure 189 Fiche remplie par l'utilisateur « ACE » avant la formation sur l'outil de quantification histologie version BMS-LM

A « ACE »

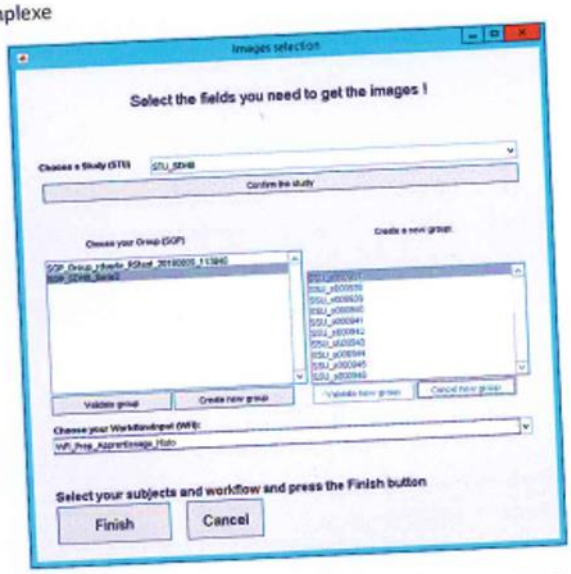
Ton avis nous intéresse ☺

1- Avis sur le temps d'exécution

a. Rapide-Intéressant
b. Normal- Rien de spécial
c. Acceptable – logique
d. Lent- Inexplicable

2- Avis sur l'interface de téléchargement des données depuis le PLM

a. Simple – Clair
b. Maitrisable
c. Complexe



3- Avis sur l'organisation des données selon la logique PLM

(il faut expliquer en un mot le trigramme (XXX) suivant)

a. STU : Study
b. SGP : Groupe
c. SSU : Subject
d. WFI : Treatment
e. DUR : Resultat
f. PUR : Resultat
g. PLM : Product Life Management cycle

+ + Amel
configuration du traitement
donnée Image
résultat de traitement

Amel Raboudi - PLMBoost - jeudi 9 août 2018

Signature de l'utilisateur :
User's signature:

Date : 09/08/2018

Témoin (prénom, nom) : Roberto DUARTE
Witness (first name, last name):

Date : 09/08/2018
Signature :

43

Figure 190 Fiche remplie par l'utilisateur « ACE » après la formation sur l'outil de quantification histologie version BMS-LM

GUIDE MEDISO2PLM V3 FOURNI AUX UTILISATEURS LORS DE LA FORMATION

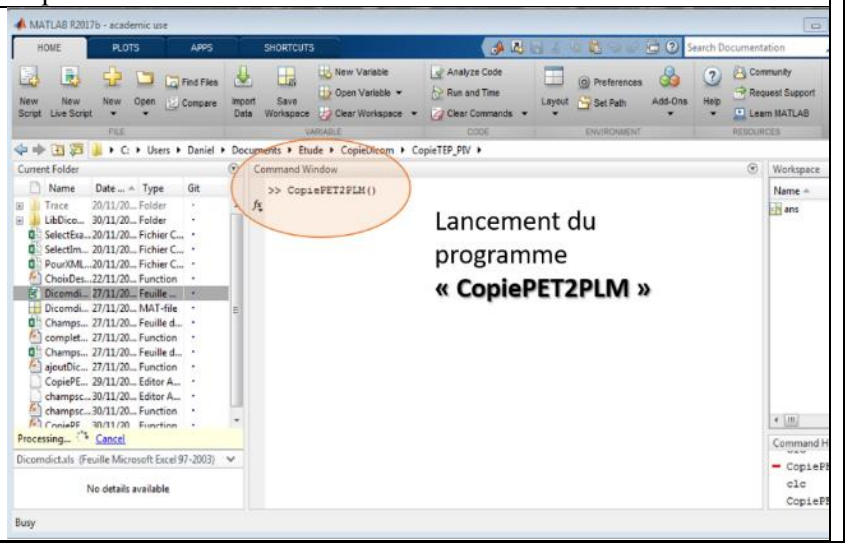
Les étapes d'utilisation sont présentées ci-après de deux manières :

1. Résumées rapidement avec une description si la tâche est manuelle ou automatisée et aussi en précisant l'outil utilisé pour l'accomplir
2. Illustrées pas à pas avec des copies d'écran

Synopsis

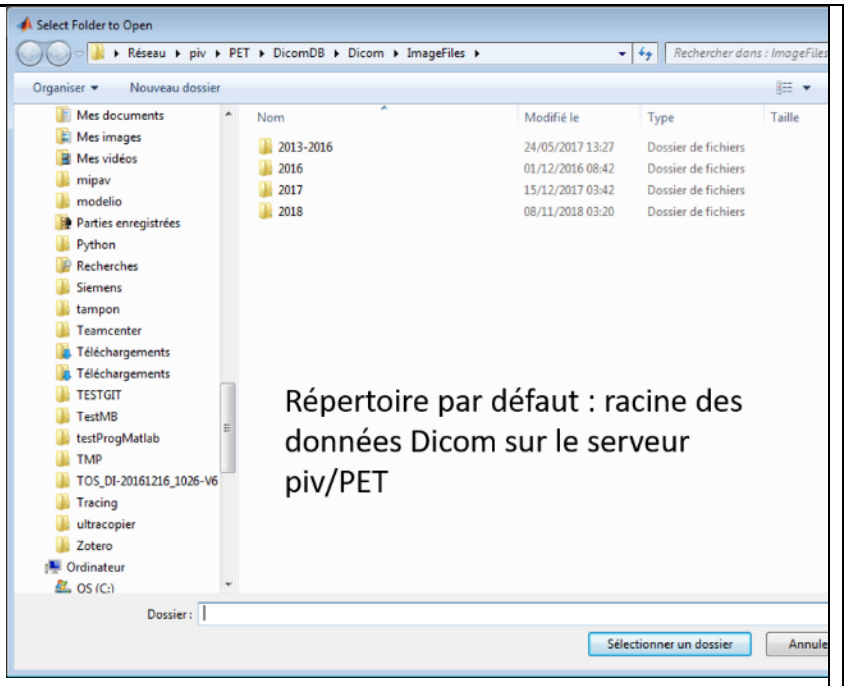
a) Manuel : Lancement de CopiePET2PLM dans Matlab [Matlab]
b) Manuel : Sélection du dossier qui contient les examens à copier [Matlab]
c) Manuel : Sélectionner parmi les examens présents ceux à copier [Matlab]
d) Manuel : Sélectionner dans les examens à copier, les types de séquences d'intérêt [Matlab]
e) Automatisé : Copie des données et d'une table de synthèse dans un espace tampon [Matlab]
f) Automatisé : Conversion de la table en fichier xml [Matlab ; .bat sur le bureau]
g) Manuel : Transfert ftp des données et du fichier xml sur le serveur du PLM [FilleZilla]
h) Automatisé : intégration de la structure et du contenu en données dans le modèle PLM [SWOMed]
i) Manuel : contrôle de la bonne intégration des données dans le PLM[SWOMed]

Mediso2PLM pas à Pas

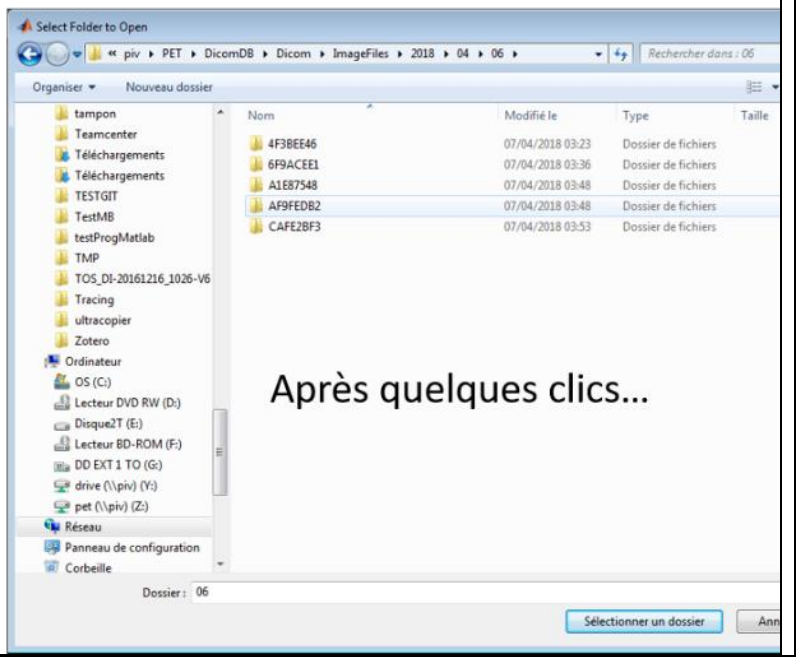
Étape	Copies d'écran
1. Lancement du programme Matlab 	

2. Sélection du dossier à analyser

➢ (Il faut avoir connecté le lecteur réseau \\piv\PET avant)

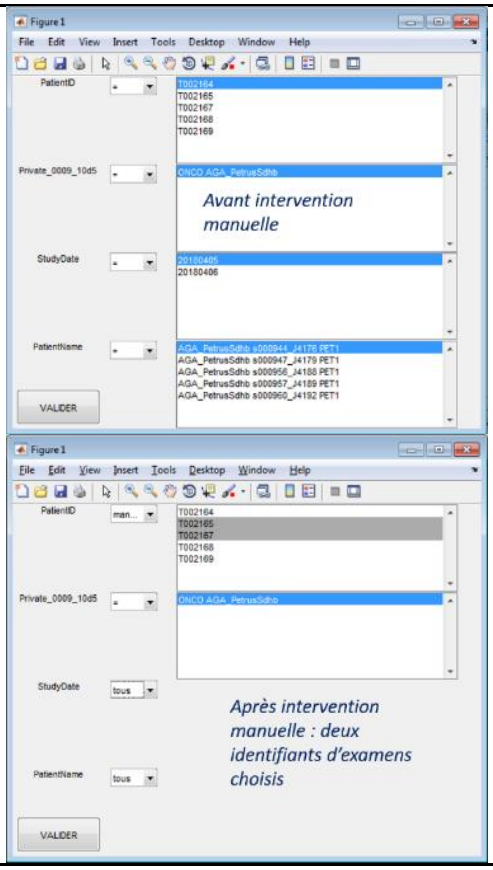


• • •



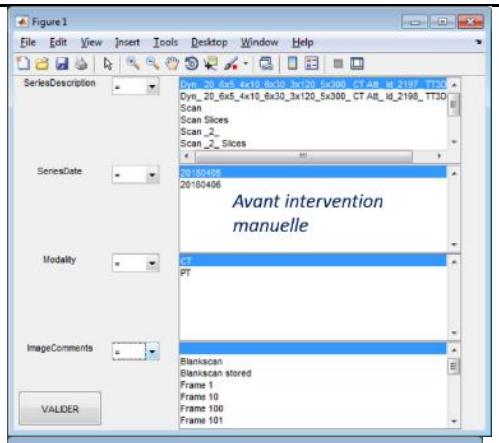
3. Sélection des EXA dans ce dossier

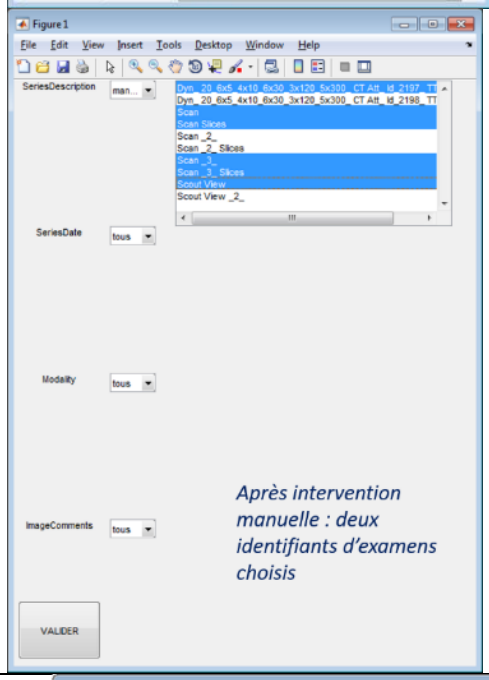
- Liste de toutes les valeurs de champs trouvées dans les examens du répertoire choisis (**à gauche**).
- Restriction aux examens d'intérêt (**à droite**)



