Interopérabilité des données médicales dans le domaine des maladies rares dans un objectif de santé publique

Meriem Maaroufi

THESE DE DOCTORAT DE
L'UNIVERSITE PIERRE ET MARIE CURIE

Spécialité
Informatique Biomédicale

ECOLE DOCTORALE PIERRE LOUIS DE SANTE PUBLIQUE A PARIS : EPIDEMIOLOGIE ET SCIENCES DE L'INFORMATION BIOMEDICALE

Présentée par

Mme Meriem Maaroufi

Pour obtenir le grade de

DOCTEUR de l'UNIVERSITÉ PIERRE ET MARIE CURIE

Sujet de la thèse :

INTEROPERABILITE DES DONNEES MEDICALES DANS LE DOMAINE DES MALADIES RARES DANS UN OBJECTIF DE SANTE PUBLIQUE

soutenue le 07/11/2016

devant le jury composé de :

Marie-Christine Jaulent, DR, INSERM, Directrice de thèse
Paul Landais, PU-PH, Université de Montpellier, Directeur de thèse
Sandra Bringay, PR, Université de Montpellier, Rapporteur
Régis Beuscart, PU-PH, Université de Lille 2, Rapporteur
Xavier Jeunemaître, PU-PH, Université Paris Descartes, Examinateur
Annick Clément, PU-PH, Université Pierre et Marie Curie, Examinateur

Université Pierre & Marie Curie - Paris 6
Bureau d'accueil, inscription des doctorants et base de données
Esc G, 2ème étage
15 rue de l'école de médecine
75270-PARIS CEDEX 06

Tél. Secrétariat : 01 42 34 68 35
Fax : 01 42 34 68 40
Tél. pour les étudiants de A à EL : 01 42 34 69 54
Tél. pour les étudiants de EM à MON : 01 42 34 68 41
Tél. pour les étudiants de MOO à Z : 01 42 34 68 51
E-mail : scolarite.doctorat@upmc.fr

REMERCIEMENTS

J'adresse mes remerciements à tous ceux qui ont contribué à la réalisation de ce projet de thèse.

En premier lieu, je remercie mes directeurs de thèse Marie-Christine Jaulent et Paul Landais. Sans leur soutien et leur confiance, je n'aurais pas été aussi loin dans ce projet. Ils m'ont appris la pensée et la méthode scientifique sans quoi mes travaux de recherche n'auraient pas abouti. Merci pour votre disponibilité.

Je remercie aussi Rémy Choquet, directeur opérationnel de la Banque Nationale des Données Maladies Rares. Il m'a donné l'opportunité de faire partie de cette merveilleuse aventure qu'est le projet BNDMR et m'a fait toucher du doigt la complexité de l'interopérabilité dans le domaine de la santé.

Je remercie les membres du jury d'avoir accepté d'évaluer mes travaux : Sandra Bringay, Régis Beuscart, Xavier Jeunemaître et Annick Clément. Merci pour votre disponibilité.

Mes remerciements vont à tous les anciens et actuels membres de la cellule opérationnelle de la BNDMR : Claude Messiaen, Yannick Fonjallaz, Céline Angin, Amélie Ruel, Jean-Philippe Necker, Minh Maccabiau et Albane de Carrara. Leur collaboration était essentielle pour certains des travaux décrits dans cette thèse. Merci surtout pour cette belle ambiance au sein de l'équipe.

Merci à tous les membres du LIMICS desquels j'ai beaucoup appris durant les staffs du laboratoire.

Je remercie enfin l'Assistance Publique des Hôpitaux de Paris de m'avoir permis de travailler sur le projet BNDMR.

SOMMAIRE

Remerciements **1**

Sommaire **2**

Chapitre I : Introduction générale **7**

1 E-santé et santé publique **7**

1.1 L'e-santé est un fait **7**

1.2 L'innovation digitale pour la santé publique **8**

2 Maladies rares et santé publique **8**

2.1 Qu'est-ce qu'une maladie rare ? **8**

2.2 Enjeux de santé publique **9**

2.3 Les maladies rares en France **10**

2.3.1 Plans nationaux maladies rares (PNMR) 10

2.3.2 Les structures maladies rares 11

2.3.3 Epidémiologie et maladies rares 14

3 La Banque Nationale de Données Maladies Rares **17**

3.1 BNDMR - Le projet **17**

3.1.1 Présentation et objectifs 17

3.1.2 Structure organisationnelle 18

3.2 BNDMR – Le système d'information pour les maladies rares **20**

3.2.1 CEMARA – La base de données historique 20

3.2.2 BaMaRa-BNDMR le nouveau système d'information pour les maladies rares 21

4 Cadre d'interopérabilité pour les maladies rares **24**

4.1 Problématique : Une hétérogénéité constatée **24**

4.2 Les trois piliers du cadre d'interopérabilité **26**

4.2.1 Identification des patients 26

4.2.2 Interopérabilité des données médicales 27
4.2.3 Gestion des flux de données 28
4.3 Propositions 28
Chapitre II : Identification des patients 30
1 Introduction 30
1.1 Contexte et objectifs 30
1.2 Contraintes 32
1.2.1 Unique 32
1.2.2 Pérenne 32
1.2.3 Anonyme 32
1.2.4 Global 32
2 Identifiant patient et chainage de données : Etat des lieux 33
2.1 Un identifiant national de santé en France 33
2.1.1 Le Numéro d'Inscription au Répertoire (NIR) 34
2.1.2 L'Identifiant National de Santé - Calculé 36
2.2 Les numéros patient pour le chainage des données 38
2.2.1 Chainage SNIIRAM PMSI 38
2.2.2 Identifiants patients dans les entrepôts de données maladies rares 39
2.2.3 Systèmes multimodaux 39
2.3 Tableau récapitulatif 41
3 L'identifiant patient maladie rare : IdMR 43
3.1 Choix des technologies et des informations pour la construction de l'IdMR 43
3.1.1 Comment garantir l'anonymat ? 43
3.1.2 Quelles données? 44
3.2 Processus de génération de l'IdMR 45
3.2.1 Les étapes du processus de génération de l'IdMR 46

3.2.2 Schéma général du processus de génération de l'IdMR....49

3.3 Un identifiant pour les fœtus: Pourquoi et comment ?....50

3.3.1 Motivations....50

3.3.2 Un IdMR pour les fœtus ?....51

4 Evaluation de l'algorithme de l'IdMR....53

4.1 Matériel....53

4.2 méthode de détection des collisions: critère d'évaluation des résultats....53

4.3 Résultats....54

5 Conclusion....56

5.1 Résumé....56

5.2 Discussion....56

5.2.1 Non intégration de la « commune de naissance »....56

5.2.2 Risque non nul pour les contraintes de départ....57

5.3 Conclusion....60

Chapitre III : L'intégration de données – du recueil standardisé à la découverte de correspondances....63

1 Introduction....63

1.1 Approche entrepôt pour l'intégration des données maladies rares....63

1.2 Les différents niveaux d'hétérogénéité....64

1.3 Propositions....65

2 Set minimal de données maladies rares et standardisation....66

2.1 Set minimal de données maladies rares....66

2.1.1 Qu'est-ce qu'un set minimal de données ?....66

2.1.2 Méthodologie de création du set minimal de données maladies rares....68

2.1.3 Résultats....72

2.2 Un format électronique interopérable pour le set minimal de données....73

2.2.1 Les standards : de la modélisation des données aux terminologies spécialisées 73
2.2.2 Standardisation du set minimal de données maladies rares 79
2.3 Discussion 83
3 Alignement et découverte de correspondances 84
3.1 Vers l'automatisation des alignements 84
3.1.1 Contexte 84
3.1.2 Problématique sémantique 85
3.1.3 Classifications des techniques d'alignement automatisé 87
3.2 Tests avec un outil d'alignement 88
3.2.1 Choix de l'outil d'alignement 89
3.2.2 Tests sur un alignement de terminologies 89
3.2.3 Test sur un alignement des schémas de données 93
3.2.4 Conclusion des tests 95
3.3 Proposition : Alignement des schémas de données et processus d'intégration de données 96
3.3.1 Besoins 96
3.3.2 Formalisation des correspondances 98
3.3.3 Les processus spécifiques d'intégration de données 102
3.4 Discussion 108
4 Articulation entre recueil standardisé et alignements 110
Chapitre IV : Apports et perspectives 112
1 Une approche globale pour l'intégration des données maladies rares 112
1.1 Complémentarité au sein du cadre d'interopérabilité maladies rares 112
1.2 De la recherche à l'opérationnel 114
1.2.1 Une approche différente des autres expériences 114
1.2.2 Migration CEMARA 114

1.2.3 Reprise de données des registres et autres bases de données 115
1.2.4 Les DPI et le cadre d'interopérabilité de l'ASIP Santé 117
1.2.5 IdMR et études transversales 117
2 Problématiques connexes 118
2.1 Qualité de données 118
2.2 Codage diagnostic des maladies rares 118
2.3 Annuaire pour les structures maladies rares 119
3 Au-delà de la BNDMR 120
Bibliographie 123
Annexe 1 : Set Minimal des Données Maladies Rares (v1.09.2) 135
Annexe 2 : Standardisation HL7 FHIR du set minimal de données maladies rares 142
Annexe 3 : Jeux de valeurs et standardisation 147
Table des illustrations 152
Table des Tableaux 154

CHAPITRE I : INTRODUCTION GENERALE

1 E-SANTE ET SANTE PUBLIQUE

1.1 L'E-SANTE EST UN FAIT

L'e-santé ou la santé électronique est l'application des technologies de l'information et de la communication au domaine de la santé et du bien-être (*Science & Santé* 2016). Le secteur de l'e-santé ne cesse de se développer et commence à peser au niveau économique. Le marché mondial de l'e-santé a été estimé à 85.44 milliards de dollars en 2014 et la tendance est à la hausse avec une croissance de 15,8% attendue pour 2022 (« eHealth Market Size & Share | Global Industry Report, 2022 » 2016). En France, Le potentiel du marché de l'e-santé a été estimé quant à lui à 2,7 milliards d'euros pour la même année (Direction Générale des Entreprises 2016). La branche de la télésanté, incluant, entre autres, la télémédecine et la m-santé, est certes en plein essor mais le segment le plus développé de l'e-santé, selon ces études, reste de loin celui des systèmes d'information (SI) de santé. Ces SI, qu'ils soient hospitaliers ou à destination des professionnels de santé libéraux, englobent de multiples outils assurant la gestion des établissements et des rendez-vous, la gestion des dossiers médicaux des patients qu'ils soient génériques ou de spécialité, l'aide à la décision pour le diagnostic et le traitement, la gestion des essais cliniques…

Les pouvoirs publics ne cessent d'investir et d'inciter au développement des systèmes d'information de santé et croient en leur potentiel d'action sur diverses problématiques :

- réduire les coûts en améliorant l'efficacité du système de soins ;
- garantir la sécurité des patients en améliorant la qualité des soins (en réduisant les erreurs) et en permettant une médecine personnalisée ;
- faciliter la mise en place de réseaux d'expertise et le partage des données des patients dans le cadre de prise en charge mutualisée ville-hôpital ;
- faciliter le partage des données de soins pour la santé publique (surveillance sanitaire et épidémiologie) et la recherche (recherche translationnelle).

Même si les chercheurs restent partagés quant à la validation de ces apports (Schneider 2015; Murphy et al. 2012; Blaya, Fraser, et Holt 2010; Glasgow 2007; Piette et al. 2012), l'e-santé demeure néanmoins un domaine de recherche en pleine expansion.

1.2 L'INNOVATION DIGITALE POUR LA SANTE PUBLIQUE

La santé publique se digitalise. Les prises de décision, définissant la politique de santé publique, se basent de plus en plus sur des études et des analyses de données numériques. Ces données sont généralement collectées par des systèmes institutionnels, tels que les systèmes de surveillance des maladies infectieuses gérés par l'Institut de Veille Sanitaire en France (InVS) (Santé publique France, InVS 2016).

D'autres systèmes, initialement non destinés à cet usage, commencent aussi à être utilisés pour faire de l'épidémiologie et de la veille sanitaire. La réutilisation des données de soin, issues des SI hospitaliers, pour l'épidémiologie et la recherche est une approche qui a été adoptée au sein de plusieurs projets (Balas, Krishna, et Tessema 2008; Geissbuhler et al. 2013; De Moor et al. 2015). Ces projets s'articulent généralement autour de deux phases. La première phase est la phase d'intégration de données de plusieurs sources telles que deux ou plusieurs établissements hospitaliers. Dans un deuxième temps, des techniques de fouille de données sont appliquées pour extraire des connaissances et répondre à des questions posées par les chercheurs (Prather et al. 1997; Mullins et al. 2006). Ces dernières années, l'utilisation de ces techniques s'est étendue aux données des réseaux sociaux ou des moteurs de recherche pour la détection précoce de phénomènes épidémiologiques (Seifter et al. 2010; Chunara, Andrews, et Brownstein 2012; Salathé et al. 2013, 9; Eysenbach 2009). Etant donné les quantités importantes de ce type de données à traiter dans ce contexte, l'utilisation des technologies Big Data est devenue indispensable. Cependant, des questions restent posées quant à l'utilité de l'utilisation des technologies Big Data sur des données issues des systèmes d'information de santé et quant à la fiabilité des études qui en découleraient (Murdoch TB et Detsky AS 2013; Khoury et Ioannidis 2014).

2 MALADIES RARES ET SANTE PUBLIQUE

2.1 QU'EST-CE QU'UNE MALADIE RARE ?

Une maladie rare est une affection dont la prévalence est faible. Il n'existe pas une seule définition universelle pour les maladies rares mais plusieurs définitions épidémiologiques avec des seuils de prévalence qui dépendent des pays. En Europe par exemple, une maladie est considérée comme rare lorsque sa prévalence ne dépasse pas 1 cas sur 2000 (« Eurordis Position Paper on the WHO Report on Priority Medicines for Europe and the World » 2004). Aux Etats-Unis, le seuil de prévalence fixé est de 1 cas sur 1500 (van Weely et Leufkens 2004).

Du fait de la multiplicité des définitions il est compliqué de dresser une liste universelle et exhaustive des maladies rares (Orphanet 2016b). Ceci est d'autant plus complexe lorsque la prévalence calculée n'est pas une prévalence à l'échelle mondiale mais sont des prévalences restreintes à des territoires donnés. Ceci implique que certaines maladies puissent être considérées comme rares dans certains pays et non rares dans d'autres. Selon la définition européenne, la drépanocytose est une maladie rare en France avec une prévalence de 1 cas sur 30.000. Selon la même définition, elle n'est plus considérée comme étant rare en Afrique Noire mais plutôt fréquente avec une prévalence de 1 cas sur 30 (« La drépanocytose » 2011).

Selon les estimations, il existerait entre 6000 et 8000 maladies rares (« Orphanet » 2012; « Définition et chiffres clés » 2016). Considérées dans leur totalité, l'impact de ces maladies devient plus évident. En France, l'ensemble de ces maladies rares affecterait entre 3 et 4 millions de personnes (« RARE Diseases: Facts and Statistics » 2012). En Europe, ce chiffre augmente pour toucher 30 millions de personnes ce qui représenterait 4% de la population européenne contre 25 millions aux Etats-Unis ce qui représenterait 8% de la population du pays (« About Rare Diseases | www.eurordis.org » 2016; Haffner, Whitley, et Moses 2002). Ces maladies sont très diverses : maladies neuromusculaires, anomalies du développement, maladies métaboliques, maladies auto-immunes, maladies infectieuses, cancers rares,... Quatre-vingt pour cent d'entre elles ont une origine génétique. Deux maladies sur trois sont graves et invalidantes. Chez un patient sur deux survient un déficit moteur, sensoriel ou cognitif. Enfin dans près d'un cas sur deux le pronostic vital est engagé.

2.2 ENJEUX DE SANTE PUBLIQUE

Une maladie rare, considérée indépendamment des autres, touche très peu de personnes. La rareté des cas expose à un manque de connaissances sur l'histoire naturelle de la maladie. Ceci

pose un problème essentiel au niveau de la prise en charge car les durées d'errance diagnostique, selon les maladies, peuvent être très conséquentes. De plus, la plupart de ces maladies rares sont des maladies orphelines; des maladies pour lesquelles il n'existe pas de traitement curatif. Les patients qui sont pris en charge bénéficient de soins visant à prolonger la durée de vie et à améliorer la qualité de vie ce qui est primordial pour des maladies qui, comme nous l'avons souligné, sont pour la plupart chroniques, invalidantes et engageant souvent le pronostic vital.

Les maladies rares, considérées dans leur totalité, constituent clairement un enjeu de santé publique. Cependant, même à cette échelle, il manque des données épidémiologiques assez précises. En particulier, dans les systèmes actuels, les maladies sont souvent mal « codées » (« WHO/OMS Vaccines and Biologicals » 2016). En effet, dans les systèmes d'information hospitaliers elles sont pour la plupart identifiées avec des codes non spécifiques et agrégatifs tels que « autres troubles endocriniens et métaboliques » de la Classification Internationale des Maladies puisque cette dernière ne répertorie pas la totalité des maladies rares. Des études épidémiologiques plus précises permettraient d'une part une meilleure structuration des réseaux de prise en charge des patients atteints de maladies rares, mais aussi une avancée en recherche clinique en facilitant la mise en place d'études permettant de mieux comprendre l'interaction génotype-phénotype, d'essais cliniques pour les nouveaux médicaments et/ou d'études interventionnelles ou observationnelles. Les maladies rares soulèvent donc des problématiques de soins et de recherche mais aussi des problématiques économiques et sociétales nécessitant une surveillance épidémiologique globale.

2.3 LES MALADIES RARES EN FRANCE

2.3.1 *PLANS NATIONAUX MALADIES RARES (PNMR)*

En collaboration avec les experts du domaine et les associations de patients, les autorités françaises ont entrepris plusieurs initiatives en faveur des maladies rares, plus spécifiquement en proposant deux plans maladies rares consécutifs et un troisième récemment annoncé par la Ministre le 15 juin 2016 (Roinet 2016). Le premier plan maladies rares (2005-2009) a favorisé la mise en place d'un réseau d'expertise en maladies rares afin de faciliter le recours aux soins et d'améliorer la prise en charge des patients (« Plan National Maladies Rares 2005-2008 » 2005). Concrètement, 131 centres de référence maladies rares ont été labellisés. Un ensemble de 501

centres de compétences ont aussi été désignés en 2008 par les agences régionales de l'hospitalisation afin de mieux couvrir l'ensemble du territoire (plus de détails dans la section suivante).

Le second plan maladies rares (2011-2016) définit notamment, dans le cadre de l'axe d'amélioration de la qualité de prise en charge des patients, la mise en œuvre d'un ensemble d'outils informatiques visant à soutenir l'activité de soin et de recherche pour les maladies rares (« Plan National Maladies Rares 2011-2014 » 2011). Le programme national RADICO, programme Investissements d'avenir, vise à offrir les outils et les services nécessaires pour la mise en place de cohortes maladies rares pour la recherche (« Radico - Rare Disease Cohorts » 2016). La Banque Nationale de Données Maladies Rares répondra quant à elle à ces principaux objectifs:

- Documenter les modes de prise en charge et leur impact afin d'en réduire les coûts en évaluant l'activité des centres labellisés et en optimisant leur financement ;
- Décrire la demande de soins et son niveau d'adéquation avec l'offre correspondante en évaluant le maillage du réseau maladies rares afin de l'adapter en conséquence ;
- Identifier les patients souffrant de maladies rares et susceptibles d'être éligibles pour des essais cliniques et des études de cohortes.

2.3.2 *LES STRUCTURES MALADIES RARES*

L'instruction de la Direction Générale de l'Offre des Soins du 11 janvier 2016 relative aux missions et périmètres des centres de référence, centres de compétences et des filières de santé dans le domaine des maladies rares reprécise la définition et les missions des différentes structures du réseau maladies rares (MR) en France (Direction Générale de l'Offre de Soins 2016).

Centre de Référence Maladies Rares (CRMR)

Un CRMR est reconnu pour son expertise autour d'une maladie rare ou un groupe de maladies rares. Il regroupe des compétences multidisciplinaires autour du soin, de la recherche et de la formation. C'est un centre de recours exerçant une attraction plus ou moins importante selon la rareté des maladies relatives à son expertise, avec un objectif d'équité en termes d'accès au diagnostic, au traitement et à la prise en charge globale des personnes malades. Les missions des CRMR sont principalement des missions de :

- Coordination, avec son réseau de filière et les associations de patients ;
- Expertise, par la production de recommandations, la participation aux réunions de concertations pluridisciplinaires et la prise en charge directe des patients ;
- Recours, puisque la prise en charge des patients ne se limite pas au bassin de son implantation, le CRMR doit exercer une attractivité et veille à l'équité de l'accès aux soins dans sa discipline ;
- Recherche, en participant à des travaux de recherche translationnelle, clinique ou organisationnelle ;
- Formation, en participant à des enseignements universitaires ou autres dans son domaine d'expertise.

En terme organisationnel, un CRMR peut être mono-site avec un seul site coordonnateur ou multi-sites avec un site coordonnateur et un ou plusieurs sites constitutifs. Un site est une unité de soins (unité fonctionnelle dans une structure hospitalière) exerçant une activité clinique. L'activité du site coordonnateur et des sites constitutifs d'un CRMR doit être en adéquation avec les missions du CRMR en apportant une complémentarité au niveau de l'expertise (spécialisation enfant ou adulte par exemple) ou de la couverture du territoire (meilleure gestion du recours aux soins).

Ces sites maladies rares peuvent faire partie de plusieurs CRMR. A titre d'exemple, l'unité fonctionnelle de génétique clinique de l'hôpital Robert Debré de l'AP-HP constitue un site coordonnateur pour le centre de référence des anomalies du développement et syndromes malformatifs - Ile de France et un site constitutif du centre de référence des déficiences intellectuelles de causes rares.

Actuellement, les centres de référence sont au nombre de 131 couvrant les diverses spécialités du champ des maladies rares et sont financés par le ministère de la santé. Le nombre de sites qui les composent, sites coordonnateurs ou sites constitutifs, est estimé à 350 unités de soins. Ces unités de soins constituent le maillage national du réseau des centres de référence.

Centre de Compétences Maladies Rares (CCMR)

Les CCMR ont pour mission la prise en charge et le suivi du patient atteint d'une maladie rare au plus près de son domicile. Ils participent au diagnostic et mettent en œuvre la thérapeutique en lien avec leurs centres de référence de rattachement. Les CCMR ont la même structuration que les

CRMR, avec l'implication d'une ou de plusieurs unités de soins avec une unité responsable. Un CCMR peut être rattaché à un ou plusieurs CRMR.

Contrairement aux CRMR, les CCMR n'ont pas de financement propre. Ils ont été désignés par les Agences Régionales de l'Hospitalisation, maintenant devenues Agences Régionales de Santé ; ils sont au nombre de 501 et contribuent au maillage régional du réseau MR.

Les Filières de Santé Maladies Rares (FSMR)

Les filières sont des réseaux composés de CRMR, de CCMR, de laboratoires de diagnostic, de laboratoires de recherche, d'associations de patients, de structures sociales et médico-sociales, d'universités et de tout autre partenaire apportant une expertise autour de la maladie rare ou du groupe de maladies rares faisant partie du champ de la filière.

Les missions de la filière sont essentiellement des missions de coordination, de mise en réseau et d'information dans un objectif d'amélioration de la prise en charge, de la recherche et de la formation autour de sa maladie d'expertise. Chaque filière doit inclure au moins 3 CRMR et est dotée d'une unité de gouvernance avec à sa tête un animateur de filière. Aujourd'hui, 23 filières ont été désignées pour une durée de 5 ans (« Les filières de santé maladies rares - Prises en charge spécialisées - Ministère des Affaires sociales et de la Santé » 2016a). A titre d'exemple nous citons la filière AnDDI-Rares (Filière de Santé Anomalies du Développement et Déficience Intellectuelle de causes Rares) qui regroupe, entre autres, 30 CRMR et 7 CCMR, en plus des divers laboratoires de génétique moléculaire et de cytogénétique, et la filière Cardiogen (filière nationale de santé maladies cardiaque héréditaires) qui regroupe notamment 3 CRMR et 22 CCMR (« AnDDI-Rares » 2016; « Cardiogen » 2016).

Les unités de soins maladies rares sont localisées dans les hôpitaux. Le réseau de soins maladies rares est donc un sous-réseau des structures hospitalières avec une nouvelle hiérarchisation. La Figure 1 illustre la superposition de ces deux organisations : organisation des établissements hospitaliers et organisation du réseau MR.

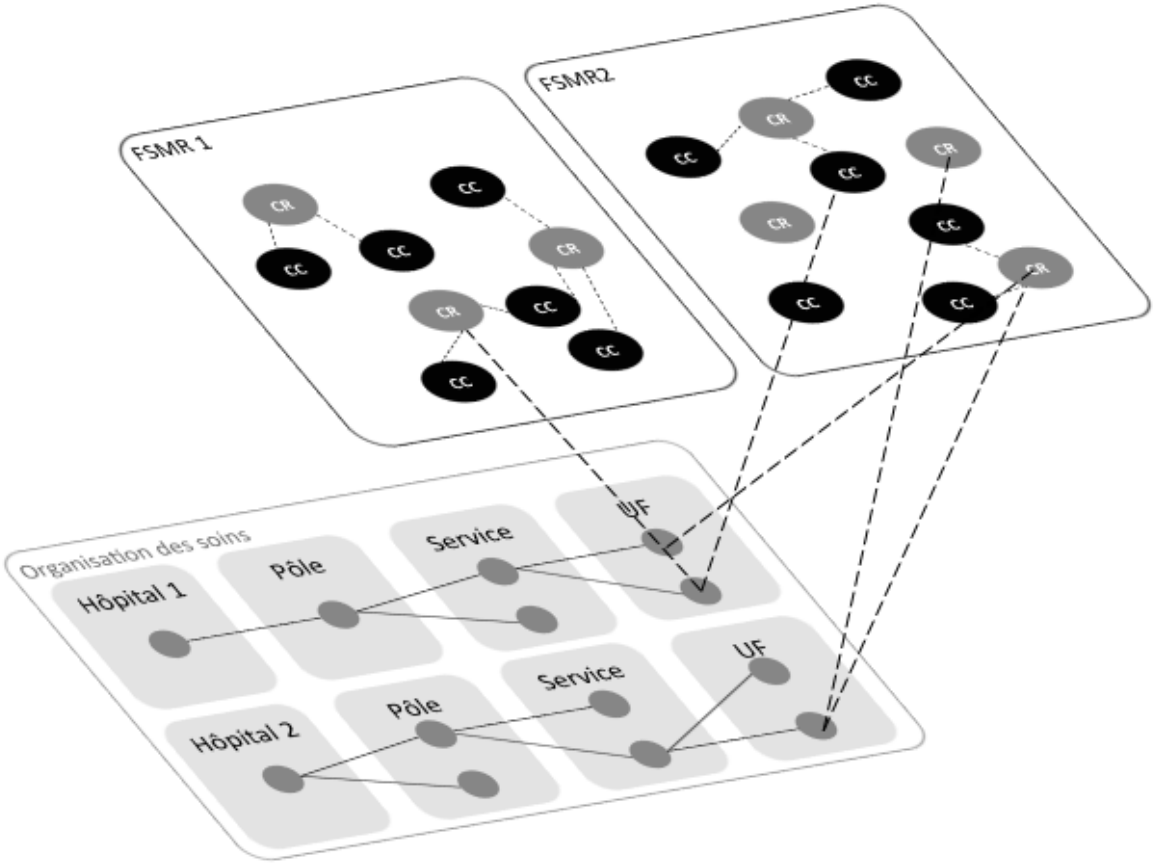

FIGURE 1 SUPERPOSITION DE L'ORGANISATION MR A L'ORGANISATION HOSPITALIERE ET COMPLEXITE DES LIENS ENTRE LES UF ET LES CRMR ET CCMR

(CR : CENTRE DE REFERENCE ; CC : CENTRE DE COMPETENCES ; FSMR, FILIERE DE SANTE MALADIES RARES ; UF : UNITE FONCTIONNELLE)

Par ailleurs, il n'existe pas actuellement un annuaire détaillé et contrôlé de toutes ces structures maladies rares. Certes, les listes des CRMR et des FSMR ont été officiellement publiées (« Les filières de santé maladies rares - Prises en charge spécialisées - Ministère des Affaires sociales et de la Santé » 2016b). Certaines organisations, telles qu'Orphanet ou des associations de patients, ont aussi créé des annuaires des CRMR et des CCMR pour les patients (« Centres de Référence - Fondation maladies rares » 2016; Orphanet 2009). Cependant, il n'existe pas d'annuaire détaillant tous les sites, coordonnateurs et constitutifs, et leurs liens multi-hiérarchiques avec les CRMR qui soit évolutif et maintenu.

2.3.3 *EPIDEMIOLOGIE ET MALADIES RARES*

2.3.3.1 **Absence de surveillance épidémiologique globale pour les maladies rares**

La mise en place d'études épidémiologiques pour les maladies rares est une tâche complexe étant donné les spécificités des maladies rares.

D'abord, leur diversité ne permet pas une surveillance globale pour toutes les maladies. Certaines études maladies spécifiques ont été mises en place, avec la création de registres qualifiés par exemple (Santé publique France 2016), mais elles restent insuffisantes pour avoir une évaluation complète de l'impact de la totalité des maladies rares.

Il est aussi difficile de systématiser la surveillance épidémiologique des maladies rares sans l'identification et l'implication directe de tous les acteurs intervenant dans la prise en charge et le suivi des patients. La labellisation des centres de référence a constitué une première structuration de ce réseau complexe. Il reste cependant à compléter car les acteurs de la médecine de ville, tels que les pédiatres, les généralistes et les spécialistes de ville, n'y sont pas intégrés.

Les outils nationaux tels que le Programme de Médicalisation des Systèmes d'Information (PMSI) ou le Système National d'Information Inter-Régimes de l'Assurance Maladie (SNIIRAM) sont quant à eux inadaptés pour cet usage.

Dans le cadre du PMSI, le codage diagnostic pour la mesure d'activité des établissements de soins est basé sur la Classification Internationale de Maladies (CIM) qui n'intègre qu'un nombre restreint de maladies rares (généralement les maladies les plus fréquentes). Ainsi, des codes non précis, faisant référence à des groupes de maladies, sont souvent utilisés pour y remédier. Dans d'autres cas, le code diagnostic maladie rare n'est pas utilisé parce que le contexte du recueil n'incite pas à son utilisation. Dans le cadre de la tarification à l'acte, pour un patient atteint de cystinose par exemple, l'acte de la dialyse est souvent indiqué pour une insuffisance rénale et non pas pour une cystinose malgré la présence du code de cette maladie dans la CIM. De plus, les consultations ne sont pas intégrées dans le cadre du PMSI. Ainsi, certains patients atteints de maladies rares, vus en consultations mais non hospitalisés, ne sont pas répertoriés dans le système. Toutes ces raisons rendent les patients atteints de maladies rares non détectables via le PMSI.

Dans le cadre du SNIIRAM, qui ne dispose pas des diagnostics des maladies, la détection des maladies rares est possible par l'intermédiaire d'un « médicament traceur ». Par exemple, la maladie de Fabry est une maladie lysosomale génétique, liée au chromosome X provoquée par un déficit enzymatique de l'alpha-galactosidase lysosomale avec accumulation de

globotriaosylcéramide et de digalactosylcéramide dans les cellules. Un traitement est constitué par l'Agalsidase-alpha ou bêta, un équivalent de synthèse fabriqué par génie génétique, qui pallie l'enzyme manquante. A partir de ce traitement le SNIIRAM peut identifier la maladie de Fabry car ce traitement est spécifique. Cette approche ne s'applique qu'au nombre très limité des maladies rares pour lesquelles un traitement spécifique est identifiable. L'agence européenne des médicaments répertorie seulement une centaine de médicaments orphelins dont l'indication n'est pas forcément exclusive à la maladie rare (« European Medicines Agency - Find medicine - European public assessment reports » 2016).

La réutilisation des données de soins dans le cadre de la recherche épidémiologique ou clinique est aussi une solution envisageable. Mais étant donné l'hétérogénéité et la complexité du paysage des systèmes d'information de santé en France cela nécessite d'importants travaux d'homogénéisation (voir section 4.1). Diverses initiatives régionales ont été lancées pour encourager le développement des SI hospitaliers. D'autres initiatives poussent au développement des nouvelles technologies pour la santé publique et la recherche. Malheureusement, ces initiatives ont été lancées indépendamment les unes des autres et ne s'intègrent pas dans une approche globalisée. Cette hétérogénéité des systèmes est donc entretenue par la diversité des acteurs, des financements et des objectifs de mise en œuvre et complexifie d'avantage le paysage. Ce manque d'interopérabilité ne permet donc pas une réutilisation des données de soins pour des études épidémiologiques nationales.

2.3.3.2 **L'épidémiologie des maladies rares parmi les priorités des institutions**

Cette absence de surveillance épidémiologique pour les maladies rares a conduit à la proposition de l'axe « Mieux connaître l'épidémiologie des maladies rares » dans le PNMR1. Cette surveillance épidémiologique est nécessaire pour obtenir de meilleures études d'incidence, de prévalence, de mortalité, de morbidité, de qualité de vie et de circuits des patients maladies rares. Cet objectif était sous la coordination de l'Institut de Veille Sanitaire (InVS) en collaboration avec les partenaires concernés. Parmi les travaux identifiés nous retrouvons la mise en place d'une nomenclature adaptée pour les maladies rares, la définition des outils nécessaires pour la création de bases de données, la collecte de données et l'application des méthodes statistiques adéquates, le croisement de ces données avec les données des décès répertoriés par le CépiDc (Centre

d'épidémiologie sur les causes médicales de décès), les données des Affections de Longue Durée déclarés ou les données du PMSI pour l'évaluation de l'activité.

Le Haut Conseil de Santé Publique (HCSP) a audité le PNMR1 et a émis les recommandations suivantes relatives aux études épidémiologiques (HCSP 2009):

- La mise en œuvre des outils logiciels servant au recueil des données ;
- Un consensus sur un ensemble commun de données ;
- L'aide aux cliniciens pour la définition et la mise en œuvre de ces outils logiciels.

Ces recommandations ont été intégrées dans le PNMR2 et plus spécifiquement dans le focus Banque Nationale de Données Maladies Rares (BNDMR).

3 LA BANQUE NATIONALE DE DONNEES MALADIES RARES

3.1 BNDMR - LE PROJET

3.1.1 *PRESENTATION ET OBJECTIFS*

Suite à ces recommandations, le projet Banque Nationale de données maladies rares a été initié en tant que projet prioritaire du PNMR2. Le projet est financé par le ministère de la santé et vise à mettre en place un outil national pour l'épidémiologie et la santé publique dans le domaine des maladies rares (« Banque Nationale de Données Maladies Rares » 2014). Cet outil doit permettre la collecte sécurisée et la centralisation anonymisée de données médicales de tous les patients atteints de maladies rares à l'échelle nationale. Ces données sont principalement produites au niveau des CRMR et CCMR et sont notamment présentes dans diverses plateformes informatisées (Dossier Patient Informatisé, applications de spécialité, registres, etc.). L'expérience de la plateforme CEMARA (voir section suivante) sera prise en compte dans la conduite du projet.

Les objectifs du projet sont les suivants :

- Mieux documenter le malade et sa maladie : offrir un support pour tracer l'histoire de la maladie chez les patients atteints de maladies rares et pris en charge dans le réseau de soin français.

- Mieux organiser les réseaux de soins : décrire la demande et son adéquation avec l'offre de soins dans le domaine en proposant, par exemple, des analyses territoriales de distance d'accès aux soins.
- Rendre visible l'activité « maladies rares » et aider au « *reporting* » national : Proposer un outil de suivi de l'activité des CRMR qui facilite la remontée de l'information vers les tutelles.
- Mieux exploiter le potentiel des grandes bases de données nationales : Permettre le chainage avec les bases de données nationales telles que le PMSI et le SNIIRAM pour la conduite d'études épidémiologiques et médico-économiques.
- Faciliter la recherche dans le domaine : en facilitant la recherche de patients éligibles pour des essais cliniques ou l'inclusion dans des cohortes.

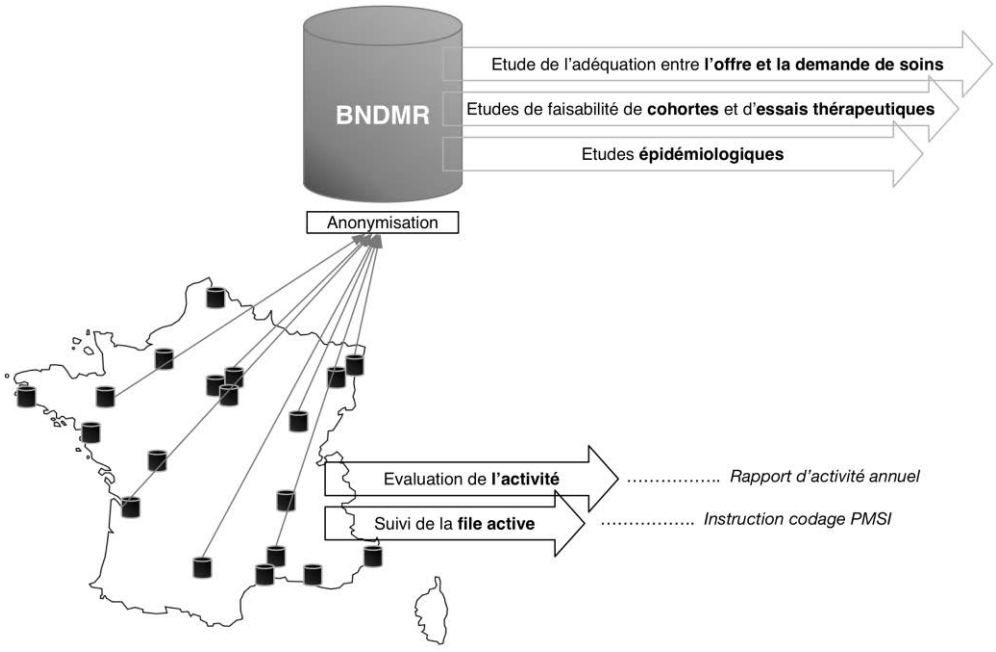

FIGURE 2 OBJECTIFS NATIONAUX DE LA BNDMR

3.1.2 *STRUCTURE ORGANISATIONNELLE*

L'équipe BNDMR est implantée au sein de l'Hôpital Necker - Enfant Malades de l'Assistance Publique des Hôpitaux de Paris (AP-HP). C'est une équipe multidisciplinaire qui a pour mission la mise en place de la BNDMR et des outils nécessaires pour le recueil d'un set minimal de données pour les maladies rares en France. Les travaux de l'équipe du projet ont été, pendant 4 années, directement supervisés par un coordonnateur national. L'équipe a essentiellement des compétences en informatique médicale et des compétences en exploitation de données.

Un **comité de pilotage** dirige et supervise la mise en place des missions de l'équipe, notamment l'organisation générale, la répartition, l'évolution et la validation des travaux et le respect du calendrier. Il valide un programme de travail annuel et le montant du budget qui en découle.

Le comité de pilotage est présidé par le Directeur Général de l'Offre de Soins ou son représentant. Il se réunit au moins deux fois par an, sur convocation de son président ou à la demande d'une majorité qualifiée des membres. Il est composé de représentants de l'Etat et d'autres représentants institutionnels et associatifs tel que le Laboratoire de recherche en Informatique Médicale et d'Ingénierie des Connaissances (LIMICS – UMR-S 1142).

Suite à la signature de la convention entre le ministère en charge de la santé et l'Assistance Publique des Hôpitaux de Paris (AP-HP) relative au pilotage de la BNDMR en septembre 2015, l'AP-HP est devenue maitre d'œuvre et d'ouvrage pour la mise en place de la BNDMR et des outils nécessaires pour le recueil du set minimal de données maladies rares. L'équipe projet est devenue la **cellule opérationnelle** de la BNDMR organisée et animée par un directeur opérationnel désigné par l'AP-HP.

Un **comité scientifique** a aussi été mis en place. Il assure le rôle de contrôle, de consultation et de proposition concernant l'exploitation de données au sein de la BNDMR. Son président est nommé par la DGOS. Il s'est réuni pour la première fois en mars 2016 et son président actuel est le Pr. Xavier Jeunemaître.

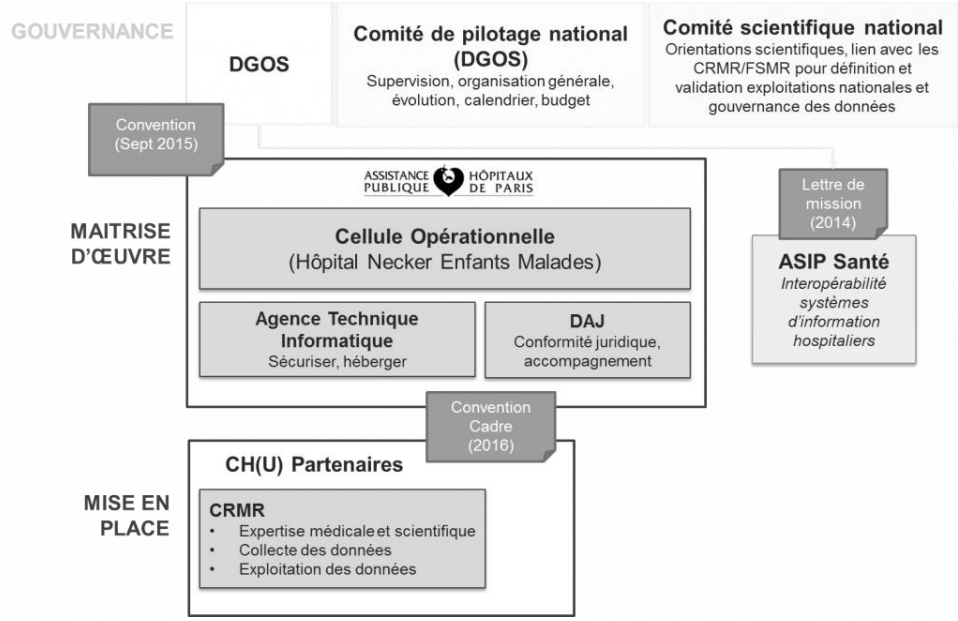

FIGURE 3 STRUCTURE ORGANISATIONNELLE DU PROJET BNDMR EN 2016

3.2 BNDMR – LE SYSTEME D'INFORMATION POUR LES MALADIES RARES

3.2.1 *CEMARA – LA BASE DE DONNEES HISTORIQUE*

Le projet CEMARA, pour les Centres Maladies Rares, a été défini en 2005 pour identifier les patients atteints de maladies rares sur le territoire français. Il s'agit d'un projet universitaire qui a La plateforme CEMARA a aussi été développée pour contribuer à la surveillance épidémiologique conduite par l'InVS durant le premier PNMR, identifier les patients éligibles pour des essais cliniques ou l'inclusion dans des cohortes et aider les CRMR à évaluer leur activité.

