UNIVERSITÉ DE LORRAINE FACULTÉ DE MÉDECINE DE NANCY
2017

THÈSE

pour obtenir le grade de

DOCTEUR EN MÉDECINE

Présentée et soutenue publiquement
dans le cadre du troisième cycle de Médecine Spécialisée
par

Maxime WACK

le 6 octobre 2017

Installation d'un entrepôt de données cliniques pour la recherche au CHRU de Nancy

Déploiement technique, intégration et gouvernance des données

Membres du jury :
Président :
M. JAY Nicolas, Professeur
Juges :
M. MARIE Pierre-Yves, Professeur
Mme. AGRINIER Nelly, Maître de Conférence
Mme. JANNOT Anne-Sophie, Maître de Conférence

UNIVERSITÉ DE LORRAINE FACULTÉ DE MÉDECINE DE NANCY
2017

THÈSE

pour obtenir le grade de

DOCTEUR EN MÉDECINE

Présentée et soutenue publiquement
dans le cadre du troisième cycle de Médecine Spécialisée
par

Maxime WACK

le 6 octobre 2017

Installation d'un entrepôt de données cliniques pour la recherche au CHRU de Nancy

Déploiement technique, intégration et gouvernance des données

Membres du jury :
Président :
M. JAY Nicolas, Professeur
Juges :
M. MARIE Pierre-Yves, Professeur
Mme. AGRINIER Nelly, Maître de Conférence
Mme. JANNOT Anne-Sophie, Maître de Conférence

4 septembre 2017

Président de l'Université de Lorraine :
Professeur Pierre MUTZENHARDT

Doyen de la Faculté de Médecine
Professeur Marc BRAUN

Vice-doyens
Pr Karine ANGIOI-DUPREZ, Vice-Doyen
Pr Marc DEBOUVERIE, Vice-Doyen

Assesseurs :
Premier cycle : Pr Guillaume GAUCHOTTE
Deuxième cycle : Pr Marie-Reine LOSSER
Troisième cycle : Pr Marc DEBOUVERIE

Innovations pédagogiques : Pr Bruno CHENUEL

Formation à la recherche : Dr Nelly AGRINIER

Affaires juridiques et Relations extérieures : Dr Frédérique CLAUDOT
Vie Facultaire et SIDES : Pr Laure JOLY
Relations Grande Région : Pr Thomas FUCHS-BUDER

Chargés de mission

Bureau de docimologie : Dr Guillaume VOGIN
Commission de prospective facultaire : Pr Pierre-Edouard BOLLAERT
Orthophonie : Pr Cécile PARIETTI-WINKLER
PACES : Dr Mathias POUSSEL
Plan Campus : Pr Bruno LEHEUP
International : Pr Jacques HUBERT

==========

DOYENS HONORAIRES

Professeur Jean-Bernard DUREUX - Professeur Jacques ROLAND - Professeur Patrick NETTER - Professeur Henry COUDANE

==========

PROFESSEURS HONORAIRES

Etienne ALIOT - Jean-Marie ANDRE - Alain AUBREGE - Gérard BARROCHE Alain BERTRAND - Pierre BEY
Marc-André BIGARD - Patrick BOISSEL – Pierre BORDIGONI - Jacques BORRELLY - Michel BOULANGE
Jean-Louis BOUTROY – Serge BRIANÇON - Jean-Claude BURDIN - Claude BURLET - Daniel BURNEL - Claude CHARDOT
Jean-François CHASSAGNE - François CHERRIER Jean-Pierre CRANCE - Gérard DEBRY - Emile de LAVERGNE
Jean-Pierre DESCHAMPS - Jean DUHEILLE - Jean-Bernard DUREUX - Gilbert FAURE - Gérard FIEVE Bernard FOLIGUET
Jean FLOQUET - Robert FRISCH - Alain GAUCHER - Pierre GAUCHER - Professeur Jean-Luc GEORGE - Alain GERARD
Hubert GERARD - Jean-Marie GILGENKRANTZ - Simone GILGENKRANTZ - Gilles GROSDIDIER - Oliéro GUERCI
Philippe HARTEMANN Gérard HUBERT - Claude HURIET Christian JANOT - Michèle KESSLER - François KOHLER
Jacques LACOSTE - Henri LAMBERT - Pierre LANDES - Marie-Claire LAXENAIRE - Michel LAXENAIRE - Alain LE FAOU
Jacques LECLERE - Pierre LEDERLIN - Bernard LEGRAS - Jean-Pierre MALLIÉ - Philippe MANGIN – Jean-Claude MARCHAL
– Yves MARTINET Pierre MATHIEU - Michel MERLE - Pierre MONIN Pierre NABET – Patrick NETTER - Jean-Pierre NICOLAS -
Pierre PAYSANT Francis PENIN - Gilbert PERCEBOIS Claude PERRIN - Luc PICARD - François PLENAT - Jean-Marie POLU
Jacques POUREL - Jean PREVOT - Francis RAPHAEL - Antoine RASPILLER – Denis REGENT - Michel RENARD
Jacques ROLAND - Daniel SCHMITT - Michel SCHMITT - Michel SCHWEITZER - Daniel SIBERTIN-BLANC - Claude SIMON
Danièle SOMMELET - Jean-François STOLTZ -Michel STRICKER - Gilbert THIBAUT - Gérard VAILLANT - Paul VERT
Hervé VESPIGNANI - Colette VIDAILHET - Michel VIDAILHET - Jean-Pierre VILLEMOT - Michel WEBER

==========

PROFESSEURS ÉMÉRITES

Professeur Etienne ALIOT - Professeur Gérard BARROCHE - Professeur Serge BRIANÇON - Professeur Jean-Pierre CRANCE
Professeur Gilbert FAURE - Professeur Bernard FOLIGUET – Professeur Alain GERARD - Professeur Gilles GROSDIDIER
Professeur Philippe HARTEMANN - Professeur François KOHLER - Professeur Alain LE FAOU - Professeur Jacques LECLERE
Professeur Yves MARTINET – Professeur Patrick NETTER - Professeur Jean-Pierre NICOLAS – Professeur Luc PICARD -
Professeur François PLENAT - Professeur Jean-François STOLTZ

==========

PROFESSEURS DES UNIVERSITÉS - PRATICIENS HOSPITALIERS

(Disciplines du Conseil National des Universités)

42ème Section : MORPHOLOGIE ET MORPHOGENÈSE

1ère sous-section : ***(Anatomie)***
Professeur Marc BRAUN – Professeure Manuela PEREZ
2ème sous-section : ***(Histologie, embryologie et cytogénétique)***
Professeur Christo CHRISTOV
3ème sous-section : ***(Anatomie et cytologie pathologiques)***
Professeur Jean-Michel VIGNAUD – Professeur Guillaume GAUCHOTTE

43ème Section : BIOPHYSIQUE ET IMAGERIE MÉDICALE

1ère sous-section : ***(Biophysique et médecine nucléaire)***
Professeur Gilles KARCHER – Professeur Pierre-Yves MARIE – Professeur Pierre OLIVIER
2ème sous-section : ***(Radiologie et imagerie médicale)***
Professeur René ANXIONNAT - Professeur Alain BLUM - Professeur Serge BRACARD - Professeur Michel CLAUDON
Professeure Valérie CROISÉ-LAURENT - Professeur Jacques FELBLINGER - Professeur Pedro GONDIM TEIXEIRA

44ème Section : BIOCHIMIE, BIOLOGIE CELLULAIRE ET MOLÉCULAIRE, PHYSIOLOGIE ET NUTRITION

1ère sous-section : ***(Biochimie et biologie moléculaire)***
Professeur Jean-Louis GUEANT - Professeur Bernard NAMOUR - Professeur Jean-Luc OLIVIER
2ème sous-section : ***(Physiologie)***
Professeur Christian BEYAERT - Professeur Bruno CHENUEL - Professeur François MARCHAL
4ème sous-section : ***(Nutrition)***
Professeur Didier QUILLIOT - Professeure Rosa-Maria RODRIGUEZ-GUEANT - Professeur Olivier ZIEGLER

45ème Section : MICROBIOLOGIE, MALADIES TRANSMISSIBLES ET HYGIÈNE

1ère sous-section : ***(Bactériologie – virologie ; hygiène hospitalière)***
Professeur Alain LOZNIEWSKI – Professeure Evelyne SCHVOERER
2ème sous-section : ***(Parasitologie et Mycologie)***
Professeure Marie MACHOUART
3ème sous-section : ***(Maladies infectieuses ; maladies tropicales)***
Professeur Thierry MAY - Professeure Céline PULCINI - Professeur Christian RABAUD

46ème Section : SANTÉ PUBLIQUE, ENVIRONNEMENT ET SOCIÉTÉ

1ère sous-section : ***(Épidémiologie, économie de la santé et prévention)***
Professeur Francis GUILLEMIN - Professeur Denis ZMIROU-NAVIER
3ème sous-section : ***(Médecine légale et droit de la santé)***
Professeur Henry COUDANE
4ème sous-section : ***(Biostatistiques, informatique médicale et technologies de communication)***
Professeure Eliane ALBUISSON - Professeur Nicolas JAY