4. Sélection des ACQ et DUR dans ces EXAs

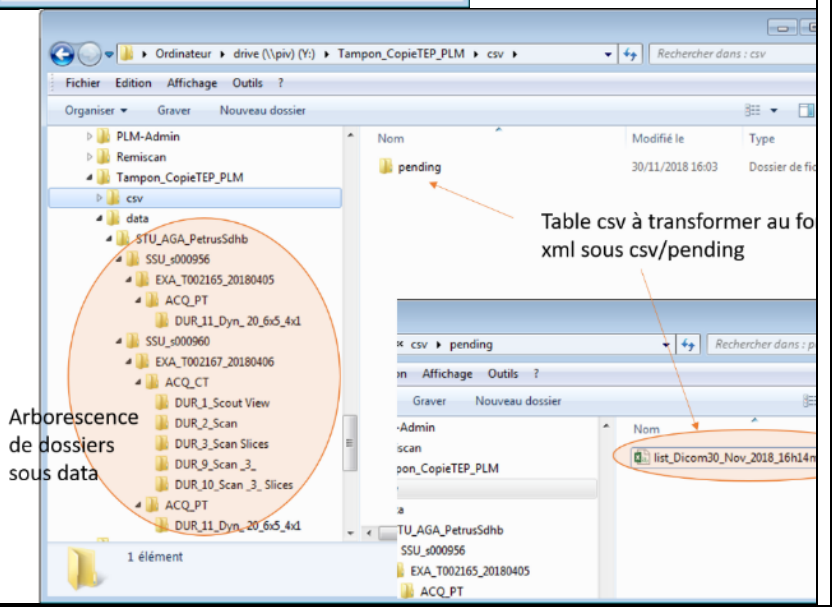
- Liste des valeurs de champs identifiant les acquisitions contenues dans les examens sélectionnés en 3 (**à gauche**).
- Restriction des acquisitions d'intérêt d'après les valeurs de champs qui permet de les identifier (**à droite**)





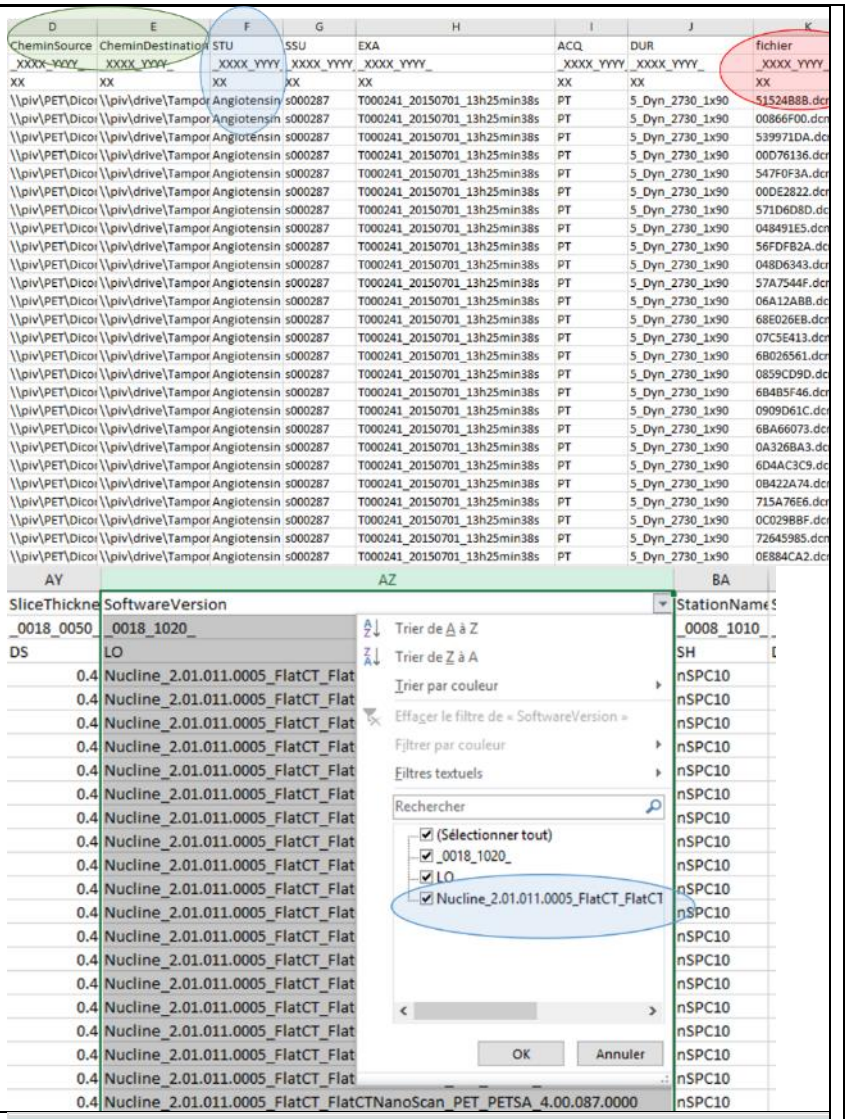
5. Génération de l'arborescence de dossiers et du CSV décrivant les DICOMs

- Généré automatiquement après la sélection
- Rangement rationnel des données avec structure et trigrammes PLM
- Dépôt de la table dans un dossier à traiter (pending)



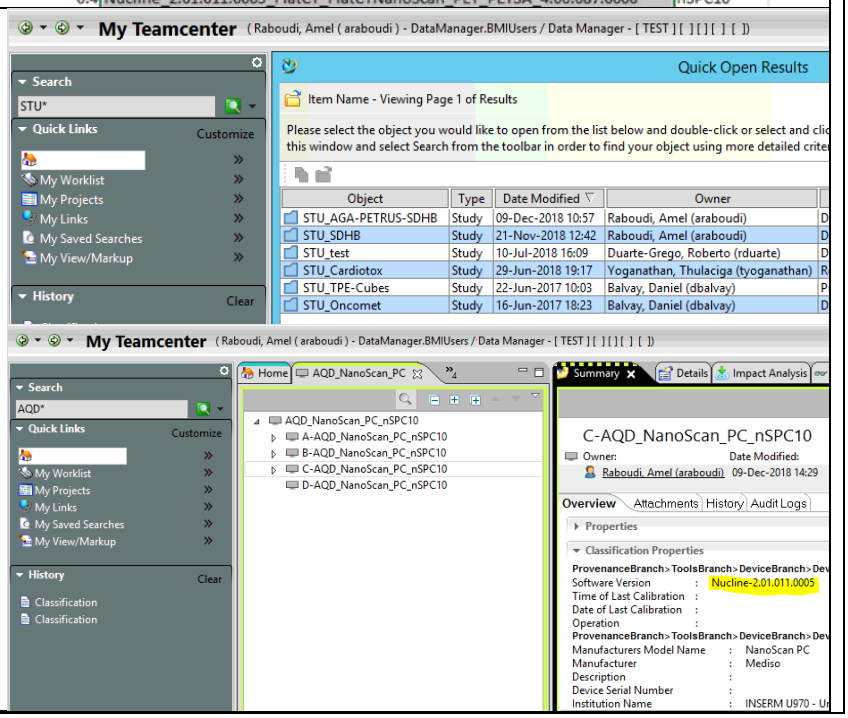
6. **Contrôle** du csv généré

- Bon déroulement de la copie et valeurs normaux dans les colonnes
- Noter les IDs d'étude (STU)
- L'emplacement des copies dans l'espace tampon
- Les fichiers avec .dcm
- Le Software Version (capture d'écran en bas)
- La table peut être utilisée indépendamment du PLM comme une table de pointage des données produites sur le PLM



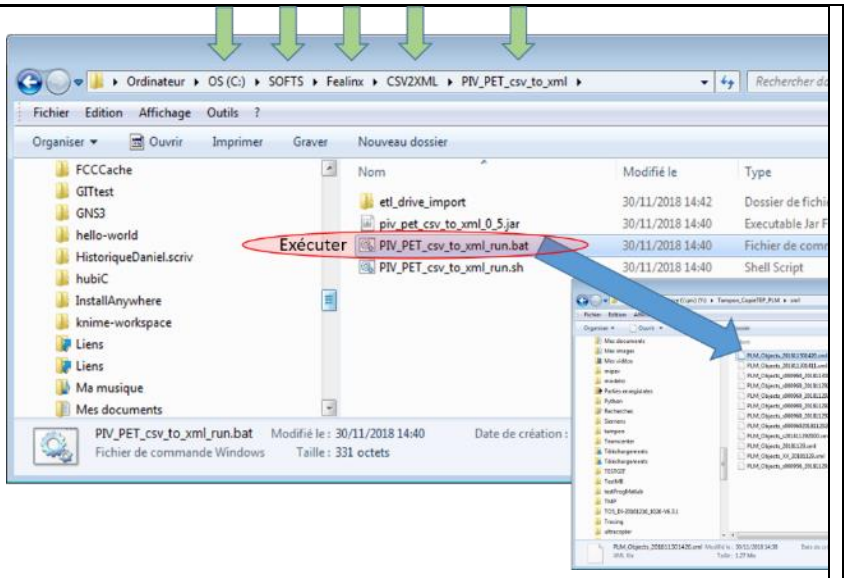
7. **Contrôle** de l'existence <u>préalable</u> de l'étude (STU) et de la version du logiciel dans le PLM : on ne peut pas créer de tels objets par l'outil Mediso2PLM.

- Contacter le référent PLM (Da) pour le faire si besoin
- Captures d'écran : gestion dans le client riche du PLM



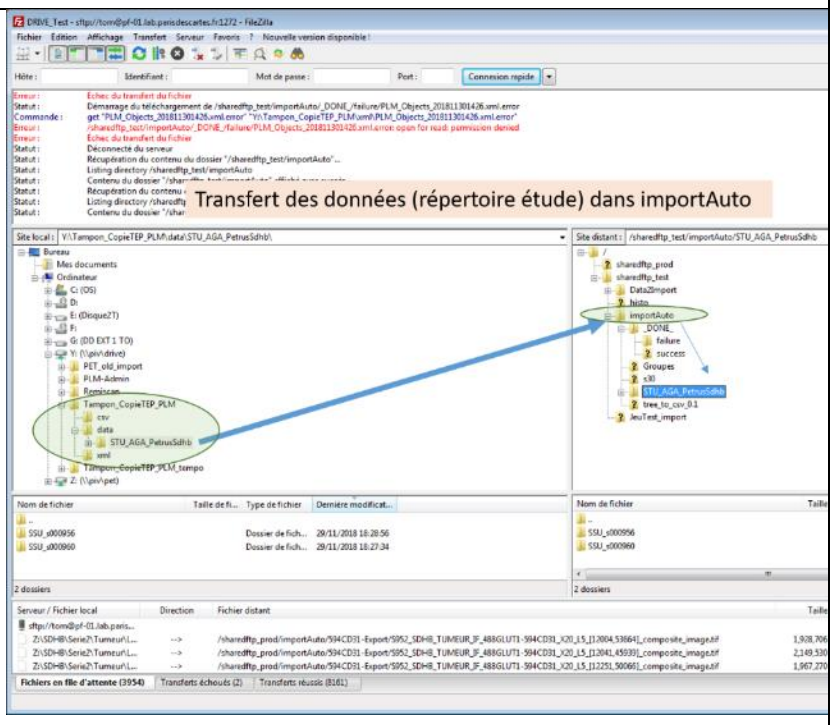
8. Génération d'un fichier xml pour le PLM