CEMARA est une plateforme Web qui a été mise à la disposition des CRMR. Cette plateforme permet la saisie et le stockage d'un ensemble de données commun à toutes les maladies rares, le tronc commun de CEMARA, permettant le suivi des patients qu'ils soient adultes, enfants ou même fœtus. L'accès aux données des patients est restreint au site ou unité de soin où le patient est suivi. Ce recueil permet aussi de générer les rapports annuels d'activité des centres de référence comme recommandé par le ministère de la santé et la Haute Autorité de Santé (HAS). En partenariat avec Orphanet, unité Inserm qui répertorie et édite l'encyclopédie des maladies rares, un vocabulaire commun pour enregistrer les diagnostics a été intégré à la plateforme. D'autres recueils de données plus spécifiques ont été proposés aux CRMR : les pétales de CEMARA. Adossé au tronc commun, un pétale constitue un recueil complémentaire spécifique à un groupe homogène de

maladies rares permettant un suivi longitudinal plus détaillé des patients, semblable à ce qui est mis en place dans le cadre de suivi de cohortes.

Le recueil de CEMARA, qui a débuté en 2007, concerne actuellement les données de 360.000 patients répartis sur une soixantaine de centres de référence. Ce réseau de CRMR représente presque 50% de l'ensemble des CRMR en France et s'étend aux départements et territoires d'Outremer.

La non pérennité de la plateforme et son manque d'interopérabilité poussent à son remplacement par une nouvelle solution.

CEMARA constituera le principal matériel de nos travaux de thèse.

3.2.2 *BAMARA-BNDMR LE NOUVEAU SYSTEME D'INFORMATION POUR LES MALADIES RARES*

3.2.2.1 **Enjeux**

Le principal objectif du projet BNDMR est de mettre en place un entrepôt de données alimenté par les données de tous les patients atteints de maladies rares en France. Ces données sont définies dans un set minimal de données qui a été élaboré selon une méthodologie que nous décrivons un peu plus loin dans ce manuscrit. Cet entrepôt de données doit tendre vers l'exhaustivité pour améliorer la fiabilité des études qui en émanent. Ce projet fait face à deux enjeux majeurs et qui peuvent être contradictoires à un certain niveau : la protection et l'interopérabilité des données de santé.

La protection des données de santé

En France, la législation est protectrice des données de santé. Les données de santé ne sont pas seulement des données personnelles, un statut qui leur procure déjà un certain niveau de protection, elles sont en plus considérées comme étant « sensibles » (CNIL 2011). Leur collecte et leur traitement sont par principe interdits, mais il existe certaines dérogations. Parmi ces dérogations figurent la collecte et le traitement des données de santé à des fins de suivi et de gestion de patients restreints à l'équipe de soins et les traitements nécessaires à la recherche dans le domaine de la santé (*Loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux*

libertés 2016). Ces traitements à des fins de recherche, notamment épidémiologiques, sont encadrés par les articles du chapitre IX de loi Informatique et Libertés.

Dans le respect de ce cadre juridique français et du cadre juridique européen (Commission Européenne 2016), les données de l'entrepôt BNDMR destiné aux études épidémiologiques doivent être non directement nominatives et traitées de telle sorte qu'on ne puisse pas remonter aux identités des patients. S'adosse à cela, d'autres mesures de sécurité de la plateforme informatique.

L'interopérabilité

Pour atteindre cet objectif d'exhaustivité, nous ne pouvons passer outre la récupération des données des systèmes qui sont d'ores et déjà utilisés par les professionnels de santé des CRMR et CCMR. Cela permet aussi de ne pas leur rendre la tâche encore plus lourde en leur imposant un autre système de saisie. Nous nous confrontons alors à un grand enjeu d'interopérabilité, dans un paysage numérique très hétérogène et un cadre juridique assez contraignant.

3.2.2.2 **Organisation du SI**

Nous avons défini une architecture à deux niveaux : niveau soin et niveau recherche. Cette architecture correspond au modèle métier avec une séparation entre la partie « suivi des patients » et gestion d'activité des sites maladies rare d'une part, et d'autre part les « traitements statistiques » à des fins épidémiologiques et médico-économiques au niveau national. Pour répondre à ces besoins, les technologies qui doivent être implémentées sont évidemment elles aussi différentes selon les fonctionnalités qui sont attendues. Cette architecture nous permet par ailleurs de remédier à des problématiques juridiques et de qualité de données.

BaMaRa – Base Maladies Rares pour le soin

BaMaRa est une application web offrant un outil sécurisé de saisie de données maladies rares et une interface permettant de les suivre et de les exploiter. C'est l'évolution de CEMARA vers une application plus conviviale, proposant de nouvelles fonctionnalités et des interfaces d'interconnexion. Elle s'adresse en priorité aux CRMR et CCMR en France. Dans le cas où ces centres ne disposent pas d'une application spécifique, l'application BaMaRa sera proposée aux professionnels de santé pour qu'ils puissent y saisir directement leurs données maladies rares. En

revanche, lorsque des systèmes d'information sont déjà en place dans ces centres et qu'ils présentent un niveau de compatibilité adéquat, un interfaçage automatisé d'envoi de données doit être possible afin d'éviter une double saisie.

BaMaRa est donc une application pour le suivi des données de soins. Les droits d'accès aux données respectent la notion d'équipe de soins. Un professionnel de santé donné n'aura en effet accès qu'aux données des patients qui sont suivis dans son unité de soins.

BaMaRa facilitera l'analyse des données des CRMR des CCMR par site et le suivi de leur activité, notamment en vue de la rédaction du rapport d'activité pour les tutelles. Elle constitue enfin un sas où des contrôles de qualité sont effectués avant la consolidation et la remontée des données vers l'entrepôt BNDMR destiné aux études nationales.

BNDMR – Entrepôt pour les études nationales

L'entrepôt de données dans le cadre duquel sont effectuées les études nationales constitue la « Banque Nationale de Données Maladies Rares ». Il contient les données du set minimal de données maladies rares issues de BaMaRa après anonymisation et consolidation. La BNDMR est la plateforme technique pour la conduite d'études statistiques qu'elles soient récurrentes et génériques à destination des tutelles, des filières ou du grand public ou plus ponctuelles ou spécifiques pour répondre à des questions de recherche épidémiologique à la demande de certains CRMR par exemple. C'est aussi avec la BNDMR que des études plus élargies seront effectuées par des méthodes de chainages de données de cohortes ou de bases nationales telles que celle du SNIIRAM. Le comité scientifique gère et valide tous les projets d'études émanant de la BNDMR.

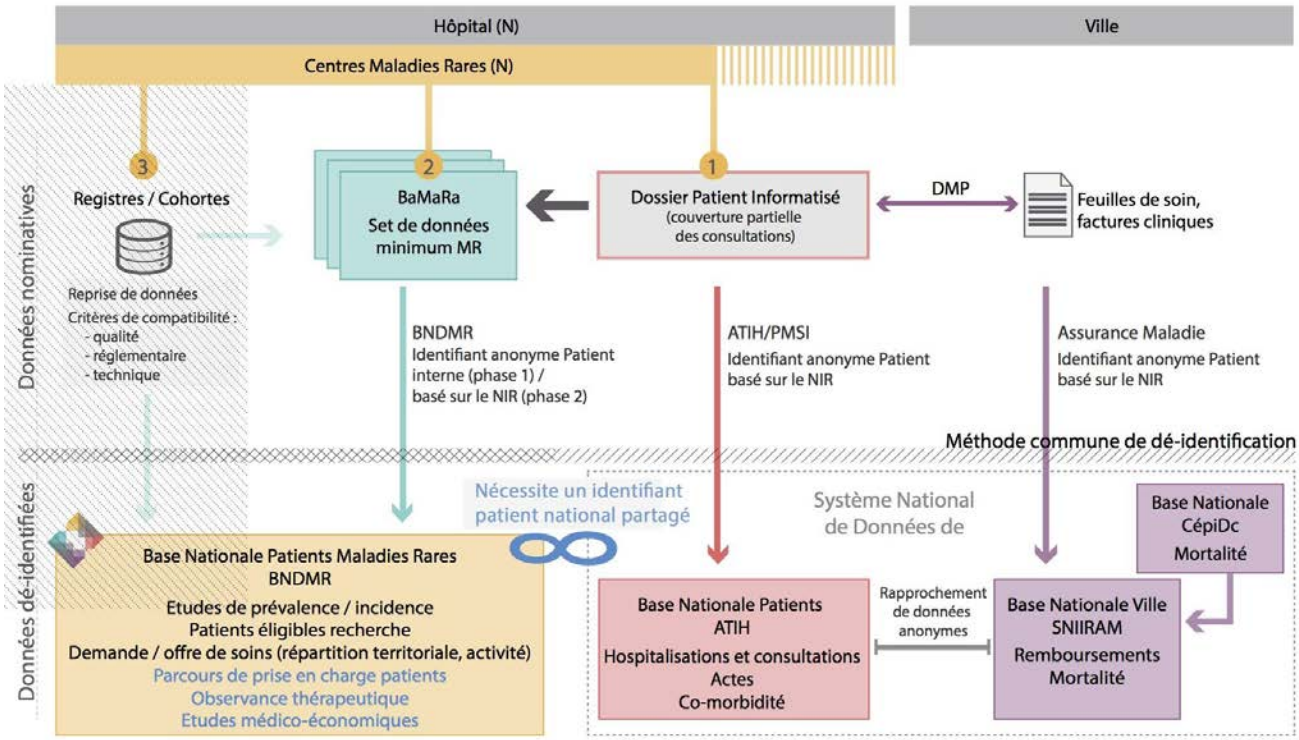

FIGURE 4 ORGANISATION DU SYSTEME D'INFORMATION BAMARA-BNDMR ET SES INTERACTIONS AVEC LES SYSTEMES NATIONAUX ET LOCAUX EXISTANTS

4 CADRE D'INTEROPERABILITE POUR LES MALADIES RARES

4.1 PROBLEMATIQUE : UNE HETEROGENEITE CONSTATEE

La cellule opérationnelle de la BNDMR a conduit une enquête nationale sur les bases de données maladies rares en France qui vise à dresser un état des lieux du nombre et du contexte d'utilisation de ces bases par les centres de référence maladies rare. Le rapport de l'enquête recense 165 bases de données outre CEMARA et ses pétales (Angin et al. 2015). Ces résultats ont été rapprochés des chiffres d'autres sources : l'annuaire Orphanet des registres de patients maladies rares (Orphanet 2016a) et la liste de registres identifiés « maladies rares » du portail Epidémiologie - France (« Portail Epidemiologie - France | Health Databases » 2016)) ce qui a permis l'identification de 91 bases de données supplémentaires. A ce grand chiffre s'adosse une hétérogénéité des objectifs et des types des données recueillies (voir Figure 5 ci-dessous).

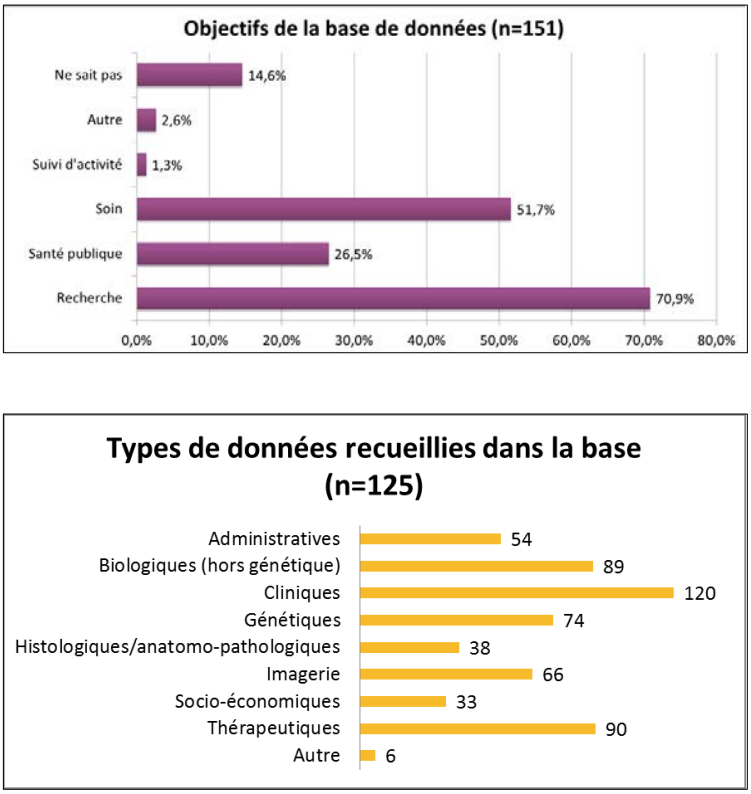

FIGURE 5 HETEROGENEITE DES OBJECTIFS ET DES TYPES DES DONNEES RECUEILLIES DES BASES DE DONNEES DECLAREES POUR L'ENQUETE AUPRES DES CRMR

Ces différentes bases de données ont été créées et maintenues soit directement par les fonds des CRMR soit par d'autres formes de financement de projets de recherche spécifiques tels que le programme hospitalier de recherche clinique ou le programme investissements d'avenir (« Le programme hospitalier de recherche clinique - PHRC - Appels à projets - Ministère des Affaires sociales et de la Santé » 2016; « Investissements d'avenir (CGI) » 2016). Des institutions publiques telle que l'Institut National du Cancer ou des associations de patients telle que l'Association Française contre les Myopathies proposent aussi le financement de projet de plateformes informatisées pour répondre à des usages spécifiques (« Institut National Du Cancer - Accueil » 2016; « L'AFM-Téléthon en bref » 2013).

Outre ces diverses bases de données et registres, les sites maladies rares, étant des unités fonctionnelles hospitalières, utilisent par ailleurs les outils qui leur sont proposés au sein du système d'information hospitalier comme le Dossier Patient Informatisé (DPI) ou l'application de spécialité déployée au sein du service auquel l'unité est rattachée. Dans le contexte du soin, les pouvoirs publics incitent à la numérisation des hôpitaux et au développement des dossiers patients

informatisés. Le programme « hôpital numérique » vise à instaurer une intégration minimale des nouvelles technologies dans les hôpitaux (« Le programme hôpital numérique - Hôpital numérique - Ministère des Affaires sociales et de la Santé » 2016). Ce plan partait du constat que le niveau d'implémentation du numérique dans les différents hôpitaux français était très variable. Certains hôpitaux fonctionnent encore avec des dossiers papier. D'autres ont commencé le développement de leur système d'information hospitalier mais, dans certains cas, le déploiement reste non généralisé à l'ensemble de l'hôpital ou les différentes composantes restent non intégrées et ne communiquent pas entre elles.

Le transfert des données de ces diverses sources vers BaMaRa ou la BNDMR suppose une interopérabilité possible entre ces différents systèmes. Une interopérabilité qu'il convient de mettre en œuvre.

4.2 LES TROIS PILIERS DU CADRE D'INTEROPERABILITE

Nous avons choisi d'aborder cette hétérogénéité selon 3 piliers d'interopérabilité (voir Figure 6) : L'identification des patients (01), l'interopérabilité des données médicales (02) et la gestion des flux de données (03).

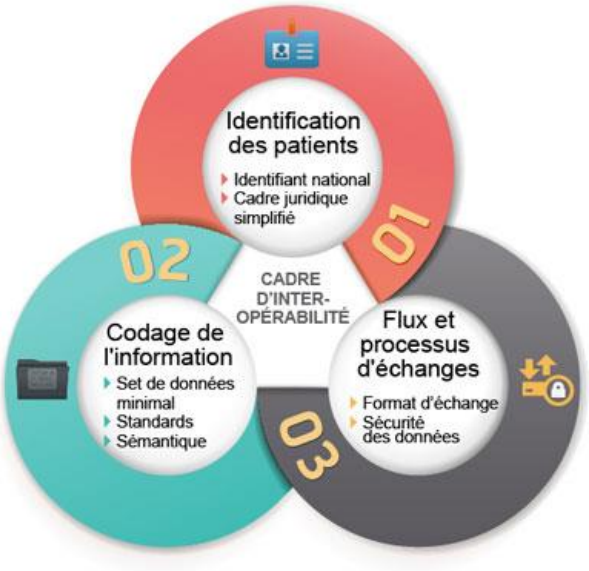

FIGURE 6 LES TROIS NIVEAUX DU CADRE D'INTEROPERABILITE POUR LES MALADIES RARES

4.2.1 *IDENTIFICATION DES PATIENTS*

Quel que soit l'objectif ou le contexte dans lequel deux systèmes souhaitent s'échanger les données, quand les objets échangés, des documents médicaux notamment, concernent des patients, il est primordial que ces deux systèmes réussissent à identifier ces patients d'une manière commune. Que cela soit à l'échelle d'une équipe, d'un établissement ou d'un réseau de soin, la mise en place d'un système unique d'identification des patients constitue le premier pas vers l'interopérabilité au sein de ce groupe et ce quel que soit le type de contenu échangé (documents en textes libres ou documents structurés et interprétables).

Dans un contexte de soin, l'identification du patient permet d'assurer la continuité des soins. Ainsi, au niveau de chaque hôpital, un identifiant patient permanent est assigné à chaque patient pour que son dossier soit unique et soit consultable et éditable lors de son parcours au sein de l'établissement et même lors d'une visite ultérieure à l'hôpital. Au niveau national, et avec la mise en place du Dossier Médical Personnel (DMP), dossier partagé entre les différents professionnels de santé qui prennent en charge le patient, un identifiant national de santé a été proposé pour permettre une identification fiable du patient.

Dans le contexte de recherche et d'épidémiologie, l'identification des patients permet de garantir la qualité des études qui sont conduites. Un identifiant patient unique et robuste permet d'éviter la génération de doublons dans la base de données de recherche. Dans le cadre de chainage de données de sources multiples, il permet de correctement juxtaposer les différentes données pour constituer un jeu de données plus riche et plus porteur d'information.

Dans le cadre du projet BNDMR, l'identification des patients s'inscrit dans ces deux contextes: niveau soin pour BaMaRa et niveau épidémiologie pour la BNDMR. Le contexte épidémiologique est d'autant plus contraint par la nature des maladies rares, qui, par définition, touchent très peu de personnes et dont les études épidémiologiques sont par conséquent très sensibles.

4.2.2 *INTEROPERABILITE DES DONNEES MEDICALES*

Une fois les patients bien identifiés, comment est interprété le contenu échangé ?

Dans certains cas, ce contenu est destiné à être uniquement consulté par des professionnels de santé. Dans le cadre de prescriptions d'examens biologiques par exemple, la prescription peut être directement interprétée par les professionnels de santé du laboratoire biologique, dans ce cas un

texte libre non structuré au format PDF peut constituer le contenu échangé. Dans d'autres cas, et pour de multiples raisons (traçabilité, gestion, analyse, etc.) le contenu échangé est destiné à être interprété par le système qui le reçoit. Cela devient possible lorsque :

- le périmètre de l'échange de données est bien défini,
- les deux systèmes partagent le même format, la même notation et la même compréhension des éléments échangés,
- les deux systèmes se sont mis d'accord sur des règles de traduction préalables si chacun possède ses propres formats, notation et compréhension des éléments.

Dans le contexte de la BNDMR, c'est un entrepôt de données à des fins épidémiologiques qui est mis en place. Le contenu recueilli et envoyé par les sites maladies rares doit être structuré puisqu'il est destiné à être étudié et analysé au niveau de la BNDMR. D'abord, pour que ce recueil soit homogène, il est nécessaire de définir un set minimal de données commun à toutes les maladies rares (Choquet et al. 2014). Les échanges de données qui s'effectuent s'inscrivent donc dans le cadre de l'alimentation de la BNDMR avec des données de patients maladies rares définies au sein du set minimal de données. Etant donné l'hétérogénéité des bases de données utilisées au niveau des sites maladies rares, ce set minimal de données doit être standardisé en utilisant des standards et des terminologies internationales afin de pointer vers une compréhension unique et partagée et vers l'implémentation des mêmes formats. Par ailleurs, pour certaines sources qui ne peuvent implémenter ces nouveaux formats, des alignements doivent être effectués pour définir les traductions et les transformations de données à effectuer avant leur envoi à la BNDMR (Maaroufi et al. 2013; Maaroufi et al. 2015).

4.2.3 *GESTION DES FLUX DE DONNEES*

Ce troisième niveau aborde la mise en place technique et opérationnelle de ces échanges de données. Etant donné la multiplicité des systèmes, il est primordial de mettre en place un système de gestion et de supervision de ces multiples flux de données. La sécurité des données et l'authentification fiable de chaque source constituent des points cruciaux.

4.3 PROPOSITIONS

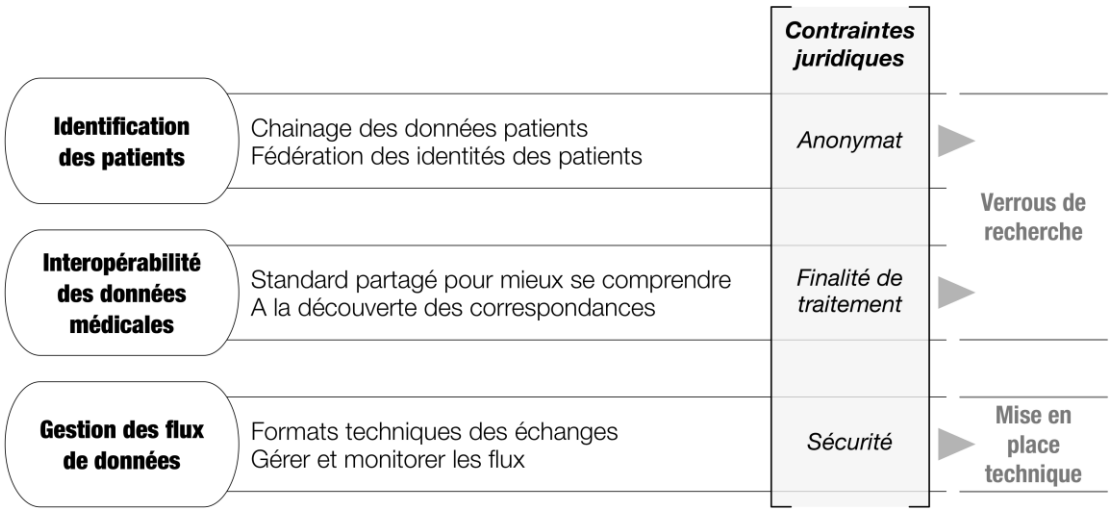

FIGURE 7 IDENTIFICATION DES VERROUS DE RECHERCHE DU CADRE D'INTEROPERABILITE POUR LES MALADIES RARES

Sur le premier pilier d'interopérabilité que nous avons défini, nous étudierons les diverses solutions d'identification des patients dans le cadre français et les différentes propositions disponibles dans la littérature internationale. Ces solutions étaient pour la plupart non applicables dans notre contexte et n'étaient pas en adéquation avec nos contraintes et nos objectifs : anonymat des données au niveau épidémiologique, identification des fœtus qui constituent une proportion non négligeable de la population des patients atteints de maladies rares. Nous proposons dans le deuxième chapitre de ce manuscrit une méthode simple et fiable d'identification de patients, notamment les fœtus, qui permette d'une part la fédération des identités des patients au sein de la BNDMR et le chainage de ces données avec d'autres données de source externes pour la conduite d'études spécifiques d'autre part.

Sur le deuxième niveau d'interopérabilité qui s'intéresse aux données échangées, nous proposons d'abord une méthodologie de définition et de standardisation du set minimal de données pour les maladies rares. Cette méthodologie se base sur un travail collaboratif pour la mise en place d'un ensemble commun de données et sur l'identification des standards et des terminologies les plus adaptés à cet ensemble de données. Nous proposons par la suite une nouvelle approche pour l'alignement semi automatisé de schémas de données hétérogènes. Cette approche vise à découvrir des correspondances entre les données en optimisant l'emploi des outils d'alignement existants et en adaptant leur utilisation à un contexte d'intégration de données. Ce deuxième niveau d'interopérabilité sera étudié dans le troisième chapitre de ce manuscrit.

CHAPITRE II : IDENTIFICATION DES PATIENTS

1 INTRODUCTION

1.1 CONTEXTE ET OBJECTIFS

Certaines maladies rares affectent de multiples systèmes et appareils, le phénotype observé devient alors complexe et requiert l'intervention de plusieurs spécialistes. Un patient atteint par la maladie de Gaucher (« Maladie de Gaucher | LORD » 2016) par exemple peut présenter simultanément une anémie, une augmentation de taille du foie et de la rate, des douleurs et une fragilité au niveau des os, des troubles psychiatriques et du comportement, une anomalie du système génital, etc. Tous ces signes ne sont que les signes les plus fréquents, la liste des symptômes que pourrait présenter un patient atteint de la maladie de Gaucher est encore plus longue. Cette présentation clinique complexe conduit le patient à être suivi par plusieurs spécialistes dans différentes unités de soins. Cette prise en charge multiple constitue un défi à relever par la Banque Nationale de Données Maladies Rares. Ce défi est d'autant plus important à relever lorsque les maladies sont très rares et affectent un nombre très réduit de patients, une incidence de 1-9 / 100 000 pour la maladie de Gaucher par exemple. Tout au long de son parcours de soins, un patient peut être vu dans des unités MR différentes. Chacune de ces unités met en place son propre dossier de soins pour ce même patient. Si le même système d'identification de patient n'est pas adopté par tous ces sites, la banque nationale risque de récolter des doublons et de ne pas pouvoir les fédérer. La qualité des analyses et des études en serait altérée et les estimations d'incidence et de prévalence seraient biaisées.

L'échange de contenu, qu'il soit sous la forme d'un document non structuré ou d'un ensemble de données structurées, entre les systèmes sources utilisés au niveau des sites maladies rares ne peut donc se faire sans une identification robuste du patient auquel ce contenu devrait être attribué. Dans le cas du patient souffrant de la maladie de Gaucher décrit précédemment, la méthode d'identification du patient devrait être commune à tous les systèmes utilisés au niveau des centres de référence ou de compétence où il est suivi et qui transmettent des données vers la BNDMR. Ainsi les données seront attribuées au même patient en évitant la création de doublons

dans la base nationale. Cette approche qui vise à fédérer les identités des patients est d'autant plus importante dans le contexte des analyses épidémiologiques qui seront opérées au niveau de la banque.

Les objectifs de ce premier travail de recherche qui vise à définir un identifiant patient pour la banque nationale de données maladies rares sont essentiellement :

- Une identification fiable et robuste afin d'attribuer correctement les données transmises par les sources aux patients appropriés dans la base nationale. A la réception de nouvelles données d'un patient, ce dernier est recherché dans BaMaRa en se basant entre autre sur cet identifiant. Si le patient est retrouvé, sa fiche est mise à jour avec les nouvelles données reçues, sinon une nouvelle fiche patient est créée.
- La fédération des identités des patients visant à limiter le nombre de doublons dans la banque nationale et garantissant ainsi la fiabilité des études épidémiologiques. Cet objectif s'inscrit dans une stratégie globale d'identitovigilance, associant d'autres approches de gestion de qualité des données, pour le traitement *a posteriori* des doublons créés dans la banque nationale.
- Le chaînage des données de la BNDMR avec d'autres collections de données issues de registres nationaux, par exemple dans le contexte d'études épidémiologiques. L'identifiant constituera donc une clé de chainage permettant de rapprocher ces données.

Etant donné le contexte épidémiologique de cette collecte de données une marge d'erreur est tolérée contrairement au contexte des soins ou de la prise en charge d'un patient où une erreur d'attribution des données peut avoir des conséquences graves sur sa santé.

Dans ce chapitre, nous décrivons nos besoins et contraintes en termes d'identification des patients et de fédération d'identité pour la banque nationale de données maladies rares. Nous explorons par la suite quelques pistes de l'existant : les identifiants nationaux et autres identifiants spécifiques. Nous expliquons les raisons qui nous prémunissent de les utiliser dans le cadre du projet. Nous présentons notre nouvelle méthode pour créer l'identifiant patient maladies rares que nous appelons IdMR, un algorithme en trois phases, se basant sur l'anonymisation des informations nominatives des patients. Nous terminons par une discussion sur notre proposition sur divers aspects : identification des risques et justification des choix.

1.2 CONTRAINTES

L'identifiant patient de la Banque Nationale de Données Maladies Rares doit répondre à un ensemble de critères que nous avons sélectionné afin de garantir la fiabilité des études émanant de la banque et ce dans le respect des contraintes éthiques et juridiques.

1.2.1 *UNIQUE*

Chaque patient doit se voir attribuer un seul et unique identifiant. Il en est ainsi, par exemple, lors de la prise en charge d'un patient dans un hôpital, un patient a le même identifiant qui permet de l'identifier dans toutes les unités et pour tous les actes de soins dont il va bénéficier. Un identifiant unique maladie rare par patient diminuera le risque de création de plusieurs fiches « maladie rare » pour le même patient et évitera ainsi le risque de création de doublons dans la BNDMR.

1.2.2 *PERENNE*

Chaque patient doit garder le même identifiant maladie rare pour tout son suivi dans le système de soin. D'un rendez à l'autre, dans la même unité de soin, le même identifiant maladie rare doit être réutilisé pour l'identifier. Cet identifiant ne doit pas être attribué à un autre patient pour éviter ainsi les collisions.

1.2.3 *ANONYME*

Les informations personnelles du patient ne doivent pas transparaitre à travers son identifiant. Cet identifiant doit être non-signifiant et anonyme. Il doit être généré :

- soit aléatoirement de telle sorte qu'il soit impossible d'établir un lien entre l'identifiant et la personne identifiée.
- Soit avec des méthodes d'anonymisation permettant de transformer d'une manière irréversible les données personnelles du patient en un identifiant anonyme.

1.2.4 *GLOBAL*

Chaque patient et tout patient doit pouvoir se faire attribuer un identifiant. Quel que soit son origine ou son âge, tant que le patient est suivi dans le système de soin français pour sa maladie rare il doit, après son accord, être répertorié dans la banque nationale et ainsi se voir doté d'un

identifiant. Ceci est d'autant plus important pour les enfants qui représentent une grande proportion de la population des patients atteints de maladies rares en France ou ailleurs. En effet 80% des maladies rares sont des maladies génétiques (« RARE Diseases: Facts and Statistics » 2012) et apparaissent le plus souvent durant l'enfance. L'identification des fœtus est un défi supplémentaire à relever puisque les tests génétiques se font de plus en plus tôt permettant ainsi de diagnostiquer les maladies rares durant la grossesse.

2 IDENTIFIANT PATIENT ET CHAINAGE DE DONNEES : ETAT DES LIEUX

2.1 UN IDENTIFIANT NATIONAL DE SANTE EN FRANCE

Dans des pays tels que la Suède où un système global d'identification des personnes existe depuis des décennies (Ludvigsson et al. 2009) la problématique d'identification des patients au niveau d'un registre maladies rares est simplifiée. En effet, les systèmes d'identification des personnes gérés par l'état, même restreints au domaine de la santé, sont généralement fiables et constituent un outil à la disposition des projets connexes tels que les projets de recherche où différentes sources de données sont croisées. En France, le projet de création d'un identifiant national de santé est en cours de concrétisation.

Suite au lancement du projet du Dossier Médical Personnel, le besoin d'un identifiant patient national pour le partage des documents médicaux entre professionnels et établissements de santé s'est fait ressentir. En 2007, l'article L1111-8-1 du code de la santé publique (*Code de la santé publique - Article L1111-8-1* 2007), stipule que dans le cadre de la prise en charge des bénéficiaires de l'assurance maladie dans les réseaux de santé, l'Identifiant National de Santé sera utilisé «[…] pour la conservation, l'hébergement et la transmission des informations de santé. Il est également utilisé pour l'ouverture et la tenue du dossier médical personnel et du dossier pharmaceutique […]»

L'utilisation du Numéro d'Inscription au Répertoire (NIR), plus communément connu sous le nom de numéro de sécurité sociale, a été étudiée et réfutée en 2007 sur avis de la Commission Nationale de l'Informatique et des Libertés (CNIL). Un INS spécifique a été défini et serait généré par des procédures d'anonymisation à partir du NIR pour garantir la non-signifiance de cet identifiant patient. Avec la nouvelle Loi de Santé 2015 (*Code de la santé publique - Article L1111-8-*

1 2016; *Projet de loi de modernisation de notre système de santé* 2015), l'utilisation du NIR en tant qu'Identifiant National de Santé est à nouveau discutée et autorisée.

2.1.1 *LE NUMERO D'INSCRIPTION AU REPERTOIRE (NIR)*

2.1.1.1 **Description et propriétés**

Toute personne née en France est enregistrée dans le Répertoire National d'Identification des Personnes Physiques (RNIPP). Un Numéro d'Inscription au Répertoire est assigné à chaque personne qui y est répertoriée. Ce numéro est essentiellement utilisé par les autorités d'assurance maladie pour l'émission de la carte vitale et est populairement connu sous le nom de numéro de sécurité sociale. Ainsi, même les personnes étrangères vivant en France se voient attribuer ce numéro suite à leur adhésion au système de sécurité sociale. Cet identifiant est unique par personne et est constitué de 13 chiffres :

- Le premier chiffre correspond au sexe de la personne : « 1 » pour un homme et « 2 » pour une femme,
- Les deux chiffres suivants correspondent aux deux derniers chiffres de l'année de naissance,
- Les deux chiffres suivants correspondent au mois de naissance,
- Les cinq chiffres suivants correspondent au lieu de naissance : code INSEE du département suivi du code INSEE de la commune pour les personnes nées en France ou « 99 » suivi du code INSEE du pays pour les personnes nées à l'étranger,
- Les trois chiffres suivants correspondent à un numéro d'ordre chronologique pour distinguer les personnes nées au même endroit dans la même période.

Ce numéro d'inscription de 13 chiffres est complété par une clé de contrôle de deux chiffres permettant de vérifier la validité et l'authenticité du NIR.

Le NIR est un numéro unique et pérenne pour chaque individu. Il est aussi considéré fiable puisqu'il est géré par l'INSEE et qu'il est généré à partir des données de l'état civil envoyées directement par les mairies. Cependant il ne s'agit pas d'un numéro anonyme ou généré au hasard, au contraire il est signifiant et des informations relatives aux personnes y sont directement lisibles.

2.1.1.2 **Restrictions d'utilisation**

Depuis sa création dans les années 40, l'utilisation du NIR s'est peu à peu généralisée (organismes de protection sociale, administration fiscale, l'éducation nationale…) et les systèmes de fichiers l'utilisant se sont centralisés. Des inquiétudes d'ordre éthique ont été éveillées et vers la fin des années 1970, l'affaire SAFARI a mis en garde contre le caractère discriminatoire du NIR. Ces inquiétudes se sont apaisées avec la loi informatique et libertés du 6 juin 1978 et la création de la Commission Nationale de l'Informatique et des Libertés (CNIL).

La CNIL a depuis restreint l'utilisation du NIR en imposant une autorisation par décret en Conseil d'État pris après avis de la Commission. Cette autorisation a été accordée aux organismes de sécurité sociale (assurance maladie, assurance vieillesse, allocations familiales…) et par la suite aux professionnels et aux établissements de santé dans un objectif d'avance ou de remboursement des frais médicaux.

Avec le lancement du projet DMP, l'avis de la CNIL a été demandé par rapport à l'utilisation du NIR en tant qu'identifiant patient, pour étendre son domaine d'utilisation non plus seulement à la sécurité sociale mais aussi à la santé. En 2007, la CNIL a donc publié ses conclusions à ce sujet (CNIL 2007).

« Toutefois, partant de la constatation que les données de santé ne sont pas des données personnelles comme les autres et qu'elles appellent une protection renforcée, notre Commission estime que le NIR, compte tenu de son usage répandu, du fait qu'il est signifiant et facile à reconstituer et des risques précédemment évoqués, ne constitue pas, aujourd'hui, un numéro adapté pour identifier le dossier médical de chacun. »

La situation s'est encore renversée avec l'adoption de la loi da santé 2015 (*Projet de loi de modernisation de notre système de santé* 2015) qui autorise désormais l'utilisation du NIR en tant qu'identifiant national de santé par l'article L. 1111-8-1. (*Code de la santé publique - Article L1111-8-1* 2016):

« I. – Le numéro d'inscription au répertoire national d'identification des personnes physiques est utilisé comme identifiant de santé des personnes pour leur prise en charge à des fins sanitaires et médico-sociales, dans les conditions prévues à l'article L. 1110-4. »

Les modalités d'utilisation restent encore à définir par un décret en Conseil d'État pris après avis de la CNIL afin de restreindre son utilisation au domaine de la santé et du médico-social.

2.1.1.3 **Le NIR candidat pour la BNDMR ?**

Le NIR a constitué un candidat en tant qu'identifiant patient dans la Banque Nationale de Données Maladies Rares. En effet, il est unique et pérenne par individu et il est fiable. Cependant, son utilisation actuelle dans les systèmes de soins répond à un objectif de prestations de l'assurance maladie, le numéro qui est donc récupéré est souvent le NIR de l'assuré. En particulier, les NIR des ayants droit, tels que les conjoints et les enfants, ne sont pas récupérés et ne sont même pas inscrits dans la carte vitale des assurés. Les professionnels de santé ne disposent pas des moyens nécessaires pour récupérer les NIR des enfants et encore moins ceux des fœtus qui n'ont pas d'existence juridique autonome en dehors de la personne de la mère. Ce mode de fonctionnement n'est pas envisageable pour la banque nationale puisque les patients ne seront pas identifiés par leur propre numéro mais par celui de l'assuré. Avec l'adoption de la nouvelle loi de santé, les mécanismes d'utilisation du NIR en tant qu'identifiant national de santé seront définis ultérieurement et plus spécifiquement les modalités de récupération du NIR des enfants. Cependant, cette solution ne sera pas disponible à court terme.

Il est d'autant plus difficile de récupérer les NIR, même ceux des assurés, depuis les registres et autres bases de données sources hors dossier patient informatisé des hôpitaux puisque ces systèmes ne sont pas interfacés avec les systèmes de lecture de cartes vitales. Enfin, et pour des raisons juridiques, le NIR est un numéro signifiant incompatible avec le caractère anonyme de la Banque Nationale de Données Maladies Rares.

2.1.2 *L'IDENTIFIANT NATIONAL DE SANTE - CALCULE*

2.1.2.1 **Description et propriétés**

L'utilisation du NIR en tant qu'identifiant national de santé ayant été écartée suite aux recommandations de la CNIL en 2007, un identifiant national de santé spécifique devait être créé pour être utilisé dans le cadre du dossier médical personnel et du dossier pharmaceutique. Afin de garantir les caractéristiques requises d'un identifiant national fiable, l'INS devait être à terme créé et attribué d'une manière centralisée et inscrit dans la puce de la carte vitale des bénéficiaires de

l'assurance maladie (« Les raisons d'être et le cadre réglementaire de l'INS | esante.gouv.fr, le portail de l'ASIP Santé » 2015). La mise en place de cet organisme national et la montée en charge à un niveau national de la distribution des identifiants est un objectif atteignable sur le long terme, assujetti à contrainte du renouvellement des décideurs et des priorités. Un INS-c, ou calculé, a été alors défini afin de répondre aux besoins actuels des systèmes de santé partagés. L'INS-c peut être calculé localement dans tout système d'information en santé. Le NIR, le prénom et le sexe de chaque patient sont récupérés grâce à un lecteur de carte vitale. Ces données sont par la suite traitées et transformées via un procédé d'anonymisation faisant intervenir la fonction de hachage SHA-256 (Gilbert et Handschuh 2003) pour générer un identifiant de 22 chiffres non signifiant sans collisions[1] ni doublons[2] tel que recommandé par la CNIL.

L'INS-c proposé par l'Agence des Systèmes d'Information Partagés en Santé (ASIP Santé) est présenté comme une solution transitoire, avant la mise en place d'un système centralisé de gestion des INS pérennes.

2.1.2.2 **L'INS-c candidat pour la BNDMR ?**

L'INS-c respecte bien certaines contraintes qui ont été définies pour l'identifiant patient à utiliser dans la BNDMR. Il est non signifiant, et les risques de doublons et de collisions ont été évalués et jugés acceptables sur une population restreinte selon le document de conception de l'INS-c fait par l'ASIP Santé en 2009 (« Dossier de conception de l'Identifiant National de Santé calculé (INS-C) | esante.gouv.fr, le portail de l'ASIP Santé » 2014). Cependant les mêmes raisons qui nous empêchaient d'utiliser le NIR nous empêchent d'utiliser l'INS-c. En effet, l'INS-c est calculé à partir du NIR de l'assuré, il sera ainsi difficile d'attribuer des identifiants aux ayants droits tels que les enfants qui n'ont pas de cartes vitales. Il s'agit d'ailleurs de la principale limite qui a été soulevée lors de la définition de la stratégie de déploiement de l'INS-c et sa généralisation. Ceci est d'autant plus vrai pour les fœtus qui n'ont pas encore une existence officielle et qui ne peuvent donc pas se faire attribuer un NIR et encore moins un INS-c. Néanmoins, l'approche de calcul de l'INS-c reste

[1] Une collision = un seul identifiant pour deux patients différents

[2] Un doublon = deux identifiants différents pour le même patient

intéressante d'un point de vue anonymisation des données personnelles des patients, et de génération locale des identifiants au niveau des systèmes sources.

2.2 LES NUMEROS PATIENT POUR LE CHAINAGE DES DONNEES

Dans cette section, nous nous intéressons aux identifiants calculés *a posteriori* afin de permettre le chainage ou l'appariement de données issues de systèmes différents.