47ème Section : CANCÉROLOGIE, GÉNÉTIQUE, HÉMATOLOGIE, IMMUNOLOGIE

1ère sous-section : ***(Hématologie ; transfusion)***
Professeur Pierre FEUGIER
2ème sous-section : ***(Cancérologie ; radiothérapie)***
Professeur Thierry CONROY - Professeur François GUILLEMIN - Professeur Didier PEIFFERT - Professeur Frédéric MARCHAL
3ème sous-section : ***(Immunologie)***
Professeur Marcelo DE CARVALHO-BITTENCOURT - Professeure Marie-Thérèse RUBIO
4ème sous-section : ***(Génétique)***
Professeur Philippe JONVEAUX - Professeur Bruno LEHEUP

48ème Section : ANESTHÉSIOLOGIE, RÉANIMATION, MÉDECINE D'URGENCE, PHARMACOLOGIE ET THÉRAPEUTIQUE

1ère sous-section : ***(Anesthésiologie-réanimation)***
Professeur Gérard AUDIBERT - Professeur Hervé BOUAZIZ - Professeur Thomas FUCHS-BUDER
Professeure Marie-Reine LOSSER - Professeur Claude MEISTELMAN

2ème sous-section : ***(Réanimation)***
Professeur Pierre-Édouard BOLLAERT - Professeur Sébastien GIBOT - Professeur Bruno LÉVY

3ème sous-section : ***(Pharmacologie fondamentale ; pharmacologie clinique ; addictologie)***
Professeur Pierre GILLET - Professeur Jean-Yves JOUZEAU

4ème sous-section : ***(Thérapeutique ; addictologie)***
Professeur François PAILLE - Professeur Patrick ROSSIGNOL – Professeur Faiez ZANNAD

49ème Section : PATHOLOGIE NERVEUSE ET MUSCULAIRE, PATHOLOGIE MENTALE, HANDICAP ET RÉÉDUCATION

1ère sous-section : ***(Neurologie)***
Professeur Marc DEBOUVERIE - Professeur Louis MAILLARD - Professeur Luc TAILLANDIER - Professeure Louise TYVAERT

2ème sous-section : ***(Neurochirurgie)***
Professeur Jean AUQUE - Professeur Thierry CIVIT - Professeure Sophie COLNAT-COULBOIS - Professeur Olivier KLEIN

3ème sous-section : ***(Psychiatrie d'adultes ; addictologie)***
Professeur Jean-Pierre KAHN - Professeur Raymund SCHWAN

4ème sous-section : ***(Pédopsychiatrie ; addictologie)***
Professeur Bernard KABUTH

5ème sous-section : ***(Médecine physique et de réadaptation)***
Professeur Jean PAYSANT

50ème Section : PATHOLOGIE OSTÉO-ARTICULAIRE, DERMATOLOGIE ET CHIRURGIE PLASTIQUE

1ère sous-section : ***(Rhumatologie)***
Professeure Isabelle CHARY-VALCKENAERE - Professeur Damien LOEUILLE

2ème sous-section : ***(Chirurgie orthopédique et traumatologique)***
Professeur Laurent GALOIS - Professeur Didier MAINARD - Professeur Daniel MOLE - Professeur François SIRVEAUX

3ème sous-section : ***(Dermato-vénéréologie)***
Professeur Jean-Luc SCHMUTZ

4ème sous-section : ***(Chirurgie plastique, reconstructrice et esthétique ; brûlologie)***
Professeur François DAP - Professeur Gilles DAUTEL - Professeur Etienne SIMON

51ème Section : PATHOLOGIE CARDIO-RESPIRATOIRE ET VASCULAIRE

1ère sous-section : ***(Pneumologie ; addictologie)***
Professeur Jean-François CHABOT - Professeur Ari CHAOUAT

2ème sous-section : ***(Cardiologie)***
Professeur Edoardo CAMENZIND - Professeur Christian de CHILLOU DE CHURET - Professeur Yves JUILLIERE
Professeur Nicolas SADOUL

3ème sous-section : ***(Chirurgie thoracique et cardiovasculaire)***
Professeur Thierry FOLLIGUET - Professeur Juan-Pablo MAUREIRA

4ème sous-section : ***(Chirurgie vasculaire ; médecine vasculaire)***
Professeur Sergueï MALIKOV - Professeur Denis WAHL – Professeur Stéphane ZUILY

52ème Section : MALADIES DES APPAREILS DIGESTIF ET URINAIRE

1ère sous-section : ***(Gastroentérologie ; hépatologie ; addictologie)***
Professeur Jean-Pierre BRONOWICKI - Professeur Laurent PEYRIN-BIROULET

3ème sous-section : ***(Néphrologie)***
Professeur Luc FRIMAT - Professeure Dominique HESTIN

4ème sous-section : ***(Urologie)***
Professeur Pascal ESCHWEGE - Professeur Jacques HUBERT

53ème Section : MÉDECINE INTERNE, GÉRIATRIE, CHIRURGIE GÉNÉRALE ET MÉDECINE GÉNÉRALE

1ère sous-section : ***(Médecine interne ; gériatrie et biologie du vieillissement ; addictologie)***
Professeur Athanase BENETOS - Professeur Jean-Dominique DE KORWIN - Professeure Gisèle KANNY
Professeure Christine PERRET-GUILLAUME – Professeur Roland JAUSSAUD – Professeure Laure JOLY

2ème sous-section : ***(Chirurgie générale)***
Professeur Ahmet AYAV - Professeur Laurent BRESLER - Professeur Laurent BRUNAUD

3ème sous-section : ***(Médecine générale)***
Professeur Jean-Marc BOIVIN – Professeur Paolo DI PATRIZIO

54ème Section : DÉVELOPPEMENT ET PATHOLOGIE DE L'ENFANT, GYNÉCOLOGIE-OBSTÉTRIQUE, ENDOCRINOLOGIE ET REPRODUCTION

1ère sous-section : ***(Pédiatrie)***
Professeur Pascal CHASTAGNER - Professeur François FEILLET - Professeur Jean-Michel HASCOET
Professeur Emmanuel RAFFO - Professeur Cyril SCHWEITZER
2ème sous-section : ***(Chirurgie infantile)***
Professeur Pierre JOURNEAU - Professeur Jean-Louis LEMELLE
3ème sous-section : ***(Gynécologie-obstétrique ; gynécologie médicale)***
Professeur Philippe JUDLIN - Professeur Olivier MOREL
4ème sous-section : ***(Endocrinologie, diabète et maladies métaboliques ; gynécologie médicale)***
Professeur Bruno GUERCI - Professeur Marc KLEIN - Professeur Georges WERYHA

55ème Section : PATHOLOGIE DE LA TÊTE ET DU COU
1ère sous-section : ***(Oto-rhino-laryngologie)***
Professeur Roger JANKOWSKI - Professeure Cécile PARIETTI-WINKLER
2ème sous-section : ***(Ophtalmologie)***
Professeure Karine ANGIOI - Professeur Jean-Paul BERROD
3ème sous-section : ***(Chirurgie maxillo-faciale et stomatologie)***
Professeure Muriel BRIX

==========

PROFESSEURS DES UNIVERSITÉS

61ème Section : GÉNIE INFORMATIQUE, AUTOMATIQUE ET TRAITEMENT DU SIGNAL
Professeur Walter BLONDEL

64ème Section : BIOCHIMIE ET BIOLOGIE MOLÉCULAIRE
Professeure Sandrine BOSCHI-MULLER - Professeur Pascal REBOUL

65ème Section : BIOLOGIE CELLULAIRE
Professeure Céline HUSELSTEIN

==========

PROFESSEUR ASSOCIÉ DE MÉDECINE GÉNÉRALE

Professeur associé Sophie SIEGRIST

==========

MAÎTRES DE CONFÉRENCES DES UNIVERSITÉS - PRATICIENS HOSPITALIERS

42ème Section : MORPHOLOGIE ET MORPHOGENÈSE
1ère sous-section : ***(Anatomie)***
Docteur Bruno GRIGNON
2ème sous-section : ***(Histologie, embryologie et cytogénétique)***
Docteure Chantal KOHLER

43ème Section : BIOPHYSIQUE ET IMAGERIE MÉDICALE
1ère sous-section : ***(Biophysique et médecine nucléaire)***
Docteur Antoine VERGER (stagiaire)
2ème sous-section : ***(Radiologie et imagerie médicale)***
Docteur Damien MANDRY

44ème Section : BIOCHIMIE, BIOLOGIE CELLULAIRE ET MOLÉCULAIRE, PHYSIOLOGIE ET NUTRITION
1ère sous-section : ***(Biochimie et biologie moléculaire)***
Docteure Shyue-Fang BATTAGLIA - Docteure Sophie FREMONT - Docteure Isabelle AIMONE-GASTIN
Docteure Catherine MALAPLATE-ARMAND - Docteur Marc MERTEN - Docteur Abderrahim OUSSALAH
2ème sous-section : ***(Physiologie)***
Docteure Silvia DEMOULIN-ALEXIKOVA - Docteur Mathias POUSSEL – Docteur Jacques JONAS (stagiaire)
3ème sous-section : ***(Biologie Cellulaire)***
Docteure Véronique DECOT-MAILLERET