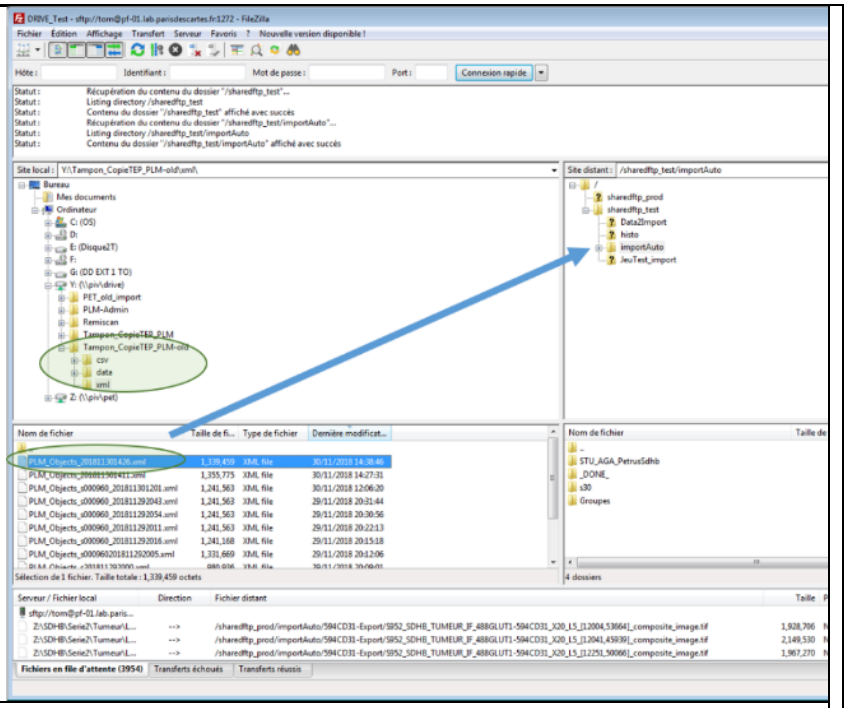
- L'exécution est automatique
- Tous les fichiers csv dans le répertoire « pending » sont convertis en fichiers xml puis déplacés dans un répertoire « historique »



9. Copie de l'arborescence et de l'xml généré par LibDICOM sur le serveur PLM via FTP

- On ouvre FileZilla qui a été préalablement configuré (pour la connexion au serveur du PLM)
- On sélectionne localement (à gauche) l'emplacement des données qui sont sur l'espace tampon : (\\piv\DRIVE\\Tampon CopieTep-PLM
- On sélectionne (à droite) l'emplacement cible : /sharedftp_test/importAuto
- On sélectionne l'étude d'intérêt dans data qu'on déplace vers importAuto (glisser-déposer) voir **grande flèche bleue (im1)**
- L'étude (STU) se copie : petite flèche bleue (im1)
- On sélectionne le fichier xml associé qu'on déplace de la même manière (les originaux sont conservés) : **flèche bleu (im2)**



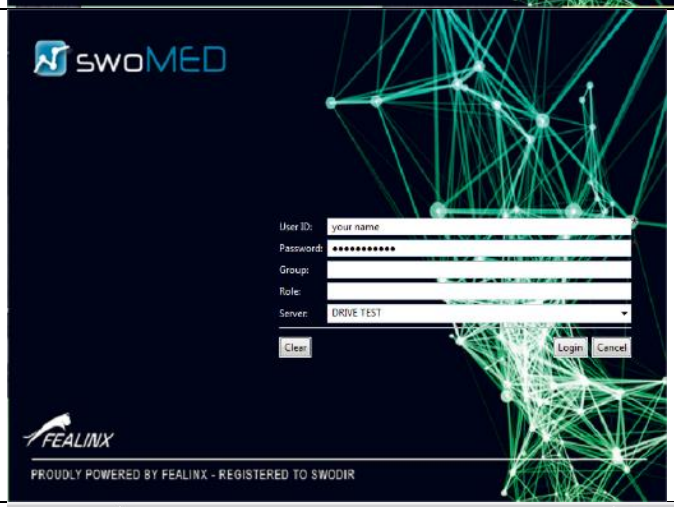
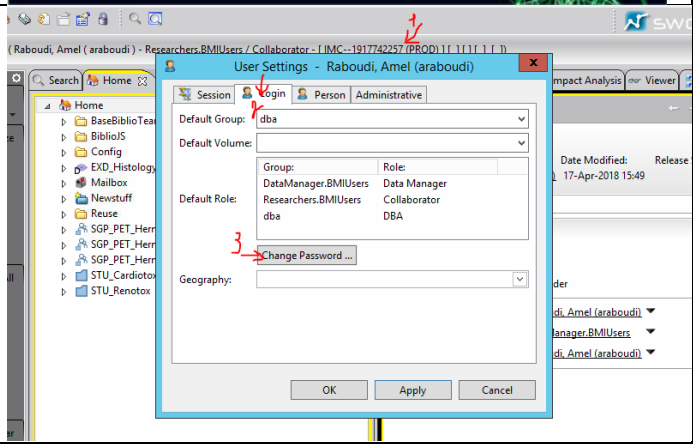
➢ Sélectionner le dernier fichier en date dans la pile	
10. Attendre la fin de l'envoi des EXAs qui peut prendre un certain temps si beaucoup de données doivent être transférées	• Deux événements peuvent indiquer la fin de l'import : o La réception d'un email de notification o Le déplacement du fichier XML envoyé dans _DONE_
11. **Contrôle** de l'intégration des données dans le PLM	Voir section suivante **CONTROLE DE L'INTEGRATION DANS LE PLM : NAVIGATION VIA CLIENT SWOMED**

Contrôle de l'intégration dans le PLM : Navigation via client SWOMed

Il y a deux options de navigation :

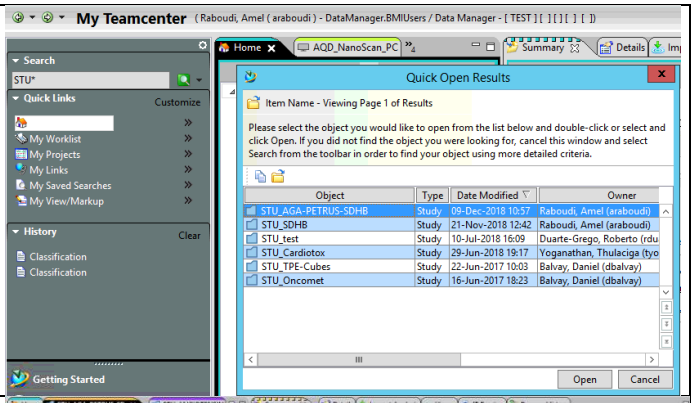
- Le navigateur web via une adresse http :
  - PROD : http://pf-01.lab.parisdescartes.fr:1327/awc/
  - TEST: http://pf-01.lab.parisdescartes.fr:2345/awc/
- Le client SWOMed sur la station de travail **SAPIN**

Ci-après, quelques éléments pour la navigation via client SWOMed.

1. Lancer SWOMed 	
2. Se connecter avec votre compte : PNOM/pnom	
3. Vous pouvez changer votre mot de passe (recommandé)	

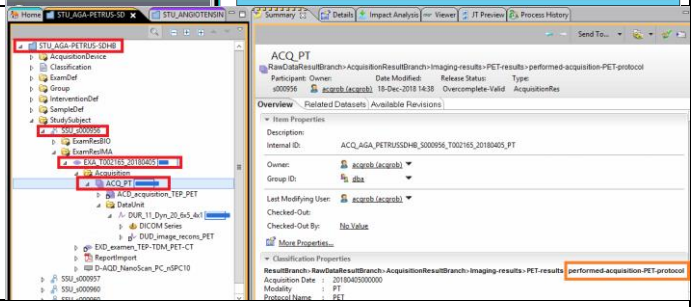
4. Recherche de l'étude copiée

- Dans la barre de **recherche rapide** (à gauche), il faut écrire STU* puis appuyer sur la touche entrée
- Sélectionner en double cliquant, l'étude concernée (dans le popup bleu), Ici STU_AGA-PETRUS-SDHB
- L'objet « dossier » correspondant s'affiche dans le panneau central (arborescence)



5. Examination de l'import

- Double cliquer sur STU_AGA-PETRUS-SDHB et naviguer dans ce qu'il y a dedans via la petite flèche (à gauche dans l'encadré rouge).
- Si vous voulez isoler un objet « dossier » pour une navigation plus claire, il faut double cliquer et ceci l'ouvre dans un nouvel onglet à côté de l'onglet actuel.



En cliquant successivement sur les objets on déplie d'arborescence (à gauche). On contrôle la présence de la structure avec les objets types qui correspondent à l'arborescence envoyé depuis l'espace tampon. On vérifie que tous les examens sont bien présents. À droite on peut considérer des informations complémentaires (encadré orange).

FICHES D'ÉVALUATION DE LA MÉTHODE D'INTÉGRATION DE DONNÉES TEP-TDM

Après la formation Mediso2PLM v3. Les utilisateurs « TVI » et « TYO » ont rempli les formulaires des figures Figure 191 et Figure 192 ci-après le 11/01/2019 par « TVI » et par le 14/01/2019 « TYO » afin d'évaluer le déroulement de la formation et l'acceptation de l'outil. L'écriture manuscrite du formulaire « TYO » est retranscrite au-dessous de la Figure 192.