2.2.1 *CHAINAGE SNIIRAM PMSI*

Dans le contexte de planification des activités et d'analyse des coûts pour les établissements de santé, le NIR est indirectement utilisé pour chaîner les données issues des établissements de soin avec ceux du soin de ville. Dans le cadre du Programme de Médicalisation des Systèmes d'Information (PMSI), le NIR de l'assuré, adossé à d'autres données personnelles (date de naissance et sexe) du patient est anonymisé une première fois au niveau de chaque établissement puis une seconde fois à un niveau centralisé. Cette méthode de double anonymisation s'appuie sur l'algorithme FOIN qui a été initialement développé par le DIM du CHU de Dijon (Trouessin et Allaert 1997). Les mêmes données identifiantes des patients sont colligées par le SNIIRAM. Elles peuvent être anonymisées de la même manière que celles du PMSI par la méthode de double anonymisation. Ainsi une correspondance des deux sources de données peut être réalisée.

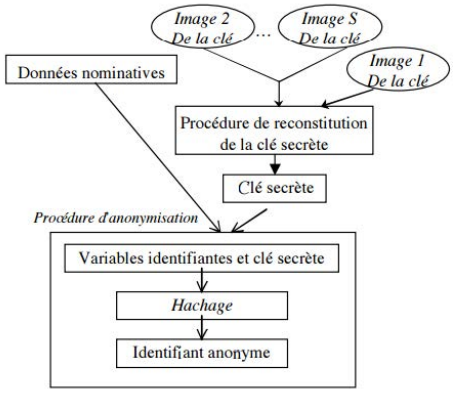

FIGURE 8 PROCEDURE FOIN (KALAM ET AL. 2004)

Ce système de chainage de données a permis la conduite de diverses études épidémiologiques à un niveau national (Foulon et al. 2015; Hanf et al. 2013; Quantin et al. 2008). Il présente néanmoins certaines limites qui sont liées à l'utilisation du NIR sur lequel est fondé le calcul du numéro

anonyme. Même si la contrainte du NIR de l'assuré, et non de celui du patient, est contournée par l'utilisation de la date de naissance et du sexe du patient, le problème subsiste dans le cas des vrais jumeaux pour qui le même identifiant sera généré. En effet les vrais jumeaux ont la même date de naissance, le même sexe et le même numéro d'assuré que celui du parent qui les a déclarés.

2.2.2 *IDENTIFIANTS PATIENTS DANS LES ENTREPOTS DE DONNEES MALADIES RARES*

Aux Etats-Unis, le programme GRDR, lancé par le NCATS (National Center for Advancing Translational Science) des NIH (National Institutes of Health), vise à construire un registre de données patients maladies rares. Parmi l'ensemble d'éléments de données communs qui a été défini figure un identifiant patient spécifique au registre. Cet identifiant patient, ou GUID (« Global Unique Identifier (GUID) » 2016), permet le suivi des patients à travers les différentes études et les différents registres maladie spécifique. Il s'agit d'une chaine de caractères unique et aléatoire assignée à chaque patient, générée par un algorithme de hachage irréversible à partir des informations personnelles suivantes : prénom, nom de famille et nom complémentaire, jour, mois et année de naissance, ville et pays de naissance et le sexe physique à la naissance (cette dernière donnée est optionnelle). Nous n'avons pas retrouvé dans la littérature de publication décrivant la méthode d'élaboration ou des résultats de validation de ce GUID.

En Europe, le projet RD-CONNECT (Thompson et al. 2014), financé par le septième programme cadre dans le contexte du consortium IRDiRC lancé en 2011 (International Rare Diseases Research Consortium), projette de proposer un identifiant patient unique pour les maladies rares. L'adoption du GRDR-GUID est actuellement en discussion. Par ailleurs, Le projet EpiRare (Taruscio et al. 2014) définit aussi un identifiant européen patient maladie rare, EU-GUID, parmi les éléments de données communs de la plateforme européenne RDR (Rare Disease patient Registration). L'EU-GUID est un code élaboré à partir des données suivantes : prénom, nom de famille, sexe, date de naissance, ville de naissance et un code national unique d'identification. Et là aussi, nous n'avons retrouvé ni une description de la méthode ni une publication de résultats d'utilisation de cet EU-GUID (dernière recherche juin 2016).

2.2.3 *SYSTEMES MULTIMODAUX*

2.2.3.1 **Chainage des registres autisme**

Une collaboration entre l'initiative SFARI (Simons Foundation Austism Research Initiative) et le projet NDAR (National Database for Autism Research) aux Etats Unis a permis la mise en place d'une méthode d'identification des patients dans l'objectif de chainer les données issues de différents registres d'autisme (Johnson et al. 2010). Un système centralisé exposant des services web permet de générer des identifiants globaux et uniques (GUID). Cinq identifiants peuvent être générés après le hachage de combinaisons différentes d'informations nominatives liées aux patients et à leurs familles :

- Identifiant national ou gouvernemental
- Prénom du patient
- Nom de naissance du patient
- Nom complémentaire du patient
- Date de naissance du patient
- Sexe à la naissance du patient
- Commune de naissance du patient
- Prénom de la mère
- Nom de naissance de la mère
- Jour et mois de naissance de la mère
- Prénom du père
- Nom de naissance du père
- Jour et mois de naissance du père

Chaque registre source ayant fait appel à ces services web sécurisés récupère l'ensemble des identifiants ayant pu être calculés selon la disponibilité des informations nominatives envoyées et renvoie par la suite les données patients anonymisées adossées aux GUID. Dans le registre centralisé, les données issues des différentes sources sont chainées en s'appuyant sur un rapprochement entre les différents GUID.

2.2.3.2 **Outil de chainage des dossiers patients à Chicago**

Dans le cadre d'un projet de recherche clinique, une méthode a été proposée et implémentée pour chainer les données issues de 6 établissements de la région de Chicago aux Etats Unis (Kho et al. 2015). Les auteurs proposent un système à deux niveaux :

- Dans chaque établissement concerné par l'étude, à un niveau local, une application développée et distribuée par les auteurs permet de générer un ensemble d'empreintes de hachage issues du hachage par la fonction sha512 de différentes combinaisons d'informations identifiantes : prénom, nom de naissance, date de naissance, numéro de sécurité sociale et sexe.
- A un niveau centralisé, des correspondances entre les patients des différents établissements sont déduites en se basant sur un système à coefficients attribués aux ensembles d'empreintes de hachage envoyés par les établissements.

Ce système a permis la détection de 2 millions de doublons réduisant ainsi le nombre total de dossiers patients de 7 millions à 5 millions de dossiers (Kho et al. 2015). Les auteurs déclarent que le système est assez performant avec une spécificité de 100% et une sensibilité de 96%.

2.2.3.3 **Inconvénients**

Dans le contexte de notre projet, le principal inconvénient de ces approches est leur complexité. D'une part, la tâche des établissements sources de données est alourdie puisqu'il leur est demandé de calculer plusieurs empreintes de hachage, générées à partir de données identifiantes qui ne sont pas forcément disponibles dans leurs systèmes, de les envoyer au système central pour récupérer en retour l'identifiant attribué après la recherche de correspondances, d'adosser cet identifiant aux dossiers et de renvoyer l'ensemble à la plateforme où les analyses vont être menées. D'autre part, le système central doit être assez performant aussi bien au niveau technique, pour opérer assez rapidement tous les traitements, qu'au niveau fonctionnel où l'algorithme de mise en correspondances doit être adapté aux données de départ.

Par ailleurs, les différentes données sélectionnées pour le calcul des empreintes de hachage ne sont pas pertinentes dans le contexte de la BNDMR. Nous insistons ainsi sur l'impossibilité de l'utilisation du numéro de sécurité sociale à court terme si nous souhaitons identifier les enfants et sur le long terme si nous souhaitons identifier les fœtus. De plus, nous ne préconisons pas l'utilisation d'autres données, telle que la commune de naissance du patient pour des raisons que nous expliciterons dans la section de discussion de ce chapitre.

2.3 TABLEAU RECAPITULATIF

TABLEAU 1 LISTE DES IDENTIFIANTS

Identifiant	Couverture et domaine d'application	Identifiant anonyme ?	Système centralisé ?	Données nominatives
NIR	National (FR) Sécurité sociale – santé (récemment)	non	oui	Sexe Lieu de naissance Mois et année de naissance
INS-c (abandonné)	National (FR) Santé	oui	Non (peut être généré localement)	NIR Prénom Sexe
Numéro Anonyme FOIN	National (FR) Epidémiologie	oui	oui	NIR Date de naissance Sexe
GUID GRDR	National (US)	oui	Non (software distribué)	Prénom Nom Date de naissance Commune de naissance Pays de naissance Sexe (optionnel) ID local dans le registre source
EU-GUID EpiRare	Européen Epidémiologie	?	?	Prénom Nom Date de naissance Commune de naissance Code national unique d'identification
GUID Autism Collections	National (USA) Epidémiologie	oui	oui	Identifiant national Prénom Nom de naissance Nom complémentaire Date de naissance Sexe Commune de naissance Prénom de la mère Nom de naissance de la mère Jour et mois de naissance de la mère Prénom du père Nom de naissance du père Jour et mois de naissance du père => 5 empreintes de hachage
EHR linking tool in Chicago	Régional (Chicago-USA) Recherche clinique	oui	oui	Prénom Nom de naissance Date de naissance Numéro de sécurité sociale Sexe => 4 empreintes de hachage

3 L'IDENTIFIANT PATIENT MALADIE RARE : IDMR

3.1 CHOIX DES TECHNOLOGIES ET DES INFORMATIONS POUR LA CONSTRUCTION DE L'IDMR

3.1.1 *COMMENT GARANTIR L'ANONYMAT ?*

Afin de garantir l'anonymat de l'identifiant patient maladie rare deux méthodes s'offraient à nous. La première était de générer un identifiant aléatoire non signifiant pour chaque nouvelle fiche patient créée ou reçue. La seconde consistait à se baser sur des méthodes d'anonymisation pour transformer les informations nominatives préalablement collectées en un identifiant anonymisé. La première méthode a été écartée puisqu'elle ne satisfaisait pas les contraintes de minimisation de création de doublons dans la banque nationale. En effet, générer un nouvel identifiant aléatoire pour chaque fiche créée ou reçue suppose la mise en place par ailleurs de méthodes de fédération d'identité *a posteriori* afin de fusionner les fiches qui ont été créées en double en se basant entre autre sur les données nominatives des patients, qui ne sont présentes qu'au niveau de BaMaRa. Cependant, les fiches des patients dans BaMaRa ont vocation à évoluer tout au long de la prise en charge et les différents suivis du patient. La fédération des identités en aval n'est donc pas en adéquation avec la nature évolutive des fiches dans BaMaRa et devrait plutôt intervenir lors de la création de la fiche pour que les mises à jour ultérieures restent cohérentes et centralisées. De plus, avec l'adoption de cet identifiant généré aléatoirement il serait impossible de fédérer les fiches qui seront reçues directement dans la BNDMR de manière anonyme.

Ainsi, afin d'assurer cet objectif de fédération d'identité, dans le respect des contraintes d'accès aux données de santé, la solution retenue a été d'utiliser un algorithme d'anonymisation certifié comme ceux qui ont été utilisés pour la génération de l'INS ou du numéro FOIN pour le chainage des données du PMSI avec le SNIIRAM. Les fonctions de hachage permettent une transformation irréversible des données en entrée en une chaine de caractères non signifiante. L'algorithme SHA-256 qui a été défini dans une publication FIPS 180-2 par le NIST (National Institute of Standards and Technology) aux Etats-Unis (Dang 2015), est recommandé en France et fait partie intégrante du Référentiel Général de Sécurité de l'Agence Nationale de Sécurité des Systèmes d'Information (« Référentiel Général de Sécurité version 2.0 - Annexe B1 Mécanismes cryptographiques Règles et

recommandations concernant le choix et le dimensionnement des mécanismes cryptographiques » 2014).

3.1.2 *QUELLES DONNEES?*

Notre choix s'est porté sur certaines données nominatives des personnes. Des informations spontanément utilisées par les personnes et les différentes structures administratives pour identifier des individus. Pour identifier les patients d'une manière univoque, un compromis doit être trouvé pour déterminer le nombre de données nominatives qui doivent être sélectionnées. Les informations nominatives recueillies ne doivent pas être nombreuses pour éviter le risque d'erreurs sur la saisie des données d'un endroit à l'autre et éviter ainsi la génération de doublons d'identité. En revanche, il est nécessaire d'en recueillir un nombre suffisant pour qu'elles soient assez informatives et permettent ainsi de bien distinguer les patients.

Pour générer l'identifiant patient maladie rare, nous avons opté pour quatre données nominatives issues de l'état civil essentiellement pour leur stabilité, la facilité de leur recueil et de leur codage en base de données. Les deux premières données sont le premier prénom du patient et son nom de naissance tels qu'ils figurent sur son acte de naissance et sur ses documents officiels. Nous avons préféré le nom de naissance, dit aussi patronymique, au nom usuel puisqu'il est stable dans le temps contrairement au nom usuel qui lui a plus tendance à évoluer suite à des changements d'état civil par exemple. La date complète de naissance du patient est aussi demandée telle qu'elle figure sur les documents officiels. Enfin, nous utilisons le sexe du patient, féminin ou masculin et la modalité sexe « indéterminé » dans le cas des fœtus (ce cas particulier sera traité plus en détail plus loin dans le document).

Notre choix est aussi appuyé par la haute disponibilité de ces données par rapport à d'autres informations personnelles. Dans (Johnson et al. 2010) les auteurs classent dans un tableau les données personnelles relatives à un patient et aux membres de sa famille selon leurs disponibilités dans les dossiers qui sont remplis par les professionnels de santé. Leur étude montre que le mois de naissance, le prénom, l'année de naissance, le jour de naissance et le nom de naissance sont disponibles dans les dossiers avec un taux de 99,6% pour une population de 2000 personnes dans le cadre de cette étude.

Nous avons par ailleurs validé notre choix avec une étude quantitative visant à déterminer quelles informations sont suffisamment discriminantes pour éviter les collisions d'identités. Pour ce faire, les données nominatives ont été ordonnées selon leur disponibilité et nous avons suivi l'évolution du nombre de collisions au fur et à mesure qu'on adossait une nouvelle donnée aux précédentes. Cette étude a été conduite sur 45 000 identités, issues de la base de données des personnes de Wikipedia (« DBpedia » 2016), qui nous ont servi comme base de test. La Figure 9 montre qu'à partir des trois premières données les plus disponibles, prénom, nom et date de naissance, aucune collision n'est générée pour une population de 45 000 personnes. Nous avons tout de même gardé le sexe du patient dans la génération de l'IdMR puisqu'il s'agit d'une information facilement récupérable et qui pourrait éviter les collisions d'homonymes ayant des prénoms mixtes comme Claude, Camille ou Dominique.

La non sélection de la donnée « commune de naissance », malgré son intégration au calcul de plusieurs GUID comme indiqué dans la section 2.2, sera discutée plus en détail dans la 5.2.1.

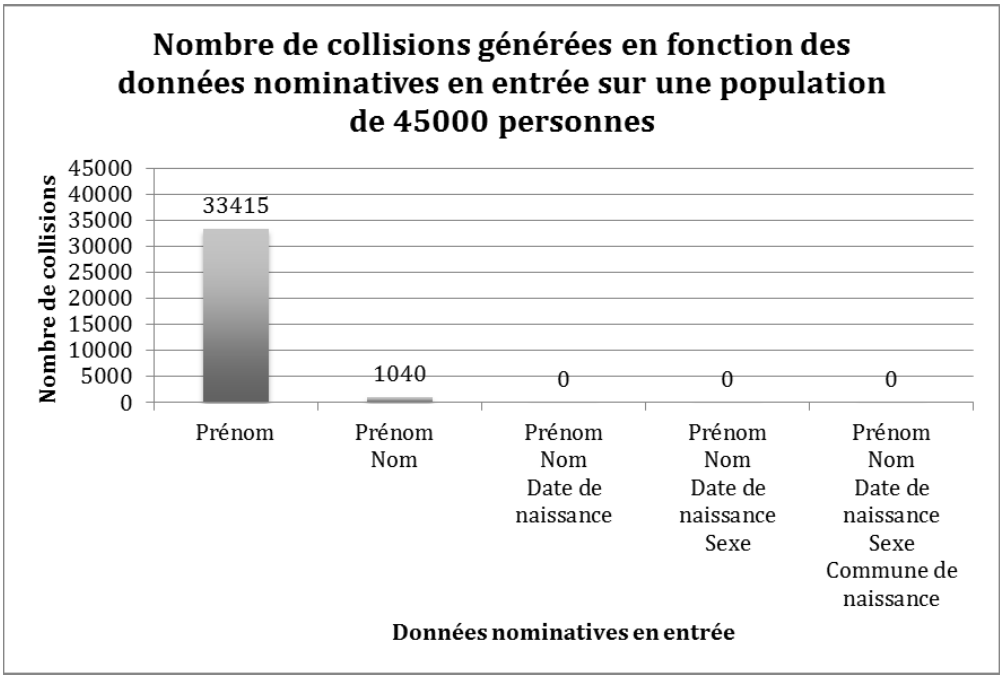

FIGURE 9 EVOLUTION DU NOMBRE DE COLLISIONS D'IDENTITES EN FONCTION DES DONNEES NOMINATIVES

3.2 PROCESSUS DE GENERATION DE L'IDMR

3.2.1 *LES ETAPES DU PROCESSUS DE GENERATION DE L'IDMR*

La transformation des données nominatives en un identifiant patient maladie rare passe par 3 étapes :

- Prétraitement des données
- Hachage
- Post-traitement de l'empreinte de hachage

3.2.1.1 *Prétraitement des données [A]*

L'objectif est d'homogénéiser les informations sur lesquelles se base le calcul de l'IdMR afin que les données d'un même patient respectent le même format quel que soit l'endroit où le patient a été vu et quel que soit le système où ses données ont été saisies.

Les données nominatives sur lesquelles se base le calcul de l'IdMR, sont issues de sources hétérogènes où elles sont potentiellement stockées sous des formats différents. La première étape consiste donc à transformer ces informations dans un format unique afin que les informations de chaque patient soient comparables.

Convertir le format d'une date est une tâche qui est techniquement assez simple et que la plupart des langages de programmation supporte. Le choix du format de la date de naissance en entrée de l'algorithme de génération de l'IdMR s'est porté sur le format ISO-8601 sans les tirets : 8 caractères numériques dont les 4 premiers représentent l'année, les deux suivants représentent le mois et les deux derniers représentent les jours (voir l'exemple dans la Figure 11).

Il est aussi assez simple d'harmoniser les éléments de valeurs du sexe qui représentent le sexe féminin, masculin et indéterminé. Ces éléments de valeurs seront codés sur un seul caractère représentant la première lettre de chaque sexe : « F », « M » et « I ».

En ce qui concerne les données *prénom* et *nom de naissance*, il n'est pas réaliste, voire impossible, de vouloir contrôler toutes les entrées possibles en essayant d'imposer un domaine de valeurs restreint qui constituerait finalement la liste de tous les prénoms et noms possibles que pourrait porter un patient donné. Cependant, ces chaines de caractères qui ont été saisies librement peuvent être traitées afin de réduire la variabilité qui aurait pu être introduite par les

erreurs de frappe, ex. « Jean-Pierreeeeeee » au lieu de « Jean-Pierre » ou les variations d'orthographe, ex. « Marie Adélaïde » ou « Marie-Adélaide ». Ainsi ces données seront traitées afin de supprimer les caractères spéciaux (ponctuation, symboles et espaces) et de ne permettre que l'usage de caractères alphanumériques (l'intégration des caractères numériques sera expliquée plus loin dans la partie fœtus). En entrée de la fonction de hachage, ces caractères alphanumériques, de A à Z et de 0 à 9 seront encodés en UTF-8. Les lettres accentuées seront remplacées par les caractères alphabétiques correspondant non accentués (voir Tableau 2). Par la suite toute la chaine de caractères sera mise en majuscule. Ces chaines de caractères doivent aussi être tronquées afin de réduire la variabilité due à des erreurs de saisie telles que dans l'exemple de « Jean-Pierreeeeeee » que nous avons déjà mentionné. Pour ce faire, nous nous sommes basés sur l'étude de la distribution des longueurs des prénoms et des noms des 280 000 patients de la base CEMARA afin de déterminer la longueur maximale autorisée de ces données. La médiane se situait aux alentours de 6,5 caractères pour les prénoms et autour de 7,1 caractères pour les noms. Nous avons donc fixé le seuil à 10 caractères pour couvrir 75% de la population étudiée : 3e quartile (voir Figure 10). Ainsi, les prénoms et les noms de naissance sont tronqués pour ne pas dépasser les 10 caractères. Lorsque ce seuil n'est pas atteint, des espaces sont rajoutés à droite pour atteindre la longueur de 10 caractères.

TABLEAU 2 TABLEAU DE SUBSTITUTION DES CARACTERES ALPHABETIQUES

Caractère à substituer	Caractère substitut
À Á Â Ã Ä Å Æ à á â ã æ ä å	A
Ç ç	C
Ð ð	D
È É Ê Ë è é ê ë	E
Ì Í Î Ï ì í î ï	I
Ñ ñ	N
Ò Ó Ô Õ Ö Ø ò ó ô õ ö ø	O
Š š	S
Ù Ú Û Ü ù ú û ü	U
Ý Ÿ ý ÿ	Y
Ž ž	Z

Œ œ	OE
ß	SS
Caractères minuscules de ‘a’ à ‘z’	Caractères majuscules de ‘A’ à ‘Z’

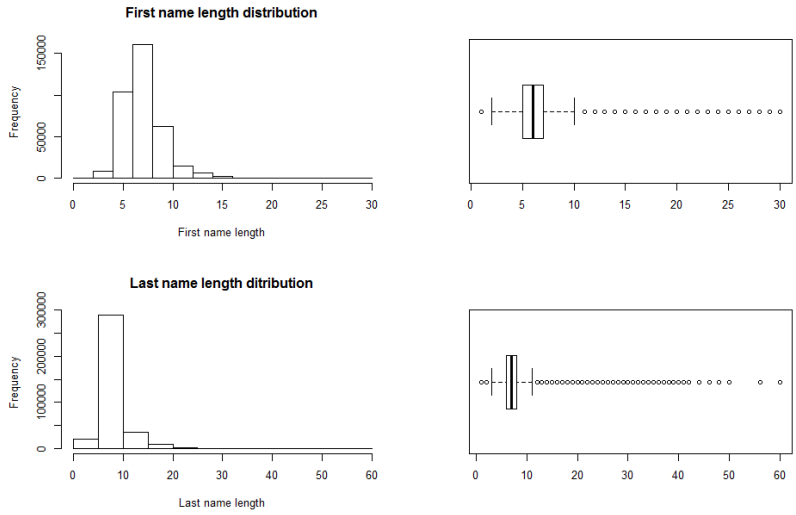

FIGURE 10 DISTRIBUTION DES LONGUEURS DES PRENOMS ET DES NOMS DES PATIENTS DANS LE BASE DE DONNEES DE CEMARA

3.2.1.2 *Hachage [B]*

L’objectif est de rendre anonyme, d’une manière irréversible, les données nominatives des patients.

Les données issues du prétraitement sont par la suite concaténées dans cet ordre : prénom, nom de naissance, date de naissance et sexe. Nous obtenons ainsi une chaine de 29 caractères qui porte encore les informations nominatives du patient.

La fonction de hachage SHA-256 est par la suite utilisée pour transformer cette chaîne de caractères en une empreinte de hachage de 256 bits. Cette empreinte est non significative, une suite de 0 et de 1 ne portant aucune information en clair, et a été générée d’une manière irréversible, les données nominatives en entrée ne peuvent être retrouvées en faisant le chemin inverse.

3.2.1.3 *Post-traitement de l’empreinte de hachage [C]*

L'objectif est de rendre l'identifiant facile d'utilisation tout en minimisant la détérioration de ses performances (point de vue risque de collisions).

L'empreinte de 256 bits est convertie en décimal. Chaque octet (suite de 8 bits) des 32 octets constituant l'empreinte est traduit en un nombre décimal (étant codé sur un octet sa valeur décimale ne dépassera pas 255). Pour que l'empreinte reste la plus discriminante possible, les zéros à gauche de chaque nombre décimal seront supprimés, ex. 25 au lieu de 025. Les 32 nombres décimaux ainsi obtenus seront concaténés en une chaine de caractères numériques qui peut atteindre une longueur de 96 caractères. Pour des raisons de praticité, nous avons essayé de raccourcir cette chaîne tout en essayant de ne pas trop augmenter les risques de collisions. Nous avons utilisé la base des 45000 individus récupérés de DBpedia afin de valider le seuil que nous devons respecter pour tronquer l'empreinte de hachage. Sur 45 000 identités, et avec une empreinte tronquée à seulement 10 caractères, nous avons détecté la survenue de 6 collisions. Avec une empreinte tronquée à 20 caractères, aucune collision n'a été détectée.

3.2.2 *Schema general du processus de generation de l'IdMR*

Le processus de génération de l'IdMR génère donc un identifiant patient de 20 caractères numériques, issu du hachage par l'algorithme SHA-256 de quatre informations personnelles nominatives : le prénom, le nom de naissance, la date de naissance et le sexe du patient. Le processus inclut aussi d'autres traitements sur les chaines de caractères afin de réduire la variabilité due aux erreurs de saisie.

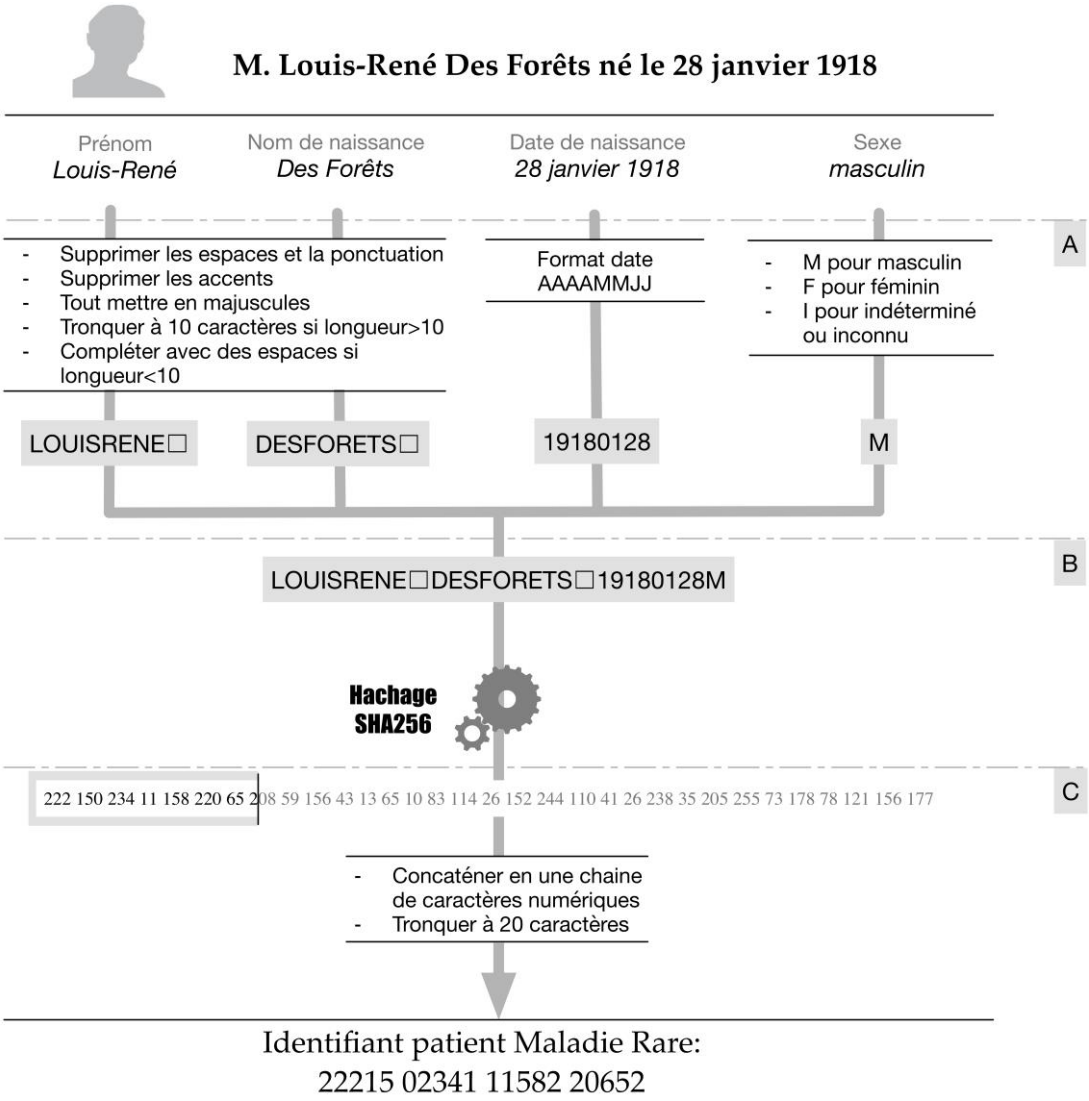

FIGURE 11 PROCESSUS GENERAL DE GENERATION DE L'IDMR

3.3 UN IDENTIFIANT POUR LES FŒTUS: POURQUOI ET COMMENT ?

3.3.1 *MOTIVATIONS*

Parmi les 7000 maladies rares estimées, 80% seraient des maladies génétiques. Qu'elles soient héréditaires ou pas, ces maladies génétiques touchent pour la plupart les cellules germinales et sont donc présentes dès le début de la vie fœtale. Ceci expliquerait la grande proportion d'enfants, 50%, parmi les personnes atteintes de maladies rares. Un objectif est de diagnostiquer ces affections de plus en plus tôt au cours de la vie, voire durant la période prénatale, pour que la prise en charge puisse être mise en œuvre le plus précocement possible.

Les fœtus, même s'ils ne bénéficient pas d'un état civil en dehors de certaines conditions précisées dans 3 décrets de la Cour de cassation et qu'ils n'aient pas d'existence juridique autonome en dehors de la personne de leur mère, ils n'en constituent pas moins des « personnes humaines en devenir » (Cour de cassation 2016; Sureau 2010; Moutel et al. 2010; aikido 2011). Ils sont considérés dans les systèmes de soins lors de la prise en charge de leurs mères. De plus, d'un point de vue de modélisation de l'information et des processus de prise en charge, ce sont les fœtus qui sont diagnostiqués, ce sont eux qui portent la maladie et c'est sur eux que les examens génétiques sont opérés. Il était donc naturel de créer des fiches maladies rares dans BaMaRa et de comptabiliser les fœtus présentant une maladie rare dans la BNDMR.

3.3.2 *UN IDMR POUR LES FŒTUS ?*

Comment correctement identifier les fœtus dans la BNDMR sachant qu'ils n'ont pas encore d'état civil, qu'ils n'ont peut-être pas encore un prénom et un nom de famille, qu'ils n'ont pas de date de naissance et que leur sexe n'a peut-être pas encore été déterminé ?

Nous nous sommes basés sur la pratique des utilisateurs de CEMARA pour proposer un ensemble d'informations sur lesquelles se baser pour identifier les fœtus. Cette proposition d'ensemble de données, qui sont le plus souvent disponibles, et la méthode d'uniformisation ont été validées auprès d'un groupe de foetopathologistes que nous remercions : Pr. Tania Attie-Bitach (Hôpital Necker - Paris), Dr. Marie Gonzales (Hôpital Trousseau - Paris), Dr. Sophie Blesson (Hôpital Bretonneau - Tours), Dr. Laurence Loeuillet (Hôpital Cochin - Paris) et Dr. Marie-Hélène Saint-Frison (Hôpital Robert Debré - Paris).

Les 4 données en entrée de l'algorithme de génération de l'IdMR sont les suivantes :

- Comme prénom nous noterons la lettre « f », pour fœtus, suivie du numéro d'ordre du fœtus dans sa fratrie dans le cas d'une grossesse gémellaire. Le prénom de la mère sera par la suite concaténé.
- Comme nom de naissance nous récupérerons le nom de naissance (de jeune fille) de la mère.
- Pour remplacer la date de naissance nous nous baserons sur la date de début de grossesse afin de différencier les grossesses et donc les fœtus d'une même mère. Sachant que cette

information n'est pas très précise nous nous contenterons de récupérer le mois et l'année du début de grossesse. Pour le calcul de l'IdMR le jour sera fixé au premier jour du mois.

- Le sexe inconnu, et donc la lettre « I », sera adopté pour la génération de l'IdMR de tous les fœtus. Même si le sexe du fœtus est déjà connu, pour des raisons de stabilité de l'information lorsque le fœtus est vu à différentes phases de son développement, et pour que son IdMR reste ainsi toujours comparable, par convention son sexe sera considéré comme « inconnu ».

Certes ces données peuvent être considérées comme non suffisantes ou pas assez précises pour bien connaître le patient et le contexte de sa maladie. Nous tenons donc à préciser que ces données telles que décrites dans ce paragraphe sont destinées à calculer l'IdMR pour le fœtus. Cela n'empêche pas un recueil plus complet des informations autour du fœtus dans la fiche maladie rare. Dans BaMaRa par exemple, le prénom du fœtus est recueilli si les parents en ont déjà choisi un. Les données du père sont récupérées lorsqu'elles sont disponibles. De même, la date de début de grossesse est plus précise. Ces informations sont donc transformées dans un but d'uniformisation et de garantie de stabilité de l'information tout en essayant de garder assez d'informations discriminantes pour pouvoir différencier les fœtus et ainsi les identifier (Tableau 3).

TABLEAU 3 CORRESPONDANCES ENTRE LES DONNEES RECUEILLIES POUR UN FŒTUS ET LES DONNEES SERVANT AU CALCUL DE SON IDENTIFIANT

Informations recueillies	Données pour le calcul de L'IdMR
Prénom du fœtus - renseigner **prénom** s'il est connu. *Ex : Robert* - sinon renseigner ou deviner le numéro d'ordre du fœtus dans la fratrie dans le cas d'une grossesse gémellaire. *Ex : fœtus 1, fœtus 2…*	Prénom patient **f1marta** **f2marta**
Prénom de la mère - Ex : Marta	
Nom de la mère - Ex : Langdon	Nom patient **Langdon**
Date de début de grossesse - Date approximative, l'année et le mois étant les données les plus importantes Ex : 2014-05-20	Date de naissance du patient **2014-05-01**
Sexe du fœtus Ex : masculin	Sexe du patient **inconnu**

L'approche que nous proposons est non seulement simple mais aussi innovante dans l'approche d'identification des fœtus dans les systèmes d'information de soins. Des approches similaires ont été décrites dans la littérature (Quantin et al. 2008) dans un but de chainage des comptes-rendus entre différents services de maternité, de pédiatrie ou de néo natalité. Ces approches permettaient le chainage des données des mères et des nouveaux nés. Ces derniers disposaient déjà de prénoms et de dates de naissance mais ne sont peut-être pas encore déclarés. Ces approches ne traitaient pas le cas des fœtus.

4 EVALUATION DE L'ALGORITHME DE L'IDMR

4.1 MATERIEL

L'IdMR a été testé sur tous les dossiers patients de CEMARA. Au premier trimestre 2015, la base de données contenait 359.339 dossiers de patients (dont notamment des apparentés avec suspicion) avec un taux de doublons estimé à 9% selon un algorithme d'identitovigilance intégrée à l'application. Les enfants et les fœtus représentent la moitié de la population étudiée : 45% d'enfants et 5% de fœtus.

La saisie des champs nominatifs, prénom, nom de naissance, date de naissance et sexe, est obligatoire dans CEMARA. Ainsi le problème de disponibilité des données de l'IdMR ne se pose pas dans notre contexte de test. Par ailleurs, d'autres aspects de qualité de données sont à considérer notamment en ce qui concerne l'exactitude des données saisies. Des cas flagrants ont pu être corrigés telle que la saisie du signe de ponctuation « ? » lorsque le prénom du patient n'est pas connu. Dans le processus de gestion de qualité de données, des retours vers les professionnels de santé sont effectués afin de vérifier les dates de naissance fixées au premier janvier ou à la date du jour de saisie. Par ailleurs, des risques subsistent quant à l'orthographe et l'exactitude des données saisies. Le nom de naissance, par exemple, peut être confondu avec le nom marital ou usuel.

4.2 METHODE DE DETECTION DES COLLISIONS: CRITERE D'EVALUATION DES RESULTATS

L'algorithme de calcul de l'IdMR peut être validé lorsque le taux de collisions introduites par l'algorithme tend vers zéro. Nous savons qu'une collision est introduite lorsque le nombre de doublons dans l'ensemble des identités en sortie, les IdMR générés dans notre cas, est plus grand

que le nombre de doublons parmi l'ensemble des identités en entrée, les données nominatives. Autrement dit, lorsque le nombre de personnes distinctes parmi l'ensemble d'entrée est plus grand que le nombre d'IdMR distincts nous pouvons déduire que des collisions d'identités sont survenues.

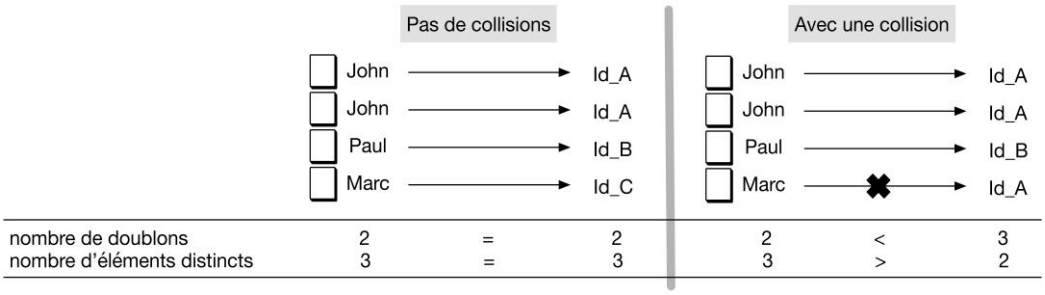

FIGURE 12 EXEMPLES DE COLLISIONS

Cette évaluation du nombre de collisions est étudiée à deux niveaux de l'algorithme :

- Collisions introduites par les phases B et C de l'algorithme : La fonction de hachage par sa nature intrinsèque peut générer des collisions dès lors que l'ensemble de départ (toutes les données possibles en entrée) a un cardinal supérieur à l'ensemble d'arrivée (empreintes de hachage) qui est égal à 2^{256} dans le cas du SHA256. Le post-traitement augmente quant à lui le risque de collisions avec la troncature des empreintes de hachage. En effet, la distinction entre deux empreintes peut n'être présente qu'au niveau des bits de poids le plus faible qui ont été supprimés (correspondants aux chiffres grisés à droite de la chaine numérique dans la Figure 11).
- Collisions introduites par la phase A : La phase de prétraitement des données réduit relativement la variabilité au niveau des données nominatives en entrée en enlevant les accents et en tronquant les noms et les prénoms à 10 caractères maximum. Ainsi, deux patients du même sexe, nés le même jour, et ayant des noms et prénoms similaires avec une légère différence au niveau de l'accentuation où dans les terminaisons d'au-delà des 10 caractères peuvent se voir attribuer la même chaine de caractère nominative prétraitée et par conséquent le même IdMR.

4.3 RESULTATS

D'abord, nous avons estimé le nombre de collisions introduites par la phase A de l'algorithme de calcul de l'IdMR durant laquelle les données nominatives des patients sont prétraitées. Le nombre

de doublons exacts parmi l'ensemble des données nominatives originales était plus petit que le nombre de doublons dans l'ensemble des données en sortie de la phase de prétraitement. Nous avons ainsi détecté 1771 collisions dues à la phase A. Nous avons étudié un échantillon tiré au sort de ces collisions et découvert qu'ils s'agissaient de doublons d'identité qui faisaient partie des 9% de doublons estimés dans la base de données CEMARA. En effet, certains patients avaient plus d'un dossier dans la base. Cela est dû à la variabilité introduite lors de la saisie des données nominatives du patient. Il s'agit généralement de différences au niveau des noms et des prénoms qui sont saisis en texte libre. Nous remarquons entre autres des différences d'accentuation, des espaces supplémentaires, l'utilisation ou non des traits d'union dans les noms composés ou par exemple la saisie d'un nom usuel composé par exemple (fictif) « Chantournais - Ponthierry » au lieu du nom de naissance « Chantournais » seul. Pour ces 1771 cas, l'algorithme de calcul de l'IdMR grâce à sa phase de prétraitement des données nominatives a assuré la fédération des identités qui étaient dupliquées dans la base de données.

Ensuite, nous avons estimé le nombre de collisions introduites par les phases B et C de l'algorithme de calcul de l'IdMR durant lesquelles les données prétraitées sont hachées puis les empreintes de hachage post-traitées. Le même nombre de doublons a été détecté parmi les données prétraitées parmi les IdMR en sortie de l'algorithme. Ainsi aucune collision n'a été introduite par les phases B et C de l'algorithme pour les données de test.

Parmi les 359.339 fiches patients distinctes dans la base de données CEMARA, 17.470 doublons exacts ont été détectés en considérant les données nominatives permettant le calcul de l'IdMR telles qu'elles étaient saisies par les utilisateurs. 1771 doublons supplémentaires ont pu être détectés grâce à la phase de prétraitement de données de l'algorithme. En total, l'algorithme de l'IdMR a permis de détecter et de fédérer 19.241 doublons d'identité. Ce chiffre représente 5,35% du nombre total de fiches patient dans la base de données CEMARA (voir Tableau 4).

TABLEAU 4 RESULTATS D'EVALUATION DE L'ALGORITHME DE GENERATION DES IDENTIFIANTS

	Nombre de doublons	Pourcentage de doublons	Conclusion
Sur les données nominatives en entrée	17.470	4.86%	4,86% des 359.339 fiches patients présentent des doublons d'identité au niveau: du prénom, nom de naissance, date de naissance et sexe.

Sur les chaines de caractères nominatives prétraitées	19.241	5.35%	1771 doublons d'identité supplémentaires ont été détectés grâce à la phase A (0.49%).
Sur les IdMR	19.241	5.35%	Aucune collision n'a été détectée en sortie des phases B et C.