45ème Section : MICROBIOLOGIE, MALADIES TRANSMISSIBLES ET HYGIÈNE
1ère sous-section : ***(Bactériologie – Virologie ; hygiène hospitalière)***
Docteure Corentine ALAUZET - Docteure Hélène JEULIN - Docteure Véronique VENARD
2ème sous-section : (Parasitologie et mycologie)
Docteure Anne DEBOURGOGNE

46ème Section : SANTÉ PUBLIQUE, ENVIRONNEMENT ET SOCIÉTÉ
1ère sous-section : ***(Epidémiologie, économie de la santé et prévention)***
Docteure Nelly AGRINIER - Docteur Cédric BAUMANN - Docteure Frédérique CLAUDOT - Docteur Alexis HAUTEMANIÈRE
2ème sous-section ***(Médecine et Santé au Travail)***
Docteure Isabelle THAON
3ème sous-section ***(Médecine légale et droit de la santé)***
Docteur Laurent MARTRILLE

47ème Section : CANCÉROLOGIE, GÉNÉTIQUE, HÉMATOLOGIE, IMMUNOLOGIE
1ère sous-section : ***(Hématologie ; transfusion)***
Docteure Aurore PERROT – Docteur Julien BROSEUS
2ème sous-section : ***(Cancérologie ; radiothérapie)***
Docteure Lina BOLOTINE – Docteur Guillaume VOGIN
4ème sous-section : ***(Génétique)***
Docteure Céline BONNET

48ème Section : ANESTHÉSIOLOGIE, RÉANIMATION, MÉDECINE D'URGENCE, PHARMACOLOGIE ET THÉRAPEUTIQUE
2ème sous-section : ***(Réanimation ; Médecine d'urgence)***
Docteur Antoine KIMMOUN
3ème sous-section : ***(Pharmacologie fondamentale ; pharmacologie clinique ; addictologie)***
Docteur Nicolas GAMBIER - Docteure Françoise LAPICQUE - Docteur Julien SCALA-BERTOLA
4ème sous-section : ***(Thérapeutique ; Médecine d'urgence ; addictologie)***
Docteur Nicolas GIRERD

50ème Section : PATHOLOGIE OSTÉO-ARTICULAIRE, DERMATOLOGIE ET CHIRURGIE PLASTIQUE
1ère sous-section : ***(Rhumatologie)***
Docteure Anne-Christine RAT
3ème sous-section : ***(Dermato-vénéréologie)***
Docteure Anne-Claire BURSZTEJN
4ème sous-section : ***(Chirurgie plastique, reconstructrice et esthétique ; brûlologie)***
Docteure Laetitia GOFFINET-PLEUTRET

51ème Section : PATHOLOGIE CARDIO-RESPIRATOIRE ET VASCULAIRE
3ème sous-section : ***(Chirurgie thoracique et cardio-vasculaire)***
Docteur Fabrice VANHUYSE

52ème Section : MALADIES DES APPAREILS DIGESTIF ET URINAIRE
1ère sous-section : ***(Gastroentérologie ; hépatologie ; addictologie)***
Docteur Jean-Baptiste CHEVAUX – Docteur Anthony LOPEZ (stagiaire)

53ème Section : MÉDECINE INTERNE, GÉRIATRIE, CHIRURGIE GÉNÉRALE ET MÉDECINE GÉNÉRALE
2ème sous-section : (Chirurgie générale)
Docteur Cyril PERRENOT (stagiaire)
3ème sous-section : ***(Médecine générale)***
Docteure Elisabeth STEYER

54ème Section : DEVELOPPEMENT ET PATHOLOGIE DE L'ENFANT, GYNECOLOGIE-OBSTETRIQUE, ENDOCRINOLOGIE ET REPRODUCTION
5ème sous-section : ***(Biologie et médecine du développement et de la reproduction ; gynécologie médicale)***
Docteure Isabelle KOSCINSKI

55ème Section : PATHOLOGIE DE LA TÊTE ET DU COU
1ère sous-section : ***(Oto-Rhino-Laryngologie)***
Docteur Patrice GALLET

==========
MAÎTRES DE CONFÉRENCES

5ème Section : SCIENCES ÉCONOMIQUES
Monsieur Vincent LHUILLIER

7ème Section : SCIENCES DU LANGAGE : LINGUISTIQUE ET PHONETIQUE GENERALES
Madame Christine DA SILVA-GENEST

19ème Section : SOCIOLOGIE, DÉMOGRAPHIE
Madame Joëlle KIVITS

64ème Section : BIOCHIMIE ET BIOLOGIE MOLÉCULAIRE
Madame Marie-Claire LANHERS - Monsieur Nick RAMALANJAONA

65ème Section : BIOLOGIE CELLULAIRE
Madame Nathalie AUCHET - Madame Natalia DE ISLA-MARTINEZ - Monsieur Jean-Louis GELLY - Madame Ketsia HESS
Monsieur Hervé MEMBRE - Monsieur Christophe NEMOS

66ème Section : PHYSIOLOGIE
Monsieur Nguyen TRAN

==========

MAÎTRES DE CONFÉRENCES ASSOCIÉS DE MÉDECINE GÉNÉRALE

Docteur Pascal BOUCHE – Docteur Olivier BOUCHY - Docteur Arnaud MASSON – Docteur Cédric BERBE
Docteur Jean-Michel MARTY

==========

DOCTEURS HONORIS CAUSA

Professeur Charles A. BERRY (1982)
Centre de Médecine Préventive, Houston (U.S.A)
Professeur Pierre-Marie GALETTI (1982)
Brown University, Providence (U.S.A)
Professeure Mildred T. STAHLMAN (1982)
Vanderbilt University, Nashville (U.S.A)
Professeur Théodore H. SCHIEBLER (1989)
Institut d'Anatomie de Würtzburg (R.F.A)
Université de Pennsylvanie (U.S.A)
Professeur Mashaki KASHIWARA (1996)
Research Institute for Mathematical Sciences de Kyoto (JAPON)

Professeure Maria DELIVORIA-PAPADOPOULOS (1996)
Professeur Ralph GRÄSBECK (1996)
Université d'Helsinki (FINLANDE)
Professeur Duong Quang TRUNG (1997)
Université d'Hô Chi Minh-Ville (VIÊTNAM)
Professeur Daniel G. BICHET (2001)
Université de Montréal (Canada)
Professeur Marc LEVENSTON (2005)
Institute of Technology, Atlanta (USA)

Professeur Brian BURCHELL (2007)
Université de Dundee (Royaume-Uni)
Professeur Yunfeng ZHOU (2009)
Université de Wuhan (CHINE)
Professeur David ALPERS (2011)
Université de Washington (U.S.A)
Professeur Martin EXNER (2012)
Université de Bonn (ALLEMAGNE)

Dédicaces

Au Président du Jury

À notre Maître et Président de Thèse

M. le professeur Nicolas Jay
Professeur de Biostatistiques, Informatique Médicale et Technologies de Communication

Nous vous remercions d'avoir accepté la direction et la présidence de cette thèse.

Vous nous avez encadré, accompagné, et soutenu pendant une majeure partie de notre internat, et fourni un environnement accueillant et propice au développement des compétences utiles à l'exercice que nous convoitons.

Tout au long de notre parcours, nous avons grandement apprécié votre aide et votre disponibilité.

Soyez assuré de notre plus profond respect et notre sincère gratitude.

Aux Juges

À notre Juge

M. le Professeur Pierre-Yves Marie
Professeur de Biophysique et Médecine Nucléaire

Nous sommes honorés que vous ayez accepté de juger ce travail.

Soyez assurée de notre reconnaissance.

À notre Maître et Juge

Mme le Docteur Nelly Agrinier
Maître de Conférence des Universités, Praticien Hospitalier

Nous vous remercions d'avoir accepté de juger ce travail.

Tout au long de notre parcours, et avant même le début de notre internat de santé publique, vous nous avez soutenu, accompagné et encouragé.

Soyez assurée de notre reconnaissance.

À notre Maître et Juge

Mme. le Docteur Anne-Sophie Jannot
Maître de Conférence des Universités, Praticien Hospitalier

Nous vous remercions d'avoir accepté de juger ce travail.

Nous espérons que notre future collaboration sera fructueuse.

Soyez assurée de notre reconnaissance.

À ma famille

À mes parents, à qui je n'ai cessé de donner des frayeurs, merci d'avoir toujours été là quand j'en ai eu besoin, et de ne pas avoir renoncé ou abandonné. Merci de m'avoir apporté un environnement idéal pour grandir en faisant ce qui me plaisait.

À mes frères et sœurs, merci pour l'animation, les bagarres, l'apprentissage de la vie, c'était cool de grandir avec vous.

À Benjamin, le lièvre après lequel j'ai longtemps couru et qui m'a appris tant de choses, je te dois en grande partie mon goût pour l'informatique. Merci à toi d'avoir ouvert la voie.

À Abel, je sais que tu iras très loin, continue comme ça !

À Fantine, Sidonie, Salomé, Imanol, Juliette (et les suivants ?), vous aussi vous irez loin je n'en doute pas, mais faut bien que je fasse un peu de favoritisme avec mon filleul !