Tho

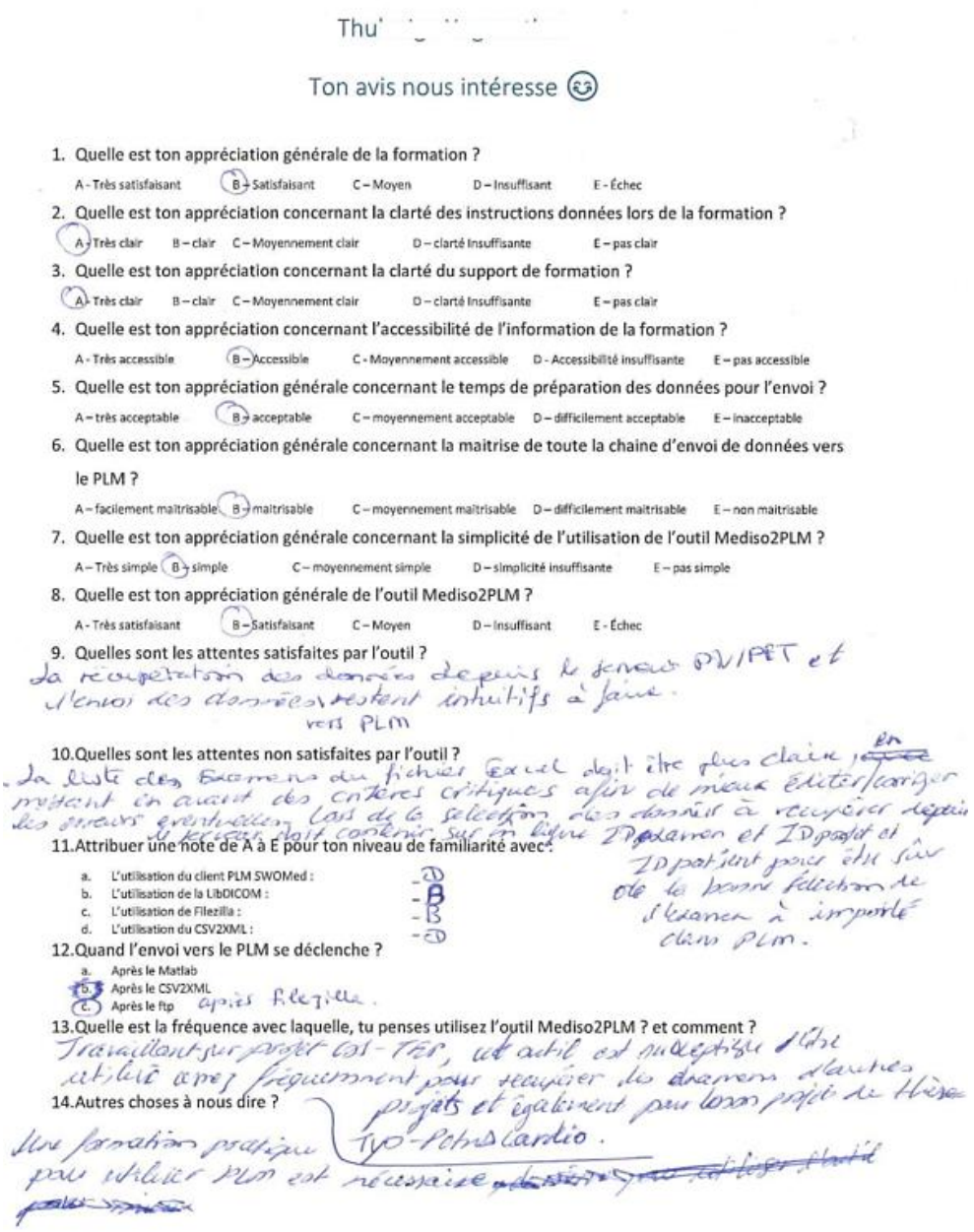
Ton avis nous intéresse ☺

1. Quelle est ton appréciation générale de la formation ?
   A - Très satisfaisant B – Satisfaisant C – Moyen D – Insuffisant E - Échec
2. Quelle est ton appréciation concernant la clarté des instructions données lors de la formation ?
   A - Très clair B – clair C – Moyennement clair D – clarté insuffisante E – pas clair
3. Quelle est ton appréciation concernant la clarté du support de formation ?
   A - Très clair B – clair C – Moyennement clair D – clarté insuffisante E – pas clair
4. Quelle est ton appréciation concernant l'accessibilité de l'information de la formation ?
   A - Très accessible B – Accessible C - Moyennement accessible D - Accessibilité insuffisante E – pas accessible
5. Quelle est ton appréciation générale concernant le temps de préparation des données pour l'envoi ?
   A – très acceptable B – acceptable C – moyennement acceptable D – difficilement acceptable E – inacceptable
6. Quelle est ton appréciation générale concernant la maitrise de toute la chaine d'envoi de données vers le PLM ?
   A – facilement maitrisable B – maitrisable C – moyennement maitrisable D – difficilement maitrisable E – non maitrisable
7. Quelle est ton appréciation générale concernant la simplicité de l'utilisation de l'outil Mediso2PLM ?
   A – Très simple B – simple C – moyennement simple D – simplicité insuffisante E – pas simple
8. Quelle est ton appréciation générale de l'outil Mediso2PLM ?
   A - Très satisfaisant B – Satisfaisant C – Moyen D – Insuffisant E - Échec
9. Quelles sont les attentes satisfaites par l'outil ?

10. Quelles sont les attentes non satisfaites par l'outil ?

11. Attribuer une note de A à E pour ton niveau de familiarité avec :
    a. L'utilisation du client PLM SWOMed : E
    b. L'utilisation de la LibDICOM : E
    c. L'utilisation de Filezilla : C
    d. L'utilisation du CSV2XML : E
12. Quand l'envoi vers le PLM se déclenche ?
    a. Après le Matlab
    b. Après le CSV2XML
    c. Après le ftp
13. Quelle est la fréquence avec laquelle, tu penses utilisez l'outil Mediso2PLM ? et comment ?
    Hebdomadaire
14. Autres choses à nous dire ?

Figure 191 Évaluation Mediso2PLM v3 avec l'utilisateur clé « TVI » dans le cadre du plan d'expérimentation Intégration_2



Thu

Ton avis nous intéresse

1. Quelle est ton appréciation générale de la formation ?
   A - Très satisfaisant (B) Satisfaisant C – Moyen D – Insuffisant E - Échec
2. Quelle est ton appréciation concernant la clarté des instructions données lors de la formation ?
   (A) Très clair B – clair C – Moyennement clair D – clarté insuffisante E – pas clair
3. Quelle est ton appréciation concernant la clarté du support de formation ?
   (A) Très clair B – clair C – Moyennement clair D – clarté insuffisante E – pas clair
4. Quelle est ton appréciation concernant l'accessibilité de l'information de la formation ?
   A - Très accessible (B) Accessible C - Moyennement accessible D - Accessibilité insuffisante E – pas accessible
5. Quelle est ton appréciation générale concernant le temps de préparation des données pour l'envoi ?
   A – très acceptable (B) acceptable C – moyennement acceptable D – difficilement acceptable E – inacceptable
6. Quelle est ton appréciation générale concernant la maitrise de toute la chaine d'envoi de données vers le PLM ?
   A – facilement maitrisable (B) maitrisable C – moyennement maitrisable D – difficilement maitrisable E – non maitrisable
7. Quelle est ton appréciation générale concernant la simplicité de l'utilisation de l'outil Mediso2PLM ?
   A – Très simple (B) simple C – moyennement simple D – simplicité insuffisante E – pas simple
8. Quelle est ton appréciation générale de l'outil Mediso2PLM ?
   A - Très satisfaisant (B) Satisfaisant C – Moyen D – Insuffisant E - Échec
9. Quelles sont les attentes satisfaites par l'outil ?
   la récupération des données depuis le serveur PIV/PET et l'envoi des données restent intuitifs à faire vers PLM
10. Quelles sont les attentes non satisfaites par l'outil ?
    la liste des Examens du fichier Excel doit être plus claire, en mettant en avant des critères critiques afin de mieux éditer/corriger les erreurs éventuelles lors de la sélection des données à récupérer depuis le serveur, doit contenir sur en ligne ID examen et ID projet et ID patient pour être sûr de la bonne sélection de l'examen à importé dans PLM.
11. Attribuer une note de A à E pour ton niveau de familiarité avec :
    a. L'utilisation du client PLM SWOMed : D
    b. L'utilisation de la LibDICOM : B
    c. L'utilisation de Filezilla : B
    d. L'utilisation du CSV2XML : D
12. Quand l'envoi vers le PLM se déclenche ?
    a. Après le Matlab
    (b.) Après le CSV2XML
    (c.) Après le ftp après Filezilla.
13. Quelle est la fréquence avec laquelle, tu penses utilisez l'outil Mediso2PLM ? et comment ?
    Travaillant sur projet COS-TEP, cet outil est susceptible d'être utilisé assez fréquemment pour récupérer des données d'autres projets et également pour mon projet de thèse TYO-PetrusCardio.
14. Autres choses à nous dire ?
    Une formation pratique pour utiliser PLM est nécessaire

Figure 192 Évaluation Mediso2PLM I3 avec l'utilisateur clé « TYO » dans le cadre du plan d'expérimentation Intégration_2

9. La récupération des données depuis PIV/PET (le serveur local) et l'envoi des données restent intuitifs à faire vers PLM (le nom courant du système BMS-LM au LRI)

10. La liste des examens des fichiers Excel doit être plus claire, en mettant en avant des critères critiques afin de mieux éditer/corriger les erreurs éventuelles lors de la sélection des données à récupérer depuis le serveur, doit contenir sur une ligne ID examen et ID projet et ID patient pour être sûr de la bonne sélection de l'examen à importer dans PLM.

13. travaillant sur le projet COS-TEP, cet outil est susceptible d'être utilisé assez fréquemment pour récupérer des données d'autres projets et également pour mon projet de thèse « TYO_PetrusCardio »

14. Une formation pratique pour utiliser PLM est nécessaire

ANNEXE E : AUTRES CAS D'APPLICATION AU LABORATOIRE LRI

Dans cette annexe, nous présentons, tout d'abord, les réalisations effectuées pour l'intégration de données TEP-TDM Mediso2PLM en sa version 4. Nous détaillons en deuxième lieu l'application du système BMS-LM pour les données en protéomique.

MEDISO2PLM VERSION 4

La version 4 du sous-projet Mediso2PLM a fait l'objet du stage de Master2 de Ismael Bakayoko et constituait une version exploratoire afin de tester si le passage à une solution 100% web et libre (sans Matlab) améliorerait ou non l'utilisabilité de l'outil de gestion de données. Lors de cette version, nous avons aussi exploré la possibilité de corriger les informations DICOM avant l'envoi au système BMS-LM.

Pour rappel, les travaux se sont déroulés en trois étapes :

4. La conversion de la bibliothèque LibDICOM Matlab (fournie à l'issue de Mediso2PLM version 3) en Python. Une formation de l'utilisateur clé « DBA » à son utilisation et maintenance a été effectuée.
5. La mise en place d'architecture web pour regrouper toutes les étapes de l'intégration « générique » dans un même outil web : du côté du laboratoire LRI, la bibliothèque LibDICOM a été encapsulée dans un serveur Django accessible via des messages URI GET depuis une interface web, et du côté du système BMS-LM, un client web personnalisé a été mis en place pour faciliter la consultation de données via l'API REST du système BMS-LM. L'interface web au LRI et le client personnalisé ont été développés en utilisant le Framework Angular.
6. La mise en place d'interfaces graphiques adaptées pour la sélection de données, leur correction, et leur envoi ainsi que leur consultation après envoi au système BMS-LM.

Les technologies choisies sont celles utilisées à l'entreprise Fealinx : Python, services web, Angular, Django). Le but étant d'effectuer un transfert d'expertises au laboratoire LRI tout en rendant l'environnement logiciel de la gestion de données plus simple et ergonomique que ce qui existait auparavant (Mediso2PLM v3 : Matlab, FTP, « Client Riche » Teamcenter, etc.). L'architecture qui a été mise en place lors de ce stage est détaillée dans la Figure 193. Nous expliquons ses composants dans les paragraphes suivants.

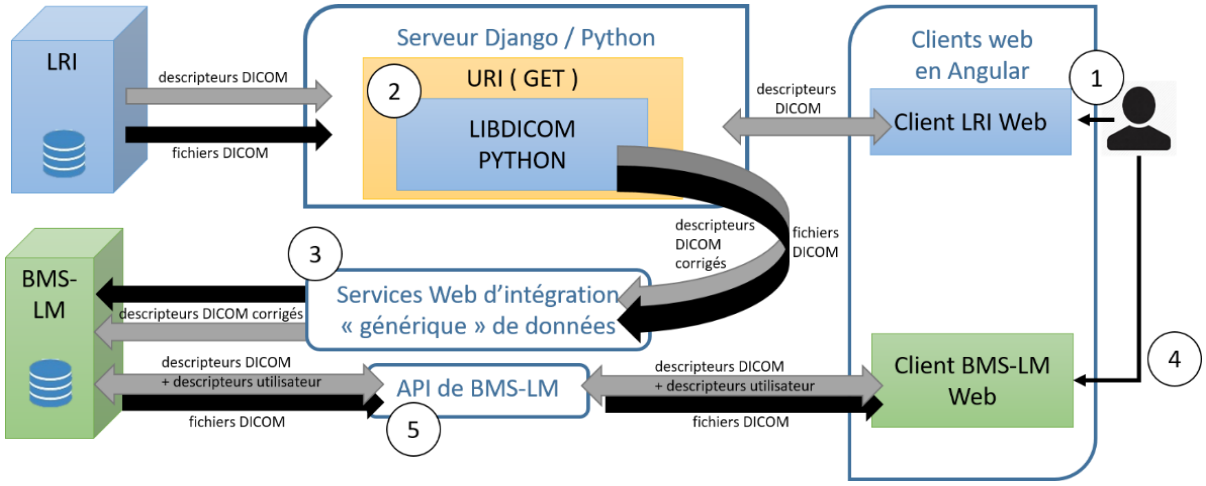


Figure 193 Architecture mise en place lors du passage au web pour l'intégration et l'exploration des données au laboratoire LRI
adaptée du rapport de stage de Ismael Bakayoko

Un scénario typique d'utilisation commence par une connexion utilisateur au stockage local du laboratoire LRI ((1) dans la Figure 193), en se connectant via le client LRI-Web. L'utilisateur a accès à la liste d'études en cours dans la base de données locale LRI (voir Figure 189) pour sélectionner les données. Il prépare ainsi les données à envoyer au système BMS-LM en faisant appel au serveur Django ((2) dans la Figure 193). Le serveur Django utilise la version Python de la bibliothèque LibDICOM. Comme dans la version Mediso2PLM v3, l'utilisateur a la possibilité de filtrer les données (voir Figure 195) qu'il souhaite envoyer au système BMS-LM. Il a, de plus, la possibilité de corriger les incohérences dans les descripteurs DICOM avant envoi. Après la sélection, l'envoi des données (flèches noires sur la Figure 193), et des descripteurs DICOMs (flèches grises dans la Figure 193) est effectué via la méthode d'intégration « générique » ((3) dans la Figure 193). Ensuite, et une fois l'e-mail de fin d'import de données reçu, l'utilisateur se connecte au serveur distant du système BMS-LM via le Client « BMS-LM Web » ((4) dans la Figure 193). Ce client personnalisé se connecte au serveur BMS-LM via l'API REST de ce dernier ((5) dans la Figure 193).