5 CONCLUSION

5.1 RESUME

Nous avons proposé une méthode pour construire un identifiant patient maladies rares, que nous nommons IdMR, permettant la fédération des identités des patients et empêchant le risque direct de ré-identification des patients. L'IdMR est construit à partir de données nominatives permettant une identification stable du patient dans le temps et l'espace : le prénom, le nom de naissance, la date de naissance et le sexe. Ces données sont prétraitées afin de réduire la variabilité pouvant être introduite par les erreurs de saisie ou les différences orthographiques. Un premier pas vers la fédération des identités est ainsi assuré par l'algorithme. Une fonction de hachage permet d'anonymiser ces données nominatives d'une manière irréversible empêchant ainsi la ré-identification directe du patient à partir de l'IdMR. Cet algorithme a été validé après un test sur les dossiers patients de CEMARA. La simplicité et la facilité d'implémentation étaient nos prérequis afin d'encourager l'adhésion des systèmes sources au projet. Cette approche a été testée et évaluée dans le contexte du projet BNDMR et pourrait être implémentée à un niveau Européen sur des cas d'usage similaires.

5.2 DISCUSSION

5.2.1 *NON INTEGRATION DE LA « COMMUNE DE NAISSANCE »*

Nous avons étudié la possibilité d'intégrer la donnée « commune de naissance » au calcul de l'IdMR. Le principal argument qui nous prémunissait de l'utiliser était la difficulté à garantir une harmonisation entre les différentes terminologies géographiques qui sont utilisées au niveau des applications sources et ce à plusieurs niveaux :

- Harmonisation au niveau du choix de la terminologie :

Un standard géographique pour les communes qui soit unique n'existe ni à l'échelle nationale (codes INSEE et codes postaux) ni à l'échelle internationale (une seule terminologie standardisée pour toutes les villes du monde).

- Harmonisation au niveau de la gestion des évolutions :

Une terminologie évolue dans le temps, par exemple deux communes peuvent fusionner. Cela pourrait impliquer l'entrée de deux données différentes au niveau de deux systèmes sources différents utilisant des versions différentes d'une même terminologie.

- Harmonisation au niveau de la manière de coder l'information :

Certaines terminologies sont hiérarchisées et leurs éléments peuvent ainsi se situer à des niveaux de granularité différents. Par exemple, certains utilisateurs peuvent saisir 75000 pour Paris et d'autres préfèreraient être plus précis en saisissant 75001 ou 75002 pour spécifier l'arrondissement de Paris.

Enfin, cette information n'est pas automatiquement disponible et les professionnels de santé n'ont pas l'habitude de la demander à leurs patients. En effet, l'argument qui est généralement avancé pour que cette donnée soit demandée est que, en France, la commune de naissance permet la vérification du statut vital du patient. C'est en effet la mairie de la commune de leur naissance qui est récipiendaire du certificat de décès de tous les français. Ce recueil devient justifié à des fins d'études épidémiologiques. Dans CEMARA par exemple, 16% des fiches patients ont des communes de naissance manquantes.

Par ailleurs, nous remarquons, que mis à part les difficultés autour du recueil et du codage des communes de naissance, cette information n'est pas indispensable pour différencier et distinguer les personnes. En effet, comme le montre la Figure 9, les quatre premières données nominatives suffisent pour distinguer les patients et l'ajout de la commune de naissance n'est pas nécessaire pour éviter les collisions d'identités.

5.2.2 *RISQUE NON NUL POUR LES CONTRAINTES DE DEPART*

Nous avons investi beaucoup d'efforts afin de répondre à nos premières contraintes : unicité, pérennité, anonymat et globalité. Cependant, dans certains cas extrêmement rares, ces contraintes pourraient ne pas être complètement satisfaites.

5.2.2.1 **Les doublons**

Le risque d'avoir des doublons, en attribuant deux identifiants différents à un même patient persiste et menace l'unicité et la pérennité de l'IdMR. Ce problème est généralement dû aux modifications ou évolutions qui affectent les données nominatives en entrée : prénom, nom de naissance, date de naissance et sexe. Deux raisons sont en cause de ces changements :

- De vraies modifications affectant les données nominatives comme un changement de prénom. Ce cas est très rare mais il est tout de même rappelé aux responsables des systèmes sources de notifier ce changement en envoyant le couple ancien IdMR-nouveau IdMR afin d'éviter la création d'une nouvelle fiche patient dans la base de données nationale.
- Des modifications peuvent être introduites à cause d'erreurs de frappe lors de la saisie des données par les utilisateurs dans les systèmes sources. Afin d'éviter cela, il est important d'intégrer des contrôles de qualité dans le processus de saisie de données et de rappeler aux utilisateurs l'importance de l'exactitude de ces données et des principes de l'identitovigilance.

5.2.2.2 **Les collisions**

Comme pour les doublons, le risque d'avoir des collisions, en attribuant le même identifiant à deux patients différents, n'est pas nul. Ce risque est dû à la nature intrinsèque de la fonction de hachage puisqu'elle génère moins de valeurs possibles d'empreintes de hachage (taille fixe de 256 bits pour le SHA256) que de valeurs possibles de données en entrée (des chaines de caractères de n'importe quelle longueur). Nous avons estimé la probabilité de survenue d'une collision qui serait due au hachage par la fonction SHA256 (Gilbert et Handschuh 2003). Cette probabilité est estimée à seulement 8,5 e^{-12} pour une population de 1,48 e^{33} individus, nés sur une période de 100 années, au sein de laquelle toutes les identités possibles seraient représentées sur les quatre données nominatives en entrée de l'algorithme de l'IdMR (voir Tableau 5). Cette probabilité baisse encore significativement si on ne considère qu'une population de 4 millions d'individus, qui représenterait la population des personnes touchées par une maladie rare, pour atteindre 6,9 e^{-65}.

TABLEAU 5 ESTIMATION DU NOMBRE MAXIMAL D'IDENTITES DIFFERENTES EN ENTREE DE L'ALGORITHME

	Description	Estimation
Nombre de tous les prénoms possibles	Toutes les combinaisons possibles des caractères de l'alphabet latin sur une longueur de 10 caractères	26^{10}
Nombre de tous les noms possibles	Toutes les combinaisons possibles des caractères de l'alphabet latin sur une longueur de 10 caractères	26^{10}
Nombre de toutes les dates de naissance possibles sur cent ans	31 jours par mois 12 mois par an 100 ans	31×12×100
Nombre de tous les genres possibles	masculin ou féminin	2
Nombre de toutes les identités possibles en entrée de l'algorithme	**Produit des estimations de chaque donnée**	**$1.48\ e^{33}$**

Des collisions peuvent aussi survenir à cause des traitements qui visent à réduire la variabilité, en tronquant les chaines de caractères en entrée ou en sortie, par exemple.

Enfin, même s'il est extrêmement rare de rencontrer ce cas dans une population estimée à 3 millions de patients atteints de maladies rares, des homonymes ou autrement dit des personnes avec exactement le même prénom, le même nom de naissance, la même date de naissance et le même sexe pourraient exister.

5.2.2.3 **L'anonymat**

Des personnes malveillantes ayant accès à des fiches patients non directement nominatives (sans données nominatives mais avec l'IdMR) pourraient essayer de reconstituer l'identité des patients. L'attaque type dictionnaire est la plus populaire. Elle se baserait sur la reconstruction de toutes les valeurs possibles en entrée et le calcul de leurs IdMR en espérant retrouver l'identité des patients en croisant les IdMR des fiches avec ceux de la nouvelle liste constituée. Cependant, construire une telle table nécessite le déploiement de moyens colossaux (des temps de calcul pouvant atteindre des dizaines d'années).

Une autre approche consisterait à utiliser les informations existantes dans la fiche patient dé-identifiée pour remonter à une identification nominative du patient, même si l'IdMR est bien une chaine de caractères anonyme et non signifiante. En effet, en disposant de l'âge, du lieu de

naissance et du diagnostic d'un patient, l'identité de ce dernier peut être facilement retrouvée. Cela est d'autant plus facile dans le périmètre des maladies rares où le nombre de patients affectés peut ne pas dépasser, pour certaines maladies, une dizaine de cas en France (« Syndrome de Mowat-Wilson | LORD » 2016; « Syndrome onycho-digito-mammaire | LORD » 2016). Dans ce cas, remplacer certaines données nominatives tels que le nom et le prénom par un identifiant dans une fiche patient n'est pas suffisant. Les autres informations contenues dans la fiche patient sont aussi sensibles et doivent être protégées. Ainsi, la sécurité du système d'information en santé doit être impérativement renforcée et notamment au niveau de la gestion des droits d'accès. D'autres bonnes pratiques sont aussi recommandées dans le domaine de la protection de données de santé tel que le passage obligé à un niveau supérieur d'agrégation de données lors de l'ouverture de l'accès à ces données pour des exploitations statistiques et la supervision des traitements incluant un chainage de données (BRAS 2013).

5.2.2.4 **Exactitude**

Quand les données sont envoyées par les dossiers patients informatisés (DPI) dans les hôpitaux, l'exactitude des données nominatives est généralement fiable. Dans les hôpitaux, les patients sont pour la plupart enregistrés en utilisant leurs cartes vitales qui contiennent entre autres les données nominatives nécessaires au calcul de l'IdMR. Cependant le problème demeure pour les enfants dont les données sont parfois saisies manuellement. Les personnes ayant moins de 16 ans n'ont pas leurs propres cartes vitales et sont souvent enregistrés, pour des raisons de remboursement des frais de soins, avec la carte de leurs tuteurs. Cette carence peut être avantageuse dans le cas des fœtus, dont les mères sont enregistrées au niveau des DPI et pour qui le calcul de l'IdMR repose essentiellement sur les données de la mère.

De plus, dans la base de données nationale, la qualité des données est vérifiée à deux niveaux. D'abord, un contrôle automatisé lors de la saisie des données via l'application BaMaRa pour vérifier la complétude et la cohérence des données. Ensuite, grâce à un travail de gestion de données en collaboration avec les utilisateurs ayant saisi les données sur un sous-ensemble de fiches patients sélectionnées aléatoirement.

5.3 CONCLUSION

Afin de garantir une fluidité dans la continuité des soins, cela ne fait aucun doute qu'un identifiant patient unique et global à un niveau national est la solution la mieux adaptée. Les états européens et américains investissent dans ce sens (« White Paper on Unique Health Identifier for Individuals » 2016; Health 2016; Wall 2016). La France a initié ce projet ambitieux avec l'Identifiant National de Santé qui sera utilisé au sein du Dossier Médical Partagé et le vote récent de la Loi de Santé 2015 qui désigne le NIR, ou plus communément numéro de sécurité sociale, comme le nouvel INS. Des questions restent cependant posées par rapport aux procédures d'utilisation du NIR notamment pour les enfants qui ne disposent pas de carte vitale et dont le numéro n'est pas inscrit actuellement sur la carte des parents. Par ailleurs les fœtus, qui représentent une proportion non négligeable de la population des patients atteints de maladies rares, n'ont pas de NIR et ne sont même pas reconnus légalement en tant que personnes juridiques autonomes.

Afin de satisfaire les besoins à court terme du projet BNDMR et vu qu'aucune des solutions d'identification de patients décrites dans la littérature ne satisfaisaient nos contraintes, nous avons construit un identifiant patient spécifique à la base de données nationale. L'IdMR permet non seulement l'identification des patients mais aussi la fédération d'identités et le chainage de données dans le cadre de projets de recherche spécifiques, et particulièrement dans le cadre des cohortes maladies rares.

Les atouts de l'IdMR consistent en :

- la simplicité de son algorithme : l'algorithme de génération de l'IdMR est simple ce qui permet de le mettre en œuvre localement, évitant ainsi le besoin d'échange d'éléments au préalable entre les différents systèmes tel que l'échange de clés de hachage. Il ne s'agit pas non plus d'un système centralisé d'identitovigilance se basant sur des méthodes complexes de rapprochement d'identités.
- L'universalité des données nominatives sur lesquelles l'IdMR est bâti et leur stabilité : Le prénom, nom, date de naissance et sexe sont des informations acquises à la naissance et n'ont pas vocation à évoluer. Ces informations sont aussi universelles puisqu'elles ne dépendent pas des contextes locaux de certains pays contrairement à certaines données telles que les identifiants nationaux et les communes de naissance qui sont codées selon les standards nationaux.

L'IdMR répond actuellement aux besoins du projet BNDMR avec un objectif de fédération des identités des patients afin de limiter le nombre de doublons dans l'entrepôt national. Dans un contexte épidémiologique, une légère marge d'erreurs reste tolérable ce qui n'est pas admissible dans un contexte de soin et de prise en charge où la sécurité des patients est impliquée. Adossé à des procédures d'identitovigilance en amont, l'algorithme de l'IdMR rend les études épidémiologiques émanant de l'entrepôt national fiables.

Nous croyons que l'utilisation de l'IdMR peut dépasser le champ initial de son application pour concerner des projets de recherche ou des études connexes. Ainsi dans le cadre d'une étude conduite par l'Association Française contre les Myopathies (AFM) cela s'est déjà concrétisé. Cette étude visait à évaluer l'impact de ses référents parcours santé. Un appariement avec les données de CEMARA a été effectué. Cet appariement anonyme, opéré avec l'IdMR, a permis à l'AFM de vérifier l'exhaustivité de leur recueil en relevant le nombre de patients inclus dans les centres de référence neuromusculaires ayant CEMARA et non répertoriés dans les bases de l'association. Nous croyons aussi que, grâce aux propriétés de l'identifiant que nous proposons, ces études pourraient être d'une plus grande envergure pour concerner notamment des projets européens ou être appliqués hors du champ des maladies rares.

CHAPITRE III : L'INTEGRATION DE DONNEES – DU RECUEIL STANDARDISE A LA DECOUVERTE DE CORRESPONDANCES

1 INTRODUCTION

1.1 APPROCHE ENTREPOT POUR L'INTEGRATION DES DONNEES MALADIES RARES

L'intégration de données regroupe les différentes techniques qui permettent d'offrir un accès uniforme à un ensemble de sources de données autonomes et hétérogènes. La BNDMR, s'inscrit bien dans ce contexte puisqu'elle vise à offrir aux institutionnels et aux chercheurs une vue homogène et unifiée d'un ensemble de données patients maladies rares hétérogènes recueillies au niveau des différents sites des CRMR et des CCMR.

La première approche d'intégration de données est l'approche de médiation et de fédération de données. Elle consiste en la traduction de requêtes à la volée pour interroger les différentes sources et apporter une réponse fédérée à la requête initiale de l'utilisateur. Dans le domaine de la santé, c'est surtout dans le contexte d'initiation de recherches cliniques (Amarouche et al. 2011; Do et al. 2007; De Moor et al. 2015) ou en génomique que cette méthodologie a été adoptée (Louie et al. 2007). La médiation de requêtes suppose la pérennité et la stabilité des systèmes interrogés et ne garantit pas la qualité des données présentes dans les différentes sources.

Une approche de type entrepôt de données est une deuxième approche d'intégration de données. Elle consiste en la centralisation et la consolidation de données hétérogènes dans une base de données unique ayant un schéma global de données. Cette présence « physique » des données dans un seul endroit permet, en termes de performances, une exécution rapide des requêtes ainsi qu'un accès direct à l'ensemble des données qui permette les contrôles de qualité. C'est pour ces diverses raisons que cette approche est souvent employée dans le domaine de la santé. Au niveau des établissements de santé, des entrepôts ont été mis en place pour intégrer les données cliniques d'un hôpital donné avec des données d'autres bases externes, données omiques par exemple, pour permettre aux professionnels de santé de conduire des études de recherche

translationnelle spécifiques (Garcelon, Salomon, et Burgun 2014). Au niveau national, cette méthode a été choisie pour monter le système de la Caisse Nationale de l'Assurance Maladie (CNAM) qui permet la gestion de l'assurance maladie et des politiques de santé: le SNIIRAM. Elle a été aussi adoptée pour monter des entrepôts nationaux « maladie spécifique » telle que la Base Nationale Alzheimer (BNA).

La BNDMR aussi a opté pour l'approche entrepôt d'intégration de données, une architecture qui permet la consolidation et la vérification de la qualité des données issues des différents sites maladies rares et qui assure de bonnes performances en termes d'exécution de requêtes et de conduite d'études.

1.2 LES DIFFERENTS NIVEAUX D'HETEROGENEITE

L'enquête sur les bases de données conduite par la cellule opérationnelle auprès des centres de référence maladies rares a montré qu'il existe une importante hétérogénéité au niveau de ce paysage numérique. Cette hétérogénéité se manifeste à plusieurs niveaux.

Hétérogénéité technique

Tous ces systèmes existants ont été développés indépendamment les uns des autres pour répondre à des objectifs spécifiques aux sites MR qui les ont conçus. Les choix technologiques ont donc été faits en complète autonomie pour répondre à des besoins locaux. Cette diversité technique comporte entre autres le choix du système de gestion de bases de données ou du système de gestion de fichiers et la manière dont il a été implémenté, le choix de l'architecture (locale ou web), le choix de l'ouverture du système (la mise en place ou non d'une API),…

Hétérogénéité structurelle

Outre l'hétérogénéité technique, les données collectées au sein de ces systèmes sont structurées différemment. Cette structuration dépend notamment de l'objectif du recueil qui a été défini. En effet, pour répondre aux besoins spécifiques des professionnels de santé et mettre en place les fonctionnalités attendues, les concepteurs de ces systèmes organisent l'information dans la base de données d'une manière adaptée à ces besoins. Cette hétérogénéité est donc perceptible au niveau des schémas de données et des éléments qui les composent.

Hétérogénéité des données

En s'intéressant plus en détail aux données et en supposant que les recueils soient comparables et répondent au même objectif, il demeure une hétérogénéité relative au format des données et aux appellations qui sont utilisées pour désigner et coder les données. Dans ce cadre, nous relevons 3 niveaux d'hétérogénéité :

- L'hétérogénéité **syntaxique** qui concerne les dénominations des éléments de données. En effet, un même concept peut être nommé différemment d'une base de données à une autre. Pour désigner le patient par exemple, certains systèmes implémentent la dénomination « patient » d'autres utilisent plutôt « sujet ».
- L'hétérogénéité **des formats** qui concerne le codage des données. Les mêmes données peuvent être codées différemment d'une base de données à une autre. L'exemple le plus courant est l'unité de certaines valeurs numériques telles que le poids qui peut être codé en grammes ou en kilogrammes.
- L'hétérogénéité **sémantique** qui concerne les vraies signifiances des dénominations des données telles qu'elles ont été énoncées par leur concepteur. En effet, derrière deux données portant le même nom et présentes dans deux systèmes différents peut se cacher deux concepts différents. La dénomination « statut du patient » peut faire référence dans un système au statut marital du patient ou au statut médical du patient (porteur sain ou non) dans un autre système.

A toutes ces couches d'hétérogénéité à traiter au niveau de chaque système, s'ajoute le grand nombre des systèmes à intégrer.

1.3 PROPOSITIONS

Dans ce contexte d'hétérogénéité, la BNDMR vise à intégrer les données issues de multiples systèmes utilisés au niveau des sites maladies rares. Nous nous confrontons ainsi à plusieurs problématiques.

Quelles données récupérer ? Il n'est certainement pas judicieux de récupérer l'ensemble des données de ces systèmes puisque l'hétérogénéité sera seulement transposée à un niveau central. Malgré la mise en place de l'identifiant patient MR qui permet le chainage des données, Les données collectées au niveau de chaque site MR concernent des populations de patients

différentes avec un recouvrement potentiel qui reste partiel. L'intégration des données des systèmes existants telles qu'elles sont, constituerait un entrepôt abritant une multitude de données de toutes sortes, non juxtaposables et inexploitables. Les données de la BNDMR doivent être communes à toutes les maladies rares et d'intérêt pour les études qui vont en découler. Nous exposons au début de ce chapitre une méthodologie collaborative de construction d'un set (ensemble) minimal de données pour les maladies rares.

Pour que ce set minimal de données soit correctement compris et recueilli par les professionnels de santé des sites maladies rares, il a été standardisé. Nous avons étudié les différents standards de santé afin de sélectionner les plus adaptés à notre cas d'usage. Cette standardisation comporte non seulement la standardisation des données, celle de leurs jeux de valeurs mais aussi celle des flux échangés. La standardisation a certes ses avantages, mais l'objectif d'interopérabilité reste non atteint lorsque le même standard n'est pas implémenté par tous les systèmes souhaitant communiquer.

A cause de l'hétérogénéité structurelle, les données du set minimal de données ne sont pas modélisées de la même manière dans les différents systèmes sources. Il est donc parfois indispensable de chercher et d'identifier les données qui correspondent au périmètre du set minimal de données. Ainsi, pour chaque base de données, nous devons étudier le recouvrement entre son ensemble de données et celui du set minimal. La recherche de correspondances entre les ensembles de données des systèmes utilisés par les sites maladies rares et le set minimal de données constitue une tâche lourde notamment lorsque ces systèmes sont multiples et lorsque les ensembles de données à explorer sont volumineux. Nous nous sommes donc tournés vers les approches automatisées d'alignement décrites dans la littérature, nous les avons étudié et cherché à améliorer leur utilisation dans le contexte d'alignement de schémas de données hétérogènes.

2 SET MINIMAL DE DONNEES MALADIES RARES ET STANDARDISATION

2.1 SET MINIMAL DE DONNEES MALADIES RARES

2.1.1 *QU'EST-CE QU'UN SET MINIMAL DE DONNEES ?*

2.1.1.1 **Définitions**

Les approches qui proposaient la définition et la construction de sets minimaux de données ont pour objectif de permettre la conduite de larges études dans des domaines déterminés à des coûts contrôlés (Pheby et Etherington 1994; Tilyard et al. 1998; Webster 1998; Bird et Farrar 2008). Un set de données est un ensemble composé d'éléments de données. Dans ses spécifications, ces éléments de données doivent être bien définis ainsi que le contexte et l'objectif de leur recueil, leurs statuts (obligatoire ou pas) et les règles de vérification de leur qualité (les bornes des valeurs possibles par exemple) (AIHW 2016). Le set minimal de données est un ensemble d'éléments de données dont le recueil est, pour la plupart, obligatoire au niveau national. Parmi les éléments de données on distingue les éléments de données communs, qui sont d'ores et déjà existants dans plusieurs recueils, et les éléments de données spécifiques, dans le recueil a été proposé pour répondre à un besoin spécifique tel que les indicateurs de santé publique.

2.1.1.2 **Expériences internationales et multidisciplinaires**

Cette approche a été adoptée par de multiples institutions à travers le monde. En Finlande, c'est pour définir les données du dossier patient national qu'un set minimal de données a été défini pour la prise en charge des patients (Häyrinen et Saranto 2005). Ce set minimal de données comporte entre autres les informations pour l'identification du patient et du professionnel de santé qui le prend en charge, les diagnostics, les examens, les traitements, les facteurs de risque… Aux Etats Unis, un ensemble d'éléments de données communs a été défini pour le cancer pour faciliter la recherche multidisciplinaire et multi-institutionnelle (Winget et al. 2003). En Australie plusieurs sets minimaux de données ont été définis pour diverses applications : pour évaluer les services aux personnes ayant un handicap (« Disability Services National Minimum Data Set (DS NMDS) collection (AIHW) » 2016), pour faciliter la recherche sur le développement de la petite enfance (« Developing the National Early Childhood Development Researchable Data Set (AIHW) » 2016) ou encore pour évaluer les services de traitement de la dépendance à l'alcool ou autres substances (« Alcohol and other drug treatment services in Australia 2013–14 (AIHW) » 2016).

Dans le domaine des maladies rares, cette approche a souvent été utilisée pour l'étude d'une seule maladie ou d'un groupe de maladies et non pas à un niveau national (Jason et al. 2012). Aux états unis l'ORDR (Office of Rare Diseases Research) a développé un ensemble de données

communes pour collecter les données à un niveau national dans l'entrepôt de données GRDR (Forrest et al. 2011).

Toutes ces expériences ne comportent malheureusement pas des descriptions sur la méthodologie adoptée. Obtenir un consensus autour d'un set minimal de données est une tâche complexe qui doit suivre une méthodologie appropriée. Dans ce sens, Svensson-Ronallo et al. proposent une méthodologie globale pour la construction de sets minimaux de données pour la clinique.

2.1.2 *METHODOLOGIE DE CREATION DU SET MINIMAL DE DONNEES MALADIES RARES*

L'approche de création du set minimal de données se décompose en 4 phases :

i. Création des groupes experts
ii. Revue systématique de la littérature
iii. Validation par les experts via une enquête
iv. Validation nationale

2.1.2.1 **Identification des groupes experts**

Plusieurs groupes ont été identifiés composés d'experts et de décideurs afin de rationaliser le processus de création du set minimal de données et arriver à un consensus national rassemblant les cliniciens maladies rares et les autorités publiques. Quatre groupes ont été créés ayant chacun des responsabilités propres :

- Les 131 centres de référence représentés par leurs coordonnateurs ou leurs représentants. Ce groupe était chargé de l'expression des besoins en termes de data management ainsi que de la sélection des éléments de données d'intérêt pour les groupes de maladies rares dont ils ont la charge.
- Un groupe de travail d'expertise nationale composé d'experts en maladies rares, des représentants des agences nationales et du ministère de la santé, des chercheurs et des représentants de l'Inserm. Ce groupe de travail avait pour responsabilité l'étude de la pertinence des éléments de données par rapport aux objectifs spécifiés par le second plan national maladies rares.

- Le comité stratégique du second plan maladies rares, regroupant entre autres des représentants de la DGOS, avait pour rôle la validation et l'approbation du set minimal de données maladies rares.
- La cellule opérationnelle de la BNDMR pour mettre en place les outils nécessaires pour la revue systématique de la littérature, pour déterminer les méthodes statistiques adéquates pour la sélection des éléments de données.

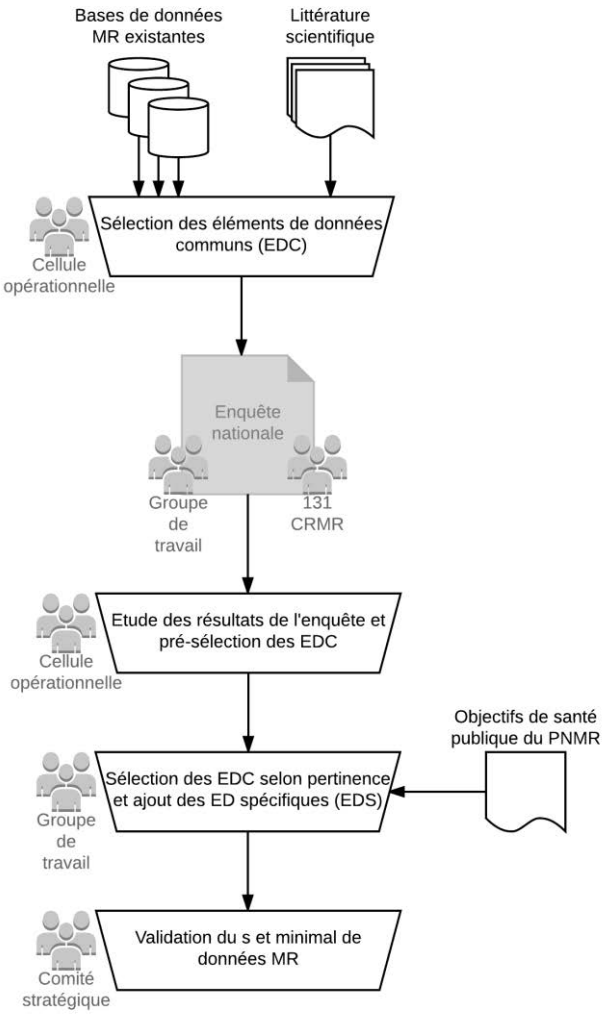

FIGURE 13 METHODOLOGIE DE MISE EN PLACE DU SET MINIMAL DE DONNEES MALADIES RARES

2.1.2.2 Revue systématique de la littérature et première proposition

La version 0 du set minimal de données regroupait des éléments de données issues de catalogues existants, de la littérature scientifique, de projets de recherche existants ainsi que des registres maladies rares existants. Les premiers éléments de données ont donc été identifiés dans

les catalogues de données de systèmes équivalents pour les maladies rares. Parmi ces catalogues nous avons consulté celui de l'ORDR, du projet EpiRare (Taruscio et al. 2014) et celui de CEMARA. Les éléments de données du projet européen ESID ont aussi été récupérés (« ESID - European Society for Immunodeficiencies » 2016). Toutes ces données ont été alignées pour construire un ensemble préliminaire d'éléments de données communs.

En parallèle, une revue systématique de la littérature a permis d'identifier un ensemble d'articles scientifiques relatifs aux sets de données dans le domaine des maladies rares. La revue a été effectuée sur PubMed en construisant requêtes spécifiques (« PubMed - NCBI » 2016). Pour construire les requêtes, l'ensemble des 7000 maladies rares et de leurs groupes, récupérés de l'ontologie OntoOrpha (Dhombres et al. 2011), ont été associés aux mots clés {Data set, Minimum data set, Data catalogue, Data model, Models of data, Common data elements, Data elements}. La recherche a d'abord été effectuée pour les maladies existantes dans MeSH (Medical Subject Headings) (« MeSH - NCBI » 2016), puis directement sur les titres et les résumés des articles pour les autres maladies.

Une liste de 2126 articles scientifiques a été proposée à un expert pour être revue. Plusieurs articles n'ont pas été retenus soit parce qu'ils étaient trop spécifiques soit parce qu'ils étaient hors du champ d'étude. Les articles décrivant des sets minimaux de données sans les méthodes de leur mise en place n'ont pas été retenus non plus. Ainsi le nombre d'articles retenus est passé à seulement six articles (R. J. Buchanan, Wang, et Ju 2002; Robert J. Buchanan et al. 2005; McCormick et al. 2005; O'Donnell et al. 2007; Adelson et al. 2012; Kerr et al. 2001). Les informations issues de ces articles ont permis de valider la qualité du set minimal de données préliminaire.

2.1.2.3 **Enquête nationale auprès des CRMR**

Une enquête nationale auprès des centres de référence a été conduite pour valider les éléments du set minimal de données. Cette enquête a été adressée à 160 personnes, les 131 coordonnateurs des CRMR et les 29 membres du groupe de travail. Cette enquête, préparée sur un outil en ligne, comportait 118 items organisés en 12 sections. La question posée était la même pour tous les items et demandait de valider l'inclusion de l'item dans le set minimal de données, et si c'était le cas, de préciser s'il devrait être obligatoire ou optionnel.

Les résultats de l'enquête ont par la suite été traités. La prise de décision était basée sur des règles quantitatives :

Si un item avait cumulé plus de « non » que de « oui », il était écarté.

Si un item avait cumulé plus de « oui » que de « non », deux options se présentent :

- L'item est directement retenu lorsque le taux de réponses à cet item dépasse la médiane des taux de réponses à toutes les autres questions. Autrement dit, il faut que le taux de participation soit significatif.
- Sinon l'item est soumis à la validation du groupe de travail.

Le pourcentage de participation à l'enquête était de 81%. La médiane des pourcentages de réponses aux questions était à 60% de réponses. Le questionnaire comportait 105 questions relatives à des éléments de données. Parmi ces éléments de données 54 ont été directement validés suite à l'enquête et 30 ont été directement écartés. Les 21 éléments de données restants ont été soumis au groupe de travail. Cette revue par le groupe de travail a aussi remis en discussion certains éléments de données qui ont été validés par les règles quantitatives.

Une étude de sensibilité a permis de voir si les participants les plus actifs et les moins actifs avaient donné des réponses similaires. Ces deux groupes étaient définis à partir des taux de réponses aux questions selon s'ils étaient supérieurs ou inférieurs à la médiane. La distribution des taux de réponses a aussi été étudiée selon l'appartenance des coordonnateurs des CRMR aux 18 groupes de maladies rares (maladies neuromusculaires, anomalies du développement, etc.). Cette distribution était équilibrée et les taux de réponses n'étaient pas affectés par l'appartenance des coordonnateurs à un groupe de maladies rares ou à un autre.

2.1.2.4 **Validation nationale du set minimal de données maladies rares**

L'ensemble des éléments de données issu de la revue de la littérature et validé suite à l'enquête nationale conduite après des CRMR a été proposé au comité stratégique du second PNMR. Les éléments de données ont été révisés selon leur pertinence par rapport aux objectifs du plan. Cette validation finale clos le travail d'identification des éléments du set minimal de données maladies rares.

2.1.3 *RESULTATS*

Le set minimal de données maladies rares est composé de 42 éléments de données communs et de 16 éléments de données spécifiques au contexte national. Ces éléments de données sont répartis en 13 groupes : consentement, identification patient, informations personnelles, informations familiales, statut vital, parcours de soins, activité de soin, histoire de la maladie, diagnostic, confirmation du diagnostic, traitement, informations anté et néo natales et participation à la recherche. L'Annexe 1 : Set Minimal des Données Maladies Rares (v1.09.2) détaille les éléments de données du set minimal et expose les motivations de leur recueil.

Parmi tous ces éléments de données 38 sont assez génériques pour qu'ils puissent être intégrés ou qu'ils soient déjà intégrés dans les DPI des hôpitaux pour la prise en charge générale des patients. Les systèmes de codage peuvent être adaptés d'un pays à l'autre. Par exemple, le diagnostic maladie rare peut être codé selon la terminologie Orphanet, OMIM ou SNOMED-CT en fonction du contexte et de la granularité souhaitée pour le codage.

Le set minimal de données a été comparé au set minimal du GRDR, l'entrepôt maladies rares américain. Un alignement à double sens a été effectué pour étudier le taux de recouvrement des éléments du GRDR par les éléments du set minimal de données de la BNDMR et vice versa. Un recouvrement de 43% des éléments du GRDR par ceux du set minimal de données de la BNDMR et un recouvrement de 33% des éléments du set minimal de données de la BNDMR par ceux du GRDR ont été remarqués (voir Figure 14).

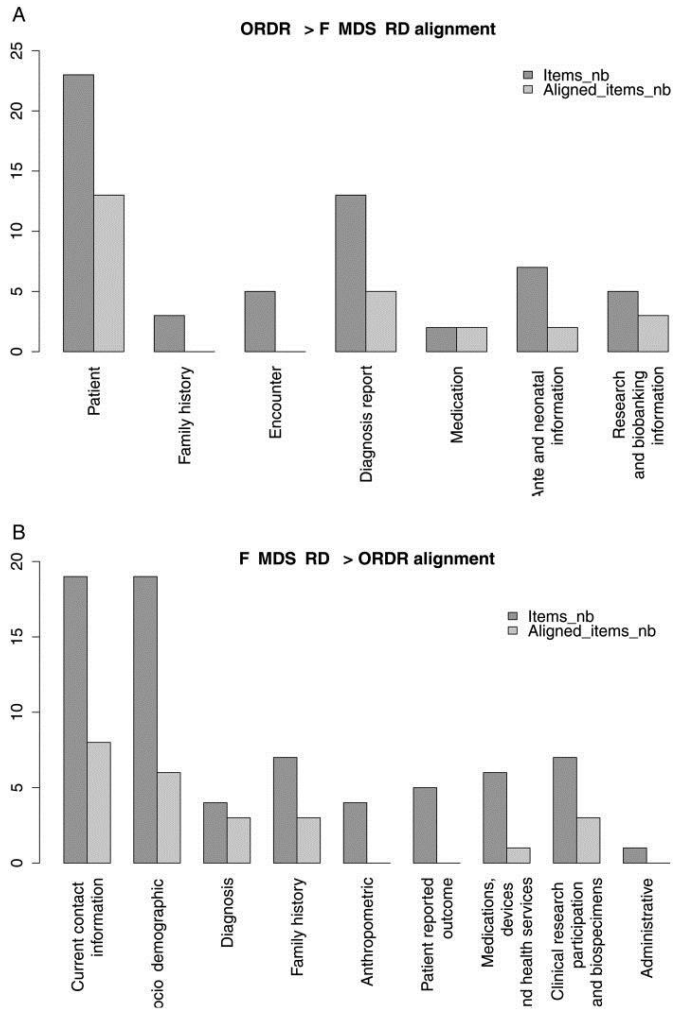

FIGURE 14 ALIGNEMENT ENTRE LES ELEMENTS DU SET MINIMAL DE DONNEES MALADIES RARES ET LES ELEMENTS DE DONNEES COMMUNS DU GRDR (CHOQUET ET AL. 2014)
A : CORRESPONDANCES TROUVEES DANS LE GRDR POUR LES GROUPES DU SET MINIMAL DE DONNEES DE LA BNDMR
B : CORRESPONDANCES TROUVEES DANS LE SET MINIMAL DE DONNEES POUR LES GROUPES DES ELEMENTS DE DONNEES DU GRDR

2.2 UN FORMAT ELECTRONIQUE INTEROPERABLE POUR LE SET MINIMAL DE DONNEES

2.2.1 *LES STANDARDS : DE LA MODELISATION DES DONNEES AUX TERMINOLOGIES SPECIALISEES*

2.2.1.1 **Définitions**

Les standards sont des référentiels, des ensembles de recommandations documentées généralement définis par un consortium ou une association d'industriels dont l'objectif est d'harmoniser l'activité de leur secteur. Les standards sont nommés « de facto standards » en anglais (standards de fait) puisqu'ils s'imposent généralement par leur usage. Les normes sont quant à elles publiées par un organisme de normalisation officiellement reconnu tels que

l'Organisation Internationale de Normalisation (ISO), le Comité Européen de Normalisation (CEN) ou l'Association Française de Normalisation (Afnor).

Dans le domaine de la santé, la standardisation est indispensable pour faire circuler les données au sein du système de santé. Que cela soit dans un objectif de gestion et de sécurisation de processus métier au sein d'un établissement de soin, dans un objectif de reporting ou d'aide à la décision, la donnée partagée doit être « recevable » et « comprise » par le système qui est censée la recevoir. Un standard constitue donc un formatage et une compréhension unique partagée par les systèmes qui l'adoptent et qui s'échangent des données.

"In response to the repeated allusions to data as "the lifeblood of medicine," I've subsequently begun to characterize "standard data" as the "Type-O Blood" of health care." (BobbyG 2016)

Les standards de santé qui sont proposés par les divers organismes interviennent à deux niveaux :

- Modélisation des données : définition des concepts et des jeux de valeurs dont les terminologies
- Modélisation des processus métiers et des flux d'échanges d'informations

Dans ce qui suit nous allons présenter les terminologies de santé séparément des autres standards de santé. Etant donné leur spécificité et l'expertise que leur définition nécessite, ils sont souvent proposés par des organismes spécialisés du domaine et pour des objectifs bien définis.

2.2.1.2 **Les standards de santé**

2.2.1.2.1 HL7

HL7 International (Health Level 7 International), fondée en 1987, est une organisation à but non lucratif qui vise à développer des standards pour l'échange, l'intégration, le partage et la récupération des informations de santé informatisées pour appuyer la pratique clinique et la gestion, la délivrance et l'évaluation des services de soins (HL7 2016). L'organisation est accréditée par l'ANSI (American National Standards Institute), l'institut de normalisation américaine, et compte plus de 1600 membres représentants plus de 50 pays.

HL7 a publié plusieurs versions de ses standards :

- HL7 V2

HL7 Version 2 est un standard de messages pour l'échange des données dans le domaine clinique (« HL7 Version 2 Product Suite » 2016, 7). Il a été originellement créé en 1987 pour supporter les processus au sein d'un établissement de soin. Actuellement, il est de loin le standard le plus implémenté avec, à titre d'exemple, à peu près 95% des établissements de santé américains qui l'implémentent. Les messages sont sous la forme de segments du type :

PID|||PATID1234^5^M11||JONES^WILLIAM^A^III||19610615|M-||C|1200 N ELM

- HL7 V3

HL7 Version 3 est une édition normative pour l'échange de données clinique qui propose une approche dirigée par le modèle (« HL7 Version 3 Product Suite » 2016; Beeler 1998). Il permet de spécifier des messages et des documents cliniques basés sur la syntaxe XML. Le modèle d'information sur lequel est basée la méthodologie est le RIM (Reference Information Model), qui a été proposé comme faisant partie intégrante de HL7 V3 (« Reference Information Model (RIM) » 2016). Ce modèle a fait l'objet de critiques scientifiques notamment à cause de sa complexité (Smith et Ceusters 2006; Schadow, Mead, et Walker 2006).

HL7 V3 inclut les spécifications des documents cliniques à échanger : les CDA (Clinical Document Architecture) (Dolin et al. 2001). Le CDA est un standard de documents à balisage qui spécifie la structure et la sémantique (dans le sens de définitions partagées) d'un document clinique tels que les résumés de sortie et les comptes rendus de consultations…

Un document CDA est composé d'un entête et d'un corps. Le corps du document peut contenir des informations structurées avec des sous-éléments XML ou non structurées tels que une image ou du son. La dernière version est le CDA R2 (Dolin et al. 2006).

```
<ClinicalDocument>
  ... CDA Header ...
  <structuredBody>
    <section>
      <text>(a.k.a. "narrative block")</text>
      <observation>...</observation>
      <substanceAdministration>
        <supply>...</supply>
      </substanceAdministration>
      <observation>
        <externalObservation>...
        </externalObservation>
      </observation>
    </section>
    <section>
        <section>...</section>
    </section>
  </structuredBody>
</ClinicalDocument>
```

FIGURE 15 EXEMPLE D'UNE DOCUMENT CDA AVEC UN CORPS STRUCTURE CONTENANT UN ELEMENT « ADMINISTRATION DE SUBSTANCE » ET UN ELEMENT « OBSERVATION »

- FHIR

FHIR ou Fast Healthcare Interoperability Resources est le dernier standard (encore en draft) d'HL7 (Bender et Sartipi 2013). Il permet la description des formats de données et des API (Application Programming Interface) pour l'échange des données de santé entre les différents dossiers patients informatisés. Inspiré des dernières versions des standards HL7 (v2 et v3), il a été néanmoins conçu pour être rapidement implémentable puisqu'il se base sur les derniers standards et technologies Web : architecture REST, XML et JSON pour la représentation des données, etc… Ses concepteurs aspirent à faciliter l'interopérabilité entre les différents systèmes et à encourager l'apparition de nouvelles solutions multiplateformes.

FHIR est basé sur des composants modulaires appelés ressources (« resources ») qui peuvent être combinés ensemble ou étendus afin de répondre à des besoins spécifiques en termes de collecte de données administratives et cliniques et de gestion de processus de soins. FHIR est rétro-compatible avec HL7 v2 et v3.