J, tu me pousses à ne cesser de m'améliorer et ne pas (trop) céder à la complaisance. Merci d'être là.

À Nao, je ne désespère pas que tu apprennes un jour à lire (si ce n'est pas déjà le cas) pour voir cette dédicace. Merci pour ta compagnie et ton affection indéfectible.

À Tomtom et Etienne, ben ouais, famille, vous êtes mes bros !

À mes amis

À mes co-internes de santé publique, Aurélie, Jonathan, Margaux, Matthieu, Marion, Diane, Benjamin et Laurie, merci pour tous les bons moments passés ensembles ! Vive le SPIN !

À Hélène, Ngoc-Ha, Arnaud, Clément, Vianney et Yoann, je compte sur vous pour continuer à bien représenter la santé publique nancéienne !

À Benjamin, Justine, Martin et Morgane, merci pour la bonne ambiance en épidémiologie.

À mes amis de Boston, Antoine, Céline, Claire et Manu, un jour on arrivera à bosser tous ensemble !

À Paul, pour avoir façonné le début de mon parcours, et offert une opportunité en or.

Scott, thank you for introducing me to the love of my life.

À mes amis d'enfance, Tomtom, Etienne, Jean, Jul', c'est avec bonheur que je partage votre amitié depuis maintenant 30 ans, et ça c'est inestimable. À Juju, je te remercierai quand tu arrêteras de me spoiler mes animés !

À mes potes de médecine, Da, Charly, Charles, Macaire, Clémence, Panda, Bobby, Thomas, Nono, Stef, Gravier, et les autres, merci de m'avoir accueilli sans question quand je suis réapparu, c'est cool.

À tous les médecins et personnels des services par lesquels je suis passé, merci pour votre accueil, et d’avoir supporté ma compagnie pas forcément toujours très calme.

SERMENT

« Au moment d'être admis à exercer la médecine, je promets et je jure d'être fidèle aux lois de l'honneur et de la probité. Mon premier souci sera de rétablir, de préserver ou de promouvoir la santé dans tous ses éléments, physiques et mentaux, individuels et sociaux. Je respecterai toutes les personnes, leur autonomie et leur volonté, sans aucune discrimination selon leur état ou leurs convictions. J'interviendrai pour les protéger si elles sont affaiblies, vulnérables ou menacées dans leur intégrité ou leur dignité. Même sous la contrainte, je ne ferai pas usage de mes connaissances contre les lois de l'humanité. J'informerai les patients des décisions envisagées, de leurs raisons et de leurs conséquences. Je ne tromperai jamais leur confiance et n'exploiterai pas le pouvoir hérité des circonstances pour forcer les consciences. Je donnerai mes soins à l'indigent et à quiconque me les demandera. Je ne me laisserai pas influencer par la soif du gain ou la recherche de la gloire.

Admis dans l'intimité des personnes, je tairai les secrets qui me sont confiés. Reçu à l'intérieur des maisons, je respecterai les secrets des foyers et ma conduite ne servira pas à corrompre les mœurs. Je ferai tout pour soulager les souffrances. Je ne prolongerai pas abusivement les agonies. Je ne provoquerai jamais la mort délibérément.

Je préserverai l'indépendance nécessaire à l'accomplissement de ma mission. Je n'entreprendrai rien qui dépasse mes compétences. Je les entretiendrai et les perfectionnerai pour assurer au mieux les services qui me seront demandés.

J'apporterai mon aide à mes confrères ainsi qu'à leurs familles dans l'adversité.

Que les hommes et mes confrères m'accordent leur estime si je suis fidèle à mes promesses ; que je sois déshonoré et méprisé si j'y manque ».

Table des matières

Introduction

L'informatisation des différentes composantes du Système d'Information Hospitalier (SIH), ainsi que du dossier patient (DPI, Dossier Patient Informatisé, ou EHR, Electronic Health Record) depuis une vingtaine d'années, a permis la collection d'un grand nombre de données relatives aux soins.

En recherche clinique, la faisabilité d'une étude, la constitution de cohortes, la sélection de patients dans les études rétrospectives, et le recueil de données, même lorsqu'elles sont issues du soin, sont des tâches coûteuses en temps, en argent et en personnel.

La collection de ces données dans un système informatique de routine présente donc un grand intérêt pour la réutilisation de ces données dans le cadre de la recherche clinique, mais aussi pour la production d'indicateurs de santé ou de pilotage d'établissement.

Les données du SIH sont généralement accessibles dans le contexte de la prise en charge clinique, à travers divers logiciels métiers, et ce de manière dite «verticale» : il est possible d'accéder à toute l'information d'un patient dans chaque logiciel, voire l'ensemble des informations d'un patient dans un unique logiciel de DPI dans le cas d'une intégration des données provenant des applications métier.

L'exploitation à des fins de recherche ou d'orientation stratégique nécessite une vue «horizontale» des données, populationnelle, permettant d'accéder à une information pour l'ensemble ou un sous-ensemble de patients.

Peu après le développement des premiers systèmes de dossier patient informatisé sont apparues les premières solutions d'entrepôt de données (EDC, Entrepôt de Données Cliniques, ou CDW, Clinical Data Warehouse) permettant d'intégrer les données provenant du dossier informatisé et de produire des requêtes sur l'ensemble de ces données.

Les problématiques soulevées par l'implémentation de ces systèmes sont nombreuses, parmi lesquelles on retrouve la représentation des données, la structure des données permettant d'effectuer des requêtes «horizontales», l'intégration des données depuis les différentes sources disponibles. Le processus d'accès aux données via une interface utilisateur est également un élément déterminant pour l'adoption de l'entrepôt par les utilisateurs.

De nombreuses solutions logicielles sont apparues au cours des années, souvent développées de manière *ad hoc* pour cibler le système d'information d'un hôpital en particulier. D'autres ont privilégié une approche agnostique permettant l'adaptation et l'utilisation avec d'autres systèmes d'information dans d'autres centres.

L'une de ces solutions est la plateforme i2b2 (Informatics for Integrating Biology and the Bedside), développée en 2004 dans le cadre du projet NCBC (Nation Center for Biomedical Computing) sur un financement du NIH, par un partenariat entre Harvard Medical School, le MIT, et les hôpitaux affiliés à HMS.

i2b2 dérive de RPDR (Research Patient Data Registry), un système développé en 2000 à l'origine pour les hôpitaux Partners Healthcare System, qui permettait de faire des requêtes pour identifier des patients dans le système hospitalier.

L'objectif de i2b2 est de fournir un système d'entrepôt de données clinique orienté pour l'exploitation pour la recherche, notamment la recherche translationnelle permettant de lier les données de laboratoire (analyses biologiques, et de nouvelle génération : tests ADN, séquençage, etc.) aux données cliniques recueillies au chevet du patient.

Cette plateforme a été choisie du fait de son ubiquité (déployé dans plus de 200 centres à travers le monde), de l'effort de recherche l'accompagnant (plusieurs centaines de publications sur pubmed faisant usage d'i2b2), et de la disponibilité de la plateforme en tant que logiciel open source permettant son utilisation sans limitation (tant financière que technique).

Ce travail décrit l'installation de la plateforme i2b2 au CHRU de Nancy, comportant le déploiement technique, la représentation des données, l'intégration des données et les aspects concernant la gouvernance des données, et le respect de la vie privée et de l'anonymat des patients.

Installation d'un entrepôt de données cliniques pour la recherche au CHRU de Nancy

Déploiement technique, intégration et gouvernance des données

Maxime Wack

Abstract

L'apparition des dossiers patient informatisés et la collection croissante de données liées aux soins a permis l'émergence d'entrepôts de données pouvant être utilisés en recherche clinique pour identifier et retrouver les données liées à un groupe de patients.
La plateforme i2b2 choisie ici est à la fois un entrepôt et un outil de requêtes, disponible sous licence open source.
L'objectif était l'étude de la faisabilité du déploiement de la plateforme i2b2 au CHRU de Nancy, explorant la représentation et l'intégration des données du SIH, et les règles de gouvernance.
Un package R (R2b2) a été développé spécifiquement, permettant de gérer l'entrepôt, manipuler et intégrer les données.
Les sources de données existantes dans le SIH ont été étudiées pour leur intégration, et un modèle de représentation des données a été choisi (ou conçu lorsque cela était nécessaire) pour organiser les concepts.
Un dispositif a été mis en place pour protéger l'accès aux données, en miroir de la responsabilité médicale au CHRU et respectant les règles de collection et d'accès aux données de soins, permettant un retour direct aux praticiens.
Un prototype fonctionnel est fourni, intégrant les données de 118330 patients lors de 247580 hospitalisations depuis début 2016, totalisant plus de 18 millions d'observations.
Le travail conduit autour de ce prototype a permis de dégager plusieurs axes de conduite pour la mise en production du projet, et les règles régissant les accès.

Introduction

Le développement des technologies de l'information a mené au développement et à l'adoption par les hôpitaux de solutions logicielles pour recueillir un nombre croissant de types de données liées aux patients et aux soins, du PMSI (Programme de Médicalisation des Systèmes d'Information) au dossier patient informatisé (DPI), en passant par la numérisation des clichés radiologiques, examens biologiques, prescriptions, et toutes autres formes de mesures et observations médicales possibles.