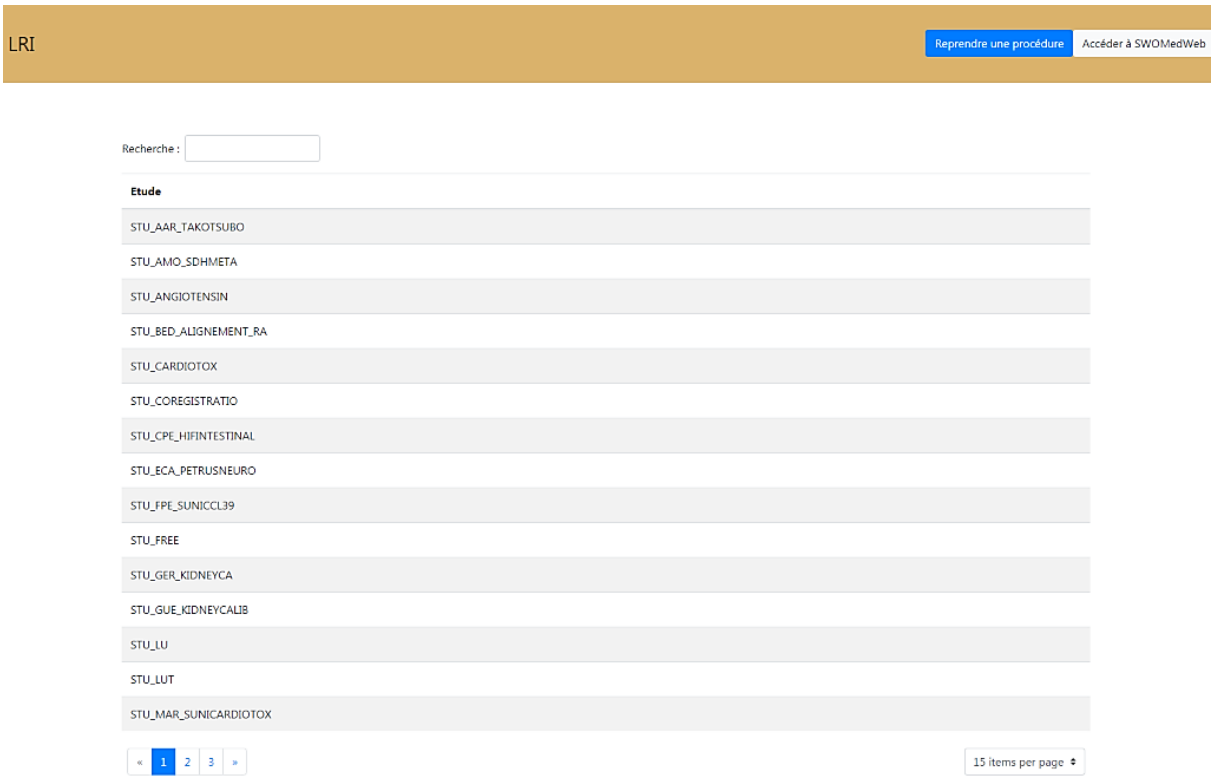


Figure 194 Interface web listant les études au laboratoire LRI

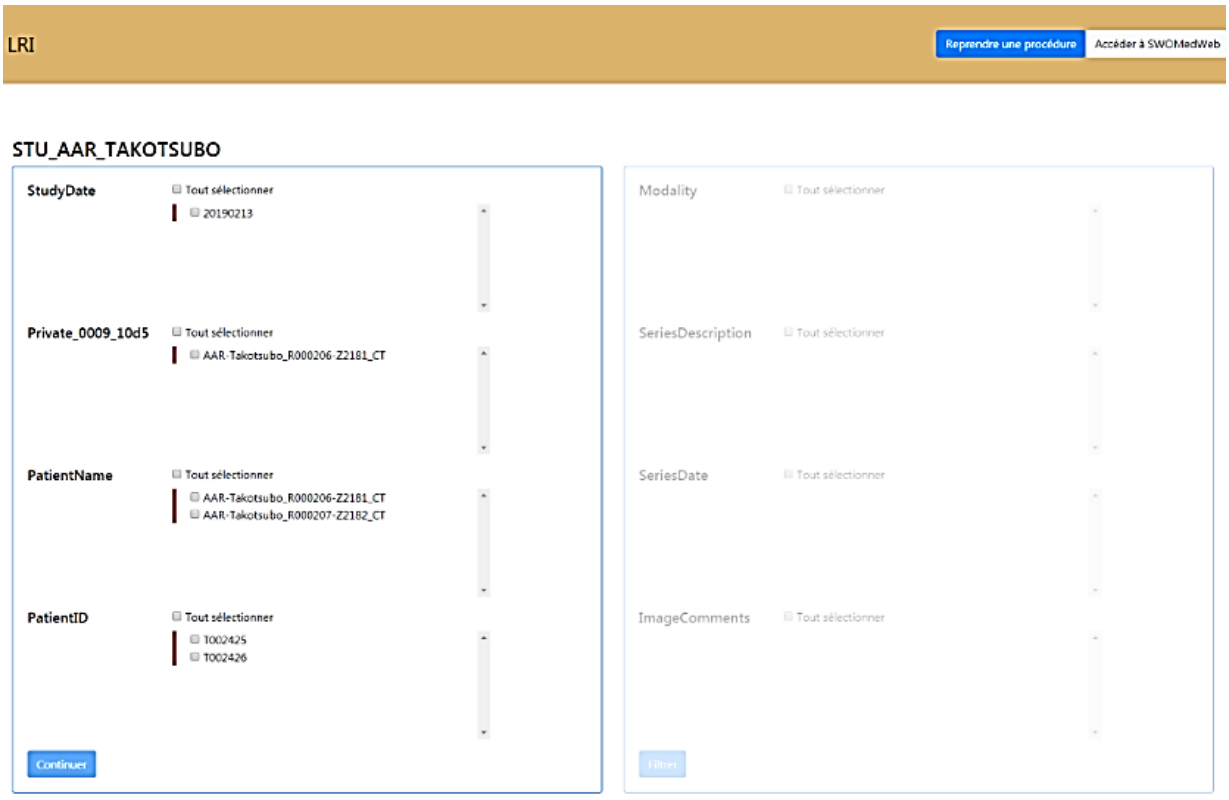


Figure 195 Interface web permettant de filtrer les données à envoyer au système BMS-LM

La consultation des données et des descripteurs envoyés est effectuée par le client web « SWOPARCC», nom choisi pour désigner le client de BMS-LM personnalisé pour le laboratoire LRI du centre de recherche PARCC. Une fois connecté au système BMS-LM (interface Figure 196), l'utilisateur peut consulter les études de recherche archivées dans le système et dont il a accès (interface Figure 197). Il aura la possibilité de télécharger les images après exploration (interface Figure 198). Cette dernière n'a pas été achevée. Le but de la refonte des interfaces et de mettre à disposition des chercheurs au laboratoire LRI, des interfaces plus adaptées et familières (utilisation via un navigateur web) pour l'exploration des données dans le système BMS-LM.

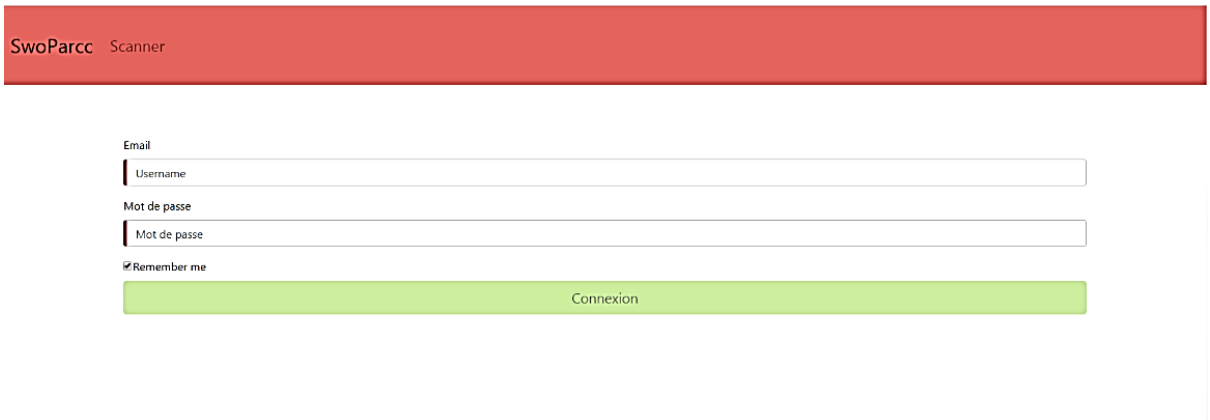


Figure 196 Interface de connexion au système BMS-LM depuis le client personnalisé « SWOPARCC »

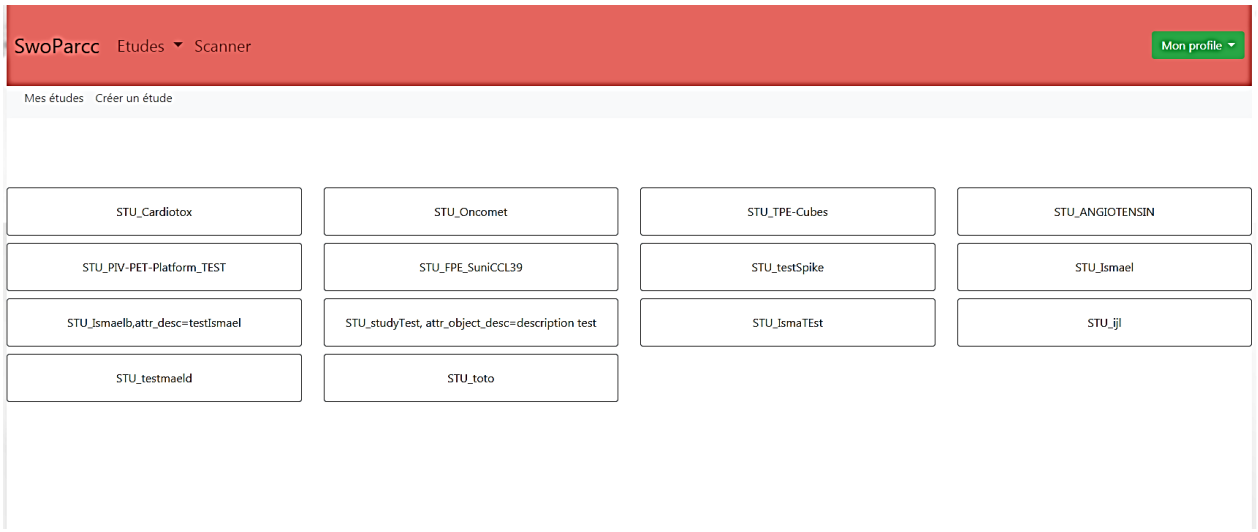


Figure 197 Interface d'accueil du client web personnalisé BMS-LM listant les études auxquelles la personne est autorisée