Les ressources FHIR se classifient en 6 grandes sections :

- Clinique : Le contenu des documents cliniques
- Identification : Identification des entités impliquées dans le processus de soins
- Workflow : Gestion du processus de soin
- Finances : Ressources gérant la facturation et le paiement des soins
- Conformité : Ressources à la disposition des développeurs et intégrateurs

- Infrastructure : Fonctionnalités générales

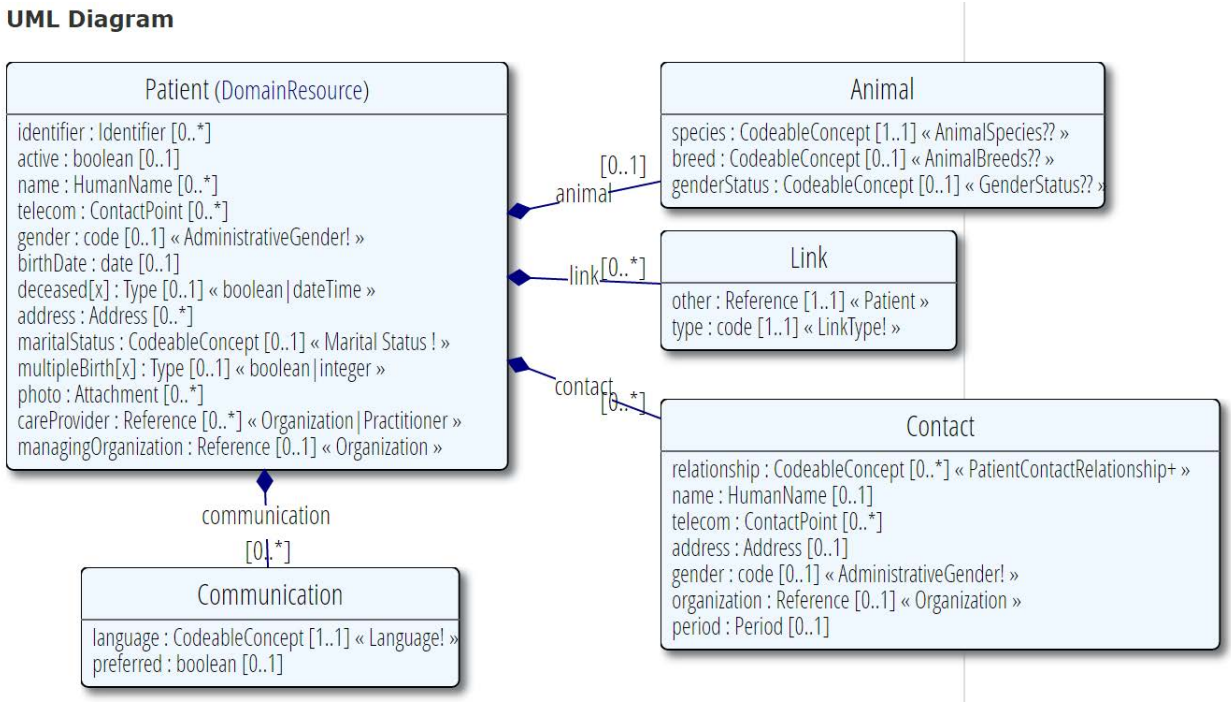

FIGURE 16 EXEMPLE DE LA RESSOURCE PATIENT FAISANT PARTIE DE LA SECTION IDENTIFICATION

2.2.1.2.2 CDISC

CDISC (Clinical Data Interchange Standards Consortium) est une organisation à but non lucratif qui vise à appuyer la collecte, l'échange et l'archivage des données de recherche clinique et de leurs métadonnées (Kuchinke et al. 2009). Depuis sa formation en 1997, le consortium a œuvré à la définition de nouveaux standards qui traitent de la recherche médicale depuis les protocoles jusqu'à l'analyse et la publication des résultats.

CDISC ODM (Operational Data Model) est l'un des plus répandus standards de CDISC qui met à disposition un ensemble de schémas XML qui permettent la constitution d'e-CRFs (electronic Case Report Forms). Un ensemble d'extensions d'ODM ont par la suite été proposés tel que Define-XML pour la définition et l'échange des métadonnées. CDISC propose aussi BRIDG qui est une modélisation du domaine de la recherche clinique permettant de définir des éléments tels que : l'investigateur, le sujet, l'étude ou l'intervention. Cette modélisation est disponible au format UML ou OWL.

2.2.1.3 **Terminologies de santé**

Une terminologie est un ensemble de termes potentiellement définis et organisés qui sont spécifiques d'une science, d'une technique, d'un domaine particulier de l'activité humaine. Les terminologies de santé sont multiples et traitent de divers domaines. Leur mise en place est généralement proposée par des organismes spécialisés pour répondre à des objectifs bien spécifiques. Nous avons dressé une liste des terminologies les plus utilisées et celles d'intérêt dans le contexte des maladies rares (Tableau 6).

TABLEAU 6 LISTE DES TERMINOLOGIES LES PLUS UTILISEES NOTAMMENT DANS LE DOMAINE DES MALADIES RARES

Terminologie	Organisme	Champs d'application	Objectifs
SNOMED-CT (SNOMED Clinical Terms)	IHTSDO (International Health Terminology Standards Development Organization)	Soins hospitaliers et descriptions cliniques	Documentation et annotation pour la coordination des soins
CIM (Classification Internationale des maladies)	OMS (Organisation Mondiale de la Santé)	Maladies de toutes spécialités et autres causes de morbidité	Etudes de morbidité et mortalité
ORPHANET	ORPHANET	Maladies rares et gènes impliqués	Classifications encyclopédiques pour les maladies rares
OMIM (Online Mendelian Inheritance in Man)	McKusick-Nathans Institute of Genetic Medicine Johns Hopkins University School of Medicine	Gènes humains et maladies génétiques	Base de données documentaire et référentiel utilisé dans la littérature scientifique du domaine
LOINC (Logical Observation Identifiers Names and Codes)	Regenstrief Institute	Observations des laboratoires médicaux	Documentation et annotation pour la coordination des soins
HPO (Human Phenotype Ontology)	Hôpital Universitaire de la Charité de Berlin	Anomalies phénotypiques	Vocabulaire standardisé pour la description des anomalies phénotypiques
MeSH (Medical Subject Headings)	NLM (United States National Library of Medicine)	Métadonnées médicales	Nomenclature pour l'indexation des articles scientifiques

2.2.1.4 **IHE : Une initiative pour faciliter l'implémentation des standards**

IHE (Integrating the Healthcare Enterprise) est une initiative lancée par un groupe de professionnels de santé et d'industriels du domaine pour améliorer les méthodes de partage et de transfert des données de santé entre différents systèmes (Bernardini et al. 2002). Cette initiative sélectionne et promeut l'utilisation de certains standards du domaine tels que les standards d'HL7

ou DICOM pour le stockage et le transfert de l'imagerie médicale. Les spécifications fournies par l'IHE sont organisées autour de profils IHE (IHE profiles). Chaque profil définit précisément comment les standards sont implémentés pour répondre à des besoins cliniques précis. Ces profils traitent des situations des domaines suivants :

- Pathologie anatomique
- Cardiologie
- Soins de l'œil
- Infrastructure IT
- Laboratoire
- Pathologie et biologie médicale
- Coordination des soins
- Dispositifs médicaux
- Pharmacie
- Qualité, recherche et santé publique
- Radio-oncologie
- Radiologie

A titre d'exemple, le profil XD-LAB (Sharing Laboratory Reports) aborde le partage des résultats de laboratoire dans un établissement de soins ou entre les membres d'un groupe de professionnels de santé. Les standards impliqués dans la définition du format numérique des résultats de laboratoire dans ce profil sont : HL7 CDA Release 2 pour spécifier la structure du document et LOINC (Logical Observation Identifiers Names and Codes) pour coder les résultats des tests (*cf. section terminologies*).

En France, c'est un groupe de travail de l'association Interop'Santé (Association qui regroupe HL7, HPRIM et IHE) qui est responsable de l'édition de l'annexe française du profil gestion administrative de patient : IHE PAM (Patient Administration Management).

2.2.2 *STANDARDISATION DU SET MINIMAL DE DONNEES MALADIES RARES*

2.2.2.1 **Choix des standards**

Afin de faciliter l'intégration du set minimal de données dans les dossiers patients informatisés et le rendre compatible avec d'autres standards, nous l'avons représenté dans le format FHIR, le dernier standard HL7. Ce choix a été motivé par la simplicité, la modularité et l'extensibilité de FHIR.

Nous sommes conscients que ce standard n'est pas encore assez mature dans sa version actuelle. Cependant le nombre de projets adoptants ce standard ne cesse de s'accroitre (Mandel et al. 2016; Wagholikar et al. 2016; Warner et al. 2016). Il a souvent été désigné, notamment par l'ONC for Health IT (Office of the National Coordinator for Health Information Technology, organisme Américain équivalent à l'ASIP Santé en France), comme la perspective dans les années à venir en terme de standardisation pour l'interopérabilité (Office of the National Coordinator for Health IT 2016).

Par ailleurs, en étudiant les autres versions des standards HL7, nous avons constaté qu'ils étaient très orientés processus de prise en charge des patients et pas aussi extensibles que FHIR. De même pour la plupart des profils IHE qui traitait des cas d'usage de prise en charge classés par spécialité. Un profil IHE avait cependant attiré notre attention qui définissait le processus d'envoi de données par les professionnels de santé vers un entrepôt de santé publique, mais ce profil était spécifique aux registres cancer (« Physician Reporting to a Public Health Repository – Cancer Registry - IHE Wiki » 2016). Dans l'approche de construction d'un ensemble de données pour la recherche, CDISC était bien adapté mais il ne permettait pas l'interopérabilité avec le système de soins puisqu'il était spécifique au domaine de la recherche clinique.

En ce qui concerne la standardisation des jeux de valeurs, nous avons en priorité cherché des correspondances dans les jeux de valeurs proposés par FHIR. Le cas échéant, nous avons utilisé des codes externes de LOINC et MeSH pour essayer de trouver malgré tout une référence externe claire et partagée. Des terminologies spécialisées ont aussi été choisies par les experts pour coder certains items, tel que Orphanet pour les maladies rares ou HPO pour les signes.

2.2.2.2 **Standardisation des éléments de données**

Nous avons donc entrepris de reconstruire le set minimal de données pour les maladies rares à partir des éléments proposés par FHIR. Nous avons pu retrouver des équivalences pour certains

éléments du MDS, pour d'autres nous étions contraints de définir des extensions des ressources auxquelles ils appartenaient. Pour d'autres, il était plus approprié de définir des questionnaires indépendants (voir Annexe 2 : Standardisation HL7 FHIR du set minimal de données maladies rares).

Nous avons sélectionné 5 ressources majeures parmi les ressources proposées pas le standard FHIR :

- La ressource patient (« Patient ») pour rassembler les données administratives et démographiques des patients
- La ressource histoire familiale (« Family History ») pour regrouper les données contextuelles relatives aux membres de la famille
- La ressource rencontre (« Encounter ») pour regrouper les informations sur les activités de soin dont le patient a bénéficié
- La ressource état (« Condition ») pour regrouper les données concernant le diagnostic, son histoire et son mode de confirmation
- La ressource médicament (« Medication ») pour regrouper les informations relatives au traitement maladie rare en cours.

Nous avons rajouté par la suite deux ressources questionnaires pour regrouper les questions anté et néonatales et les questions de recherche.

Les résultats de cette standardisation nous montrent que 42% d'éléments du set minimal de données ont été directement retrouvés dans les éléments prédéfinis de FHIR. Les ressources qui s'en sortent le mieux sont les ressources administratives : Patient et activité de soins (Tableau 7[1]).

[1] Dans sa version électronique à standardiser, le set minimal contient 62 éléments de données, et non 58 tel qu'il a été initialement défini. Cela est dû au rajout de 4 éléments de données de précision.

TABLEAU 7 DISTRIBUTION DES ELEMENTS DE DONNEES ET DES EXTENSIONS CREEES SELON LES RESSOURCES FHIR SELECTIONNEES

Ressource	Nombre total d'éléments	Nombre d'extensions
"Patient"	21	7
"Family history"	4	3
"Condition"	14	9
"Medication"	2	1
"Encounter"	8	2
"Questionnaire"	13	13

2.2.2.3 **Standardisation des jeux de valeurs**

Nous avons par ailleurs étudié la concordance entre les types et les jeux de valeurs que nous avons défini pour le set minimal de données et ceux proposés pour les éléments FHIR. Selon la notation FHIR, un élément est considéré comme un concept codable (« codeable concept ») lorsqu'il prend ses valeurs d'un système de codage public, tel que LOINC, ou d'un système de codage interne défini au sein d'un jeu de valeurs FHIR. Nous avons donc aligné les jeux de valeurs définis pour le set minimal de données avec les jeux de valeurs FHIR en priorité. Ceux pour lesquels nous n'avons pas trouvé de correspondances, des jeux de valeurs internes ont été créés tout en spécifiant les correspondances avec les éléments de systèmes de codage externes quand elles existent. 17% des jeux de valeurs restent exclusivement internes. Le Tableau 8 détaille les concepts codables et la standardisation de leurs jeux de valeurs (voir aussi Annexe 3 : Jeux de valeurs et standardisation).

TABLEAU 8 LISTE DES ELEMENTS DE DONNEES CODABLES ET LEURS JEUX DE VALEURS OU TERMINOLOGIES DE REFERENCE

Elément de donnée	Jeux de valeurs FHIR	Standard externe (terminologie)
Sexe du patient	✔	-
Pays	-	code ISO 3166-1 alpha-2
Communes	-	code INSEE
Cause du décès	-	CIM10
Lien de parenté avec le propositus	✔	LOINC
Contexte et objectif de l'activité	-	MeSH
Profession du participant à l'activité	-	Nomenclature des emplois hospitaliers
Age au diagnostic/aux premiers signes	-	-

Signes	-	HPO, CIM10
Appréciation du diagnostic à l'entrée	-	LOINC
Statut du diagnostic	✔	LOINC
Diagnostic	-	Orphanet
Mode de confirmation du diagnostic	-	MeSH
Méthode de confirmation	-	-
Hérédité	-	-
Médicament	-	EMA: Orphan drugs

2.3 DISCUSSION

Le second PNMR, a encouragé la mise en place d'un set minimal de données national pour les maladies rares pour tous les CRMR et CCMR. L'objectif était d'homogénéiser et de standardiser la collecte des données tout en évitant la multiple saisie pour les professionnels de santé. Le set minimal de données regroupe des éléments de données communs ainsi que des éléments de données spécifiques définis pour répondre aux objectifs du plan. Ce set minimal de données a été validé au niveau national et spécifie un recueil national obligatoire pour les maladies rares.

Le set minimal de données regroupe moins d'éléments que le GRDR. Contrairement à ce dernier, il ne comporte pas les informations démographiques, les informations de contact ou de qualité de vie, ou la possibilité de transplantation. En revanche, le set minimal de données recueille plus d'informations sur les structures de soins, sur les activités de soins, sur l'histoire et la confirmation du diagnostic et comporte aussi les données anté et néo-natales. Cela est en lien direct avec les objectifs du PNMR 2 en ce qui concerne l'évaluation de l'offre et de la demande de soins.

Cette harmonisation du set minimal de données à un niveau national est importante pour les futures coopérations avec les registres de recherche. Le maintien des alignements à jour entre les différentes terminologies de codage constitue aussi une condition sine qua non. Orphanet par exemple aligne ses concepts avec ceux de SNOMED-CT, OMIM et la CIM. Cependant, le recouvrement n'est pas total et reste difficile à maintenir. Sur cette problématique, d'importants efforts sont requis pour une meilleure gestion et harmonisation entre ces différentes terminologies et systèmes de codage.

La standardisation du set minimal de données est indispensable pour avoir un référentiel non seulement national mais aussi international. La standardisation permet, en effet, d'avoir des concepts avec des définitions stables et partagées et facilite par ailleurs l'interopérabilité. Un bon nombre d'éléments de données étaient présents dans les ressources FHIR. La création d'extension était cependant nécessaire pour d'autres. Nous avons cherché à expliquer cela, les principales raisons sont :

- Le format du set minimal de données se rapproche beaucoup du format d'un questionnaire. Plusieurs éléments de données demandés sont sous la forme de questions auxquelles on répond par « oui » ou « non », notamment dans les sections anté et néo-natales et la section participation à la recherche.
- Le set minimal de données comporte des éléments de données spécifiques répondant aux objectifs du PNMR. Cela explique pourquoi certains jeux de valeurs sont d'ores et déjà agrégés et s'inscrivent directement dans une approche épidémiologique. L'exemple des dates aux premiers signes et au diagnostic est très parlant avec un jeu de valeurs {anténatal, à la naissance, postnatal, je ne sais pas} ou bien la question pour savoir si le décès du patient était dû à la maladie rare ou non.

L'intégration effective du set minimal de données dans les recueils initiaux, tels que les DPI, les registres ou les autres bases de données, nécessite encore quelques efforts. Afin d'éviter la double saisie aux professionnels de santé, il est nécessaire de considérer et d'étudier les recueils qui sont d'ores et déjà mis en place au niveau des sites maladies rares. D'abord, il faut étudier le recouvrement de ces recueils avec le set minimal de données, ensuite il faut les encourager à intégrer les données manquantes lorsque le recouvrement n'est pas suffisant. Ces études de recouvrement consistent en l'alignement des schémas de données de ces différents systèmes avec le set minimal de données.

3 ALIGNEMENT ET DECOUVERTE DE CORRESPONDANCES

3.1 VERS L'AUTOMATISATION DES ALIGNEMENTS

3.1.1 *CONTEXTE*

Que cela soit dans un objectif de chargement d'entrepôt de données, de médiation de schémas de données ou d'échanges simple de données, le recours à l'alignement est souvent indispensable. L'alignement consiste en la recherche de correspondances entre les différents éléments de deux ensembles hétérogènes ou plus. Ces ensembles d'éléments peuvent être des terminologies, des ontologies ou des schémas de base de données. Il suffit que le même standard ne soit pas utilisé par deux systèmes souhaitant communiquer pour avoir recours à l'alignement de données. Cette tâche, souvent très lourde lorsque les ensembles traités sont volumineux, a longtemps été faite manuellement.

Avec la multiplication des systèmes informatisés et la croissance de la volumétrie de leurs données, l'automatisation des alignements est devenue indispensable et a fait émerger diverses techniques d'alignement automatisées ou semi-automatisées dans divers domaines d'application.

3.1.2 *PROBLEMATIQUE SEMANTIQUE*

Parmi les premiers dossiers que nous avions à étudier figurait celui de l'échange éventuel de données avec la Banque Nationale Alzheimer (Le Duff et al. 2010). Ce projet qui est équivalent à la BNDMR mais dans le domaine des maladies neurodégénératives a été lancé dans le cadre du Plan National Alzheimer 2008-2012. La BNA collecte des données démographiques, cliniques et diagnostiques recueillies par les Centres Mémoire et les Centres Mémoire de Ressources et de Recherche à des fins épidémiologiques. Certaines maladies diagnostiquées au niveau de ces centres sont des formes rares de maladies neurodégénératives et ces unités de soin portent souvent une « double casquette » en tant que Centre de Référence Maladies Rares et Centre Mémoire. Les données de certains patients maladies rares sont donc déjà collectées au niveau de la BNA d'où l'intérêt d'une interconnexion avec la BNDMR.

Nous avons donc aligné le Corpus d'Information Minimum maladie d'Alzheimer (CIMA) de la BNA et le MDS de la BNDMR. Suite à un alignement manuel, nous avons obtenu moins de 50% de recouvrement. L'étude des résultats a permis d'identifier plusieurs types de correspondances liant les différents éléments du CIMA et du MDS. Le premier type englobe les correspondances exactes entre les éléments :

- **Correspondance exacte** : Les éléments sont équivalents et leurs codages correspondent parfaitement.

Ex : L'élément « nom de naissance » est lié à l'élément « nom patronymique » sans transformation de codage.

Les autres types de correspondances que nous avons pu identifier illustrent bien la problématique sémantique à laquelle nous nous confrontons lorsque nous opérons un alignement. Ainsi nous avons identifié plusieurs types distincts de correspondances :

- **Correspondance partielle** : les éléments sont équivalents par leurs définitions mais leurs codages ne correspondent que partiellement. Cette différence est due à la différence des besoins en termes de collecte d'information dans chaque domaine d'application.

Ex : L'élément « patient envoyé par » est lié à l'élément « patient adressé par » mais leurs listes de codage se correspondent pas complètement. L'élément « Patient envoyé par » de la CIMA contient, entre autres, {Médecin généraliste, Neurologue, Psychiatre…} alors que l'élément « patient adressé par » du set minimal de données contient notamment {Venu de lui-même, Association de patient, Généraliste…}.

- **Correspondance conditionnée** : les éléments sont équivalents lorsqu'une certaine condition est vérifiée.

Ex : L'élément « nom marital » est lié à l'élément « nom d'usage » seulement lorsque ce dernier est différent du « nom de naissance ».

- **Agrégation** : Deux ou plusieurs éléments d'un ensemble sont liés à un élément du deuxième ensemble.

Ex : Les éléments « code département de naissance » et « code commune de naissance » sont agrégés pour donner l'élément « code pays de naissance ».

- **Eclatement** : un élément d'un ensemble est lié à deux ou plusieurs éléments du deuxième ensemble.

Ex : L'élément « type de l'acte » est lié aux trois éléments de la cible « contexte de l'activité », « objectif de l'activité » et « profession du personnel réalisant l'activité ».

La problématique sémantique rend les travaux d'alignement encore plus complexes. Une connaissance profonde des deux ensembles à aligner, de leurs contextes et de leurs objectifs devient nécessaire. De plus, la recherche de correspondances ne se limite plus à la recherche d'équivalences entre les éléments mais toutes les pistes qui aideraient à inférer des correspondances doivent être explorées.

3.1.3 *CLASSIFICATIONS DES TECHNIQUES D'ALIGNEMENT AUTOMATISE*

Différentes classifications des techniques d'alignement automatisé ont été proposées dans la littérature (Rahm et Bernstein 2001) (Euzenat et Shvaiko 2013). Toutefois, nous pouvons identifier aisément les grandes classes de ces techniques.

- Techniques linguistiques :

La première approche consiste à comparer à un niveau linguistique les éléments des différents schémas. Que cela soit par détection d'une égalité entre les chaines ou les sous-chaines de caractères ou par mesure de similarité plus complexe entre les libellés des éléments et de leurs descriptions. Cette approche est souvent enrichie par l'utilisation de ressources externes telles que des dictionnaires pour reconnaître les synonymes ou des bases de données enregistrant les anciennes correspondances.

- Techniques basées sur la structure :

La deuxième approche opère à un niveau structurel. Elle compare les combinaisons des éléments qui donnent des structures complexes. Par exemple, si deux éléments ont les mêmes sous-classes, ils pourraient être mis en correspondance.

- Techniques basées sur les contraintes :

La troisième approche est basée sur les contraintes qui définissent par exemple les types de données ou les domaines de valeurs. Combinée à d'autres approches, cette technique permet de détecter les correspondances erronées ou de confirmer celles qui sont correctes, mais elle reste

incapable de détecter des correspondances par elle-même. En effet, deux éléments de données peuvent avoir le même type et la même plage de valeurs sans pour autant faire référence à un même concept ; tel est le cas de la date de naissance d'un patient et de la date de l'acte médical dont il a bénéficié.

- Techniques basées sur les instances :

La quatrième approche est basée sur les instances. Elle est particulièrement utile lorsque des données semi structurées sont traitées et que l'information sur la structure des schémas n'est pas suffisante. Par exemple, la récurrence de l'instance « Cystinose » dans chacun des éléments « Maladie » et « Diag » de deux schémas de données différents peut inférer une correspondance entre les deux.

Chacune de ces techniques est plus ou moins adaptée à un type de données. En utilisant une technique basée sur les instances par exemple, on admet implicitement que les domaines de valeurs des éléments de données que nous alignons sont similaires (des domaines de valeurs basés sur la même référence ou partageant certains éléments de valeurs comme la nomenclature Orphanet et OMIM). Par ailleurs, plus le domaine de valeurs est grand plus il devient difficile de détecter les correspondances. Par exemple, il est difficile pour ces outils de détecter une similarité entre des éléments de données contenant des identifiants patients puisqu'il n'y a pas assez de redondance pour un identifiant donné. Les techniques basées sur les instances peuvent aussi inférer des correspondances erronées lorsque les éléments traités contiennent des valeurs booléennes.

Diverses approches et stratégies d'alignement ont par la suite été proposées. Des outils hybrides faisant intervenir ces différentes techniques ont été mis en place pour répondre à des besoins plus ou moins spécifiques d'alignement de données. Les études comparatives qui ont été conduites montrent que la technique linguistique est la principale technique intégrée à ces outils (Rahm et Bernstein 2001). Par ailleurs, la majorité de ces outils supporte principalement l'alignement des ontologies et non des schémas de bases de données classiques tels que les schémas des bases relationnelles ou les schémas XML.

3.2 TESTS AVEC UN OUTIL D'ALIGNEMENT

3.2.1 *CHOIX DE L'OUTIL D'ALIGNEMENT*

Dans nos diverses expérimentations nous avons utilisé l'outil OnAGUI (Ontology Alignment Graphical User Interface) Open Source développé en 2009 (Mazuel et Charlet, s. d.). Il permet la visualisation des ontologies et de leur alignement et de les éditer. Il permet surtout d'effectuer des alignements automatisés d'ontologies (formalisées SKOS ou OWL) basées sur des techniques linguistiques. L'outil propose donc l'utilisation d'algorithmes de mesure de similarité entre les différents concepts telle que la mesure de la distance de Levenshtein.

La distance de Levenshtein est utilisée pour mesurer la différence entre deux chaines de caractères. Il s'agit du nombre minimal d'opérations à effectuer pour transformer la première chaine de caractère en la deuxième ou inversement. Ces opérations sont :

- La substitution d'un caractère par un autre : aller – allée
- L'ajout d'un caractère : allé – allée
- La suppression d'un caractère : allée - allé

Pour effectuer un alignement entre les deux sources de données, l'utilisateur est invité à fixer un seuil de similarité au-dessous duquel les résultats d'alignement ne seraient pas affichés. Ce seuil est compris entre une 0 (les deux concepts sont complétement différents) et 1 (les deux concepts sont équivalents).

Nous avons fait le choix de cet outil parce qu'il était bien adapté à nos besoins et que nous n'avions pas besoin d'outil plus complexe. Cet outil proposait des techniques de mesures de similarité entre les chaines de caractères ; les techniques les plus adoptées et les plus utilisées pour effectuer des alignements. Par ailleurs, nous n'avions pas besoin d'autres techniques qui mesureraient les distances sémantiques entre les concepts puisque la totalité de nos bases de test ne sont pas formalisées sous la forme d'ontologies.

3.2.2 *TESTS SUR UN ALIGNEMENT DE TERMINOLOGIES*

3.2.2.1 **Alignement entre deux versions d'une même terminologie**

La terminologie Orphanet a été utilisée depuis le lancement de CEMARA pour le codage des diagnostics maladies rares dans l'application. Durant le deuxième trimestre 2013 et dans la perspective de la migration des données de CEMARA vers BaMaRa, une analyse des écarts entre la version de CEMARA (datant de 2007) et la dernière version de la terminologie devait être opérée pour lancer les travaux de redressement du codage diagnostique. Cette mise au point était nécessaire pour pouvoir correctement réattribuer les codes Orpha adéquats aux patients dont les fiches allaient migrer.

En 5 ans, la terminologie Orphanet avait en effet évolué de deux manières complètement décorrélées. Suite aux découvertes de nouvelles maladies et/ou aux découvertes de nouveaux liens entre les maladies existantes, les équipes d'Orphanet et leurs partenaires ont essayé de reporter chaque nouvelle évolution des connaissances dans le domaine des maladies rares. Ainsi, la terminologie a été (et est toujours) en perpétuelle évolution avec une moyenne de 20 nouveaux concepts par mois. Par ailleurs, certains concepts ont été supprimés et d'autres modifiés. Ces suppressions et modifications ont été le plus souvent dues à des réorganisations des arborescences du thésaurus. Dans CEMARA, des concepts ont aussi été ajoutés à la demande des utilisateurs. Ces concepts correspondaient soit à des nouveaux concepts d'Orphanet mais dont les codes Orpha correspondants n'avaient pas été renseignés, soit à des concepts spécifiques qui répondaient directement aux besoins des utilisateurs.

Nous avons effectué un alignement à trois niveaux entre la dernière version d'Orphanet et celle qui est utilisée dans CEMARA. Dans un premier temps, le travail a consisté à chercher des correspondances entre les identifiants, appelés « pat_id », qui sont proposés en interne par Orphanet afin de répertorier les maladies. Les « pat_id » avaient été intégrés dans CEMARA avec leur définition de l'époque, lorsque la terminologie avait été mise en place dans l'application. Ensuite, une recherche de correspondances sur les codes Orpha, appelés aussi « numéros Orpha », a été effectuée. Sur ces deux premiers niveaux, il s'agissait d'une recherche d'égalité simple entre les numéros. Au troisième niveau, ce sont les libellés qui ont été alignés grâce à l'outil OnAGUI. La recherche de correspondances se faisait non seulement sur les libellés principaux des maladies mais aussi sur l'ensemble de leurs synonymes.

D'un point de vue technique, une transformation des tables de CEMARA contenant la terminologie ainsi que des fichiers d'Orphanet au format OWL a été nécessaire. La méthodologie et les résultats sont reportés sur le schéma ci-dessous.

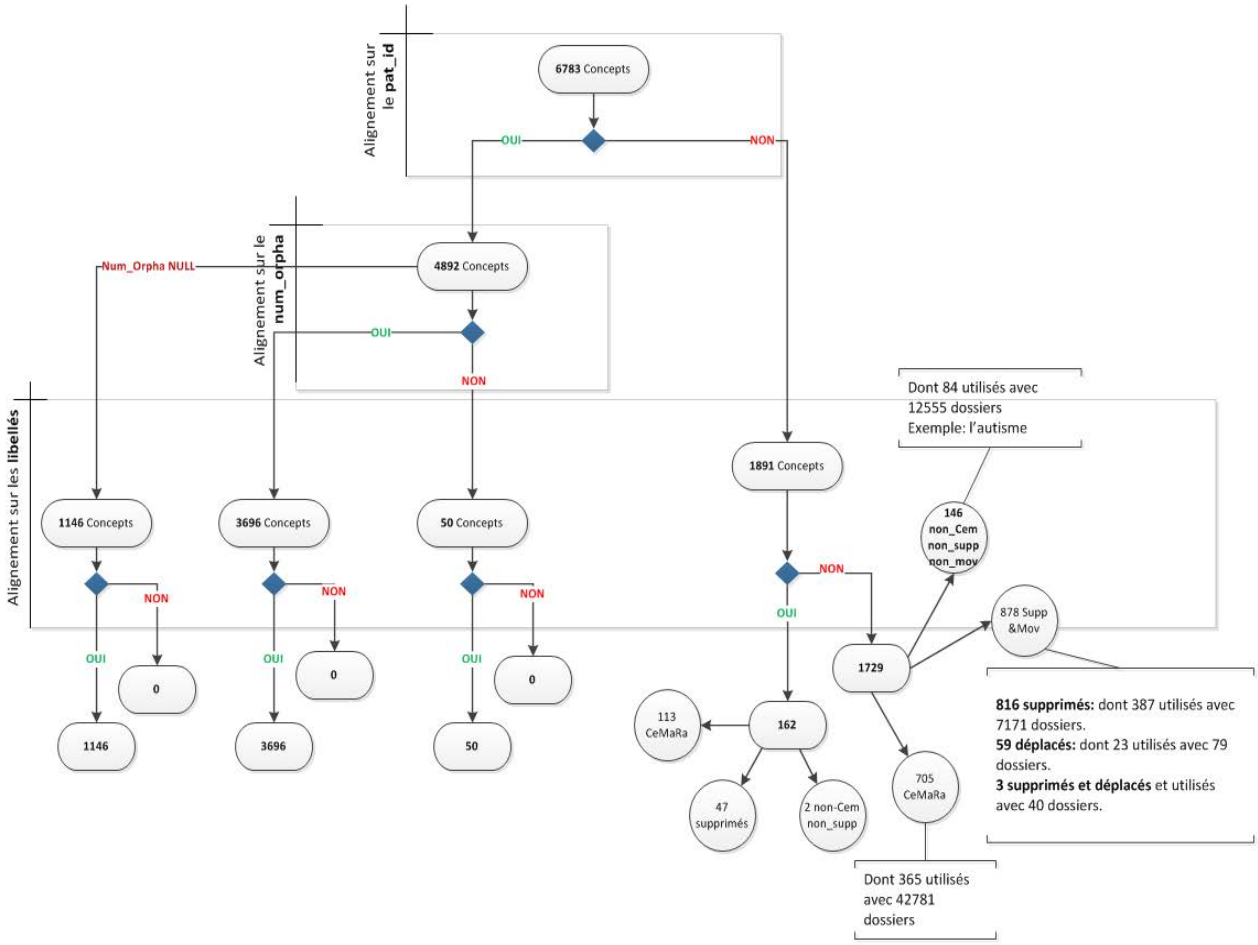

FIGURE 17 L'ALIGNEMENT A TROIS NIVEAUX ENTRE LES 2 VERSIONS D'ORPHANET

Comme le montre le schéma, utiliser un outil d'alignement a été nécessaire pour opérer un alignement au niveau des libellés répondant à deux objectifs : d'une part, valider les correspondances qui ont été découvertes auparavant au niveau des identifiants et des numéros Orpha (3696 concepts), d'autre part, rechercher des correspondances pour les concepts qui n'ont pas été alignés sur les identifiants et les codes Orpha.

Ainsi la terminologie Orphanet a pu être mise à jour dans CEMARA tout en assurant une remise à niveau du codage diagnostique dans les fiches patients. Ceci a été possible selon trois actions :

- attribuer des codes Orpha adéquats aux concepts qui n'en avaient pas et en réattribuer à des concepts qui en avaient mais qui étaient devenus obsolètes ;
- attribuer les codes Orpha de substitution, indiqués dans les fichiers d'Orphanet, pour les concepts qui avaient été supprimés d'Orphanet ou qui avaient été déplacés ;
- enfin, attribuer les codes Orpha des concepts les plus proches pour remplacer les codes de ceux qui n'ont pas pu être alignés.

3.2.2.2 **Alignement entre deux terminologies différentes**

A la demande du groupe national des foetopathologistes, qui a participé à la mise en place de l'identifiant pour les fœtus, nous avons aligné la terminologie Orphanet avec la codification des lésions développée par l'Association du Développement de l'Informatique en Cytologie et en Anatomie Pathologique. Avec le même outil OnAGUI nous avons réussi à aligner 245 concepts parmi les 737 concepts d'ADICAP qui nous ont été transmis, ce qui donne 33% de termes alignés. Nous nous attendions à ce qu'il n'y ait pas un grand recouvrement entre les deux terminologies puisque l'une est dédiée uniquement aux maladies rares et l'autre aux champs de la pathologie en général.

La plupart des correspondances ont été trouvées au sein du chapitre « Anomalie rare du développement embryonnaire » d'Orphanet. Cependant, certaines correspondances n'ont pas été détectées à cause d'une différence d'organisation et de granularité entre les différentes terminologies. Comme nous pouvons le voir dans la Figure 18, certains concepts très proches ne peuvent être mis en correspondance à cause d'une différence de granularité dans les terminologies. Ce type de correspondances peut être détecté lorsque le seuil de la mesure de similarité utilisée est abaissé.

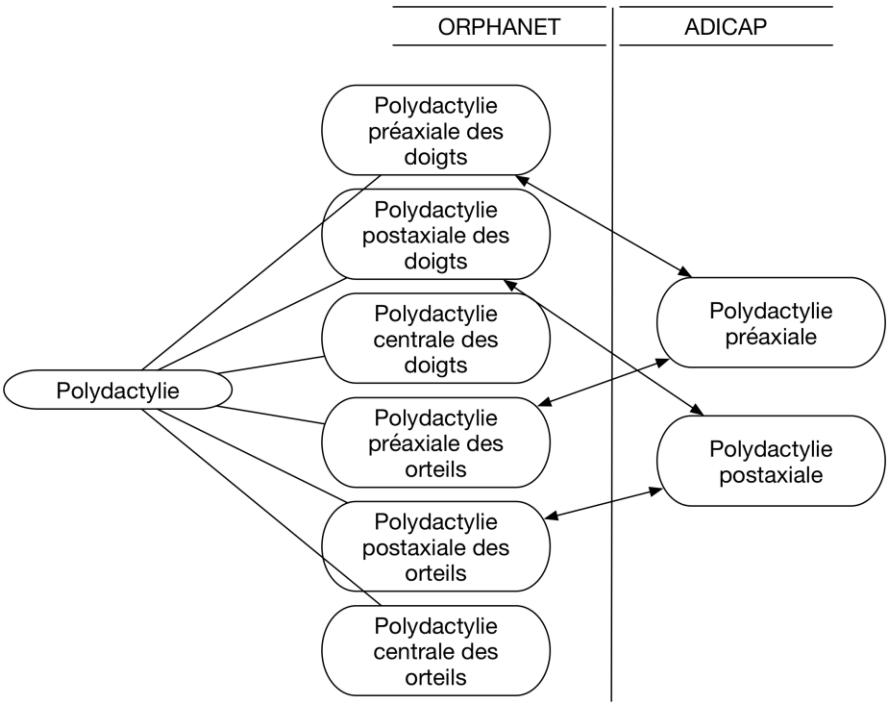

FIGURE 18 EXEMPLE SUR LA DIFFERENCE DE GRANULARITE ENTRE LES TERMINOLOGIES ORPHANET ET ADICAP

3.2.3 *TEST SUR UN ALIGNEMENT DES SCHEMAS DE DONNEES*

3.2.3.1 **Méthode**

La migration de la base de données CEMARA est la première priorité du projet BNDMR. En effet, CEMARA est considérée comme étant la base nationale historique des maladies rares et sa notoriété rend sa disparition inenvisageable. Le premier alignement de schémas de données que nous avions à opérer était donc un alignement entre le schéma de données de CEMARA et celui de BaMaRa.

Pour cet alignement nous avons utilisé l'outil OnAGUI présenté plus haut et nous l'avons associé à un enrichissement sémantique des schémas de données de CEMARA et de BaMaRa en ajoutant des synonymes et des traductions en anglais des éléments. Nous avons appliqué l'algorithme de Levenshtein avec un seuil de similarité fixé à 0.8, seuil au-dessous duquel le nombre de faux positifs dépasse le nombre de vrais positifs.

Comme format de représentation des schémas de données nous avons utilisé le format XSD (XML Schema Definition) qui est basé sur la syntaxe hiérarchique XML. Il permet de définir les

éléments d'un schéma en précisant leurs noms et leurs types et en leur attribuant des annotations en différentes langues.

Exemple :

```
<xs:element name="deceasedDatetime" type="xs:date">
    <xs:annotation>
        <xs:documentation xml:lang="fr">date de décès du patient</xs:documentation>
        <xs:documentation xml:lang="en">patient's date of death</xs:documentation>
    </xs:annotation>
</xs:element>
```

Parmi les données du tronc commun de CEMARA, nous n'avons récupéré que celles qui sont relatives aux patients, les données des annuaires et les tables de liaison n'étaient pas d'intérêt pour nous et ne rentraient pas dans le scope du set minimal de données. A partir de ces tables nous avons généré un fichier XSD. De la même manière, pour le schéma de données de BaMaRa, nous avons créé une version XSD du set minimal de données. Ces schémas XSD ont été par la suite enrichis avec les synonymes et les définitions.

L'outil OnAGUI, n'acceptant que le format SKOS ou OWL, nous avons mis en place des transformations XSLT pour transformer ces schémas XSD au format OWL.

3.2.3.2 **Résultats**

Nous avons construit un gold standard en effectuant un alignement manuel entre les deux schémas de données de CEMARA et de BaMaRa. Les correspondances issues de l'alignement automatisé ont donc été validées par comparaison aux correspondances du gold standard. Nous avons aussi pu identifier les fausses correspondances et les correspondances manquées. Le tableau ci-dessous récapitule l'évaluation des résultats.

TABLEAU 9 RESULTATS DE L'ALIGNEMENT AUTOMATISE DES SCHEMAS DE DONNEES DE CEMARA ET DE BAMARA

Alignement CEMARA-BaMaRa Total : 68 = 62 items MDS + 6 nœuds parents		**Gold Standard**	
		Alignés	Non alignés
Résultats de l'alignement	Alignés (33)	18	15
	Non alignés (35)	19	16

Parmi les correspondances qui n'ont pas été détectées nous retrouvons par exemple « le mode de confirmation du diagnostic » et « l'objectif de l'activité ». Ces informations ne sont pas codées de la même manière dans les différentes bases de données. Dans CEMARA, plusieurs éléments de données de type booléen ont été définis pour coder l'information, par exemple les éléments « confirmation clinique » et « confirmation par imagerie » sont codables en « oui » ou « non ». Dans BaMaRa, l'élément « mode de confirmation du diagnostic » est codable avec une liste de termes spécifiant le mode.

Parmi les correspondances erronées nous retrouvons les éléments « statut patient » de CEMARA et « statut diagnostic » de BaMaRa où la similarité syntaxique a induit une fausse correspondance.

3.2.4 *CONCLUSION DES TESTS*

Utiliser un outil d'alignement intégrant une technique linguistique, comme OnAGUI, était une approche adaptée pour la recherche de correspondances entre les éléments de différentes terminologies tel que l'alignement entre les 2 versions d'Orphanet ou l'alignement entre Orphanet et ADICAP. Le premier objectif était d'obtenir des résultats assez rapidement et cet objectif a été atteint. La découverte automatique de correspondances a permis de traiter en un clic 75% de la terminologie Orphanet pour sa mise à jour dans CEMARA par exemple. La présentation des résultats était adaptée au cas d'usage avec une liste de couples de concepts et les mesures de similarité qui leur ont été attribuées. Une validation manuelle était tout de même nécessaire pour traiter les propositions de correspondances qui étaient proches du seuil minimal de similarité. Par ailleurs, l'intervention d'expert était aussi nécessaire pour traiter les concepts qui n'ont pas été mis en correspondance par l'outil afin de leur trouver les éléments de substitution les plus proches. Cette tâche qui consiste à trouver des liens sémantiques hormis les équivalences exactes aurait pu être allégée avec l'utilisation dans un deuxième temps d'une autre technique telle qu'une technique basée sur la structure.