La génération rapide d'un grand volume de données par chacun de ces systèmes a rapidement mené à la conception et au développement d'entrepôts de données cliniques, inspirés des solutions de gestion d'entreprise, permettant la réutilisation des données générées par le soin pour l'utilisation en épidémiologie et en recherche clinique, la création de systèmes d'aide à la décision, ou encore le pilotage d'établissement de santé.

De nombreuses solutions d'entrepôt de données cliniques ont été développées à travers le monde, comme à l'hôpital universitaire rattaché à SNU (Seoul National University) (1), dans les hôpitaux rattachés à l'université de Stanford (2), à New York (3), ou même en France au CHU de Rennes (4).

Cependant, si un grand nombre de projets ont fleuri au cours de la première décennie du XXIe siècle (6), le projet développé initialement pour les hôpitaux Partners HealthCare System affiliés à Harvard Medical School, intitulé RPDR (Research Patient Data Repository) (10), a donné lieu quelques années plus tard au développement de la plateforme i2b2 (Informatics for Integrating Biology and the Bedside), un projet financé par le NIH et conjoint entre Harvard Medical School et le MIT (12). Cette plateforme est à la fois un entrepôt de données, gérant la représentation des données et leur stockage, organisé de manière à pouvoir accueillir des modèles de données arbitraires, et un outil de requête permettant d'interroger les données contenues dans l'entrepôt pour identifier des patients répondant à certains critères, et extraire les données les concernant.

La plateforme est distribuée avec une licence open source permettant son utilisation gratuitement, favorisant son adoption par de nombreux centres (15, 16). C'est cette plateforme qui a été retenue pour ce travail.

L'objectif était l'étude de la faisabilité du déploiement d'un entrepôt de données cliniques pour la recherche au CHRU de Nancy, prenant en considération les axes suivants :

- aspects techniques du déploiement
- intégration et représentation des données provenant des diverses sources du SIH
- gouvernance des données et respect de la confidentialité

S'agissant d'un projet de faisabilité avec l'optique d'une mise en production et pérennisation de l'installation, une attention particulière a été donnée à la production d'une méthode reproductible et documentée.

Pour ce faire, un package R(17), R2b2 (https://github.com/MaximeWack/R2b2), a été développé avec pour but de faciliter la mise en œuvre de chacun de ces axes. Il est distribué sous licence libre GPL v3. Un objectif secondaire était de rendre ce package générique et non spécifique aux données et à l'infrastructure actuels du CHRU de Nancy afin de favoriser son extensibilité et sa réutilisation.

Matériel et Méthode

Déploiement technique

La plateforme retenue comme entrepôt de données cliniques, i2b2 (Informatics for Integrating Biology & the Bedside), a été téléchargée dans sa version 1.7.08b (13 décembre 2016) sous forme d'une machine virtuelle VMWare depuis www.i2b2.org.

Un partenariat avec la DSI (Direction des Systèmes Informatiques) a permis l'obtention d'un serveur pour héberger la machine virtuelle. Celle-ci est fournie pré installée, et pré configurée avec des données et des utilisateurs de démonstration.

Un script a été développé automatisant l'installation du package R2b2 et de ses dépendances, depuis la machine virtuelle telle que téléchargée.

Le package contient des fonctions de réinitialisation de l'instance, avec suppression des ontologies, données et utilisateurs présents à des fins de démonstration. Ces fonctions sont exécutées lors de la première installation du package, produisant une machine fonctionnelle vierge à partir de l'image officielle.

Ceci permet de s'affranchir d'une installation complète manuelle de la plateforme i2b2. En revanche, cela contraint le choix du système d'exploitation utilisé (CentOS 6) et du système de base de données (PostgreSQL 9.1).

L'architecture logicielle d'i2b2 est basée sur des applets java, chacune chargée d'une tâche particulière :

- **Hive** : orchestration des services entre eux
- **PM** (Project Management) : gestion des utilisateurs et projets
- **Metadata** (Ontology Management) : organisation de la représentation des données
- **IM** (Identity Management) : gestion de l'anonymisation des patients
- **CRC** (Clinical Research Chart / Data Repository) : hébergement et requêtes sur les données
- **Work** (Workflow Framework) : gestion des requêtes

Chaque applet, ou *cell* dans le langage i2b2 possède une base de données propre, et expose un service web permettant la communication par messages XML, comme illustré dans la **Figure 1.**

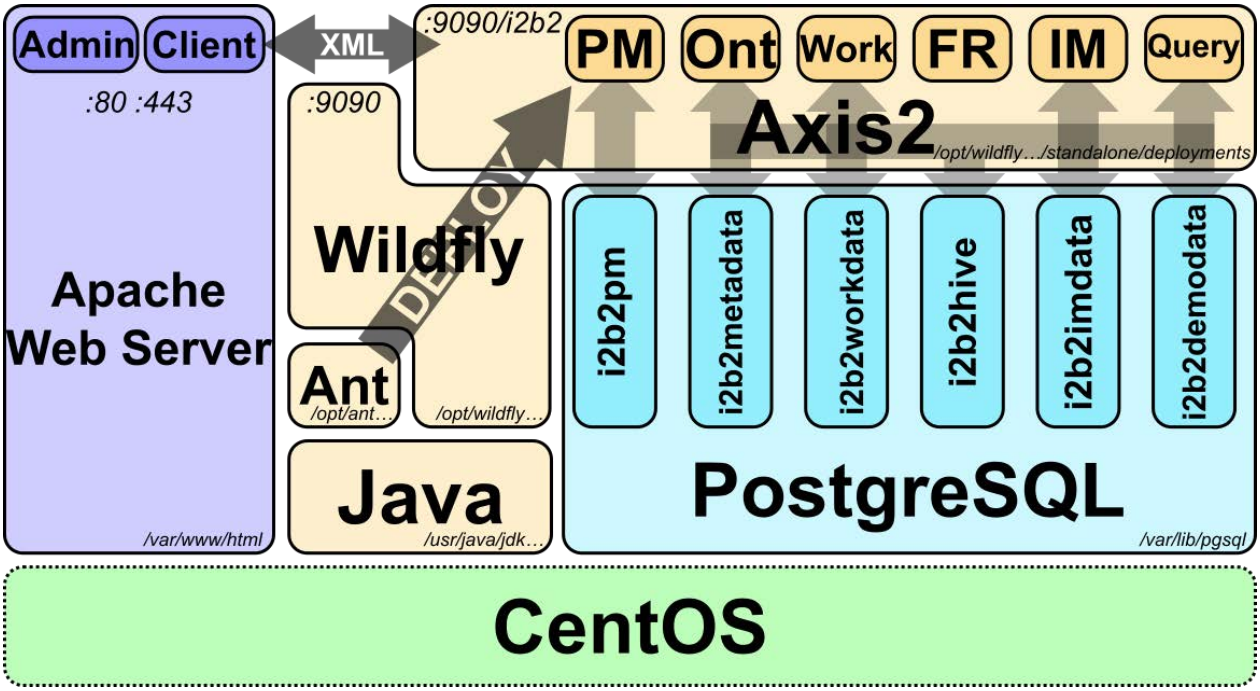


Figure 1: Architecture logicielle d'i2b2 et de ses cells

R2b2 interagit avec l'instance i2b2 majoritairement par un accès direct aux bases de données sous-jacente (via RPostgreSQL), mais implémente également une interface de communication XML avec les différentes *cells* (via xml2 et httr).

L'équipe i2b2 fournit deux outils d'administration de la machine i2b2 : une interface web d'administration pour la gestion des options, des projets, et des utilisateurs ainsi que de leurs droits ; et l'application i2b2 Workbench permettant d'explorer une instance de la même manière que l'interface web, et d'interagir avec le contenu de la machine : création et modification des ontologies, et chargement interactif de données à l'aide d'un plug-in dédié.

L'outil web d'administration est limité dans ses fonctionnalités et la capacité de personnalisation des diverses options de la plateforme, et occasionne des corruptions de la base de données. L'ajout de projets et utilisateurs ainsi que la gestion des droits se font manuellement, sans option de traitement automatisé ou de masse, rendant l'administration fastidieuse et source d'erreurs.

La création d'ontologies et l'import de données à l'aide de i2b2 Workbench se font également uniquement de manière interactive. La création d'ontologies et le chargement de données à l'aide des outils fournis ne proposent également qu'une représentation générique des données dans la base de données, qui n'est pas en mesure d'exploiter la flexibilité du modèle de représentation des données dans i2b2. (13)

R2b2 a été conçu comme une collection de fonctions de configuration de la machine par une interface programmable, permettant la création de scripts de déploiement et de gestion d'instances complexes. La création de projets (avec leur base de données respectives, expliqué plus en détail dans la section sur la gouvernance des données), la création d'utilisateurs et la gestion de leurs droits d'accès, l'ajout de types de données et l'intégration de données peuvent être gérées entièrement par le biais de R2b2.

L'ensemble des fonctions du package R2b2 est documenté en accord avec les règles de publication d'un package R, et un manuel d'administration d'une machine i2b2 à l'aide de R2b2 a été rédigé pour assurer la continuité de ce projet (**Annexe 1**).