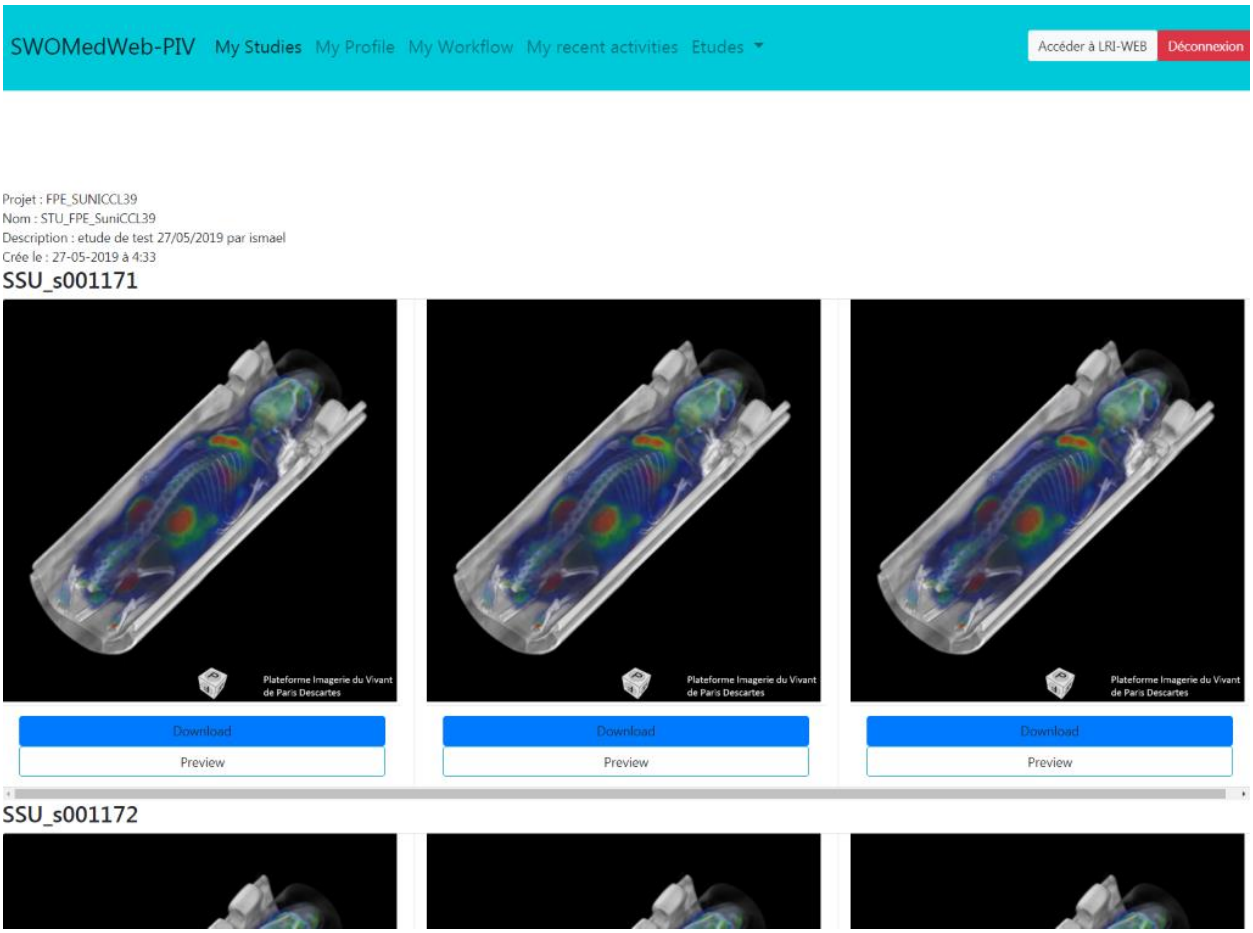


Figure 198 Interface web pour le téléchargement d'images d'intérêt depuis le système BMS-LM

INTÉGRATION ET RÉUTILISATION DES DONNÉES EN PROTÉOMIQUE

Les données de protéomique représentent un cas d'application pertinent pour l'intégration de données et de calculs scientifiques dans le système BMS-LM. En effet, aucun expert en données acquises via Spectrométrie de Masse (MS) n'a été identifié ni dans le laboratoire LRI, ni dans l'entreprise Fealinx. Pourtant ces données sont acquises dans le cadre de plusieurs projets de recherche internes au LRI (Cardiotox, SDHB, Takutsobo, etc..). Les connaissances qui y sont liées doivent être bien archivées en

vue de leur réutilisation à long terme pour d'autres projets de recherche. L'intégration de ces données et des calculs correspondants a fait l'objet d'une présentation détaillée lors de la conférence SMMAP2017.

Intégration des données brutes en protéomique

Nous avons suivi la méthode d'intégration présentée (§IV.2.1). Après collecte de données, nous avons interviewé le 22/06/2017, une personne compétente de la plateforme 3P5 de l'Hôpital Cochin, qui a réalisé les examens de protéomique. L'entretien nous a permis de collecter les informations nécessaires en regard de la provenance et du cycle de vie des données. Nous avons ainsi préparé le tableau Excel des descripteurs pour l'annotation des données d'entrée. Après, ce tableau est transformé en format XML via l'ETL de transformation de données. Les données de protéomique ont été les premières données intégrées en utilisant la méthode d'intégration « générique », le schéma du tableau d'entrée et de l'ETL a beaucoup évolué depuis le 09/2017 (période où l'import des données protéomique a été réalisé).

Les principales informations collectées dans le tableau des descripteurs sont : l'identifiant de l'étude (interne : Cardiotox, externe : OICL151130), les identifiants des souris (interne : S000248, externe : ctrl15248), l'identifiant de l'échantillon et tous les identifiants d'objets nécessaires à la traçabilité de la provenance, l'enzyme utilisée, les paramètres d'oxydation, les chemins d'accès aux fichiers à intégrer, le groupe de souris correspondant. Dans la Figure 199 ci-après présente une capture d'écran d'un fichier XML généré à la fin de l'ETL de transformation de données.

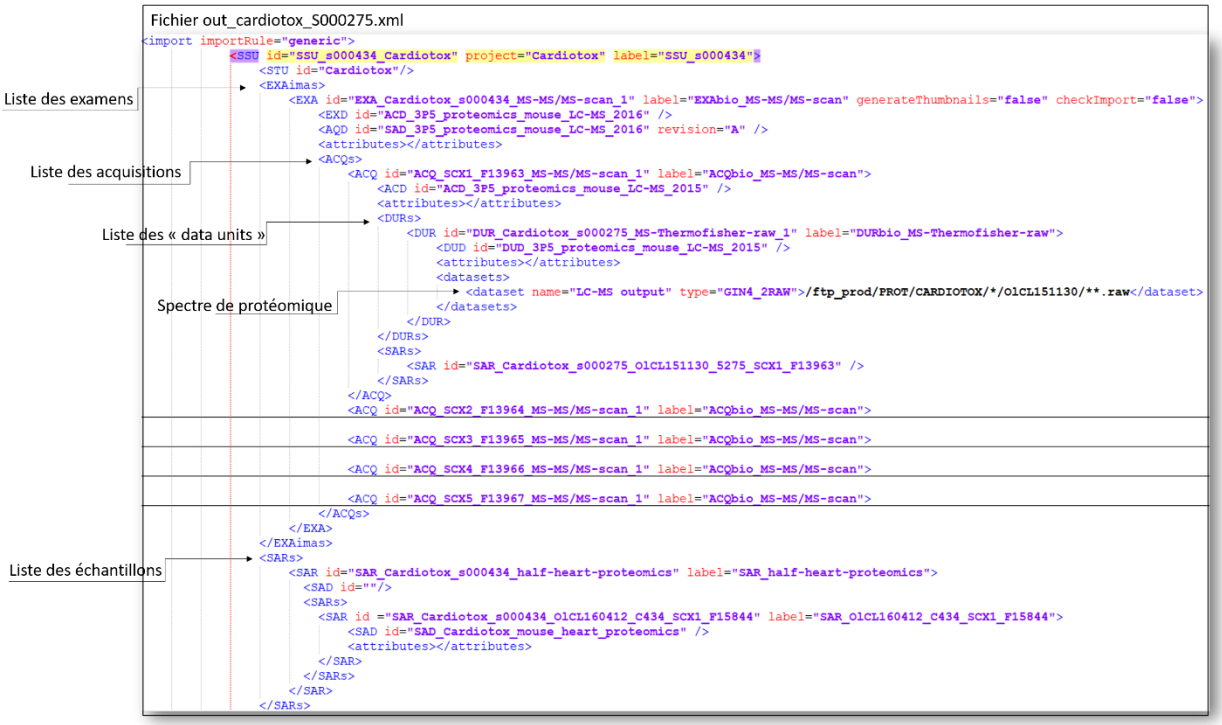


Figure 199 Exemple de fichier XMLs générés lors de la procédure d'intégration des données de protéomique

En parallèle, nous avons proposé une structuration des données ainsi qu'une liste de descripteurs pour préparer le système BMS-LM « MDD+Classification » à l'intégration de données de protéomique. Pour ce faire, nous avons collecté les informations sur les groupes de souris, les machines d'acquisition et les échantillons utilisés en plus des données sur l'examen réalisé. (voir Figure 200 et Figure 201 suivantes).

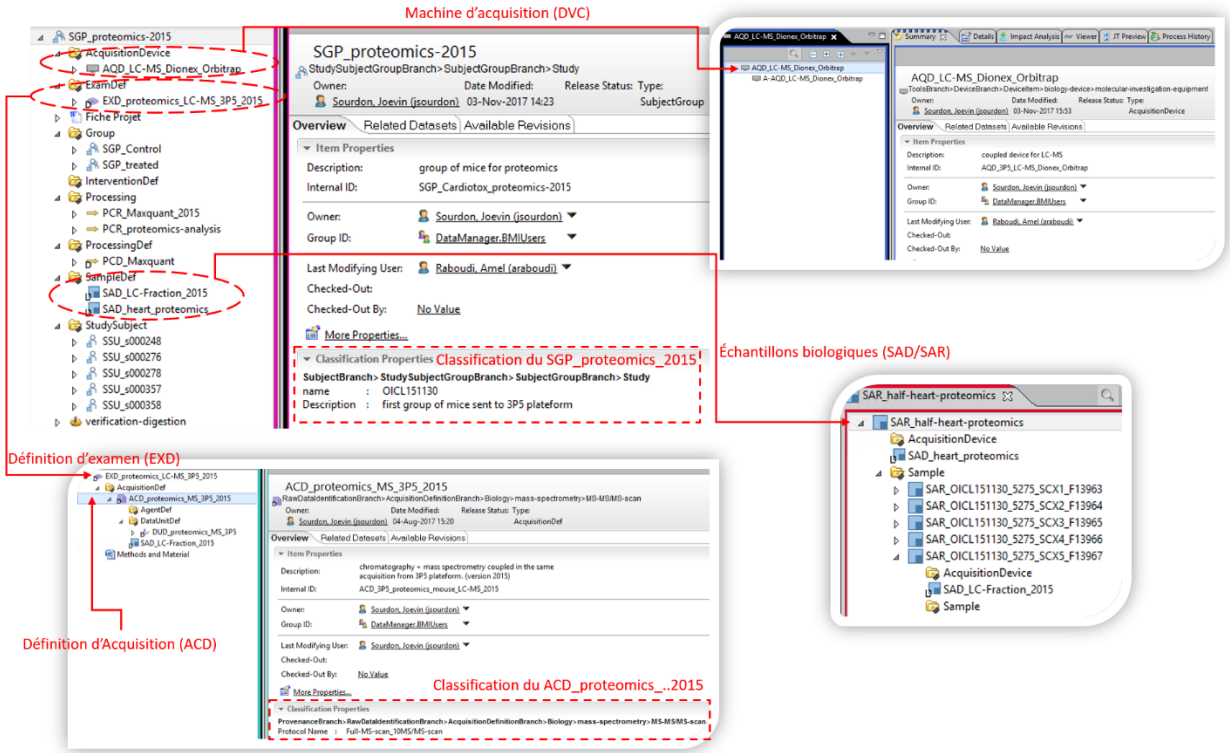


Figure 200 Objets ajoutés dans le système BMS-LM afin de le préparer à l'intégration des données BMS-LM