Dans le cas d'alignement des schémas de données, cette approche n'était pas la plus adaptée. D'abord les résultats étaient présentés dans un format que nous estimons non adapté à l'alignement de schémas de données. En effet, les correspondances proposées le sont dans un format simple liant deux éléments avec une certaine mesure de similarité. Concrètement, si nous

devons préparer un envoi de données d'une base vers une autre, devons-nous faire un simple *copier-coller* pour récupérer la donnée telle quelle, ou tenir compte d'une différence de codage?

Par ailleurs, certaines correspondances complexes de type éclatement/agrégation liant des éléments booléens de CEMARA à un élément multi-valué de BaMaRa par exemple, n'ont pas été détectées. Nous pouvons donner deux raisons à cela car :

- Ni les types des éléments, ni leurs jeux de valeurs, n'étaient pris en compte ;
- Les liens sémantiques hors équivalences entre les éléments ne pouvaient être pris en compte avec des descriptions de schémas de données aussi minimalistes tel que les schémas XSD.

Ce dernier point n'est pas un levier sur lequel nous pouvons agir. En effet, nous cherchons à faciliter et à minimiser l'intervention des partenaires. La demande de création d'ontologies, sur lesquelles les techniques d'alignement automatisées seraient sémantiquement plus efficaces, pour les schémas de données va à l'encontre de cet objectif. Nous ne pouvons qu'imposer un minimum de formalisation et de structuration des schémas de données. Dans la section suivante nous proposons une nouvelle approche agissant sur les autres leviers pour optimiser l'utilisation des techniques d'alignement automatisé et les incorporant dans un processus complet d'intégration de données.

3.3 PROPOSITION : ALIGNEMENT DES SCHEMAS DE DONNEES ET PROCESSUS D'INTEGRATION DE DONNEES

3.3.1 *BESOINS*

3.3.1.1 **Des correspondances mieux définies**

Tel qu'expliqué dans la section précédente, l'objectif des outils d'alignement automatisé consiste souvent en l'alignement d'ensembles de concepts. En utilisant certaines mesures et méthodes, ils détectent des couples de concepts et leur assignent des niveaux de similarité. Les

correspondances qui résultent de ces alignements sont sous la forme suivante : $\{C_1, C_2, s(C_1,C_2)\}$[1], par exemple *{Cystinose, Cystine, 0.8}*.

Ce type de résultats n'est pas suffisant pour définir toutes les transformations qui devraient être opérées au sein d'un processus d'intégration de données. En effet, dans ces outils, il n'est pas tenu compte des valeurs que prennent les éléments de données. Certes elles sont utilisées dans les techniques d'alignement basées sur les instances où les valeurs sont comparées, mais ceci est réalisé seulement dans le but de détecter des correspondances entre leurs éléments de données. Elles ne sont pas intégrées dans la définition des correspondances dans aucun des outils décrits dans la littérature qui a été revue. Cela est cependant primordial afin de définir les transformations nécessaires pour adapter les données au format de la cible et procéder directement au transfert. Ces transformations incluent notamment:

- La détection de correspondances entre les jeux de valeurs qu'ils soient des listes localement prédéfinies ou des thésaurus externes.
- Des opérations arithmétiques pour se conformer à un changement d'unités par exemple.
- Des concaténations simples sur des chaines de caractères.

En effet, les scripts ou les processus qui sont mis en place au sein des outils ETL ou EAI pour la migration ou le transfert de données, sont basés sur des procédures qui transforment les données de la source en données de la cible. Des valeurs sont affectées aux données de la base cible selon des conditions qui dépendent des valeurs prises par les données de la base source. Notre objectif a donc été de définir d'une manière plus fine les correspondances en prenant en compte les valeurs prises par les éléments de données, ceci afin que celles-ci puissent être directement exploitables et intégrables aux scripts et procédures de transfert et d'intégration de données.

3.3.1.2 **Processus d'intégration adaptés aux données**

Les auteurs des différentes études comparatives d'outils d'alignement automatisé s'accordent à dire que l'efficacité de ces techniques dépend des types de données traitées et du contexte

[1] s : similarité

d'application. Nous avons donc essayé de repenser les méthodes d'intégration de données afin d'optimiser le rôle des techniques d'alignement automatisé dans la détection des correspondances.

Ces processus d'intégration de données devraient se baser sur une analyse préalable des données en se basant sur les descriptions des schémas de données des bases source afin d'identifier les éléments de données, leurs types et leurs valeurs. Ils devraient intégrer des stratégies, des processus et des algorithmes impliquant une ou plusieurs techniques d'alignement de schémas, à mettre en œuvre pour chaque groupe de données homogènes créés, en se basant sur les types de données et la nature des domaines de valeurs. En sortie, ces processus d'intégration de données devraient générer des correspondances finement définies dans le format que nous proposons dans la section suivante.

3.3.2 *FORMALISATION DES CORRESPONDANCES*

3.3.2.1 **Le couple donnée/valeur**

En informatique, la donnée est souvent définie en tant que couple élément de donnée/élément de valeur où l'élément de donnée est le conteneur et l'élément de valeur est le contenu. En bases de données relationnelles par exemple, l'élément de donnée est représenté par la colonne et un élément de valeur est représenté par l'instance de cet élément de donnée sur une ligne.

Selon la norme ISO/IEC 11179-3 définissant les registres de métadonnées, le terme élément de donnée signifie une unité de donnée élémentaire qui a les attributs suivants:

Un identifiant

L'identifiant peut être le nom de l'élément (« Nom de naissance »), un code inspiré par le nom de l'élément (« NomNaiss ») ou un identifiant non signifiant (« DE_02 »).

Une définition

Une définition claire et des synonymes pour mieux comprendre le contenu de cette donnée.

Un type

Un des types de données conventionnels définit les valeurs que peuvent prendre la donnée et les opérateurs qui peuvent lui être appliqués :

- Type « String » : chaine ou séquence de caractères
- Type « Integer » : un entier signé
- Type « Float » : nombre réel écrit en virgule flottante
- Type « Boolean » : les valeurs « vrai » ou « faux » (« true » et « false »)
- Type « Date » : c'est un type composé permettant de représenter une date selon divers formats internationaux.
- Type « Enuméré » : Les valeurs prises font partie d'un jeu de valeurs interne ou d'une terminologie externe. Ces jeux de valeurs ou terminologies peuvent être codés ou pas.

Un domaine de valeurs

Le domaine de valeurs restreint la donnée à ne prendre que certaines valeurs. Pour une donnée de type énuméré son domaine de valeurs est restreint aux éléments de valeurs qui ont été définis dans le jeu de valeurs. Par exemple, le sexe du patient prend ses valeurs dans le jeu de valeurs {féminin, masculin, indéterminé, inconnu}. Un domaine de valeurs d'une donnée de type entier peut être défini par un intervalle. Par exemple, le périmètre crânien à la naissance devrait avoir une valeur comprise entre 30 et 40 cm. Les domaines de valeurs peuvent aussi être soit finis (un ensemble fini de valeurs), soit infinis (un ensemble infini de valeurs).

Les éléments de valeurs sont, quant à eux, des instances d'un élément de donnée qui appartiennent au domaine de valeurs de cet élément et qui respectent son type. Pour l'élément de donné « sexe du patient », « féminin » est un élément de valeur faisant partie de son domaine de valeurs.

Comme indiqué précédemment, dans le cadre d'alignement de schémas de données il est insuffisant de définir une correspondance seulement au niveau des éléments de données. En effet une correspondance doit inclure les éléments de valeur puisqu'entre deux éléments de données équivalents il peut y avoir une différence entre les types et/ou les domaines de valeurs. De plus, deux éléments de données non équivalents peuvent avoir des liens au niveau de leurs éléments de valeurs.

Notation : Nous noterons « E » les éléments de données et « e » les éléments de valeurs.

3.3.2.2 **Les règles de correspondances**

Une correspondance n'est pas une bijection, elle a un sens. Cela est dû, entre autres, à la différence de niveau de granularité des deux domaines de valeurs des éléments de données qui sont mis en correspondance. Par exemple, parmi les données d'un ensemble A, il existe un élément de donnée appelé « mode de confirmation du diagnostic » contenant un élément de valeur « cytogénétique » tandis que pour un ensemble de données B on retrouve un élément de donnée appelé aussi « mode de confirmation du diagnostic » contenant un élément de valeur noté « génétique ». Pour cet exemple, nous pouvons définir une correspondance spécifiant que lorsque le « mode de confirmation du diagnostic » est « cytogénétique » dans l'ensemble A, nous pouvons déduire que dans le set B « mode de confirmation du diagnostic » est « génétique ». La réciproque n'est pas forcément vraie, il s'agit d'une implication et non d'une équivalence. Il est donc essentiel de définir l'ensemble source et l'ensemble cible.

Une correspondance peut donc être écrite sous la forme d'une règle :

***SI** hypothèse **ALORS** conclusion*

Les hypothèses décrivent les valeurs prises par les éléments de données de l'ensemble source et les conclusions décrivent les valeurs qui doivent être prises, en conséquence, par les éléments de données de l'ensemble cible.

Notation : Nous noterons S l'ensemble de données source (« Source » en anglais), et T l'ensemble de données cible (« Target » en anglais).

Une correspondance (ou un mapping) n'est pas un simple lien d'équivalence entre deux éléments de données. C'est une relation complexe décrivant les transformations nécessaires pour le passage d'une donnée d'un ensemble source vers un ensemble cible.

3.3.2.3 **Formalisation des correspondances : notation**

Dans ce qui suit nous proposons une formalisation des correspondances qui prend en compte le niveau élément de donnée, le niveau élément de valeur, et la relation exacte liant les éléments de la source et de la cible.

Soient :

Le set de données source

$$S = \{ E^S_i ; i=1..n ; n=card(S) \}^1$$

Les n E^S_i sont les éléments de données du set de données source.

Le set de données cible

$$T = \{ E^T_j ; j=1..m ; m=card(T) \}$$

Les m E^T_j sont les éléments de données du set de données cible.

Les éléments de valeurs source

e^S_{ik} sont les éléments de valeurs de l'élément de donnée E^S_i

Si le domaine de valeurs de E^S_i est fini, l'ensemble de ses éléments de valeurs est le suivant :

$$\{ e^S_{ik} ; k=1..p ; p=card(E^S_i) \}$$

Si le domaine de valeurs de E^S_i est infini, l'ensemble des éléments de valeurs e^S_{ik} est aussi infini ; nous noterons e^S_{ik} pour désigner une instance donnée.

Les éléments de valeurs cible

e^T_{jl} sont les éléments de valeurs de E^T_j

Si le domaine de valeurs de E^T_j est fini, l'ensemble de ses éléments de valeurs est le suivant :

[1] card : cardinal

$$\{ e^T_{jl}; l=1..q ; q=card(E^T_j) \}$$

Si le domaine de valeurs de E^T_j est infini, l'ensemble des éléments de valeurs e^T_{jl} est aussi infini, nous noterons e^T_{jl} pour désigner une instance donnée.

E^S_i peut ainsi être lié à E^T_j par une ou plusieurs relations binaires e^S_{ik}-e^T_{jl}. Chaque relation binaire e^S_{ik}-e^T_{jl} est définie par une règle R de l'ensemble de départ S vers l'ensemble d'arrivée T. Une correspondance est donc définie pour chaque paire d'éléments de valeurs et non pour chaque paire d'éléments de données.

Pour résumer, une correspondance de *S* à *T* peut être caractérisée par le triplet

$$\{ E^S_i\text{-}E^T_j ; e^S_{ik}\text{-}e^T_{jl} ; R_{S->T} \}$$

- Une relation binaire E^S_i-E^T_j liant l'élément de donnée source à l'élément de donnée cible.
- Une relation binaire e^S_{ik}-e^T_{jl} liant un élément de valeur source de E^S_i à un élément de valeur cible de E^T_j.
- Une règle exprimée dans le format « si… alors… ».

TABLEAU 10 EXEMPLES DE CORRESPONDANCES FORMALISEES

$\mathbf{E^S_i - E^T_j}$	$\mathbf{e^S_{ik} - e^T_{jl}}$	$\mathbf{R_{S->T}}$
"décès" – "statut vital"	"oui" - "oui"	si $\mathbf{e^S_{ik}}$ ="oui" alors $\mathbf{e^T_{jl}}$ ="oui"
"décès" – "statut vital"	"non" - "non"	si $\mathbf{e^S_{ik}}$ ="non" alors $\mathbf{e^T_{jl}}$ ="non"
"venu CPC" – "patient adressé par"	"oui" - "CPC"	si $\mathbf{e^S_{ik}}$ ="oui" alors $\mathbf{e^T_{jl}}$ ="CPC"
"nom d'usage" – "nom marital"	string – string	si $\mathbf{e^S_{ik}} \neq \mathbf{e^S_{ck}}$ alors $\mathbf{e^T_{jl}} = \mathbf{e^S_{ik}}$ (E^S_c= "nom de naissance")

3.3.3 *LES PROCESSUS SPECIFIQUES D'INTEGRATION DE DONNEES*

3.3.3.1 **Présentation générale**

Afin d'optimiser l'utilisation des techniques d'alignement automatisé dans la cadre d'alignement de schémas de données, nous pensons qu'il est nécessaire de repenser les processus d'intégration

de données. Comme nous l'avons signalé dans les sections précédentes, l'efficacité de ces techniques d'alignement dépend des types de données et des domaines d'application. Nous proposons la création de processus spécifiques proposant des traitements adaptés à chaque type de données. Ces processus auront comme données en entrée les éléments de données et de valeurs des schémas de données source et cible. En sortie, ils traiteront chaque groupe homogène de données d'une manière spécifique et génèreront des correspondances bien définies respectant la formalisation décrite dans la section précédente.

Les outils automatisés d'alignement interviennent dans la deuxième étape d'un processus d'intégration de données à 4 phases :

- Première phase : Sélection des données

Comme décrit dans plusieurs études, certaines approches d'alignement sont considérées comme appropriées ou non appropriées en fonction des types de données traitées. Cette première phase vise à ne sélectionner que les données dont les types sont adressés dans le processus en cours : les données numériques, les chaines de caractères, les booléens, les listes et les énumérations. Cette sélection peut être plus restrictive en prenant en compte quelques contraintes telles qu'une longueur des chaines de caractères à ne pas dépasser, les limites et les unités de certaines données numériques. Il est aussi important de spécifier le contexte et le domaine de l'étude en cours afin d'adapter les ressources externes qui vont être utilisées : terminologies et dictionnaires.

- Deuxième phase : détection des correspondances

A cette étape, les outils d'alignement automatisé sont utilisés d'une manière adaptée aux données sélectionnées durant la première phase. Le processus peut faire appel à un seul outil d'alignement ou à plusieurs avec un cheminement des données plus ou moins complexe. Afin d'enrichir sémantiquement les schémas de données, des ressources externes peuvent être sollicitées pour détecter des synonymes des éléments de données, par exemple.

- Troisième phase : validation des correspondances

Certes, l'application de certaines approches d'alignement spécifiquement à un ensemble de données présélectionnées augmente la fiabilité des correspondances générées. Cependant, une

validation humaine reste tout de même nécessaire pour ne pas affecter la qualité des données dans la base de données cible en y injectant des données issues de fausses correspondances. Cette validation est généralement assurée par des personnes familiarisées avec les schémas de données source et cible.

- Quatrième phase : génération du code

Une fois les correspondances validées, elles sont traduites et intégrées au programme assurant la transformation des données et leur export vers la base de données cible. Ces correspondances devraient être sous un format compatible avec les langages de programmation.

3.3.3.2 **Preuve de concept**

3.3.3.2.1 Description du processus

Nous avons testé cette méthodologie de définition de processus spécifiques en utilisant comme schéma de données source le schéma de CEMARA. Nous présentons ci-dessous la description, sur chaque phase, de la preuve de concept que nous avons réalisée.

- Sélection des données

Le premier processus que nous avons défini permet de détecter des correspondances simples entre les éléments de la source et de la cible pour les données de type booléen et énuméré (type énuméré avec jeux de valeurs localement défini). Parmi les éléments de données de la source, 39 sont de type booléen et 15 sont de type énuméré. Le schéma de données cible contient quant à lui 16 données booléennes et 15 éléments de données de type énuméré.

Pour une donnée booléenne, ne pouvant contenir que les valeurs « vrai » ou « faux », considérer le niveau élément de valeur ne rajoute pas d'information. Par exemple, pour l'élément de donnée « Le patient est décédé », il suffit de considérer le niveau élément de donnée (nom de l'élément et sa définition) pour essayer de détecter une correspondance. En effet, inclure « Le patient est décédé » dans une correspondance revient à l'inclure avec son élément de valeur « vrai ». Même si à ce stade, les éléments de valeurs ne sont pas considérés pour les données de type booléen, ils seront intégrés dans la formalisation des correspondances durant la phase finale.

Par ailleurs, il est primordial de considérer le niveau élément de valeur pour les données de type énuméré. En effet, nous avons vu parmi les exemples présentés dans les sections précédentes que certains éléments de valeur peuvent être mis en correspondance alors que leurs éléments de données étaient sémantiquement éloignés. Par exemple, l'élément de donnée source « Type d'acte » et l'élément de donnée cible « Profession du participant » représentent deux concepts différents, cependant leurs éléments de valeurs respectifs « Intervention infirmier » et « Infirmier » sont très liés. Ainsi, en considérant le niveau élément de valeur pour le type énuméré, le nombre d'éléments à aligner passe de 15 éléments de données à 86 éléments de valeurs pour la base de données source CEMARA et de 15 éléments de données à 106 éléments de valeurs au niveau de la base de données cible BNDMR.

- Détection des correspondances

Nous avons opté pour un alignement croisé entre les données de type booléen et de type énuméré. En effet, une même information peut être représentée différemment d'un schéma à l'autre et nous avons remarqué, par expérience, que les données de type énuméré, et plus spécifiquement les listes à choix multiples, sont souvent représentées sous la forme de plusieurs données booléennes en base et inversement.

Nous avons utilisé les mêmes outils d'alignement que ceux utilisés pour le test de la section 3.2.3 qui représente notre test de référence. Deux sortes d'alignements ont ainsi été effectuées :

- Alignements opérant au même niveau : liant seulement des éléments de donnée entre eux ou seulement des éléments de valeur entre eux ;
- Alignement opérant à des niveaux différents : liant des éléments de données d'un côté à des éléments de valeurs de l'autre.

Cet alignement croisé dépend des types des données à aligner comme le montre la figure ci-dessous.

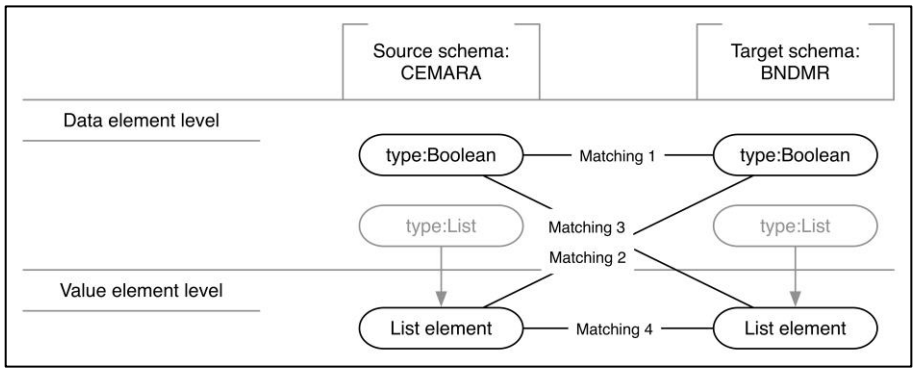

FIGURE 19 ALIGNEMENT CROISE ENTRE LES ELEMENTS DE DONNEES DE TYPE BOOLEEN ET LES ELEMENTS DE VALEURS DE TYPE ENUMERE

- Validation des correspondances

Afin de valider le résultat de cet alignement automatisé, nous avons effectué au préalable un alignement manuel que nous avons fait valider par deux personnes familiarisées avec les bases de données source et cible : le responsable d'exploitation des données de CEMARA et le médecin initiateur des travaux sur le minimum data set de la BNDMR. C'est en comparant les correspondances issues de l'alignement croisé automatisé à cet alignement manuel fiable que nous avons pu valider les résultats. Ces résultats et leur validation sont disponibles dans la section Résultats.

- Génération du code

Quatre types de correspondances ont pu être générés par ce processus spécifique liant les éléments de valeurs des types booléen et énuméré.

TABLEAU 11 LES EXPRESSIONS DE CORRESPONDANCES GENEREES

	Éléments source		Éléments cible		Correspondances détectées
	Type	Niveau	Type	Niveau	
Alignement 1	Booléen	DE*	Booléen	DE	Si E^S_i=true alors E^T_j=true
Alignement 2	Enuméré	VE*	Booléen	DE	Si E^S_i= e^S_{ik} alors E^T_j=true
Alignement 3	Booléen	DE	Enuméré	VE	Si E^S_i=true alors E^T_j=e^T_{jl}
Alignement 4	Enuméré	VE	Enuméré	VE	Si E^S_i= e^S_{ik} alors E^T_j=e^T_{jl}

* DE: data element, VE: value element

Parmi les exemples des correspondances générées :

- SI PropLink = propositus [source] ALORS Propositus = true [cible]
- SI ConfCyto = true [source] ALORS ConfirmationMode = cytogenetic [cible]

3.3.3.2.2 Résultats

Le comptage a été effectué par rapport au nombre d'éléments de la cible, l'objectif étant de peupler la base de données de BaMaRa:

- Nombre de vrais positifs : Combien d'éléments de la cible ont été correctement mis en correspondance ?
- Nombre de faux positifs : Combien d'éléments de la cible ont été inclus dans de fausses correspondances ?
- Nombre de faux négatifs : Combien d'éléments de la cible n'ont pas été mis en correspondance alors qu'ils le devaient ?
- Nombre de vrais négatifs : Combien d'éléments de la cible n'ont correctement pas été mis en correspondance ?

		Gold standard	
		Vrai	Faux
Test	Positif	Vrais positifs VP	Faux positifs FP
	Négatif	Faux négatifs FN	Vrais négatifs VN

Rappel = VP / (VP + FN) : (mesure utilisée en informatique équivalente à la mesure de sensibilité en épidémiologie) sur toutes les correspondances possibles combien ont été détectées par le test ? Un faible rappel reflète les mauvaises performances de l'algorithme à trouver des correspondances.

Précision = VP / (VP + FP) : sur toutes les correspondances détectées par le test combien sont correctes ? Une faible précision reflète une grande proportion de fausses correspondances.

Dans l'Alignement 1 et l'Alignement 2, l'algorithme est utilisé pour chercher les correspondances des 16 éléments de données cible de type booléen avec les éléments de la source. Dans

l'Alignement 3 et l'Alignement 4, l'algorithme est utilisé pour chercher les correspondances des 106 éléments de valeur cible de type énuméré avec les éléments de la source. Le tableau ci-dessous résume les résultats obtenus.

TABLEAU 12 RESULTATS DES QUATRE ALIGNEMENTS ET VALIDATION

	VP	FP	FN	VN	Total	Rappel	Précision
Al_1	3	0	4	9	16	0,43	1,00
Al_2	1	0	0	15	16	1,00	1,00
Al_3	22	4	4	76	106	0,85	0,85
Al_4	35	8	25	38	106	0,58	0,81

Nous avons comparé les résultats obtenus par cette expérimentation aux résultats obtenus par le test de référence de la section 3.2.3. Dans le tableau ci-après nous comparons le nombre de bonnes correspondances détectées par chaque test. Afin que les résultats soient comparables, nous avons sélectionné parmi les résultats du test de référence seulement ceux qui intégraient les données des types booléen et énuméré. Il faut noter que dans le test de référence seul le niveau élément de donnée est pris en compte. Le comptage est toujours fait par rapport au nombre d'éléments de données dans le test de référence alors qu'il est fait par rapport au nombre d'éléments de valeurs dans l'alignement 3 et 4 de l'expérimentation.

TABLEAU 13 EVALUATION COMPARATIVE: NOMBRE DE BONNES CORRESPONDANCES DETECTEES PAR LE TEST DE REFERENCE ET PAR L'EXPERIMENTATION

	Test de référence	Expérimentation
Al_1	3	3
Al_2	0	1
Al_3	1 (2 manquées)	22
Al_4	6 (2 manquées)	35

Nous remarquons que grâce à la prise en compte du niveau élément de valeur dans notre expérimentation, plus de correspondances ont pu être détectées. Par ailleurs, ces correspondances sont formalisées et sont plus complètes comparées à celles qui sont issues du test de référence.

3.4 DISCUSSION

Quel que soit le contexte d'intégration de données, l'alignement est la phase initiale nécessaire à la mise en place des traductions des vocabulaires et des transformations de données. La grande volumétrie des bases des données à aligner et leur grand nombre nous poussent à automatiser ce lourd processus.

Des techniques d'alignement automatisé ont été proposées dans la littérature. L'efficacité de ces techniques dépend fortement des types de données et peuvent relever des défis sémantiques seulement dans le cas où on aligne des ontologies. La construction d'une ontologie est un travail complexe (qui peut constituer le sujet d'une thèse) qui demande une collaboration étroite entre les experts métier et les experts en ingénierie des connaissances. Etant donné le nombre important de systèmes source, notamment des applications de spécialité, construire une ontologie pour chacun de ces recueils constitue un travail compliqué, un effort disproportionné par rapport aux objectifs que l'on souhaite atteindre. Comment donc améliorer les résultats de ces techniques d'alignement pour des schémas de données simples de type XSD par exemple ?

Par ailleurs, le type des correspondances généralement proposées par ces outils est simple et décrit des équivalences, plus ou moins exactes, avec une mesure de similarité. Ce type de résultat est adapté à l'alignement de terminologies et non à l'alignement des schémas de données.

Nos objectifs sont les suivants :

- Générer des correspondances finement définies adaptées aux schémas de données pouvant être directement exploitables ;
- Optimiser l'utilisation des différentes techniques d'alignement dans les processus d'intégration de données avec considération des types des éléments de données.

La formalisation des correspondances que nous proposons est basée sur une description à deux niveaux des relations entre les éléments de deux schémas de données à aligner. Il s'agit d'une formalisation en triplets : paire d'éléments de données, paire d'éléments de valeurs et règle.

Par ailleurs, nous proposons une nouvelle méthode de modélisation des processus d'intégration de données. Les processus définis selon cette méthode optimisent l'utilisation des techniques d'alignement automatisé en les adaptant aux types des données et génère des correspondances respectant la formalisation que nous avons définie précédemment.

Il est à noter qu'adopter cette nouvelle méthodologie n'améliorera pas l'efficacité des techniques d'alignement appliquées aux données. Nous n'introduisons pas non plus une nouvelle technique d'alignement. Notre proposition vise à optimiser l'utilisation des techniques d'alignement existantes en les adaptant au contexte d'alignement de schémas de données. La prise en compte du niveau des éléments de valeurs lors de l'alignement et dans la formalisation des correspondances est une approche indispensable dans le cadre d'intégration de bases de données.

Dans la pratique, aligner les schémas en utilisant des approches automatisées reste une tâche supervisée, non seulement pour valider les similarités proposées mais aussi pour choisir le « bon » résultat. En effet, les utilisateurs ont l'habitude d'appliquer les différentes approches à toutes les données, de comparer les résultats et de pondérer selon les spécificités de leurs données. La méthodologie que nous proposons permet la réutilisation des processus ayant prouvé leur efficacité pour un groupe spécifique de données, et introduit une certaine confiance dans les travaux qui ont été réalisés ainsi qu'une confiance dans les résultats proposés.

4 ARTICULATION ENTRE RECUEIL STANDARDISE ET ALIGNEMENTS

Définir un recueil national consensuel, le standardiser et accompagner les professionnels de santé pour l'implémenter dans leurs systèmes telle est notre approche globale.

Le set minimal de données est un ensemble de données commun à toutes les maladies rares. Il a été proposé et mis en place selon une méthode collaborative entre les différents experts du domaine et initié par la prise en compte de l'existant et des diverses expériences décrites dans la littérature. Ce recueil a été validé au niveau national et est obligatoire pour tous les sites de soins faisant partie des centres de référence ou des centres de compétences maladies rares. Cette méthodologie a permis de simplifier la participation de ces sites à la BNDMR puisqu'elle a défini un recueil minimal et commun.

Pour définir le format technique du set minimal de données et le rendre interopérable nous l'avons standardisé. Les standards internationaux sont des références qui définissent les données, leur modélisation et leurs échanges. La plupart des standards n'étaient pas adaptés à notre contexte, i.e. l'alimentation d'un entrepôt épidémiologique national pour les maladies rares. De

plus ils étaient plutôt orientés processus de prise en charge dans le cadre du soin ou orienté recherche clinique. Nous avons tout de même pu standardiser le recueil à l'aide d'une sélection de ressource FHIR d'HL7. Cette standardisation a permis de fixer les définitions et les formats des éléments du set minimal de données selon un référentiel international. Elle permet aussi une certaine compatibilité avec les autres standards d'HL7, très implémentés dans les systèmes d'information hospitaliers. La standardisation et le choix de ce standard constituent une approche sur le long terme visant à intégrer le set minimal de données dans le DPI de l'hôpital, au plus près du patient lors de sa prise en charge par le professionnel de santé.

Afin d'éviter la double saisie aux professionnels de santé, et de réutiliser au mieux les recueils existants dans les divers systèmes utilisés au niveau des sites maladies rares, il est indispensable d'étudier le recouvrement de ces recueil avec le set minimal de données standardisé. L'alignement et la découverte de correspondances permettent sur le court terme la mise en place des multiples connecteurs pour l'alimentation de la BNDMR via les différents registres et base de données maladies rares.

CHAPITRE IV : APPORTS ET PERSPECTIVES

1 UNE APPROCHE GLOBALE POUR L'INTEGRATION DES DONNEES MALADIES RARES

1.1 COMPLEMENTARITE AU SEIN DU CADRE D'INTEROPERABILITE MALADIES RARES

Le projet BNDMR vise à mettre en place un entrepôt national pour l'épidémiologie des maladies rares. Pour alimenter cet entrepôt, et permettre des études épidémiologiques sur un recueil exhaustif, nous avons mis en place un cadre d'interopérabilité qui permet de faciliter l'interconnexion avec les systèmes qui sont utilisés au niveau des sites maladies rares. Outre la mise en place technique, ce cadre d'interopérabilité aborde deux problématiques majeures : l'identification des patients et l'interopérabilité des données médicales. L'une ne va pas sans l'autre lorsqu'il s'agit d'interopérabilité et de partage de données de patients.

Intégrer des bases de données de livres nécessite un système unique d'identification des livres. Pour ce domaine-là, la solution a déjà été mise en place : le numéro international normalisé du livre ou ISBN (International Standard Book Number). Devant la non généralisation de l'utilisation d'un tel numéro pour les personnes ou les patients, et devant la complexité du système d'identification des patients au niveau national, nous avons cherché à mettre en place un système simple à implémenter pour l'identification des patients dans le réseau maladies rares. Cet identifiant est construit à partir de données patients nominatives stables et discriminantes à l'aide d'un algorithme de hachage qui permet de garantir l'anonymat de l'identifiant. Cet identifiant a aussi été défini pour les fœtus, qui représentent une proportion non négligeable de la population des patients atteints de maladies rares. Cet identifiant assure trois fonctions principales :

- Alimenter correctement les fiches patients dans BaMaRa : En introduisant moins d'erreurs et de variabilité que les données nominatives en texte libre, il permet, en association avec d'autres éléments tels que les identifiants patients locaux des systèmes source, de correctement attribuer les données transmises aux patients appropriés. Dans le cadre su soin, l'IdMR ne peut

assurer seul cette fonction d'identifiant de soin tel qu'il a été discuté dans la conclusion du chapitre II.

- Minimiser les biais dus aux doublons dans les études épidémiologiques de la BNDMR : Dans l'entrepôt, et contrairement à BaMaRa où la politique des droits d'accès est basée sur les sites, les données patients sont décloisonnées et la totalité des données nationales sont interrogeables. L'identifiant que nous proposons permet la fédération des identités des patients visant à limiter le nombre de doublons intersites dans la banque nationale. Cet objectif s'inscrit dans une stratégie globale d'identitovigilance, associant d'autres approches de gestion de qualité des données, pour le traitement *a posteriori* des doublons créés dans la banque nationale. Sur les données de CEMARA, l'algorithme de l'identifiant a permis de détecter 1771 doublons d'identité.
- Permettre d'autres projets de recherche : L'identifiant pourrait constituer une clé de chainage permettant de rapprocher les données de la BNDMR avec d'autres collections de données issues de registres nationaux.

L'identifiant patient est essentiel pour faciliter l'interopérabilité dans le cadre de la BNDMR. Organiser le recueil des données et le rendre interopérable constitue le deuxième pilier d'interopérabilité que nous proposons dans le CIMR. Pour faciliter l'intégration des données nationales maladies rares dans la BNDMR nous avons opéré en trois phases :

- Construire le set minimal de données, tel qu'énoncé par le second plan national maladies rares, en adoptant une méthodologie collaborative et inspirée de l'existant. Ce set minimal national de données maladies rares, est issu d'un consensus entre les différents experts du domaine. Un consensus qui facilite l'adhésion à ce recueil.
- Construire une version numérique standardisée du set minimal de données. La standardisation permet d'avoir une référence commune internationale quant aux définitions et aux formats des données ainsi que leurs jeux de valeurs. Elle facilite par ailleurs l'interopérabilité avec les systèmes adoptant le même standard ou des standards compatibles tels que les dossiers patients informatisés des hôpitaux.
- Accompagner les travaux d'interconnexion par alignement de schémas de données. Nous avons proposé une approche de découverte automatisée de correspondances adaptée aux schémas de données. Cette approche permet de construire des processus d'intégration qui génèrent des

correspondances bien définies pouvant exprimer des liens sémantiques plus complexes que de simples équivalences.

1.2 DE LA RECHERCHE A L'OPERATIONNEL

1.2.1 *UNE APPROCHE DIFFERENTE DES AUTRES EXPERIENCES*

Si nous observons de plus près les autres expériences de mise en place d'entrepôts de données nationaux nous remarquons que l'approche que nous avons adoptée pour la BNDMR est plus ouverte et interopérable.

La Plan National Alzheimer 2008-2012 a lancé la création de La Banque Nationale Alzheimer (BNA), un entrepôt de données national pour les maladies neurodégénératives (Anthony et al. 2014). La BNA collecte des données démographiques, cliniques et diagnostiques recueillies par les Centres Mémoire et les Centres Mémoire de Ressources et de Recherche à des fins épidémiologiques. Un Corpus d'Informations Minimales Alzheimer (CIMA) a été défini comme étant le recueil national pour cet entrepôt. Le CIMA n'a pas été standardisé et aucune proposition d'interconnexion n'a été faite aux centres participants. Le recueil du CIMA se fait à travers une seule application Calliope CIMA (« Kappa Santé - Calliope » 2016).

De la même manière, l'entrepôt de données américain pour les maladies rares a mis en place un recueil d'éléments de données communs bien défini et standardisé ainsi qu'un identifiant patient global (GUID) que nous avons discuté dans le chapitre II. Cependant, le GRDR ne propose pas d'interconnexion avec les systèmes existants dans le but de peupler l'entrepôt. Le recueil de ses éléments de données communs se fait à travers une plateforme en ligne, « ORDR open source patient registry template », qui permet la saisie les données requises et de leur transmission au GRDR et de définir, si l'utilisateur le souhaite, un recueil complémentaire pour un usage spécifique.

1.2.2 *MIGRATION CEMARA*

Les travaux entrepris ont permis en premier lieu de faciliter la migration de la base historique CEMARA. En effet, les utilisateurs du tronc commun de CEMARA ont vocation à devenir utilisateurs de BaMaRa et leurs données précédemment saisies dans CEMARA, 360000 dossiers patients, doivent migrer dans la nouvelle application. Pour ce faire, un alignement entre le schéma de données de CEMARA et celui du set minimal de données recueilli dans BAMARA a été effectué. Cet

alignement a permis la définition des correspondances entre les deux ensembles de données, des correspondances décrivant exactement les transformations à opérer sur les données pour qu'elles soient adaptées au schéma de données cible.

Cette migration a été développée sur l'outil Talend, un outil ETL (Extract Transform and Load) qui permet à travers une interface de modélisation graphique et d'autres outils, de générer du code java exécutant le processus de migration de données qui a été défini (« Talend Open Studio : ETL open source et intégration de données » 2016). Les correspondances ont donc été directement implémentées sur cet outil. Un test de migration d'un site test a d'ores et déjà été effectué en 2015. Ce test concernait 15000 dossiers et le temps de traitement a duré 30 minutes. L'ensemble des dossiers du site a bien été migré et aucun conflit avec le format cible n'a été remonté.

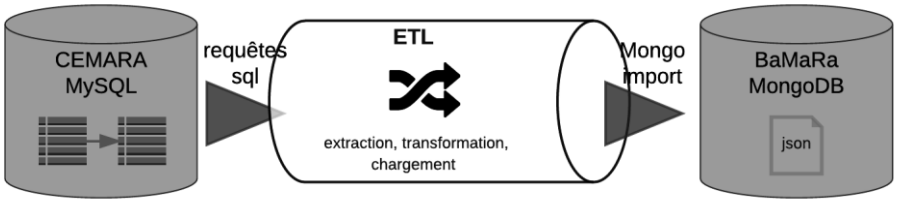

FIGURE 20 SCHEMA GENERAL MIGRATION DE CEMARA VERS BAMARA

Par ailleurs, l'alignement entre la version d'Orphanet datant de 2013 et celle qui a été intégrée à CEMARA en 2007 a permis de mettre à jour le codage diagnostique qui était fait dans CEMARA et de préparer les dossiers patients à la migration. Cette mise à jour avait pour objectif de rendre les codes Orphanet contenus dans l'item diagnostic patient compatible avec la dernière version d'Orphanet qui est implémenté dans BaMaRa. Ainsi, ces diagnostics, précédemment obsolètes par rapport à cette nouvelle version, redeviennent exploitables. Ce processus de mise à jour sera ré-exécuté en amont de la phase de migration.

1.2.3 *REPRISE DE DONNEES DES REGISTRES ET AUTRES BASES DE DONNEES*

Une quarantaine de demandes spontanées d'interconnexion nous est parvenue pour la reprise des données de certains registres ou autres bases de données utilisés au niveau des sites maladies rares. L'utilisation de ces systèmes était pour certains cloisonnée au niveau du site maladies rares, et était pour d'autres partagée au niveau de tous les sites d'un centre ou d'une filière maladies rares. A titre d'exemple nous pouvons citer : la plateforme pour les maladies neuromusculaires de

l'AFM (« Bases de données génétiques et cliniques pour l'AFM | Epiconcept » 2016), l'application partagée NHEMO pour les patient hémophiles, l'application e-Respirare pour les maladies respiratoires rares et d'autres registres de maladies tel que la sclérose latérale amyotrophique (SLA), la sclérodermie, les maladies bulleuses auto-immunes, …

Tous ces flux de données seront intégrés avec un outil EAI (Enterprise Application Integration). Cet outil permet de mettre en place des flux entre différentes applications et de les superviser. Les flux de données reçus respectent le format électronique standardisé du set minimal de données et alimentent l'application BaMaRa via son API. Par ailleurs, nous accompagnons ces partenaires en amont de la mise en place du flux sur les travaux d'alignement des données.

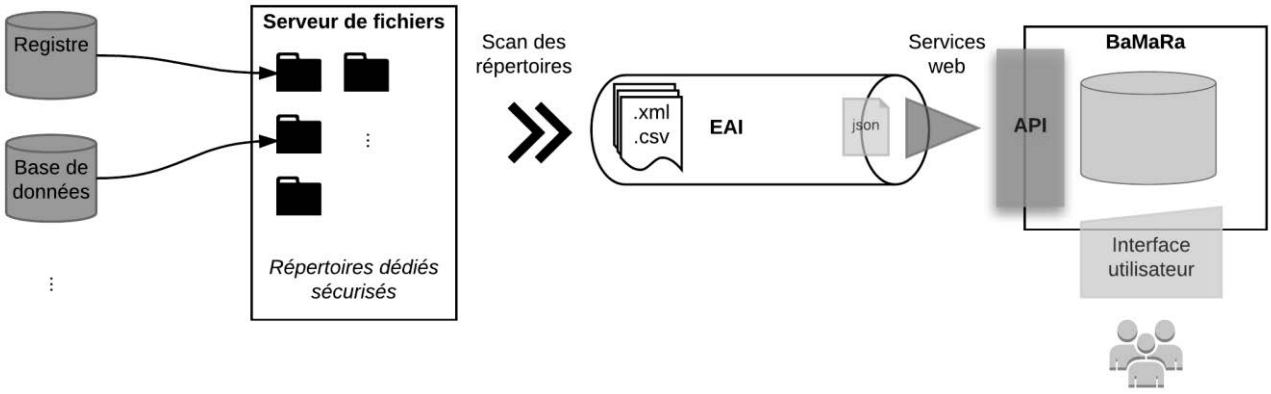

FIGURE 21 SCHEMA GENERAL DE LA REPRISE DES DONNEES DES REGISTRES ET AUTRES BASES DE DONNEES

La majorité de ces demandes concernent une alimentation périodique de la BNDMR via des fichiers CSV ou XML. L'interconnexion en temps réel n'a pas été demandée puisque l'objectif reste épidémiologique avec des études arrêtées à un temps T et non pas un objectif de soin qui nécessite le transfert instantané des données.

L'interconnexion avec ces différents systèmes est certes d'intérêt du point de vue de l'exhaustivité de la BNDMR. Cependant, des questions restent posées quant à la qualité de ces données, la pérennité de ces systèmes et le cadre juridique dans lequel ils s'inscrivent. Pour ces diverses raisons, un processus décisionnel a été mis place en amont du lancement de chaque projet d'interconnexion. Ce processus décisionnel prend en compte des critères juridiques, techniques et de qualité de données pour l'évaluation des demandes par les différents comités du projet.