Sources de données

La production des données médico-administratives et de soins au CHRU de Nancy sont à attribuer à plusieurs acteurs, avec des périmètres de responsabilité différents.

Les données médico-administratives du PMSI sont organisées par RUM (Résumé d'Unité Médicale), et rattachées à une unité médicale responsable du patient (qui peut être différente de l'unité médicale d'hébergement). Le DIM (Département d'Information Médicale) produit (via les TIM, Techniciens d'Information Médicale) les données de diagnostic CIM-10 (Classification Internationale des Maladies), et les praticiens des services produisent le codage des actes CCAM (Classification Commune des Actes Médicaux).

Les données recueillies au cours du soin dans le service sont organisées par RUM, le découpage des séjours dans le DPI suivant la logique PMSI pour les fins de traitement normal de l'information médicale, et rattachées à l'unité médicale responsable du patient. Les membres de l'unité médicale produisent cette information (principalement textuelle par les observations cliniques et comptes-rendus, mais aussi mesures morphométriques).

Les données de biologie, anatomopathologie et d'imagerie sont organisées par venue (équivalent RSS (Résumé Standardisé de Sortie) dans le cas d'une hospitalisation), et rattachées dans le DPI au découpage correspondant à un RUM en utilisant l'unité fonctionnelle de demande de l'examen. L'information est produite par les équipes des laboratoires et services d'imagerie.

L'informatisation du SIH (Système d'Information Hospitalier) au CHRU de Nancy est assurée par les logiciels suivants.

Les données de codage PMSI (diagnostics et actes) sont recueillies via l'application **WebPIMS V2** de WEB100T, avec des données disponibles depuis 2004.

Les données concernant les patients, mouvements et venues sont gérées par **GAM-GML**, et versées dans **WebPIMS** pour agrémenter la base de données.

La solution de DPI utilisée est la plateforme **DxCare** de Medasys, depuis 2015.

Les fonctionnalités implémentées dans **DxCare** sont la saisie des observations et courriers médicaux, la saisie des actes médicaux, et la visualisation des résultats d'examens biologiques. La saisie des actes est faite par les praticiens dans **DxCare** puis versée vers **WebPIMS.**

La prescription informatisée se fait via le logiciel **Pharma** de Computer Engineering, et les données de prescription ne sont, à l'heure actuelle, pas intégrées à **DxCare.**

La visualisation des clichés d'imagerie et les comptes-rendus associés sont gérés par les applications **OSA** de J4care et **Xplore**, et ne sont pas non plus encore intégrés à **DxCare.**

Les résultats d'examens de biologie (biologie «standard», microbiologie, biologies spécialisées) sont gérées en interne entre les différents laboratoires par le logiciel **GLIMS** de CliniSys Group. Les résultats sont ensuite versés dans **DxCare** pour consultation par les praticiens.

Les bases de données **WebPIMS** et **DxCare** sont accessibles via des «univers» de requête dans le logiciel **Business Objects XI (BO)**, qui expose les données via une organisation facilitant la découverte et l'exploration de ces bases. Toutes les données présentes dans les bases sous-

jacentes ne sont cependant pas nécessairement exposées (notamment, pour le DPI, les documents textuels de compte-rendu).

Il est prévu, pour la mise en production, une association avec la DSI et les responsables des diverses applications métier alimentant le SIH pour créer des flux d'intégration de données directs depuis les sources vers l'entrepôt.

Le prototype utilise des extractions de ces bases de données via **BO**. Les fichiers de requête fournis avec ce prototype serviront à définir les requêtes SQL pour interroger directement les bases Oracle de **BO** et mettre en place des flux d'intégration continus.

Les données de mouvement étant versées dans **WebPIMS** et les résultats d'analyses biologiques étant reproduits dans **DxCare**, il a été possible d'intégrer ces données par l'interrogation de ces seules bases.

Les sources de données intégrées dans ce prototype sont donc les suivantes :

- structure des services du CHRU
- données démographiques des patients (âge à la venue, sexe, date de naissance, date de décès, statut vital)
- données de diagnostics et actes PMSI
- résultats des examens biologiques

Intégration des données

Les données d'un entrepôt i2b2 résident dans la *cell* CRC dont l'organisation décrite dans (14) est un schéma en étoile (**Figure 2**).

La table centrale est celle contenant les observations uniques (un diagnostic, un acte, une mesure), liées chacunes à un patient sujet de l'observation, une visite/rencontre pendant laquelle a eu lieu l'observation, un concept décrivant l'observation, un intervenant ayant produit l'observation, et une date (accompagnée éventuellement d'une date de fin pour une observation non ponctuelle). Chacun de ces identifiants relie l'observation à d'autres tables contenant des informations détaillées pour chaque identifiant lié : les données propres au patient, à la visite/rencontre, etc.

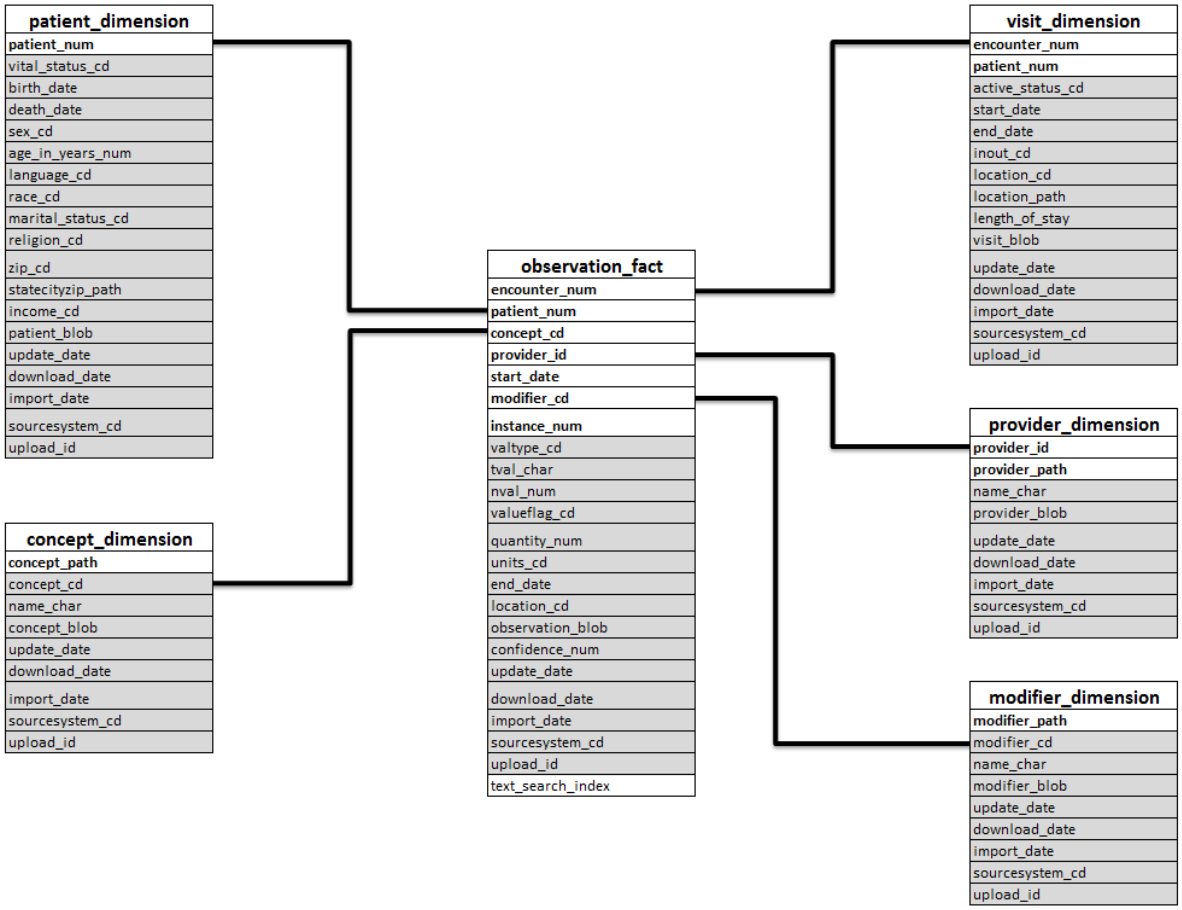


Figure 2: Schéma «en étoile» i2b2. En gras les champs contenus dans la clé primaire, en grisé les champs optionnels

Dans le cadre de l'implémentation au CHRU de Nancy, le concept de visite/rencontre (*visit/encounter*) pour i2b2 correspond à l'ensemble d'une venue (soit un RSS en ce qui concerne le PMSI, lors d'une hospitalisation), et l'intervenant (*provider_id*) au numéro d'unité médicale de prise en charge du patient. Dans une optique PMSI, un RUM isolé reste identifiable à l'aide des dates définissant la période d'observation d'un diagnostic CIM-10, l'intervenant lié renseignant l'unité médicale de ce RUM.

Ainsi les informations de codage PMSI sont enregistrées avec respect de leur granularité, avec associée pour chaque diagnostic CIM-10 une période d'application (correspondant au RUM), et pour chaque acte CCAM la date précise de réalisation de l'acte.