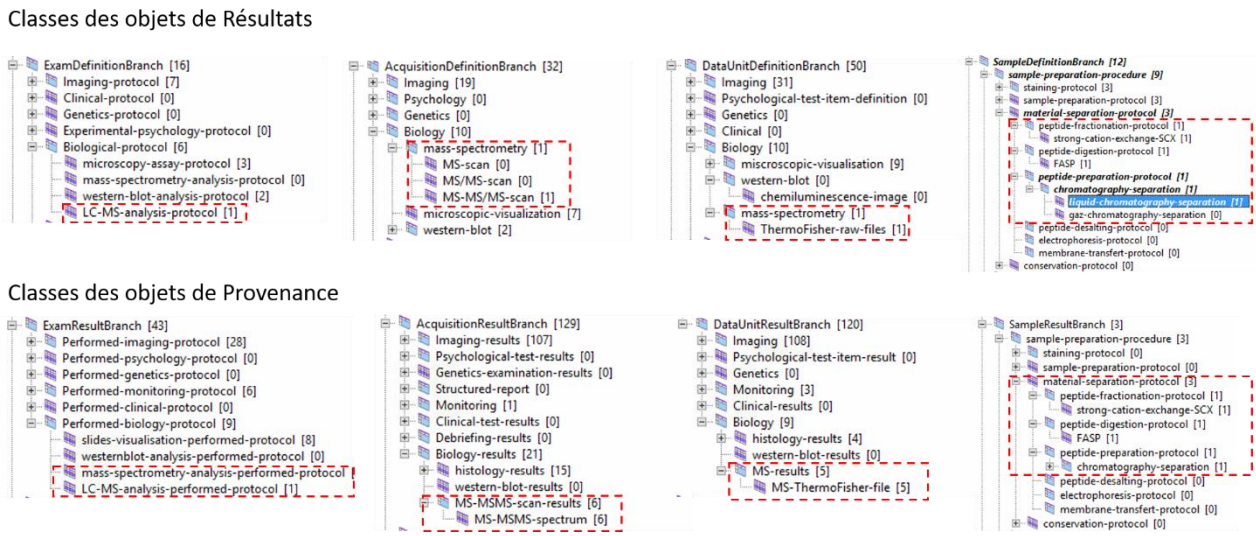


Figure 201 Classes de « Classification » ajoutées en vue d'intégration des données de protéomique

Les données archivées dans le système BMS-LM sont au nombre de 24 examens pour 24 souris avec et 24 échantillons de cœurs, qui ont donné lieu à 31 fractions de peptides et de 116 spectres de spectrométrie de masse, bien contextualisés et stockés pour une réutilisation ultérieure (calcul effectué en utilisant « l'outil de recherche rapide » sur le serveur de PROD DRIVE le 14/12/2020).

Avant l'intégration de données dans le système BMS-LM, nous avions la liste des spectres de MS dans un dossier où tout est listé en vrac. Avec notre méthode d'intégration de données, nous avons contextualisé ces fichiers, inconnus au départ, et nous avons ajouté des éléments de provenance pour les décrire. Ci-après, dans la Figure 202, un exemple du résultat de l'intégration des données de protéomique. Nous avons identifié les éléments clé via des annotations illustratives ajoutées à la capture d'écran Figure 202 . Il faut naviguer dans l'arborescence à gauche, en découvrant les différents examens

passés par un SSU et ses différents échantillons biologiques. Pour chaque examen, et avant de trouver le fichier résultant, l'utilisateur prend connaissance de la version de la machine utilisée « A-AQD_LC-MS_Dionex_Orbitrap » et des paramètres de l'acquisition.

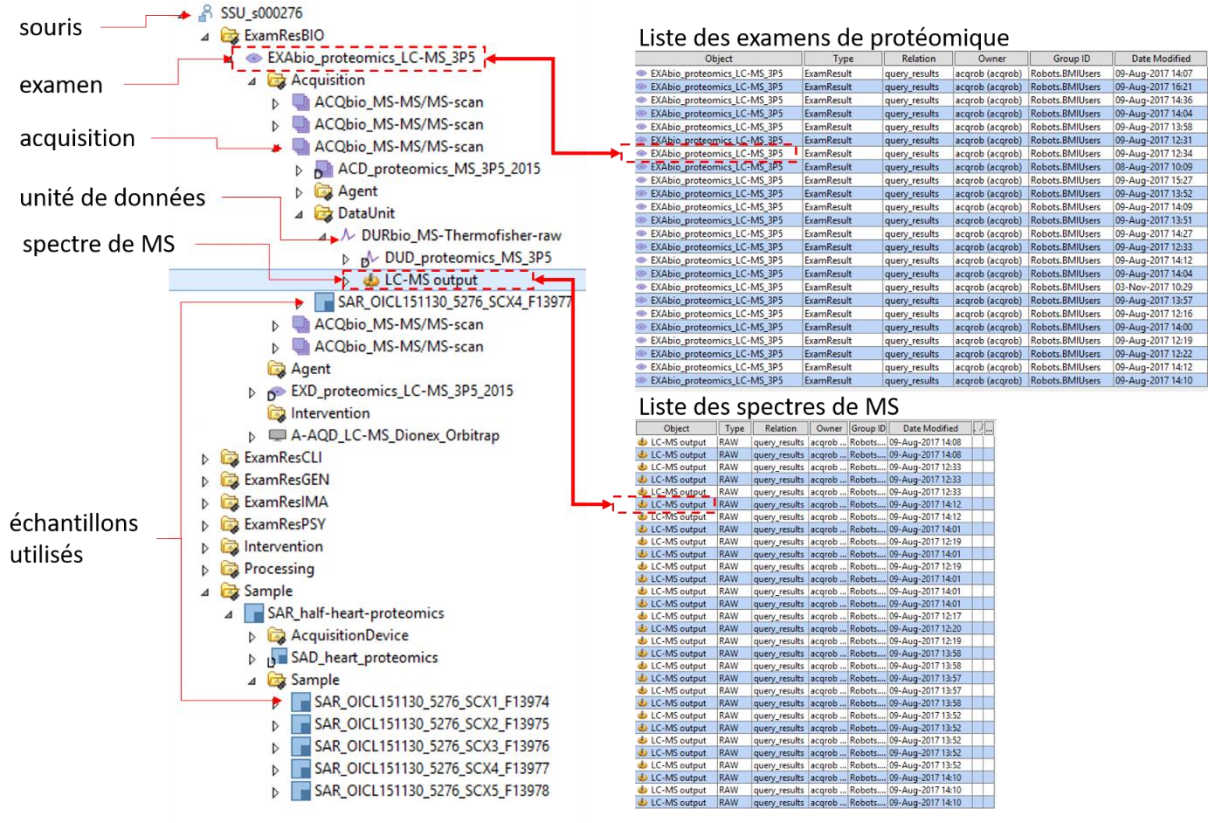


Figure 202 Exemple de données protéomiques dans le système BMS-LM

Réutilisation des données pour une intégration « totale » d'un calcul scientifique en protéomique

Comme indiqué préalablement, l'analyse protéomique est effectué par un laboratoire partenaire (plateforme 3P5) et donc le LRI ne peut pas exploiter les données spectrales directement, mais doit passer par des tableaux Excel d'analyses fournis par ce laboratoire partenaire.

Nous avons réutilisé les données brutes préalablement intégrées dans le système BMS-LM pour mettre à disposition des chercheurs au LRI un nouveau calcul d'analyse protéomique qui utilise les standards du domaine. Ce calcul permet l'identification et la quantification des protéines dans un tissu, via l'analyse de leur signature spectrale détectée par spectrométrie de masse. Il analyse les spectres automatiquement sans passer par les fichiers Excels de la plateforme 3P5. Ce développement a été élaboré en collaboration avec l'ingénieur en data science Dr Pierre-Yves Hervé de l'entreprise Fealinx.

Nous présentons ce calcul avant et après son intégration dans le système BMS-LM, afin de mieux expliciter l'intégration et la traçabilité qui y sont associées. Les étapes du workflow sont :

1. La conversion des spectres en entrée, initialement au format RAW Thermofisher, au format MzXML, format standard en protéomique.
2. L'analyse via l'outil OpenMS des spectres en entrée pour l'identification des peptides
3. L'analyse via OpenMS des peptides pour l'identification des protéines

Le calcul a été implémenté via l'outil Knime d'automatisation des workflows scientifiques. Il représente un outil de développement de scripts de calcul scientifique à partir d'une bibliothèque de nœuds

réutilisables. Les nœuds Knime utilisés pour l'étape 2 d'identification des peptides sont présentés dans la Figure 203 suivante.

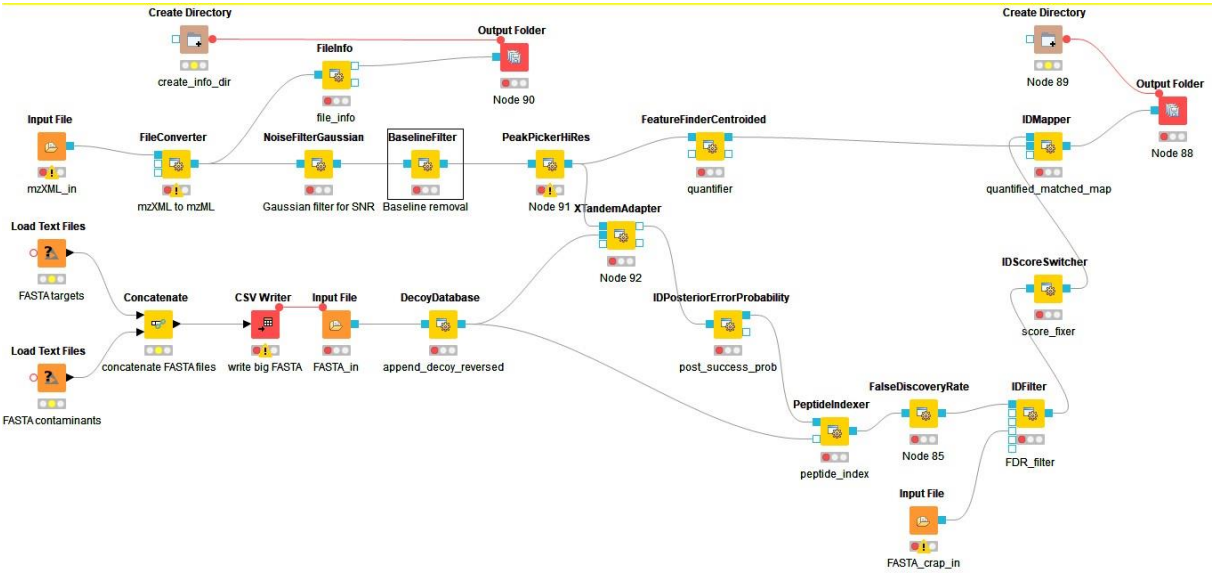


Figure 203 Capture d'écran du script KNIME pour le calcul scientifique utilisé dans la quantification des peptides

Pour pouvoir intégrer ce calcul dans le système BMS-LM, il faut préparer ce dernier à son exécution, via sa modélisation dans le système BMS-LM (voir Figure 204). Les objets de provenance sont modélisés via les concepts noyaux (MDD) et spécifiés via les concepts de domaine (« Classification ») comme dans la capture d'écran Figure 204 ci-après. En effet, le « script Knime de quantification et identification de peptides – étape 2 » est modélisé comme une chaîne de traitement via les objets de provenance : « Processing Definition (PCD) » et de « Processing Unit Definition » (PUD) ». Chaque objet est spécifié via une classe appartenant à la « Classification ». Par exemple, « PCD_ima_proteomics_peptides » est doublement spécifié via les classes « peptide-identification » et « peptides-quantification » (voir Figure 204) pour indiquer son rôle.