1.2.4 *LES DPI ET LE CADRE D'INTEROPERABILITE DE L'ASIP SANTE*

Contrairement aux demandes d'interconnexion des registres et autres bases de données qui émanent directement des sites, des centres ou des filières maladies rares, l'initiation d'interconnexion avec les dossiers patients informatisés des hôpitaux est un peu plus complexe. Elle nécessite une volonté nationale partagée par les divers acteurs concernés : communauté maladies rares, directions générales et directions techniques des hôpitaux, les éditeurs des logiciels…

Une initiative lancée par les centres de référence de l'hôpital Bicêtre a permis de lancer un projet d'interconnexion avec le DPI de l'AP-HP. Ce groupe de CRMR avait demandé à intégrer un recueil spécifique au DPI pour le suivi quotidien des patients maladies rares. Bien évidemment, ce recueil devait être compatible avec le set minimal de la BNDMR. Ainsi, et suite à la mise en place de ce recueil, nous avons entrepris des travaux d'interconnexion avec l'équipe de la direction du système d'information de l'AP-HP, et ce recueil sera progressivement généralisé à tous les hôpitaux de l'AP-HP.

Afin de suivre cette stratégie sur le long terme d'interconnexion des DPI des CHU de France, l'ASIP Santé a été missionné par le ministère de la santé pour accompagner la cellule opérationnelle de la BNDMR dans les travaux d'interopérabilité et d'urbanisation nationale avec les divers CHU. Ainsi le cadre d'interopérabilité pour les maladies rares que nous avons défini est en cours d'intégration dans CI-SIS (Cadre d'Interopérabilité des Systèmes d'Information de Santé) de l'ASIP Santé. L'étude d'urbanisation vise à donner un état des lieux des situations des CRMR et de leurs sites pour définir la meilleure stratégie d'interconnexion à mettre en place. Dans ce sens, la stratégie que nous avons proposée a été appuyée. Une stratégie qui priorise l'interconnexion avec les DPI pour des raisons de pérennité et de qualité de données. Ces interconnexions ne pouvant être opérationnelles que sur le long terme, il reste primordial d'interconnecter les autres bases de données et registres pour des raisons d'exhaustivité sur le court terme.

1.2.5 *IDMR ET ETUDES TRANSVERSALES*

Au-delà de l'objectif d'interconnexion pour l'intégration des données dans la BNDMR, l'identifiant patient que nous proposons, l'IdMR, a permis la mise en place de certaines études transversales.

L'IdMR a été utilisé dans le cadre d'une étude conduite par l'Association Française contre les Myopathies (AFM) qui vise à évaluer l'impact de ses référents parcours santé, où un appariement avec les données de CEMARA a été effectué. Cet appariement anonyme, opéré avec l'IdMR, a permis à l'AFM de vérifier l'exhaustivité de leur recueil en relevant le nombre de patients inclus dans les centres de référence neuromusculaires ayant CEMARA et non répertoriés dans les bases de l'association.

2 PROBLEMATIQUES CONNEXES

2.1 QUALITE DE DONNEES

Afin de faciliter l'interopérabilité, nous encourageons l'implémentation du set minimal de données pour les maladies rares dans les dossiers patients informatisés des hôpitaux et les autres bases de données. Nous privilégions tout de même la saisie de ces données directement par les médecins lors de la prise en charge des patients pour une meilleure qualité des données. Nous restons tout de même conscients que c'est une approche complexe à globaliser étant donné la surcharge de travail des médecins. Par ailleurs, pour pousser l'adhésion des professionnels de santé à ce recueil, garantir un retour sur investissement en termes d'études épidémiologiques et de recherche est très important et la communauté maladies rares est d'ores et déjà sensibilisée à ce sujet. Les professionnels de santé cherchent aussi à éviter la double saisie et sont extrêmement vigilants quant à la sécurité des données de leurs patients. Malheureusement, même lorsque les données sont présentes dans les DPI, elles sont difficilement réutilisables. La première problématique reste liée aux systèmes de codage, tel qu'expliqué dans l'introduction générale, avec une qualité de codage non adaptée à la recherche.

2.2 CODAGE DIAGNOSTIC DES MALADIES RARES

La nomenclature Orphanet pour les maladies rares a été implémentée dans CEMARA depuis son lancement. Par ailleurs, l'Union Européenne et les autorités françaises recommandent son utilisation par les systèmes d'information nationaux pour l'identification des patients maladies rares. Suite aux différentes expériences d'implémentation et d'utilisation, il a été reporté que la nomenclature est trop large et que sa gestion est complexe. Cette complexité ne peut garantir une

homogénéité dans le système du codage des diagnostics maladies rares en France au sein des sites maladies.

Le ministère de la santé a missionné, en octobre 2014, un groupe d'experts pour l'évaluation de l'adéquation de cette ressource avec l'objectif de codage national pour les maladies à travers l'expérience CEMARA et les besoins terrains des professionnels de santé. La principale recommandation de ce groupe de travail était d'inclure les centres de référence dans la mise en place de la nomenclature adéquate pour le codage diagnostic. Les CRMR participeront à la définition de la granularité nécessaire sur laquelle il faut s'arrêter dans les classifications des maladies rares d'Orphanet. Par conséquent, ils établiront les listes des codes adéquats au codage diagnostique. Ils participeront aussi à la mise en place des instructions de codage et du plan d'exploitation de leurs données dans la BNDMR.

Les représentants des centres de référence ont aussi remonté le fait que le double codage CIM10 pour le PMSI et Orphanet pour les maladies rares était assez lourd et qu'il leur était difficile d'adopter d'autres systèmes de codage qui seraient pourtant d'intérêt pour certain domaines. L'implémentation d'un système de codage génétique, tel que OMIM, serait intéressant pour les généticiens par exemple mais ces systèmes en perpétuelle évolution sont d'autant plus complexes à gérer. Le Collège National de l'Information Médicale a entrepris un effort dans ce sens pour proposer un transcodage entre la CIM10 et Orphanet.

Les travaux nationaux qui sont actuellement menés sont les suivants :

- Evaluation du codage diagnostic des patients maladies rares avec les maladies « feuilles » (terminaison des classifications) proposées par Orphanet
- La revue par les filières maladies rares des classifications Orphanet pour l'utilisation du niveau des groupes de maladies dans les études statistiques d'épidémiologie nationale
- La revue du transcodage CIM10 Orphanet pour le codage PMSI

2.3 ANNUAIRE POUR LES STRUCTURES MALADIES RARES

Le réseau maladies rares est un réseau complexe qui se base sur des sites physiques implantés dans tout le réseau de soins français. Ces sites se structurent dans un premier niveau de centres de références maladies rares et de centres de compétences maladies rares. Ces derniers se structurent

dans un deuxième niveau de filières maladies rares. Ces sites peuvent porter plusieurs casquettes dans ce réseau. Par ailleurs nous retrouvons d'autres sites de soins considérés comme des « sites experts » qui collaborent avec les sites reconnus ou encore des sites de consultations externes où les médecins des centres de référence se déplacent pour effectuer des consultations délocalisées.

Ce réseau complexe se superpose à une organisation complètement différente des structures hospitalières de soins tel que présenté dans l'introduction de ce manuscrit. Dans ce contexte, nous faisons face à deux problématiques majeures dans la mise en place de la BNDMR :

- La mise ne place d'un annuaire unique officiel de tous les sites maladies rares : Si nous nous basons sur les déclarations faites par les centres de référence à travers leurs coordonnateurs, tel que cela a été effectué pour les rapports d'activité, nous disposerons seulement des visions cloisonnées des centres et il sera complexe d'identifier les sites en doublons (ceux qui participent à plusieurs centres). Si nous nous basons sur les déclarations des sites nous aurons un problème d'exhaustivité (Comment pousser directement tous les sites à se déclarer sans avoir auparavant leur liste) et de qualité (la validation des centres officiellement labellisés reste nécessaire).
- Maintenir un alignement entre les structures hospitalières et les structures maladies rares : Cette problématique se pose surtout lors de l'interconnexion avec les SIH. Lors de l'interconnexion avec le DPI de l'AP-HP pour récupérer les données de l'hôpital de Bicêtre, la synchronisation de l'annuaire de l'hôpital (UFs, Services…) avec l'annuaire des sites maladies rares de Bicêtre et leurs centres d'appartenance était une tâche assez complexe. Ce travail a été simplifié grâce à l'existence d'une plateforme maladie rare à l'hôpital qui a constitué pour nous un interlocuteur unique faisant l'interface avec tous les sites de l'hôpital. Une autre difficulté réside dans le décomptage de l'activité du site maladies rares des autres unités fonctionnelles voisines auxquelles les médecins des sites participent.

3 AU-DELA DE LA BNDMR

L'approche globale que nous proposons a donné des résultats satisfaisants dans le cadre du projet BNDMR. Cette approche peut cependant être appliquée à d'autres contextes outre les maladies rares. En effet, un cadre d'interopérabilité similaire peut être construit pour alimenter un

entrepôt qu'il soit dédié à un type particulier de maladies ou à un type particulier de données. Par ailleurs, les propositions que nous avons faites peuvent être prises indépendamment les unes des autres et employées dans des contextes plus génériques.

L'identifiant patient que nous avons défini peut être implémenté au niveau de n'importe quelle base de donnée de patients. Il permet un certain niveau d'identitovigilance en assurant la fonction de détection de doublons et de fédération d'identités. Il peut par ailleurs être utilisé dans le cadre d'initiation d'études de recherche en étant employé en tant que moyen d'étude de recouvrement entre deux sources de données. Il permet ainsi d'éviter des efforts qui seraient déployés pour la mise en place de l'étude alors que la population commune aux deux sources n'est pas assez représentative. Dans le cas où le recouvrement est satisfaisant, l'identifiant sert comme clé de chainage pour lier les données issues des deux sources. Toutes ces différentes utilisations de l'IdMR peuvent dépasser le contexte français pour s'appliquer à des études transfrontalières européennes. En effet les données sur lesquelles est construit l'identifiant sont universelles et non spécifiques au cadre Français.

En termes d'évolutions, et afin d'assurer une sécurité plus importante autour de l'utilisation de cet identifiant, il serait peut être judicieux, dans certains contextes, d'intégrer dans le calcul de l'identifiant une clé partagée entre les parties concernées par l'étude. Avec l'introduction de cette clé, une partie tierce, ayant accès aux données de l'étude, ne pourra pas croiser les identités des patients de l'étude avec des identités qu'elle aurait générées par ailleurs via l'algorithme de l'IdMR sans connaissance de cette clé.

La nouvelle méthode de définition de processus d'intégration de données que nous avons proposé peut très bien être adopté au sein d'autres projets d'intégration de données quel que soit le domaine d'étude. Les techniques d'alignements, et la formalisation de correspondances s'appliquent aux données de tous types et de n'importe quel domaine. Il suffirait d'adapter les ressources externes d'enrichissement sémantique en choisissant des ressources du domaine.

Par ailleurs, la proposition de cette approche s'inscrit dans un projet plus ambitieux qui vise à construire une plateforme d'intégration de données. Cette plateforme collaborative permettrait à ses utilisateurs de créer des processus spécifiques tels que nous les avons définis dans le chapitre III. Tous ces processus spécifiques, équivalents à celui que nous avons défini pour les types booléen

et énuméré, viendraient enrichir au fur et à mesure la bibliothèque des processus de la plateforme pour couvrir l'ensemble des types de données à traiter. Au final, l'utilisateur souhaitant envoyer des données d'une source vers une cible ou intégrer des données de plusieurs sources vers une cible n'aurait qu'à appliquer l'ensemble des processus spécifiques adaptés aux types de ses données pour générer des correspondances formalisées, prenant en compte l'aspect sémantique en définissant des relations plus complexes que les équivalences simples, prêtes à être intégrées à ses scripts et exécutées.

Le travail présenté dans ce manuscrit a abordé une problématique complexe d'hétérogénéité. Il a posé des bases pour faciliter l'interopérabilité nécessaire à la mise en place de la Banque Nationale de Données Maladies Rares notamment par la mise en place d'un identifiant patient maladie rare, la définition d'un recueil de données standardisé et la proposition d'une nouvelle approche de découverte de correspondances entre des systèmes hétérogènes. Ce travail doit être poursuivi notamment pour renforcer la sécurité de la solution d'identification des patients et étendre son utilisation à d'autres projets et pour l'initiation du projet de mise en place de la plateforme collaborative de découverte de correspondances.

BIBLIOGRAPHIE

« About Rare Diseases | www.eurordis.org ». 2016. Consulté le mars 30. http://www.eurordis.org/about-rare-diseases.

Adelson, P. David, Jose Pineda, Michael J. Bell, Nicholas S. Abend, Rachel P. Berger, Christopher C. Giza, Gillian Hotz, Mark S. Wainwright, et Pediatric TBI Demographics and Clinical Assessment Working Group. 2012. « Common Data Elements for Pediatric Traumatic Brain Injury: Recommendations from the Working Group on Demographics and Clinical Assessment ». *Journal of Neurotrauma* 29 (4): 639‑53. doi:10.1089/neu.2011.1952.

AIHW. 2016. « National minimum data sets ». http://www.aihw.gov.au/national-minimum-data-sets/.

aikido, Publié par. 2011. « Médecine légale & expertises médicales: Statut de l'embryon et du foetus ». http://medecine-legale.blogspot.com/2011/03/statut-de-lembryon-et-du-foetus.html.

« Alcohol and other drug treatment services in Australia 2013–14 (AIHW) ». 2016. Consulté le juin 15. http://www.aihw.gov.au/publication-detail/?id=60129551120.

Amarouche, Idir Amine, Djamal Benslimane, Mahmoud Barhamgi, Michael Mrissa, et Zaia Alimazighi. 2011. « Electronic Health Record Data-as-a-Services Composition Based on Query Rewriting ». In *Transactions on Large-Scale Data- and Knowledge-Centered Systems IV*, édité par Abdelkader Hameurlain, Josef Küng, Roland Wagner, Christian Böhm, Johann Eder, et Claudia Plant, 95‑123. Lecture Notes in Computer Science 6990. Springer Berlin Heidelberg. http://link.springer.com/chapter/10.1007/978-3-642-23740-9_5.

« AnDDI-Rares ». 2016. Consulté le juillet 11. http://www.anddi-rares.org/.

Angin, Céline, Amélie Ruel, Claude Messiaen, Rémy Choquet, et Paul Landais. 2015. « Rapport d'enquête nationale sur les bases de données maladies rares en France, Etat des lieux auprès des centres de référence maladies rare ». http://bndmr.fr.

Anthony, Sabine, Christian Pradier, Roland Chevrier, Julie Festraëts, Karim Tifratene, et Philippe Robert. 2014. « The French National Alzheimer Database: A Fast Growing Database for Researchers and Clinicians ». *Dementia and Geriatric Cognitive Disorders* 38 (5‑6): 271‑80. doi:10.1159/000360281.

Balas, E. Andrew, Santosh Krishna, et Tsigeweini A. Tessema. 2008. « eHealth: Connecting Health Care and Public Health ». *Studies in Health Technology and Informatics* 134: 169‑76.

« Banque Nationale de Données Maladies Rares ». 2014. http://www.bndmr.fr/.

« Bases de données génétiques et cliniques pour l'AFM | Epiconcept ». 2016. Consulté le juin 17. http://www.epiconcept.fr/fr/afm-bdd.

Beeler, George W. 1998. « HL7 Version 3—An object-oriented methodology for collaborative standards development1 ». *International Journal of Medical Informatics* 48 (1–3): 151‑61. doi:10.1016/S1386-5056(97)00121-4.

Bender, D., et K. Sartipi. 2013. « HL7 FHIR: An Agile and RESTful approach to healthcare information exchange ». In *2013 IEEE 26th International Symposium on Computer-Based Medical Systems (CBMS)*, 326‑31. doi:10.1109/CBMS.2013.6627810.

Bernardini, A, Alonzi M, Campioni P, Vecchioli A, et Marano P. 2002. « IHE: Integrating the Healthcare Enterprise, towards Complete Integration of Healthcare Information Systems. » *Rays* 28 (1): 83‑93.

Bird, Sheila M., et Jeremy Farrar. 2008. « Minimum Dataset Needed for Confirmed Human H5N1 Cases ». *Lancet (London, England)* 372 (9640): 696‑97. doi:10.1016/S0140-6736(08)61126-5.

Blaya, Joaquin A., Hamish S. F. Fraser, et Brian Holt. 2010. « E-Health Technologies Show Promise In Developing Countries ». *Health Affairs* 29 (2): 244‑51. doi:10.1377/hlthaff.2009.0894.

BobbyG. 2016. « The KHIT Blog: Syntactic and Semantic Interoperababble 2016 ». Consulté le juillet 5. http://regionalextensioncenter.blogspot.com/2016/02/syntactic-and-semantic-interoperababble.html.

BRAS, Pierre-Louis. 2013. « Rapport sur la gouvernance et l'utilisation des données de santé - Rapports - Ministère des Affaires sociales et de la Santé ». Rapport public. http://drees.social-sante.gouv.fr/etudes-et-statistiques/publications/recueils-ouvrages-et-rapports/rapports/article/rapport-sur-la-gouvernance-et-l-utilisation-des-donnees-de-sante.

Buchanan, R. J., S. Wang, et H. Ju. 2002. « Analyses of the Minimum Data Set: Comparisons of Nursing Home Residents with Multiple Sclerosis to Other Nursing Home Residents ». *Multiple Sclerosis (Houndmills, Basingstoke, England)* 8 (6): 512‑22.

Buchanan, Robert J., Raymond A. Martin, Linda Moore, Suojin Wang, et Hyunsu Ju. 2005. « Nursing Home Residents with Multiple Sclerosis and Dementia Compared to Other Multiple Sclerosis Residents ». *Multiple Sclerosis (Houndmills, Basingstoke, England)* 11 (5): 610‑16.

« Cardiogen ». 2016. Consulté le juillet 11. http://www.filiere-cardiogen.fr/.

« Centres de Référence - Fondation maladies rares ». 2016. Consulté le juin 7. http://fondation-maladiesrares.org/centres-de-reference2.

Choquet, Rémy, Meriem Maaroufi, Albane de Carrara, Claude Messiaen, Emmanuel Luigi, et Paul Landais. 2014. « A Methodology for a Minimum Data Set for Rare Diseases to Support National Centers of Excellence for Healthcare and Research ». *Journal of the American Medical Informatics Association*, juillet, amiajnl-2014-002794. doi:10.1136/amiajnl-2014-002794.

Chunara, Rumi, Jason R. Andrews, et John S. Brownstein. 2012. « Social and News Media Enable Estimation of Epidemiological Patterns Early in the 2010 Haitian Cholera Outbreak ». *The American Journal of Tropical Medicine and Hygiene* 86 (1): 39‑45. doi:10.4269/ajtmh.2012.11-0597.

CNIL. 2007. « Conclusions de la Commission Nationale de l'Informatique et des Libertés sur l'utilisation du NIR comme identifiant de santé ». CNIL.

———. 2011. « Guide Professionnels de santé ». https://www.cnil.fr/sites/default/files/typo/document/CNIL-Guide_professionnels_de_sante.pdf.

Code de la santé publique - Article L1111-8-1. 2007. *Code de la santé publique*. Vol. L1111-8‑1. https://www.legifrance.gouv.fr/affichCodeArticle.do;jsessionid=BB71AC9CE2D9A3E9C5FBA1BE2B18E208.tpdila23v_1?idArticle=LEGIARTI000017841975&cidTexte=LEGITEXT000006072665&categorieLien=id&dateTexte=20160127.

———. 2016. *Code de la santé publique*. Vol. L1111-8‑1. https://www.legifrance.gouv.fr/affichCodeArticle.do;jsessionid=BB71AC9CE2D9A3E9C5FBA1BE2B18E208.tpdila23v_1?idArticle=LEGIARTI000031931997&cidTexte=LEGITEXT000006072665&categorieLien=id&dateTexte=.

Commission Européenne. 2016. « Réforme sur la protection des données: le Parlement approuve de nouvelles règles adaptées à l'ère numérique ». *Parlement européen*. avril. http://www.europarl.europa.eu/news/fr/news-room/20160407IPR21776/R%C3%A9forme-sur-la-protection-des-donn%C3%A9es-des-r%C3%A8gles-adapt%C3%A9es-%C3%A0-l'%C3%A8re-num%C3%A9rique.

Cour de cassation. 2016. « Communiqué relatif aux arrêts 06-16.498, 06-16.499 et 06-16.500 du 6 février 2008 de la première chambre civile ». Consulté le juillet 5. https://www.courdecassation.fr/jurisprudence_2/premiere_chambre_civile_568/arrets_06_11171.html.

Dang, Quynh H. 2015. « Secure Hash Standard ». NIST FIPS 180-4. National Institute of Standards and Technology. http://nvlpubs.nist.gov/nistpubs/FIPS/NIST.FIPS.180-4.pdf.

« DBpedia ». 2016. http://wiki.dbpedia.org/.

De Moor, Georges, Mats Sundgren, Dipak Kalra, Andreas Schmidt, Martin Dugas, Brecht Claerhout, Töresin Karakoyun, et al. 2015. « Using Electronic Health Records for Clinical Research: The Case of the EHR4CR Project ». *Journal of Biomedical Informatics* 53 (février): 162‑73. doi:10.1016/j.jbi.2014.10.006.

« Définition et chiffres clés ». 2016. *Alliance Maladies Rares*. Consulté le juillet 11. http://www.alliance-maladies-rares.org/les-maladies-rares/definition-et-chiffres-cles/.

« Developing the National Early Childhood Development Researchable Data Set (AIHW) ». 2016. http://www.aihw.gov.au/publication-detail/?id=60129549619.

Dhombres, Ferdinand, Pierre-Yves Vandenbussche, Ana Rath, Annie Olry, Marc Hanauer, Bruno Urbero, Rémy Choquet, et Jean Charlet. 2011. « OntoOrpha : an ontology to support edition and audit of rare diseases knowledge in Orphanet ». In *ResearchGate*. https://www.researchgate.net/publication/240114364_OntoOrpha_an_ontology_to_support_edition_and_audit_of_rare_diseases_knowledge_in_Orphanet.

Direction Générale de l'Offre de Soins. 2016. *INSTRUCTION N° DGOS/PF4/2016/1*. http://circulaire.legifrance.gouv.fr/pdf/2016/01/cir_40460.pdf.

Direction Générale des Entreprises. 2016. « Etude e-Santé : faire émerger l'offre française en répondant aux besoins des acteurs de santé ». *Syntec Numérique*. février 10. http://www.syntec-numerique.fr/publication/etude-e-sante-faire-emerger-loffre-francaise-repondant-aux-besoins-acteurs-sante.

« Disability Services National Minimum Data Set (DS NMDS) collection (AIHW) ». 2016. http://www.aihw.gov.au/disability/disability-services-nmds-collection/.

Do, Nhan V., Fola Parrish, Omar Bouhaddou, Pradnya Warnekar, et Nancy Orvis. 2007. « The Use of UMLS to Establish a Mediation Terminology for Exchanging Patients' Allergy Profiles between Two U.S. Federal Agencies' Electronic Health Records ». *Medinfo 2007: Proceedings of the 12th World Congress on Health (Medical) Informatics; Building Sustainable Health Systems*, 2405.

Dolin, Robert H., Liora Alschuler, Calvin Beebe, Paul V. Biron, Sandra Lee Boyer, Daniel Essin, Elliot Kimber, Tom Lincoln, et John E. Mattison. 2001. « The HL7 Clinical Document Architecture ». *Journal of the American Medical Informatics Association* 8 (6): 552‑69. doi:10.1136/jamia.2001.0080552.

Dolin, Robert H., Liora Alschuler, Sandy Boyer, Calvin Beebe, Fred M. Behlen, Paul V. Biron, et Amnon Shabo (Shvo). 2006. « HL7 Clinical Document Architecture, Release 2 ». *Journal of the American Medical Informatics Association* 13 (1): 30‑39. doi:10.1197/jamia.M1888.

« Dossier de conception de l'Identifiant National de Santé calculé (INS-C) | esante.gouv.fr, le portail de l'ASIP Santé ». 2014. mai 3. http://esante.gouv.fr/services/referentiels/identification/dossier-de-conception-de-l-identifiant-national-de-sante.

« eHealth Market Size & Share | Global Industry Report, 2022 ». 2016. Consulté le juin 10. http://www.grandviewresearch.com/industry-analysis/e-health-market.

« ESID - European Society for Immunodeficiencies ». 2016. Consulté le juin 15. http://esid.org/About-ESID.

« European Medicines Agency - Find medicine - European public assessment reports ». 2016. Consulté le juillet 11. http://www.ema.europa.eu/ema/index.jsp?curl=pages%2Fmedicines%2Flanding%2Fepar_search.jsp&mid=WC0b01ac058001d124&searchTab=searchByAuthType&alreadyLoaded=true&isNewQuery=true&status=Authorised&status=Withdrawn&status=Suspended&status=Re

fused&keyword=Enter+keywords&searchType=name&taxonomyPath=Diseases&treeNumber=&searchGenericType=orphan&genericsKeywordSearch=Submit.

« Eurordis Position Paper on the WHO Report on Priority Medicines for Europe and the World ». 2004. http://www.eurordis.org/fr/publication/world-health-organization-who-report-priority-medicines.

Euzenat, Jérôme, et Pavel Shvaiko. 2013. « Classifications of Ontology Matching Techniques ». In *Ontology Matching*, 73 - 84. Springer Berlin Heidelberg. http://link.springer.com/chapter/10.1007/978-3-642-38721-0_4.

Eysenbach, Gunther. 2009. « Infodemiology and Infoveillance: Framework for an Emerging Set of Public Health Informatics Methods to Analyze Search, Communication and Publication Behavior on the Internet ». *Journal of Medical Internet Research* 11 (1). doi:10.2196/jmir.1157.

Forrest, Christopher B., Ronald J. Bartek, Yaffa Rubinstein, et Stephen C. Groft. 2011. « The Case for a Global Rare-Diseases Registry ». *Lancet (London, England)* 377 (9771): 1057 - 59. doi:10.1016/S0140-6736(10)60680-0.

Foulon, S., A. Weill, G. Maura, M. Dalichampt, M. Debouverie, et T. Moreau. 2015. « Prévalence de la sclérose en plaques en France en 2012 et mortalité associée en 2013 à partir des données du Sniiram-PMSI ». *Revue d'Épidémiologie et de Santé Publique*, XXVIIIe Congrès national Émois, Nancy, 26 et 27 mars 2015, 63, Supplement 1 (mars): S17 - 18. doi:10.1016/j.respe.2015.01.037.

Garcelon, Nicolas, Rémi Salomon, et Anita Burgun. 2014. « Enrichissement sémantique associé à la détection de la négation et des antécédents familiaux dans un entrepôt de données hospitalier. » *JFIM*. http://ceur-ws.org/Vol-1379/paper-08.pdf.

Geissbuhler, A., C. Safran, I. Buchan, R. Bellazzi, S. Labkoff, K. Eilenberg, A. Leese, et al. 2013. « Trustworthy reuse of health data: A transnational perspective ». *International Journal of Medical Informatics* 82 (1): 1-9. doi:10.1016/j.ijmedinf.2012.11.003.

Gilbert, Henri, et Helena Handschuh. 2003. « Security Analysis of SHA-256 and Sisters ». In *Selected Areas in Cryptography*, édité par Mitsuru Matsui et Robert J. Zuccherato, 175-93. Lecture Notes in Computer Science 3006. Springer Berlin Heidelberg. http://link.springer.com/chapter/10.1007/978-3-540-24654-1_13.

Glasgow, Russell E. 2007. « eHealth Evaluation and Dissemination Research ». *American Journal of Preventive Medicine*, Critical Issues in eHealth ResearchCritical Issues in eHealth Research 2005, 32 (5, Supplement): S119-26. doi:10.1016/j.amepre.2007.01.023.

« Global Unique Identifier (GUID) ». 2016. *National Center for Advancing Translational Sciences*. http://www.ncats.nih.gov/grdr/guid.

Haffner, Marlene E., Janet Whitley, et Marie Moses. 2002. « Two Decades of Orphan Product Development ». *Nature Reviews. Drug Discovery* 1 (10): 821-25. doi:10.1038/nrd919.

Hanf, Matthieu, Catherine Quantin, Paddy Farrington, Eric Benzenine, N. Mounia Hocine, Michel Velten, Pascale Tubert-Bitter, et Sylvie Escolano. 2013. « Validation of the French National Health Insurance Information System as a Tool in Vaccine Safety Assessment: Application to Febrile Convulsions after Pediatric Measles/mumps/rubella Immunization ». *Vaccine* 31 (49): 5856‑62. doi:10.1016/j.vaccine.2013.09.052.

Häyrinen, Kristiina, et Kaija Saranto. 2005. « The Core Data Elements of Electronic Health Record in Finland ». *Studies in Health Technology and Informatics* 116: 131‑36.

HCSP. 2009. « Évaluation du Plan national maladies rares 2005-2008 ». Paris: Haut Conseil de la Santé Publique. http://www.hcsp.fr/explore.cgi/avisrapportsdomaine?clefr=65.

Health, Australian Government Department of. 2016. « Healthcare Identifiers Service ». Australian Government Department of Health. Consulté le juillet 5. http://www.health.gov.au/internet/main/publishing.nsf/Content/pacd-ehealth-consultation.

HL7. 2016. « About Health Level Seven International ». Consulté le juin 15. http://www.hl7.org/about/index.cfm?ref=common.

« HL7 Version 2 Product Suite ». 2016. Consulté le juin 15. http://www.hl7.org/implement/standards/product_brief.cfm?product_id=185.

« HL7 Version 3 Product Suite ». 2016. Consulté le juin 15. https://www.hl7.org/implement/standards/product_brief.cfm?product_id=186.

« Institut National Du Cancer - Accueil ». 2016. Consulté le juin 7. http://www.e-cancer.fr/.

« Investissements d'avenir (CGI) ». 2016. *Gouvernement.fr*. Consulté le juin 7. http://www.gouvernement.fr/investissements-d-avenir-cgi.

Jason, Leonard A., Elizabeth R. Unger, Jordan D. Dimitrakoff, Adam P. Fagin, Michael Houghton, Dane B. Cook, Gailen D. Marshall, Nancy Klimas, et Christopher Snell. 2012. « Minimum Data Elements for Research Reports on CFS ». *Brain, Behavior, and Immunity* 26 (3): 401‑6. doi:10.1016/j.bbi.2012.01.014.

Johnson, Stephen B., Glen Whitney, Matthew McAuliffe, Hailong Wang, Evan McCreedy, Leon Rozenblit, et Clark C. Evans. 2010. « Using Global Unique Identifiers to Link Autism Collections ». *Journal of the American Medical Informatics Association: JAMIA* 17 (6): 689‑95. doi:10.1136/jamia.2009.002063.

Kalam, Anas Abou El, Yves Deswarte, Gilles Trouessin, et Emmanuel Cordonnier. 2004. « Gestion des données médicales anonymisées : problèmes et solutions », Mons, Belgique, 9-11 octobre 2004.

« Kappa Santé - Calliope ». 2016. Consulté le juin 16. http://www.kappasante.com/fr/e-sante-calliope.

Kerr, A. M., Y. Nomura, D. Armstrong, M. Anvret, P. V. Belichenko, S. Budden, H. Cass, et al. 2001. « Guidelines for Reporting Clinical Features in Cases with MECP2 Mutations ». *Brain & Development* 23 (4): 208‑11.

Kho, Abel N., John P. Cashy, Kathryn L. Jackson, Adam R. Pah, Satyender Goel, Jörn Boehnke, John Eric Humphries, et al. 2015. « Design and Implementation of a Privacy Preserving Electronic Health Record Linkage Tool in Chicago ». *Journal of the American Medical Informatics Association: JAMIA* 22 (5): 1072‑80. doi:10.1093/jamia/ocv038.

Khoury, Muin J., et John P. A. Ioannidis. 2014. « Big data meets public health ». *Science (New York, N.Y.)* 346 (6213): 1054‑55. doi:10.1126/science.aaa2709.

Kuchinke, Wolfgang, J. Aerts, S. C. Semler, et C. Ohmann. 2009. « CDISC Standard-Based Electronic Archiving of Clinical Trials ». *Methods of Information in Medicine* 48 (5): 408‑13. doi:10.3414/ME9236.

« La drépanocytose ». 2011. *Encyclopédie Orphanet Grand Public*. https://www.orpha.net/data/patho/Pub/fr/Drepanocytose-FRfrPub125v01.pdf.

« L'AFM-Téléthon en bref ». 2013. *AFM-Téléthon*. septembre 1. http://www.afm-telethon.fr/association/afm-telethon-bref-631.

Le Duff, Franck, Nicolas Duport, Sébastien Gonfrier, Pierre Lafay, Nathalie Texier, Stéphane Schück, Christian Pradier, et Philippe Robert. 2010. « Plan national Alzheimer 2008-2012 - Mesure 34 Mise en place du recueil épidémiologique national et premières tendances ». *La Revue de gériatrie* 35 (8): 575‑82.

« Le programme hôpital numérique - Hôpital numérique - Ministère des Affaires sociales et de la Santé ». 2016. Consulté le juin 7. http://social-sante.gouv.fr/systeme-de-sante-et-medico-social/e-sante/sih/hopital-numerique/Hopital-Numerique.

« Le programme hospitalier de recherche clinique - PHRC - Appels à projets - Ministère des Affaires sociales et de la Santé ». 2016. Consulté le juin 7. http://social-sante.gouv.fr/systeme-de-sante-et-medico-social/recherche-et-innovation/appels-a-projets/article/le-programme-hospitalier-de-recherche-clinique-phrc.

« Les filières de santé maladies rares - Prises en charge spécialisées - Ministère des Affaires sociales et de la Santé ». 2016a. Consulté le juin 7. http://social-sante.gouv.fr/soins-et-maladies/prises-en-charge-specialisees/article/les-filieres-de-sante-maladies-rares.

———. 2016b. Consulté le juin 7. http://social-sante.gouv.fr/soins-et-maladies/prises-en-charge-specialisees/article/les-filieres-de-sante-maladies-rares.

« Les raisons d'être et le cadre réglementaire de l'INS | esante.gouv.fr, le portail de l'ASIP Santé ». 2015. mars 4. http://esante.gouv.fr/services/referentiels/identification/les-raisons-d-etre-et-le-cadre-reglementaire-de-l-ins.

Loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés. 2016. Consulté le juin 9.

Louie, Brenton, Peter Mork, Fernando Martin-Sanchez, Alon Halevy, et Peter Tarczy-Hornoch. 2007. « Data integration and genomic medicine ». *Journal of Biomedical Informatics*, Bio*Medical Informatics, 40 (1): 5‑16. doi:10.1016/j.jbi.2006.02.007.

Ludvigsson, Jonas F., Petra Otterblad-Olausson, Birgitta U. Pettersson, et Anders Ekbom. 2009. « The Swedish personal identity number: possibilities and pitfalls in healthcare and medical research ». *European Journal of Epidemiology* 24 (11): 659‑67. doi:10.1007/s10654-009-9350-y.

Maaroufi, Meriem, Rémy Choquet, Paul Landais, et Marie-Christine Jaulent. 2013. « Formalizing Mappings to Optimize Automated Schema Alignment: Application to Rare Diseases. » *Studies in Health Technology and Informatics* 205 (décembre): 283‑87.

———. 2015. « Towards Data Integration Automation for the French Rare Disease Registry ». *AMIA ... Annual Symposium Proceedings / AMIA Symposium. AMIA Symposium* 2015: 880‑85.

« Maladie de Gaucher | LORD ». 2016. http://lord.bndmr.fr/#disorders/355.

Mandel, Joshua C., David A. Kreda, Kenneth D. Mandl, Isaac S. Kohane, et Rachel B. Ramoni. 2016. « SMART on FHIR: A Standards-Based, Interoperable Apps Platform for Electronic Health Records ». *Journal of the American Medical Informatics Association: JAMIA*, février. doi:10.1093/jamia/ocv189.

Mazuel, Laurent, et Jean Charlet. s. d. « OnAGUI ». *http://onagui.sourceforge.net/*.

McCormick, Jonathan, Erika J. Sims, Michael W. Green, Gita Mehta, Frank Culross, et Anil Mehta. 2005. « Comparative Analysis of Cystic Fibrosis Registry Data from the UK with USA, France and Australasia ». *Journal of Cystic Fibrosis: Official Journal of the European Cystic Fibrosis Society* 4 (2): 115‑22. doi:10.1016/j.jcf.2005.01.001.

« MeSH - NCBI ». 2016. Consulté le juin 15. http://www.ncbi.nlm.nih.gov/mesh.

Moutel, Grégoire, Michèle Goussot-Souchet, Isabelle Plu, Marion Pierre, Thomas Leclercq, Jean-Christophe Coffin, et Nathalie Duchange. 2010. « [Fetuses born lifeless: new representations and new rights.] ». *médecine/sciences* 26 (8‑9): 772‑78.

Mullins, Irene M., Mir S. Siadaty, Jason Lyman, Ken Scully, Carleton T. Garrett, W. Greg Miller, Rudy Muller, et al. 2006. « Data mining and clinical data repositories: Insights from a 667,000 patient data set ». *Computers in Biology and Medicine* 36 (12): 1351 ‑ 77. doi:10.1016/j.compbiomed.2005.08.003.

Murdoch TB, et Detsky AS. 2013. « THe inevitable application of big data to health care ». *JAMA* 309 (13): 1351‑52. doi:10.1001/jama.2013.393.

Murphy, Shawn N, Anil Dubey, Peter J Embi, Paul A Harris, Brent G Richter, Fran Turisco, Griffin M Weber, James E Tcheng, et Diane Keogh. 2012. « Current state of information technologies for the clinical research enterprise across academic medical centers. » *Clinical and translational science* 5 (3): 281‑84. doi:10.1111/j.1752-8062.2011.00387.x.

O'Donnell, J. L., V. R. Stevanovic, C. Frampton, L. K. Stamp, et P. T. Chapman. 2007. « Wegener's Granulomatosis in New Zealand: Evidence for a Latitude-Dependent Incidence Gradient ». *Internal Medicine Journal* 37 (4): 242‑46. doi:10.1111/j.1445-5994.2006.01297.x.

Office of the National Coordinator for Health IT. 2016. « Interoperability Standards Advisory: Best available standards and implementation specifications ». https://www.healthit.gov/sites/default/files/2016-interoperability-standards-advisory-final-508.pdf.

Orphanet. 2009. « Centres de référence labellisés et centres de compétences désignés pour la prise en charge d'une maladie rare ou d'un groupe de maladies rares ». *Cahiers d'Orphanet - Série Politique de Santé*, n° n°2. http://social-sante.gouv.fr/IMG/pdf/Liste_des_centres_de_competences_et_de_references_par_groupe.pdf.

———. 2012. http://www.orpha.net/.

———. 2016a. « Rare Disease Registries in Europe ». *Orphanet Report Series*, janvier. http://www.orpha.net/orphacom/cahiers/docs/GB/Registries.pdf.

———. 2016b. « Prévalence des maladies rares : Données bibliographiques Prévalence, incidence ou nombre publié de cas classés par ordre alphabétique des maladies ». *Les Cahiers d'Orphanet* Série Maladies Rares (1).

Pheby, D. F. H., et D. J. Etherington. 1994. « Improving the Comparability of Cancer Registry Treatment Data and Proposals for a New National Minimum Dataset ». *Journal of Public Health* 16 (3): 331‑40.

« Physician Reporting to a Public Health Repository – Cancer Registry - IHE Wiki ». 2016. Consulté le 15 juin. http://wiki.ihe.net/index.php/Physician_Reporting_to_a_Public_Health_Repository_%E2%80%93_Cancer_Registry.

Piette, John D., K. C. Lun, Moura Jr, Lincoln A, Hamish SF Fraser, Patricia N. Mechael, John Powell, et Shariq R. Khoja. 2012. « Impacts of e-health on the outcomes of care in low-and middle-income countries: where do we go from here? » *Bulletin of the World Health Organization* 90 (5): 365‑72. doi:10.2471/BLT.11.099069.

« Plan National Maladies Rares 2005-2008 ». 2005. http://social-sante.gouv.fr/IMG/pdf/plan_national_maladies_rares_2005-2008.pdf.

« Plan National Maladies Rares 2011-2014 ». 2011. http://social-sante.gouv.fr/IMG/pdf/plan_national_maladies_rares_2011-2014.pdf.

« Portail Epidemiologie - France | Health Databases ». 2016. https://epidemiologie-france.aviesan.fr/ezfind/research?sources=all&search=maladie+rare&facet=catalogue&sort=score%7Cdesc&hsearch=.

Prather, J. C., D. F. Lobach, L. K. Goodwin, J. W. Hales, M. L. Hage, et W. E. Hammond. 1997. « Medical data mining: knowledge discovery in a clinical data warehouse. » *Proceedings of the AMIA Annual Fall Symposium*, 101‑5.

Projet de loi de modernisation de notre système de santé. 2015. *Code de la santé publique*. http://www.legifrance.gouv.fr/affichLoiPreparation.do;jsessionid=F3A3D61446B1F56786C3A823483D2972.tpdila22v_3?idDocument=JORFDOLE000029589477&type=contenu&id=2&typeLoi=proj&legislature=14.

« PubMed - NCBI ». 2016. Consulté le juin 15. http://www.ncbi.nlm.nih.gov/pubmed.

Quantin, Catherine, Cyril Ferdynus, Paul Avillach, Béatrice Gouyon, Paul Sagot, Jean-Bernard Gouyon, Catherine Quantin, et al. 2008. « Using Discharge Abstracts to Evaluate a Regional Perinatal Network: Assessment of the Linkage Procedure of Anonymous Data, Using Discharge Abstracts to Evaluate a Regional Perinatal Network: Assessment of the Linkage Procedure of Anonymous Data ». *International Journal of Telemedicine and Applications, International Journal of Telemedicine and Applications* 2009, 2009 (décembre): e181842. doi:10.1155/2009/181842, 10.1155/2009/181842.

« Radico - Rare Disease Cohorts ». 2016. Consulté le juillet 11. http://www.radico.fr/fr/.

Rahm, Erhard, et Philip A. Bernstein. 2001. « A Survey of Approaches to Automatic Schema Matching ». *The VLDB Journal* 10 (4): 334‑50. doi:10.1007/s007780100057.

« RARE Diseases: Facts and Statistics ». 2012. *Global Genes*. janvier 1. http://globalgenes.org/rare-diseases-facts-statistics/.

« Reference Information Model (RIM) ». 2016. Consulté le juin 15. http://www.hl7.org/implement/standards/rim.cfm.

« Référentiel Général de Sécurité version 2.0 - Annexe B1 Mécanismes cryptographiques Règles et recommandations concernant le choix et le dimensionnement des mécanismes cryptographiques ». 2014. ANSSI.