Les résultats de biologie intégrés sont, dans la table d'observations, liés à l'unité médicale responsable du patient ayant effectué la demande, et non le laboratoire ayant réalisé la mesure. Le numéro de visite/rencontre est celui de la venue du patient.

Seules les données d'hospitalisation sont intégrées dans ce prototype, les données recueillies en consultation et utilisables dans l'entrepôt dans son état actuel étant limitées aux seuls résultats biologiques, dont l'attribution à une unité médicale pourra être sujette à caution. Cependant, le choix de représenter une hospitalisation comme visite/rencontre permettra l'inclusion des consultations sous le même modèle (les consultations possédant également un numéro de venue, mais pas de numéro de RUM).

Un processus d'ETL (Extract, Transform and Load) a été développé pour intégrer les données du SIH à l'entrepôt.

L'extraction, dans le cadre de ce prototype, a été manuelle à partir des univers **BO** exposant les données. Les requêtes ont été conçues pour être facilement paramétrables et traduisibles ensuite vers des requêtes SQL planifiables.

Les requêtes sont parcimonieuses pour limiter le temps d'extraction, et de sorte à ce que leur résultat soit au plus proche du format de données accepté par i2b2, afin d'optimiser le processus d'intégration (par exemple, résultat biologique sans valeur, résultats non conformes, sont exclus dès l'extraction). De manière générale, les opérations de filtrage de données sont appliquées le plus tôt possible dans la chaîne de traitement afin d'optimiser la taille des fichiers et l'ensemble des traitements suivants (reformatage, alignement de terminologie, etc.). Les requêtes contiennent également une série de filtres garantissant que les données extraites pourront être intégrées à i2b2 (par exemple, contrainte de non nullité sur les champs correspondant à des variables obligatoires dans les schémas des tables i2b2).

Chaque type et source de données a donné lieu à une requête différente, afin de rendre le processus d'extraction modulaire et permettre facilement son extension pour de nouvelles sources de données, ou en cas de modification de changement de solution logicielle pour un type de donnée.

Les différentes requêtes créées à ce jour concernent donc l'extraction de :

- la liste des patients
- la liste des hospitalisations
- les données démographiques propres à chaque patient (date de naissance, sexe)
- les données patient propres à chaque hospitalisation (mensurations et âge du patient à l'hospitalisation)
- les diagnostics PMSI
- les actes PMSI
- les résultats d'examens biologiques

Ces requêtes étant spécifiques au SIH du CHRU de Nancy, elles n'ont pas été intégrées au package. Elles restent cependant disponibles et sous contrôle de version dans un dépôt annexe.

La transformation des données extraites vers un format compatible avec i2b2 étant également spécifique aux données du SIH du CHRU du Nancy, ces fonctions n'ont pas non plus été intégrées à R2b2, et sont disponibles dans le même dépôt annexe.

Ces fonctions ont été conçues pour garantir la qualité, la cohérence, et l'intégrité des données incluses dans l'entrepôt. Au niveau technique par la compatibilité avec les schémas de tables de données i2b2 (noms des variables, types de données, étiquetage des concepts avec le schéma d'ontologie lié (plus de détails sur cet étiquetage dans la section suivante sur la représentation des données)). Et au niveau du contenu des données par la vérification de la cohérence des informations (venue et patient présents dans la base de données, dates compatibles, non duplication des résultats).

Le chargement des données est assuré par trois fonctions : ajout de patients, ajouts de venues, et ajout d'observations.

La fonction d'ajout de patients se charge de remplir la table **patient_dimension,** certains champs étant construits à la volée (statut vital à partir de la date de décès, âge actuel à partir de la date de naissance et la date de décès si existante). La fonction gère également de manière transparente la transformation des identifiants de chaque patient (table **patient_mapping**) vers un identifiant i2b2 unique.

Le champ identifiant i2b2 (*patient_num*) étant contraint en base de données à un entier signé 32 bits, l'utilisation de fonctions de hachage cryptographiques pour l'anonymisation des identifiants patients était limitée. Deux solutions de transformation des identifiants pour pseudonymisation ont été implémentées : l'identité (qui autoriserait l'affichage directement dans l'interface des identifiants de patients pour les utilisateurs ayant le droit d'accès aux données détaillées), et la renumérotation des identifiants à partir de 1.

La fonction d'ajout de venues se charge de remplir la table **visit_dimension**. De la même manière que l'ajout de patients, la fonction gère la transformation d'identifiants de visite (table **encounter_mapping**) et la construction d'un champ à la volée (statut actif de la visite). Les numéros de venue au CHRU de Nancy étant parfois représentés par des entiers supérieurs à $2^{31} - 1$ (borne maximale d'un entier 32 bits signé), ou optionnellement contenir des caractères non numériques, il était impossible d'utiliser la fonction identité et la fonction de renumérotation a été utilisée.

La fonction d'ajout d'observations insère les observations dans la table **observation_fact**. La fonction gère la transformation des numéros de venue et identifiants de patients vers leur identifiants uniques i2b2. En cas de réel doublon (même patient, même venue, même concept, même intervenant, même date), la valeur de **instance_num** est incrémentée.

Afin d'optimiser les insertions en base de données et décharger R de la tâche de vérifier l'unicité des clés primaires, les données sont d'abord insérées en masse dans une table temporaire avec l'opération COPY. L'insertion définitive dans la table fait usage du moteur de PostgreSQL et implémente l'opération «d'upsert» (**up**date or in**sert**), inexistante par défaut dans postgreSQL 9.1. Cette fonction permet de gérer automatiquement la mise à jour de champs de la base de données sans créer de doublons ou violer les contraintes d'unicité des clés.

Ces fonctions permettent l'intégration des données avec vérification de leur cohérence avant insertion, la mise à jour incrémentielle des informations patients et l'ajout d'observations sans

création de doublon et prévenant la survenue des erreurs les plus communes lors de l'insertion en base de données.

Une stratégie d'intégration continue de données a été conçue utilisant ces mécanismes pour alimenter l'entrepôt avec les données les plus récentes tout en garantissant leur intégrité :

- import quotidien des nouvelles venues et des informations sur les hospitalisations en cours (sélection de données des hospitalisations dont la date d'entrée est inférieure ou égale à la date du jour, et la date de sortie supérieure ou égale à la date du jour)
- correction rétroactive hebdomadaire/mensuelle/annuelle : suppression de toutes les données liées aux venues de la semaine/mois/année précédente (suppression de l'ensemble des données des tables **visit_dimension** et **observation_fact** pour les venues comprenant au moins un jour de la semaine/mois/année passée), puis nouvelle extraction couvrant cette période et import des données consolidées.

Représentation des données

La représentation des données est gérée par la *cell* **Metadata** d'i2b2.

Pour chaque type de données intégré à la plateforme il faut fournir une «ontologie» représentant l'organisation des données, qui servira à l'utilisateur final à sélectionner les paramètres de ses requêtes.

Ces ontologies implémentent les relations taxonomiques (hyponymie, hyperonymie) et les synonymies.

L'intégration d'ontologies (préexistantes ou *ad hoc*) dans la plateforme i2b2 peut se faire à l'aide d'outils de traduction de format, manuellement via l'interface i2b2 Workbench, ou encore directement via la base de données liée à la *cell* **Metadata**.

Le format pour représenter ces ontologies dans la base de données est complexe et nécessite de préciser de nombreuses informations pour chaque concept.

La structure minimale d'une table de la base **i2b2metadata** décrivant une ontologie répond au format décrit dans la **Table 1**.

Table 1: Structure minimale de la table ***i2b2metadata***

Colonne	Usage	Valeur par défaut
c_hlevel	Profondeur hiérarchique	
c_fullname	Nom complet avec chemin d'accès	
c_name	Nom du concept	
c_visualattributes	Type de concept (catégorie ou concept)	
c_basecode	Code du concept	
c_metadataxml	Utilisé pour les données numériques	
c_facttablecolumn	Colonne à sélectionner dans **observation_fact**	concept_cd
c_tablename	Table contenant la colonne à sélectionner	concept_dimension
c_columnname	Colonne de référence	concept_path
c_operator	Opérateur de comparaison à utiliser	LIKE
c_dimcode	Valeur à laquelle comparer	chemin complet

i2b2 utilise les informations contenues dans cette table pour construire la requête de sélection de patients, de la manière suivante :

```
SELECT DISTINCT (patient_num)
  FROM observation_fact
  WHERE c_facttablecolumn IN
    SELECT c_facttablecolumn
      FROM c_tablename
      WHERE c_columnname c_operator c_dimcode
```

La manière la plus simple de définir une ontologie avec des concepts catégoriels est de remplir les colonnes *c_facttablecolumn*, *c_tablename*, *c_column*, *c_operator* et *c_dimcode* avec les

valeurs par défaut décrites dans la **Table 1**, permettant de requêter des concepts décrits dans la table **concept_dimension** (cf. **Figure 2**). Il est cependant possible, à l'aide des autres colonnes liant les tables entre elles dans le schéma en étoile, de créer des concepts interrogeant des informations présentes dans ces tables périphériques.