Une fois le système BMS-LM configuré, il faut exécuter la chaîne de traitement. Les objets et données (PCR, PUR, ...) résultants suivent la structure des objets de provenance (PCD, PUD...) spécifiée Figure 204. Pour lancer l'exécution, il faut sélectionner le WFI (Figure 204) et le SGP qui regroupent les données d'entrée (Figure 200) et les envoyer au lancement du workflow à la suite d'une commande clavier CTRL+P.

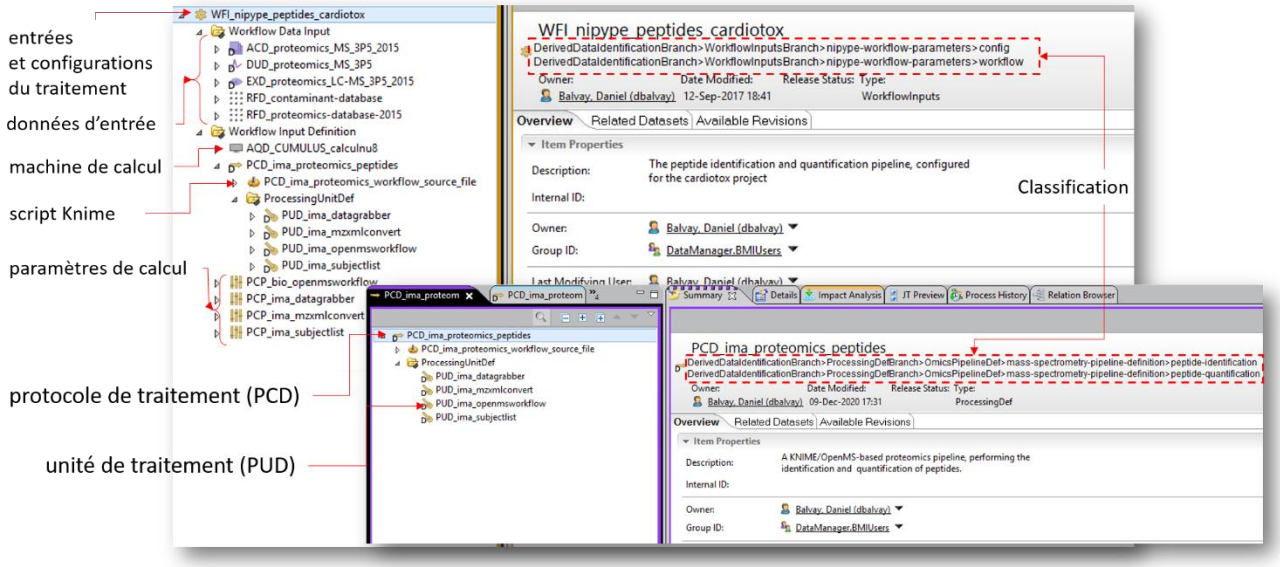


Figure 204 Objets du MDD spécifiés via les classes de classification pour la modélisation du script Knime

La Figure 205 suivante détaille les différents objets qui ont été instanciés dans le système BMS-LM en tant que résultat d'exécution de la chaîne de traitement. Par exemple, les données d'entrée à l'unité de traitement « PUR_proteomics_peptides.mzXML.Convert » sont les spectres « LC-MS output » sous format « .RAW Thermofisher » et qui seront convertis en « mzXML ». Plus généralement, dans un résultat d'une unité de traitement (PUR), il faut renseigner toutes les informations de provenance et de contexte qui y sont liés. Dans la capture d'écran (Figure 205, voir la version en ligne pour les couleurs), les liens du PUR « PUR_proteomics_peptides.openMSWorkflow » sont déployés afin de présenter les relations avec les objets associés, listés ci-après.

- mzXML_out : un « *derived data unit* » résultant de l'unité de traitement (PUR) de conversion des spectres mzXMLConvert. Il est en entrée au PUR d'identification des peptides.
- STL_Knime_peptide_identification_quantification : le script logiciel qui a été exécuté lors de ce PUR.
- RFD_proteomics-database-2015 et RFD_contaminant-database : des « *reference data* » en entrée, indispensable pour l'exécution de l'unité de traitement. Ils représentent une liste de protéines et de contaminants qui peuvent être présents dans les échantillons.
- PCR_proteomics_peptide : l'élément modélisant la chaîne de traitement (PCR) à laquelle appartient le PUR.
- PCP_bio_openmsworkflow : la liste des paramètres de cette unité de traitement (PUR).
- Identified_quantified_peptides : un « *derived data unit* », résultat de ce PUR.
- Identified_peptides : un « *derived data unit* », résultat de ce PUR.
- mzML_information : un « *derived data unit* », résultat de ce PUR

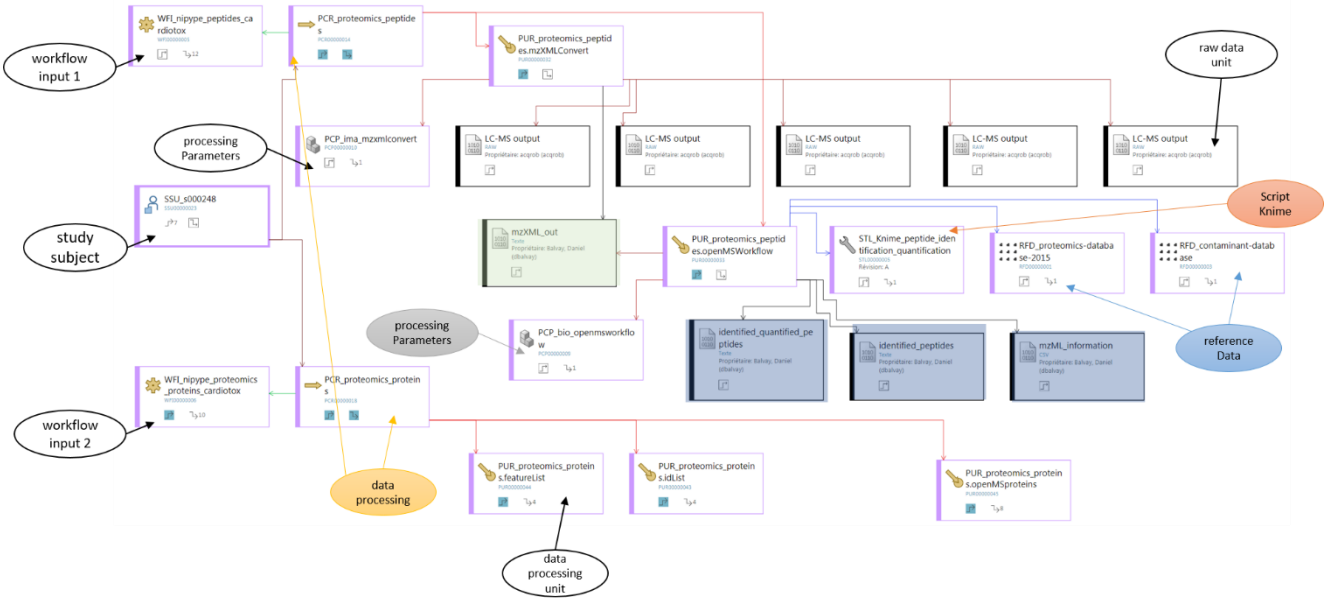


Figure 205 Objets résultants de l'intégration totale d'un workflow d'analyse protéomique

L'exécution du workflow Knime, encapsulé dans le système BMS-LM, a été réalisée sur cinq échantillons d'une même souris : la souris « SSU_S000248 » et a donné lieu à une liste conséquente de protéines identifiées en se basant sur les données de références en entrée « RDF_proteomics-database-2015 » et en retirant celles identifiées comme contaminant dans la base « RFD_contaminant-database ». La Figure 206 présente, à titre informatif, les protéines les plus présentes dans ces échantillons d'entrée.

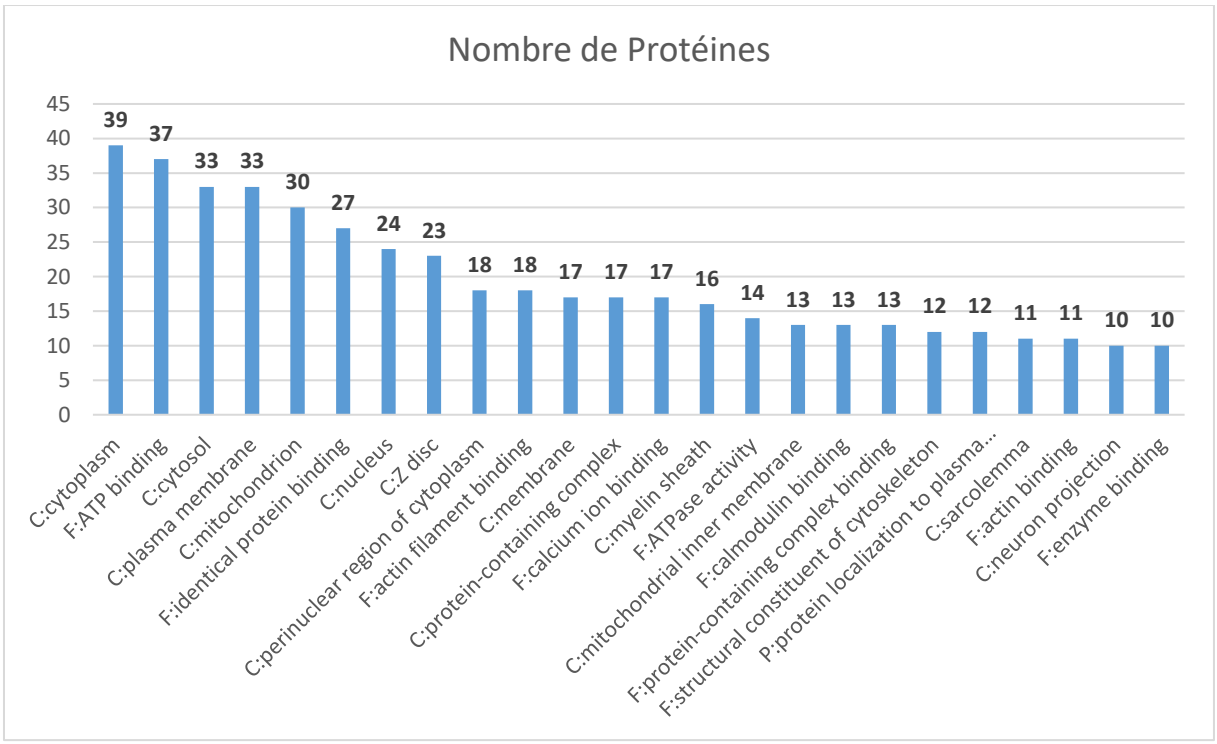


Figure 206 Les protéines les plus présentes dans les échantillons de la souris « SSU_S000248 »