Roinet, Marie. 2016. « Annonce d'un 3ème Plan Maladies Rares : UN PLAN OUI, MAIS… UN PLAN INTERMINISTERIEL ET CO-CONSTRUIT ! » *Alliance Maladies Rares*. juin 16. http://www.alliance-maladies-rares.org/annonce-dun-3eme-plan-maladies-rares-un-plan-oui-mais-un-plan-interministeriel-et-co-construit/.

Salathé, Marcel, Clark C. Freifeld, Sumiko R. Mekaru, Anna F. Tomasulo, et John S. Brownstein. 2013. « Influenza A (H7N9) and the Importance of Digital Epidemiology ». *The New England Journal of Medicine* 369 (5): 401‑4. doi:10.1056/NEJMp1307752.

Santé publique France. 2016. « Mise en place du Comité d'évaluation des registres (CER) auprès de l'InVS, l'Inserm et l'INCa / Comité d'évaluation des registres / Espace professionnels / Accueil ». Consulté le juillet 5. http://www.invs.sante.fr/Espace-professionnels/Comite-d-evaluation-des-registres/Mise-en-place-du-Comite-d-evaluation-des-registres-CER-aupres-de-l-InVS-l-Inserm-et-l-INCa.

Santé publique France, InVS. 2016. « Maladies infectieuses ». Consulté le juin 13. http://invs.santepubliquefrance.fr/%20fr/Dossiers-thematiques/Maladies-infectieuses.

Schadow, Gunther, Charles N. Mead, et D. Mead Walker. 2006. « The HL7 Reference Information Model under Scrutiny ». *Studies in Health Technology and Informatics* 124: 151‑56.

Schneider, Henning. 2015. « [Electronic patient record as the tool for better patient safety] ». *Bundesgesundheitsblatt, Gesundheitsforschung, Gesundheitsschutz* 58 (1): 61 ‑ 66. doi:10.1007/s00103-014-2081-6.

Science & Santé. 2016. « La e-santé, de quoi parle-t-on? », février 1.

Seifter, Ari, Alison Schwarzwalder, Kate Geis, et John Aucott. 2010. « The Utility of "Google Trends" for Epidemiological Research: Lyme Disease as an Example ». *Geospatial Health* 4 (2): 135‑37. doi:10.4081/gh.2010.195.

Smith, Barry, et Werner Ceusters. 2006. « HL7 RIM: An Incoherent Standard ». *Studies in Health Technology and Informatics* 124: 133‑38.

Sureau, Claude. 2010. « Médecine de l'embryon et du fœtus : le désarroi des idéologies ». In *Traité de bioéthique*, par Emmanuel Hirsch, 608. ERES. http://www.cairn.info/traite-de-bioethique-2--9782749213064-page-608.htm.

« Syndrome de Mowat-Wilson | LORD ». 2016. Consulté le juillet 5. http://lord.bndmr.fr/#disorders/2152.

« Syndrome onycho-digito-mammaire | LORD ». 2016. Consulté le juillet 5. http://lord.bndmr.fr/#disorders/238744.

« Talend Open Studio : ETL open source et intégration de données ». 2016. Consulté le juin 17. https://fr.talend.com/products/talend-open-studio.

Taruscio, Domenica, Emanuela Mollo, Sabina Gainotti, Manuel Posada de la Paz, Fabrizio Bianchi, et Luciano Vittozzi. 2014. « The EPIRARE Proposal of a Set of Indicators and Common Data Elements for the European Platform for Rare Disease Registration ». *Archives of Public Health = Archives Belges De Santé Publique* 72 (1): 35. doi:10.1186/2049-3258-72-35.

Thompson, Rachel, Louise Johnston, Domenica Taruscio, Lucia Monaco, Christophe Béroud, Ivo G. Gut, Mats G. Hansson, et al. 2014. « RD-Connect: An Integrated Platform Connecting Databases, Registries, Biobanks and Clinical Bioinformatics for Rare Disease Research ». *Journal of General Internal Medicine* 29 (S3): 780‑87. doi:10.1007/s11606-014-2908-8.

Tilyard, M. W., N. Munro, S. A. Walker, et S. M. Dovey. 1998. « Creating a General Practice National Minimum Data Set: Present Possibility or Future Plan? » *The New Zealand Medical Journal* 111 (1072): 317‑18, 320.

Trouessin, G., et F. A. Allaert. 1997. « FOIN: A Nominative Information Occultation Function ». *Studies in Health Technology and Informatics* 43 Pt A: 196‑200.

van Weely, Sonja, et Hubert G.M. Leufkens. 2004. « Background Paper 6.19 Rare Diseases ». http://www.who.int/medicines/areas/priority_medicines/BP6_19Rare.pdf.

Wagholikar, Kavishwar B., Joshua C. Mandel, Jeffery G. Klann, Nich Wattanasin, Michael Mendis, Christopher G. Chute, Kenneth D. Mandl, et Shawn N. Murphy. 2016. « SMART-on-FHIR Implemented over i2b2 ». *Journal of the American Medical Informatics Association: JAMIA*, juin. doi:10.1093/jamia/ocw079.

Wall, Martin. 2016. « New health identification number for every individual ». *The Irish Times*. Consulté le juillet 5. http://www.irishtimes.com/news/health/new-health-identification-number-for-every-individual-1.2135560.

Warner, Jeremy L., Matthew J. Rioth, Kenneth D. Mandl, Joshua C. Mandel, David A. Kreda, Isaac S. Kohane, Daniel Carbone, et al. 2016. « SMART Precision Cancer Medicine: A FHIR-Based App to Provide Genomic Information at the Point of Care ». *Journal of the American Medical Informatics Association: JAMIA*, mars. doi:10.1093/jamia/ocw015.

Webster, D. 1998. « A Minimum Dataset for Newborn Screening ». *Journal of Medical Screening* 5 (2): 109‑109. doi:10.1136/jms.5.2.109.

« White Paper on Unique Health Identifier for Individuals ». 2016. Consulté le juillet 5. https://epic.org/privacy/medical/hhs-id-798.html.

« WHO/OMS Vaccines and Biologicals ». 2016. Consulté le juin 6. http://archives.who.int/prioritymeds/report/index.htm.

Winget, Marcy D., John A. Baron, Margaret R. Spitz, Dean E. Brenner, Denise Warzel, Heather Kincaid, Mark Thornquist, et Ziding Feng. 2003. « Development of Common Data Elements: The Experience of and Recommendations from the Early Detection Research Network ». *International Journal of Medical Informatics* 70 (1): 41‑48.

ANNEXE 1 : SET MINIMAL DES DONNEES MALADIES RARES (V1.09.2)

Groupe d'items	Objectif(s) d'exploitation(s)	N° item	Item	Définition de l'item	Codage du contenu (type ou valeurs que pourront prendre les items)
1. Consentement	Information légale nécessaire	1.1	Consentement patient	Le (la) patient(e) donne-t-il(elle) son consentement pour que des informations soient enregistrées dans un système de gestion informatique de données ?	Oui - LA33-6
		1.2	Non-opposition du patient pour une réutilisation des données	Le (la) patient(e) a-t-il (elle) été dûment informé(e) qu'une partie des données, préalablement anonymisées, le (la) concernant, sera utilisée dans le cadre d'analyses de santé publique; et qu'il (elle) ne s'y est pas opposé(e) ?	Oui - LA33-6 Non - LA32-8
		1.3	Consentement par tutelle	Le consentement est-il donné par la tutelle du (de la) patient(e) ?	Oui - LA33-6 Non - LA32-8
2. Identification patient	Identitologie	2.1	Identifiant national MR	GUID (Global Unique Identifier) permettant l'identification unique de patients entre BaMaRa et la BNDMR (anonyme).	Chaîne de caractères (généré automatiquement)
		2.1 bis	Identifiant anonyme national (IdMR)	Identifiant anonyme national permettant l'identification unique du (de la) patient(e) entre BaMaRa et la BNDMR.	Chaîne de caractères (généré automatiquement)
		2.2	Identifiant National de Santé	Identifiant du patient soumis à l'appréciation de la CNIL : identifiant national unique permettant l'interconnexion future avec le dossier médical personnalisé (DMP).	Chaîne de caractères (généré automatiquement)
		2.3	Identifiant local du patient	Identifiant local du (de la) patient(e) à l'hôpital.	Chaîne de caractères
3. Informations personnelles	Informations nécessaires pour identifier le patient	3.1	Nom de naissance du patient	Le nom de naissance du (de la) patient(e), appelé auparavant nom patronymique, s'appelle dorénavant nom de famille, c'est le nom figurant sur l'acte de naissance.	Chaîne de caractères
		3.2	Nom d'usage du patient	Deuxième nom : c'est le nom utilisé dans la vie courante (nom d'usage) lorsqu'il diffère du nom de famille : double nom (nom de ses parents ou nom des deux époux accolés), nom de son époux ou nom de son épouse.	Chaîne de caractères
		3.3	Prénom du patient	Prénom du (de la) patient(e) tel que renseigné sur son acte de naissance ou carte d'identité.	Chaîne de caractères
		3.4	Date de naissance du patient	La date de naissance du (de la) patient(e) telle qu'enregistrée dans le certificat de naissance.	Date

		3.5	Sexe du patient	Sexe du (de la) patient(e).	Féminin - LA3-6 Masculin - LA2-8 Indéterminé - LA18959-9 Inconnu (sur le fœtus) - LA4489-6
		3.6	Fœtus	Dans le cas d'information enregistrée pour un fœtus.	Oui - LA33-6 Non - LA32-8
	Informations nécessaires pour identifier le patient et pour vérifier le statut vital du patient auprès de l'Insee.	3.7	Commune de naissance	Commune de naissance du (de la) patient(e).	Code Commune
		3.8	Pays de naissance	Pays de naissance du (de la) patient(e).	Code Pays
	Informations nécessaire pour les études de distance domicile/lieu de prise en charge et pour les rapports PIRAMIG.	3.9	Commune de résidence	Commune de résidence du (de la) patient(e).	Code Commune
		3.10	Pays de résidence	Pays de résidence du (de la) patient(e).	Code Pays
4. Informations familiales	Etudes sur les familles	4.1	Propositus	Premier patient enregistré dans un centre MR pour une même famille?	Oui - LA33-6 Non - LA32-8
		4.1 bis	Identifiant du propositus	Identifiant patient maladie rare (IdMR) du premier patient enregistré dans un centre MR pour une même famille.	Chaîne de caractères
		4.2	Lien de parenté avec le propositus	Permet de décrire le lien de parenté entre un(e) patient(e) et le propositus.	Frère - LA10415-0 Sœur - LA10418-4 […] Grand-père maternel - LA10412-7 Grand-mère maternelle - LA10413-5
	Informations nécessaires pour des éléments de soins	4.3	Consanguinité	Le (la) patient(e) est-il (elle) issu(e) d'une union entre apparentés ?	Oui - LA33-6 Non - LA32-8 Inconnu - LA4489-6
5. Statut vital	Informations nécessaires pour les études de survie et de prévalence	5.1	Statut vital du patient	Le (la) patient(e) est-il(elle) décédé(e) ?	Oui - LA33-6 Non - LA32-8
		5.2	Date de décès du patient	Date à laquelle le (la) patient(e) est décédé(e).	Date
	Informations nécessaires pour les études de survie	5.3	Décès dû à la maladie rare	Le décès est-il dû à la maladie rare ?	Oui - LA33-6 Non - LA32-8 Inconnu - LA4489-6
		5.4 cond(5.3)	Cause principale du décès	Si le décès n'est pas lié à la maladie rare, cause principale du décès ?	Code CIM10

6. Parcours de soins	Information nécessaire pour apprécier l'attractivité des CRMR et CCMR auprès des professionnels référant des patients et apprécier le parcours de soin	6.1	Patient adressé par	Le (la) patient(e) peut avoir été adressé(e) par un professionnel de santé. Apprécie l'attractivité des CRMR auprès des professionnels référant des patients.	Venu de lui-même -PR Association de patients -PR Généraliste - PR Pédiatre (ville) - PR Pédiatre (hôpital) - DR Centre de protection maternelle et infantile (PMI) - DR Généticien - DR Gynéco/obstétricien - DR Autre spécialiste (ville/hôpital) - DR Centre de prise en charge (CAMSP, CMPP, SESSAD, …) - DR Centre pluridisciplinaire de diagnostic prénatal - DR Centre de dépistage prénatal - DR Autre
		6.2	Date d'inclusion du patient dans le centre	Date à laquelle le (la) patient(e) a été inclus dans le CRMR.	Date
	Information nécessaire pour le droit d'accès du patient	6.3	Nom du médecin référent maladie rare	Nom du médecin ayant la charge du (de la) patient(e) dans le site (eg. Dans le centre Maladies Rares)	Chaîne de caractères
7. Activité de soins	Information nécessaire pour la constitution du rapport d'activité annuel PIRAMIG des CRMR et l'incidence (et toute étude sur le suivi du patient)	7.1	Date de l'activité réalisée pour la maladie rare	Date à laquelle l'activité renseignée pour la maladie rare considérée a été effectuée.	Date
	Information nécessaire pour la constitution du rapport d'activité annuel PIRAMIG des CRMR	7.2	Contexte de l'activité MR	Contexte permettant d'identifier le type d'activité réalisée. (Information utile pour la constitution du rapport d'activité) annuel des CRMR).	Consultation - D009819 Consultation pluridisciplinaire Hôpital de jour - D003631 Hospitalisation traditionnelle - D006760 Avis sur dossier en consultation Avis d'expertise sur un dossier - D005104 Avis en salle (dans un autre service) Téléconsultation - D019114 Autre contexte - LA4732-9
		7.3	Objectifs de l'activité MR	Objectifs de l'activité déclarée.	Diagnostic - D003933 Mise en place de la prise en charge - D019468 Suivi - D003266 Conseil génétique - D005817 Diagnostic prénatal - D011296 Diagnostic préimplantatoire - D019836 Prise en charge en urgence - D004638 Acte médical Protocole de recherche - D035843 Education thérapeutique - D010353

		7.4	Profession du personnel réalisant l'activité	Type du personnel réalisant l'activité déclarée.	Assistante sociale - 10F40 Diététicien(ne) - 05I10 Ergothérapeute - 05I60 Kinésithérapeute - 05I30 Psychologue - 05O10 Psychomotricien(ne) - 05I70 Conseiller(e) en génétique - 05O20 Infirmier(e) - 05C10 Orthophoniste - 05I20 Enseignant(e) spécialisé(e) Médecin Autre professionnel(le) - LA4732-9
		7.5	Personnel réalisant l'activité	Nom du professionnel réalisant l'activité renseignée.	Chaîne de caractères
8. Histoire de la maladie	Information nécessaire pour calculer le délai au diagnostic, l'errance du diagnostic	8.1	Age aux premiers signes	Age auquel les premiers symptômes sont apparus ?	Anténatal A la naissance A l'âge de XX an(s) et/ou XX mois Non déterminé
		8.1 bis	Précision de l'âge aux premiers signes	Age auquel les premiers symptômes sont apparus ?	Numérique
		8.2	Appréciation du diagnostic à l'entrée du centre	Le diagnostic du (de la) patient(e) à son arrivée dans le centre Maladies Rares est-il approprié ?	Absent - LA4489-6 Non-approprié - LA9045-1 Approprié - LA15290-2
		8.3	Age au diagnostic	Age au diagnostic ?	Anténatal A la naissance A l'âge de XX an(s) et/ou XX mois Non déterminé
		8.3 bis	Précision de l'âge au diagnostic	Age au diagnostic ?	Numérique
9. Diagnostic	Information nécessaire pour toute étude épidémiologique ou d'identification de patients MR pour la recherche clinique	9.1	Statut actuel du diagnostic	Quel est le statut du diagnostic ?	En cours - LA9040-2 Probable - LA12746-6 Confirmé - LA15290-2 Non déterminé - LA4489-6 Inclassable
		9.2	Diagnostic de la maladie rare	Diagnostic du (de la) patient(e), évalué dans le Centre MR.	Code Orphanet
		9.3	Signes complémentaires particuliers ou inhabituels associés à la MR	Quel(s) signe(s) complémentaire(s) ou inhabituel(s) associé(s) au diagnostic de la maladie rare ? (évalué par le professionnel)	Code HPO
		9.4	Cas sporadique ou familial	Le cas est-il isolé ou familial au moment de l'observation (évalué par le professionnel de santé) ?	Sporadique Familial

10. Confirmation du diagnostic	Informations nécessaires pour apprécier les techniques utilisées pour établir des diagnostics MR	10.1	Mode de confirmation du diagnostic	Type(s) de méthode(s) de confirmation du diagnostic utilisée(s).	Clinique - D010808 Génétique moléculaire - D008967 Cytogénétique - D020732 Biochimique - D001671 Biologique - D001695 Imagerie - D003952 Autre - LA4732-9
		10.2 cond(10.1)	Méthode biologique sur laquelle repose le diagnostic	Préciser la méthode biologique sur laquelle repose le diagnostic (si applicable).	Chromosomique (caryotype, FISH) Array-CGH Séquençage ciblé (Sanger) Séquençage de nouvelle génération (NGS) Autres méthodes
	Information nécessaire pour toute étude épidémiologique ou d'identification de patients MR pour la recherche clinique	10.3 cond(10.1)	Mutation	Quelle(s) est (sont) la (les) mutation(s) en cause ?	Nomenclature prenant en compte le génotype chez l'individu (Mutnomen : http://www.hgvs.org/mutnomen/)
		10.4 cond(10.1)	Sujet apparemment sain	Le sujet, porteur de la mutation, est-il apparemment sain ? (porteur sain)	Oui - LA33-6 Non - LA32-8
11. Traitement	Appréciation de l'utilisation de traitements orphelins spécifiques ou patients inclus dans protocoles de recherche	11.1	Un traitement spécifique à la MR est-il en cours?	Un traitement spécifique à la maladie rare est-il en cours ? Nota bene : Les traitements dits "de confort" ne sont pas pris en compte ici.	Oui - LA33-6 Non - LA32-8
		11.2 cond(11.1)	Traitement en cours pour la MR	Nom du traitement spécifique en cours pour la maladie rare. Seuls les traitements maladies rares sont pris en compte ici.	Travail en cours
12. Anté et néonatales	Elements nécessaires pour l'établissement d'études nationales néonatales	12.1	Assistance médicale à la procréation	Le (la) patient(e) est-il (elle) né(e) suite à un programme d'assistance médicale à la procréation ?	Oui - LA33-6Non - LA32-8
		12.2	Présence de malformation anténatale	Le (la) patient(e) présentait-il (elle) une malformation anténatale ?	Non - LA32-8 Unique Multiple - D000015
		12.3	Né à terme?	Le (la) patient(e) est-il (elle) né(e) au terme de la grossesse ?	Oui - LA33-6 Non - LA32-8
		12.3 bis	Précision du terme (le cas échéant)	Préciser le terme en cas d'accouchement avant le terme prévu	Numérique
		12.4	Taille à la naissance	Taille du (de la) patient(e) à la naissance.	Numérique
		12.5	Poids à la naissance	Poids du (de la) patient(e) à la naissance.	Numérique
		12.6	Périmètre crânien à la naissance	Périmètre crânien du (de la) patient(e) à la naissance.	Numérique
		12.7	Fœtopathologie	Un examen foetopathologique a-t-il été réalisé ?	Oui (avec ou sans autopsie) - LA33-6 Non - LA32-8

13. Recherche	Informations générales concernant la recherche	13.1	Patient participant à un protocole	Le (la) patient(e) participe-t-il(elle) actuellement à un protocole de recherche (cohorte, essai thérapeutique,…)?	Oui - LA33-6 Non - LA32-8
		13.2	Accord pour être contacté pour un protocole	Le (la) patient(e) donne-t-il (elle) son accord pour être contacté(e) dans le cadre de la mise en œuvre d'un protocole de recherche ?	Oui - LA33-6 Non - LA32-8
		13.3	Patient ayant précédemment donné un échantillon biologique pour la recherche MR	Le (la) patient(e) a-t-il (elle) déjà donné un échantillon biologique pour la recherche ?	Oui - LA33-6 Non - LA32-8
		13.4	Patient ayant précédemment donné un échantillon biologique pour le diagnostic moléculaire	Le (la) patient(e) a-t-il (elle) déjà donné un échantillon biologique pour un diagnostic moléculaire ?	Oui - LA33-6 Non - LA32-8
		13.5	Lien avec une biobanque	*Un travail est en cours avec la plateforme nationale Biobanque pour définir les modalités d'échange entre les deux bases.*	
Annuaire **Les données suivantes seront entrées par l'administration du système**					
A. Réseau de soins maladies rares	Annuaire interne constitué en amont pour identifier les CRMR, CCMR et les filières ainsi que les responsables	A.1	Nom filière	Nom de la filière maladies rares.	Chaîne de caractères
		A.2	Nom du coordonnateur de la filière	Nom du coordonnateur de la filière maladies rares.	Chaîne de caractères
		A.3	Groupe de maladies se rapportant à la filière	Nom des différentes maladies rares se rapportant à la filière.	Liste de maladies rares
		A.4	Nom du centre de référence	Nom du centre de référence maladies rares.	Chaîne de caractères
		A.5	Nom(s) du(des) coordonnateur(s) du CRMR	Nom du ou des coordonnateurs du centre de référence maladies rares.	Chaîne de caractères
B. Réseau soins hospitalier		B.1	Nom de l'unité de soins	Nom de l'unité de soins (une unité de soin peut être représentée par un pôle, un département, un service, une UF au sein d'un établissement hospitalier)	Chaîne de caractères
		B.2	Rattachement de l'unité au réseau MR	Une unité de soin peut être rattachée en tant que centre de compétence auprès d'un ou plusieurs centres de référence ou bien, en tant que unité de soin membre d'un centre de référence. Un laboratoire de diagnostic est rattaché à un ou plusieurs centres de référence dans le cadre de la filière.	-
		B.2.1	Type de l'unité de soins	Nature de l'unité de soins (référence, compétence, laboratoire).	Référence Compétence Laboratoire de diagnostic

		B.2.2	Centre(s) de référence d'appartenance	Nom du (des) centre(s) de référence de rattachement.	Chaîne de caractères
		B.3	Nom du correspondant de l'US	Nom du correspondant de l'unité de soins.	Chaîne de caractères
		B.4	Pays de l'US	Pays dans lequel l'unité de soins est implantée.	Code Pays
		B.5	Commune de l'US	Commune dans laquelle l'unité de soins est implantée.	Chaîne de caractères
		B.6	Adresse de l'US	Adresse postale de l'unité de soins.	Chaîne de caractères
		B.7	Année d'entrée de l'US dans la banque de données	Année d'entrée de l'unité de soins dans BaMaRa.	Année

ANNEXE 2 : STANDARDISATION HL7 FHIR DU SET MINIMAL DE DONNEES MALADIES RARES

Patient resource: Demographics and other administrative information about a person or animal receiving care or other health-related services.

Extension?	Element	Use	Sub-element(s)	Definition	Coded?	Code	Data Type	Value Set	Code System
x	consent						boolean	boolean	
x	nonOpp						boolean	boolean	
x	tutorConsent						boolean	boolean	
	identifier	usual		The identifier recommended for display and use in real-world interactions.			string	string	
	identifier	official		The identifier considered to be most trusted for the identification of this item.			string	string	
	identifier						string	string	
	identifier	secondary		An identifier that was assigned in secondary use - it serves to identify the object in a relative context, but cannot be consistently assigned to the same object again in a different context.			string	string	
	name	maiden	family	A name used prior to marriage. Marriage naming customs vary greatly around the world. This name use is for use by applications that collect and store "maiden" names. Though the concept of maiden name is often gender specific, the use of this term is not gender specific. The use of this term does not imply any particular history for a person's name, nor should the maiden name be determined algorithmically.			string	string	
	name	usual	family	Known as/conventional/the one you normally use.			string	string	
	name	usual	given	Given name. Aliases first name; middle name			string	string	
	birthdate			The date and time of birth for the individual.			dateTime	dateTime	
	gender			Administrative Gender - the gender that the patient is considered to have for administration and record keeping purposes.			Codeable Concept	VS_35	HL7 administrative gender
x	fœtus						boolean	boolean	
x							Codeable Concept	INSEE city code (liste des communes)	INSEE cities

x	birthCountry						Codeable Concept	ISO code 3166-1 alpha-2	ISO countries
	address		city	Addresses for the individual. The name of the city, town, village or other community or delivery center.			Codeable Concept	INSEE city code	INSEE cities
	address		country	Addresses for the individual. Country - a nation as commonly understood or generally accepted.			Codeable Concept	ISO code 3166-1 alpha-2	ISO countries
	deceasedBoolean			Indicates if the individual is deceased or not.			boolean	boolean	
	deceasedDatetime			Indicates if the individual is deceased or not.			dateTime	dateTime	
x	deceasedRD						Codeable Concept	VS_Xbool	
x	deceasedCause						Codeable Concept	CIM10	ATIH
x	sentBy						Codeable Concept	VS_61	
x	inclusionDate						dateTime	dateTime	
	careProvider		reference	Patient's nominated care provider.			string	string	
	managingOrganization			Organization that is the custodian of the patient record.			string	string	

Family History resource: Significant health events and conditions for people related to the subject relevant in the context of care for the subject.

Extension?	Element	Use	Sub-element(s)	Definition	Coded?	Code	Data Type	Value Set	Code System
x	propositusBoolean						boolean	boolean	
x	propositusId						string	string	
	relation		relationship	The type of relationship this person has to the patient (father, mother, brother etc.).			Codeable Concept	VS_42	HL7 Role Code
x	inbreeding						Codeable Concept	VS_Xbool	

Encounter resource: An interaction between a patient and healthcare provider(s) for the purpose of providing healthcare service(s) or assessing the health status of a patient.

Extension?	Element	Use	Sub-element(s)	Definition	Coded?	Code	Data Type	Value Set	Code System
	identifier	temp		Identifier(s) by which this encounter is known.					

	period		start	The start and end time of the encounter. / The start of the period. The boundary is inclusive.			dateTime	dateTime	
	type			Specific type of encounter (e.g. e-mail consultation, surgical day-care, skilled nursing, rehabilitation).			Codeable Concept	VS_72	
x	objective						Codeable Concept	VS_73	
x	participantProfession						Codeable Concept	VS_74	
	participant		individual/reference	Persons involved in the encounter other than the patient.			string	string	
	location		location	The location where the encounter takes place.			string	string	
	serviceProvider			Department or team providing care.			string	string	

Condition resource: Use to record detailed information about conditions, problems or diagnoses recognized by a clinician. There are many uses including: recording a Diagnosis during an Encounter; populating a problem List or a Summary Statement, such as a Discharge Summary.

Extension?	Element	Use	Sub-element(s)	Definition	Coded?	Code	Data Type	Value Set	Code System
	Identifier	temp		This records identifiers associated with this condition that are defined by business processed and/ or used to refer to it when a direct URL reference to the resource itself is not appropriate					
x	firstsSignsOcc						Codeable Concept	VS_Age	
x	ageSigns						integer	integer	
x	earlyDiagnosisSatus						Codeable Concept	VS_82	
x	onset						Codeable Concept	VS_Age	
	onsetAge			Estimated or actual date the condition began, in the opinion of the clinician. Age is generally used when the patient reports an age at which the Condition began to occur.			integer	integer	
	status			The clinical status of the condition.			Codeable Concept	VS_91	
	code			Identification of the condition, problem or diagnosis.			Codeable Concept	ORPHA code	LORD
	evidence		code	A manifestation or symptom that led to the recording of this condition.			Codeable Concept	HPO, CIM10	
x	heredity						Codeable Concept	VS_94	

x	confirmationMode						Codeable Concept	VS_101	
x	biologicMethod						Codeable Concept	VS_102	
x	mutation						string		
x	carrierStatus						boolean	boolean	

Medication resource: Primarily used for identification and definition of Medication, but also covers ingredients and packaging.

Extension?	Element	Use	Sub-element(s)	Definition	Coded?	Code	Data Type	Value Set	Code System
x	treatmentRD						boolean	boolean	
	code			A code (or set of codes) that identify this medication. Usage note: This could be a standard drug code such as a drug regulator code, RxNorm code, SNOMED CT code, etc. It could also be a local formulary code, optionally with translations to the standard drug codes.			Codeable Concept	Orphan drugs	

Questionnaire resource: A structured set of questions and their answers. The Questionnaire may contain questions, answers or both. The questions are ordered and grouped into coherent subsets, corresponding to the structure of the grouping of the underlying questions.

Extension?	Element	Use	Sub-element(s)	Definition	Coded?	Code	Data Type	Value Set	Code System
	question	name			c	MAP	boolean	boolean	
	question	name			c	AM	Codeable Concept	VS_122	
	question	name			c	BAT	boolean	boolean	
	question	name			c	term	integer	integer	
	question	name			c	8302-2	float	float	
	question	name			c	83141-9	float	float	
	question	name			c	8287-5	float	float	
	question	name			c	foetopath	boolean	boolean	
	question	name			c	inProtocol	boolean	boolean	
	question	name			c	contactAllowed	boolean	boolean	

	question	name			c	sampleResearch	boolean	boolean	
	question	name			c	sampleMolDiagnosis	boolean	boolean	
	question	name			c	biobankLink	abandoned		

ANNEXE 3 : JEUX DE VALEURS ET STANDARDISATION

VS_35 – Sexe du patient [un seul choix possible]

Value set HL7 FHIR: http://hl7.org/fhir/vs/administrative-gender

Code	Libellé	Lien externe	Ressource externe
F	féminin	LA3-6	LOINC
M	masculin	LA2-8	LOINC
UN	indéterminé	LA18959-9	LOINC
UNK	inconnu	LA4489-6	LOINC

VS_Xbool – Booléen étendu [un seul choix possible]

Code	Libellé	Lien externe	Ressource externe
Y	oui	LA33-6	LOINC
N	non	LA32-8	LOINC
UNK	inconnu	LA4489-6	LOINC

VS_61 – Patient adressé par [plusieurs choix possible]

Code	Libellé	Lien externe	Ressource externe
VLM	venu de lui-même	---	---
ASS	association de patients	---	---
GNR	généraliste	---	---
PEV	pédiatre ville	---	---
PEH	pédiatre hôpital	---	---
PMI	centre de protection maternelle et infantile	---	---
GEN	généticien	---	---
GYO	gynéco/obstétricien	---	---
SPE	autre spécialiste	---	---
CPC	centre de prise en charge	---	---
CPD	centre pluridisciplinaire de diagnostic prénatal	---	---
CDP	centre de dépistage prénatal	---	---
A	autre	---	---

VS_42 – Lien de parenté avec le propositus [un seul choix possible]

Value set HL7 FHIR: http://hl7.org/fhir/v3/RoleCode

Code	Libellé	Lien externe	Ressource externe
NBRO	Frère	LA10415-0	LOINC
NSIS	Soeur	LA10418-4	LOINC
NFTH	Père	LA10416-8	LOINC
NMTH	Mère	LA10417-6	LOINC
SIGOTHR	Conjoint Conjointe	D018454	MESH
SON	Fils	LA10426-7	LOINC
DAU	Fille	LA10405-1	LOINC
GRNDSON	Petit-fils	LA10407-7	LOINC
GRNDDAU	Petite-fille	LA10406-9	LOINC
HBRO	Demi-frère	LA10408-5	LOINC
HSIS	Demi-sœur	LA10409-3	LOINC
PUNCLE	Oncle Paternel	LA10425-9	LOINC
MUNCLE	Oncle Maternel	LA10414-3	LOINC
PAUNT	Tante paternelle	LA10421-8	LOINC
MAUNT	Tante maternelle	LA10410-1	LOINC
PCOUSN	Cousin(e) paternel(le)	LA10422-6	LOINC
MCOUSN	Cousin(e) maternel(le)	LA10411-9	LOINC
NEPHEW	Neveu	LA10419-2	LOINC
NIECE	Nièce	LA10420-0	LOINC
PGRFTH	Grand-père paternel	LA10423-4	LOINC
PGRMTH	Grand-mère paternelle	LA10424-2	LOINC
MGRFTH	Grand-père maternel	LA10412-7	LOINC
MGRMTH	Grand-mère maternelle	LA10413-5	LOINC

VS_72 – Contexte de l'activité [un seul choix possible]

Code	Libellé	Lien externe	Ressource externe
CON	consultation	D009819	MESH
CPD	consultation pluridisciplinaire	---	---
HDJ	hôpital de jour	D003631	MESH

HTR	hospitalisation traditionnelle	D006760	MESH
ASD	avis sur dossier en consultation	---	---
EXP	avis d'expertise sur un dossier	D005104	MESH
AES	avis en salle	---	---
TLC	téléconsultation	D019114	MESH
A	autre contexte	LA4732-9	MESH

VS_73 – Objectif de l'activité [plusieurs choix possibles]

Code	Libellé	Lien externe	Ressource externe
DIA	diagnostic	D003933	MESH
PEC	mise en place de la prise en charge	D019468	MESH
SUI	suivi	D003266	MESH
CGE	conseil génétique	D005817	MESH
DPR	diagnostic prénatal	D011296	MESH
DPI	diagnostic préimplantatoire	D019836	MESH
URG	prise en charge en urgence	D004638	MESH
ACT	acte médical	---	---
PRO	protocole de recherche	D035843	MESH
EDU	Education thérapeutique	D010353	MESH

VS_74 – Profession du personnel réalisant l'activité [plusieurs choix possibles]

Code	Libellé	Lien externe	Ressource externe
ASS	Assistante sociale	10F40	NMHPNM[6]
DIE	Diététicien(ne)	05I10	NMHPNM
ERG	Ergothérapeute	05I60	NMHPNM
KIN	Kinésithérapeute	05I30	NMHPNM
PSY	Psychologue	05O10	NMHPNM
PSM	Psychomotricien(ne)	05I70	NMHPNM
CEG	Conseiller(e) en génétique	05O20	NMHPNM

[6] Nomenclature des Métiers Hospitaliers – Personnels Non Médicaux

INF	Infirmier	05C10	NMHPNM
ORT	Orthophoniste	05I20	NMHPNM
ESP	Enseignant(e) spécialisé(e)	---	---
MED	Médecin	---	---
A	Autre professionnel(le)	LA4732-9	LOINC

VS_Age [un seul choix possible]

Code	Libellé	Lien externe	Ressource externe
ANA	anténatal	---	---
NAI	à la naissance	---	---
PNA	à l'âge de	---	---
UNK	non déterminé	---	---

VS_82 - Appréciation du diagnostic à l'entrée du centre [un seul choix possible]

Code	Libellé	Lien externe	Ressource externe
ABS	absent	LA4489-6	LOINC
NAP	non approprié	LA9045-1	LOINC
APP	approprié	LA15290-2	LOINC

VS_91 - Statut actuel du diagnostic [un seul choix possible]

Value set HL7 FHIR: http://hl7.org/implement/standards/fhir/condition-status.html

Code	Libellé	Lien externe	Ressource externe
ONG	en cours	LA9040-2	LOINC
PRO	probable	LA12746-6	LOINC
CON	confirmé	LA15290-2	LOINC
UNK	non déterminé	LA4489-6	LOINC
UNC	non classable	---	---

VS_94 - Cas sporadique ou familial [un seul choix possible]

Code	Libellé	Lien externe	Ressource externe
SPO	sporadique	---	---
FAM	familial	---	---

VS_101 – Mode de confirmation du diagnostic [plusieurs choix possibles]

Code	Libellé	Lien externe	Ressource externe
CLI	clinique	D010808	MESH
GMO	génétique moléculaire	D008967	MESH
CYG	cytogénétique	D020732	MESH
BCH	biochimique	D001671	MESH
BIO	biologique	D001695	MESH
IMA	imagerie	D003952	MESH
A	autre	LA4732-9	LOINC

VS_102 - Méthode biologique sur laquelle repose le diagnostic [plusieurs choix possibles]

Code	Libellé	Lien externe	Ressource externe
CHR	chromosomique	---	---
ARR	array-CGH	---	---
SCI	séquençage ciblé	---	---
NGS	séquençage de nouvelle génération (NGS)	---	---
A	autre	---	---

VS_122 - Présence de malformation anténatale [un seul choix possible]

Code	Libellé	Lien externe	Ressource externe
NON	non	LA32-8	LOINC
UNI	unique	---	---
MUL	multiple	D000015	MESH

TABLE DES ILLUSTRATIONS

Figure 1 Superposition de l'organisation MR à l'organisation hospitalière et complexité des liens entre les UF et les CRMR et CCMR.....14

Figure 2 Objectifs nationaux de la BNDMR.....18

Figure 3 Structure organisationnelle du projet BNDMR en 2016.....20

Figure 4 Organisation du système d'information BaMaRa-BNDMR et ses interactions avec les systèmes nationaux et locaux existants.....24

Figure 5 Hétérogénéité des objectifs et des types des données recueillies des bases de données déclarées pour l'enquête auprès des CRMR.....25

Figure 6 Les trois niveaux du cadre d'interopérabilité pour les maladies rares.....26

Figure 7 Identification des verrous de recherche du cadre d'interopérabilité pour les maladies rares.....29

Figure 8 Procédure FOIN (Kalam et al. 2004).....38

Figure 9 Evolution du nombre de collisions d'identités en fonction des données nominatives....45

Figure 10 Distribution des longueurs des prénoms et des noms des patients dans le base de données de CEMARA.....48

Figure 11 Processus général de génération de l'IdMR.....50

Figure 12 Exemples de collisions.....54

Figure 13 Méthodologie de mise en place du set minimal de données maladies rares.....69

Figure 14 Alignement entre les éléments du set minimal de données maladies rares et les éléments de données communs du GRDR (Choquet et al. 2014).....73

Figure 15 Exemple d'une document CDA avec un corps structuré contenant un élément « administration de substance » et un élément « observation »....76

Figure 16 Exemple de la ressource Patient faisant partie de la section Identification....77

Figure 17 l'alignement à trois niveaux entre les 2 versions d'Orphanet....91

Figure 18 Exemple sur la différence de granularité entre les terminologies ORPHANET et ADICAP 93

Figure 19 Alignement croisé entre les éléments de données de type booléen et les éléments de valeurs de type énuméré....106

Figure 20 Schéma général migration de CEMARA vers BaMaRa....115

Figure 21 Schéma général de la reprise des données des registres et autres bases de données 116

TABLE DES TABLEAUX

Tableau 1 Liste des identifiants....42

Tableau 2 Tableau de substitution des caractères alphabétiques 47

Tableau 3 Correspondances entre les données recueillies pour un fœtus et les données servant au calcul de son identifiant....52

Tableau 4 Résultats d'évaluation de l'algorithme de génération des identifiants....55

Tableau 5 Estimation du nombre maximal d'identités différentes en entrée de l'algorithme....59

Tableau 6 Liste des terminologies les plus utilisées notamment dans le domaine des maladies rares 78

Tableau 7 Distribution des éléments de données et des extensions créées selon les ressources FHIR sélectionnées....82

Tableau 8 Liste des éléments de données codables et leurs jeux de valeurs ou terminologies de référence....82

Tableau 9 Résultats de l'alignement automatisé des schémas de données de CEMARA et de BaMaRa....94

Tableau 10 Exemples de correspondances formalisées 102

Tableau 11 Les expressions de correspondances générées 106

Tableau 12 Résultats des quatre alignements et validation....108

Tableau 13 Evaluation comparative: nombre de bonnes correspondances détectées par le test de référence et par l'expérimentation 108

Interopérabilité des données médicales dans le domaine des maladies rares dans un objectif de santé publique

Résumé

La santé se digitalise et de multiples projets d'e-santé ne cessent de se développer. Dans le contexte des maladies rares, un champ qui est devenu parmi les priorités de la stratégie de santé publique en France, l'e-santé pourrait constituer une solution pour améliorer les connaissances sur l'épidémiologie des maladies rares et pour proposer par la suite une meilleure prise en charge des patients. La Banque Nationale de Données Maladies Rares (BNDMR) propose de centraliser la conduite de ces études épidémiologiques pour toutes les maladies rares et tous les patients, atteints de ces maladies, suivis dans le système de soin français. La BNDMR doit se développer au sein d'un paysage numérique dense et hétérogène. Développer l'interopérabilité de la BNDMR constitue l'objectif des travaux de cette thèse.

Comment identifier les patients, incluant les fœtus ? Comment fédérer les identités des patients pour éviter les doublons ? Comment chainer des données de patients pour permettre la conduite des études ? En réponse à ces questions, nous proposons une méthode universelle d'identification des patients et qui respecte les contraintes de protection des données de santé.

Quelles données doivent être recueillies dans la BNDMR ? Comment améliorer et faciliter la mise en place d'une interopérabilité entre ces données et celles qui sont issues du large éventail des systèmes existants ? En réponse à ces questions, nous proposons d'abord de standardiser le recueil d'un set minimal de données pour toutes les maladies rares. L'implémentation de standards internationaux assure ainsi un premier pas vers l'interopérabilité. Nous proposons par la suite d'aller à la découverte de correspondances entre les données de sources hétérogènes. Minimiser l'intervention humaine en adoptant des techniques d'alignement automatisé et rendre fiables et exploitables les résultats de ces alignements ont constitué les principales motivations de notre proposition.

Mots clés : Interopérabilité, intégration de données, identification patient, maladies rares, standardisation, alignement automatisé, découverte de correspondances

Interoperability of medical data for the rare diseases field in a public health objective

Abstract

The digitalization of healthcare is on and multiple e-health projects are unceasingly coming up. In the rare diseases context, a field that has become a public health policy priority in France, e-health could be a solution to improve rare diseases epidemiology and to propose a better care for patients. The national data bank for rare diseases (BNDMR) offers the centralization of these epidemiological studies conduction for all rare diseases and all affected patients followed in the French healthcare system. The BNDMR must grow in a dense and heterogeneous digital landscape. Developing the BNDMR interoperability is the objective of this thesis' work.

How to identify patients, including fetuses? How to federate patients' identities to avoid duplicates creation? How to link patients' data to allow studies' conduction? In response to these questions, we propose a universal method for patients' identification that meets the requirements of health data protection.

Which data should be collected in the national data bank? How to improve and facilitate the development of interoperability between these data and those from the wide range of the existing systems? In response to these questions, we first propose the collection of a standardized minimum data set for all rare diseases. The implementation of international standards provides a first step toward interoperability. We then propose to move towards the discovery of mappings between heterogeneous data sources. Minimizing human intervention by adopting automated alignment techniques and making these alignments' results reliable and exploitable were the main motivations of our proposal.

Key words: Interoperability, data integration, patient identification, rare diseases, standardization, automated alignment, mappings detection