R2b2 implémente des fonctions d'importation d'ontologies utilisant un format simplifié, automatisant la création de la majorité des champs requis par la base de données : création des codes définissant l'arborescence, remplissage de la profondeur hiérarchique et des types, création du code de concept, etc.

Le format simplifié de base ne nécessite que de fournir les concepts finaux à utiliser avec leur chemin. Le package se charge de créer les tables et entrées dans **i2b2metadata** pour gérer cette ontologie, ainsi que le schéma «standard» de concepts renseignés dans la table **concept_dimension** de la *cell* **CRC**, avec les valeurs par défaut présentées dans la **Table 1.**

Chaque fichier d'ontologie peut être accompagné d'un fichier de modificateurs, qui sont utilisés pour préciser l'utilisation d'un concept (ex : spécifier si un diagnostic CIM-10 a été renseigné lors du codage PMSI comme Diagnostic Principal, Diagnostic Relié, Diagnostic Associé Significatif, ou Diagnostic Associé Documentaire), et d'un fichier de correspondances pour transformer des codes présents dans les fichiers sources vers des codes de l'ontologie.

Il est également possible pour chaque concept de spécifier le contenu des colonnes de base de données qui sont par défaut remplies automatiquement, pour obtenir des comportements plus complexes autorisés par i2b2 : lien vers des données présentes dans les tables périphériques, sélection selon les valeurs des concepts numériques (âge, valeurs biologiques)

Les terminologies utilisées comme ontologies i2b2 pour représenter les données sont les suivantes :

- **CIM 10** pour les diagnostics PMSI, avec des modificateurs pour le type de diagnostic renseigné
- **CCAM** pour les actes médicaux
- structure des pôles, services et unités médicales au CHRU de Nancy telle que définie institutionnellement
- terminologie locale pour représenter les données démographiques des patients
- terminologie locale pour représenter les résultats d'examens biologiques

Les données issues du PMSI sont des données existant déjà sous une forme structurée, utilisant la Classification Internationale des Maladies dans sa 10ème version (CIM 10) pour décrire les diagnostics et symptômes, et la Classification Commune des Actes Médicaux (CCAM) pour décrire les actes médicaux.

La CIM (ou ICD, International Classification of Diseases) est définie par l'OMS et utilisée internationalement. Tous les pays n'utilisent pas systématiquement la même version de l'ICD (par exemple les États-Unis ont utilisé l'ICD 9 jusqu'à octobre 2015), et la plupart des pays la modifient (en sus de la traduction) pour l'adapter à des particularités de fonctionnement, notamment liées à la facturation.

Ainsi, plutôt que d'utiliser des fichiers intégrables à i2b2 disponibles pour l'ICD 10 en version internationale, un script de conversion des fichiers fournis par l'ATIH vers le format accepté par

R2b2 a été développé. L'ensemble des niveaux hiérarchiques définis dans la CIM sont reproduits dans l'ontologie (par exemple, `C00-D48 Tumeurs / D10-D36 Tumeurs bénignes`).

Le fichier d'ontologie pour la CIM-10 est accompagné d'un fichier décrivant les modificateurs pour spécifier le type de diagnostic (DP, DR, DAS, DAD).

La CCAM est une nomenclature strictement française utilisée pour codifier la réalisation d'actes médicaux. De la même manière que pour la CIM-10, un script a été créé permettant de générer un fichier d'ontologie au format R2b2 à partir des fichiers officiels fournis par l'ATIH.

Ceci permettra la mise à jour de ces ontologies lors de la parution des nouvelles versions et modifications.

La structure des unités médicales au CHRU de Nancy est décrite par Pôles (cliniques et médicotechniques), Services, Unités Médicales (niveau de granularité pour le codage de l'information médicale), et Unités Fonctionnelles. Cette structure peut être représentée sous la forme d'un arbre enraciné.

L'ontologie décrivant la structure est définie directement à partir des fichiers de structure institutionnels, en incluant les niveaux de granularité du Pôle à l'UM (soit Pôle, Service, UM). Les concepts de la structure sont importés dans la table **provider_dimension**, et la table correspondante dans la *cell* **Metadata** utilise le mécanisme décrit ci-dessus pour référencer la table **provider_dimension** au lieu de la table **concept_dimension**, évitant la duplication des concepts.

Deux ontologies ont été créées pour accéder à une sélection de concepts renseignés dans les tables **patient_dimension** (sexe et statut vital), et **visit_dimension** (type de venue, hospitalisation en cours ou terminée). Cette dernière ontologie contient également des références vers la table **concept_dimension** pour les constantes renseignées au début à chaque hospitalisation (âge, poids, taille, et IMC à l'entrée de l'hospitalisation).

Les résultats d'examens biologiques ne sont pas représentés dans le SIH à l'aide d'une terminologie existante (comme c'est le cas pour les diagnostics et les actes), mais chacun des laboratoires a défini ses propres concepts. Les résultats étant extraits de **DxCare** et non du logiciel interne de gestion des résultats, ils sont organisés par leur représentation sous forme de «feuille de résultats».

L'exploration des bases de données a permis d'identifier 2302 concepts utilisés pour présenter les résultats biologiques, accompagnés de leur label, organisation et unité.

Une ontologie a été spécifiquement conçue dans le cadre de ce travail pour héberger ces concepts et les organiser. Le premier niveau hiérarchique de cette ontologie décrit le milieu source du prélèvement (sang, urine, moelle, LCR, liquides divers, etc.). Le second niveau hiérarchique décrit le type de bilan (hémogramme, ionogramme, exploration fonctionnelle, etc.). Lorsque c'est nécessaire pour préciser certaines catégories (par exemple le système exploré dans les explorations fonctionnelles), des niveaux supplémentaires de hiérarchie ont été créés.

Un travail d'harmonisation a été effectué pour unifier les concepts représentant des examens identiques (milieu, objet, et unité de mesure identiques), mais réalisés par différents laboratoires, résultant en un fichier d'alignement de terminologie associant 392 termes source à 241 termes

destination. Les concepts ne représentant pas des données directement liées au patient (données d'assurance qualité du matériel de mesure par exemple), et les concepts exprimés à l'aide de texte libre (résultats d'antibiogramme entre autres) ont été exclus.

Au total, après exclusion et alignement, 1320 termes représentant des résultats d'examen biologique ont été produits et organisés.

Cette ontologie est faiblement systématisée, et construite *ad hoc* pour les données présentes dans le SIH. Elle n'est pas une représentation systématique universelle des résultats d'examen biologiques. Elle reste cependant extensible pour accueillir de nouveaux types de résultats, et ayant été conçue à partir de la présentation des résultats telle qu'existant dans DxCare pour les praticiens, rend familier l'accès au résultat.

Gouvernance des données et respect de la confidentialité

Le cadre législatif actuel, s'il définit les règles entourant la conduite de recherches biomédicales sur les patients (que ce soit une recherche interventionnelle ou non interventionnelle, avec modification ou non de la prise en charge normale au cours du soin) et les règles de partage de données entre les professionnels de santé dans le cadre du soin, ne donne pas d'encadrement spécifique quant à la constitution d'un entrepôt de données et l'accès aux données par essence transversales présentes dans l'entrepôt.

La conduite de recherche non interventionnelle telle que décrite au 3° de l'article L1121-1 du CSP (21) (loi Jardé (22)), nécessite un avis favorable du Comité de Protection des Personnes (23). Les projets de recherche des utilisateurs de l'entrepôt devront ainsi être validés scientifiquement dès qu'il y aura extraction de données, même anonymes. Ce circuit (qui concerne également les thèses et mémoires des internes en poste dans les services) est déjà en place au CHRU de Nancy avec le soutien de la PARC (Plateforme d'Aide à la Recherche Clinique) et les autres services du pôle S2R (Structures de Soutien à la Recherche) pour la conduite de recherches biomédicales sur les patients.

L'article L1110-12 du Code de Santé Publique (CSP) (18) définit «l'équipe de soin» comme «l'ensemble de professionnels qui participent directement au profit d'un même patient à la réalisation d'un acte diagnostique, thérapeutique […] exerçant dans le même établissement de santé». Les membres de cette équipe de soin peuvent consulter et partager les informations concernant le patient pris en charge, au terme de l'article L1110-4 du CSP. (19) Ces textes ne s'appliquent néanmoins qu'à la finalité de prise en charge du patient dans le cadre du soin.

Le choix a été fait, lors des demandes ponctuelles d'extractions de données du SIH, de suivre ce même cadre d'équipe de soin pour la protection des données : le chef de service est chargé d'assurer la protection de l'identité des données produites dans son service, et peut les partager avec les membres de l'équipe de soin qui a participé à leur production.

Le traitement des données PMSI et les extractions de données du SIH sont confiées au DIM. La procédure en place lors d'une demande d'extraction de données du SIH nécessite que le médecin demandeur (ci-après «investigateur») produise une demande formelle contresignée par le responsable médical du périmètre des données demandées.

Pour une requête concernant les données produites dans un unique service, c'est le chef de service qui doit donner son accord. Pour une requête concernant les données produites dans plusieurs services d'un même pôle, on demande l'accord du chef de pôle. Dans le cas d'une requête s'étendant sur l'ensemble du CHRU, c'est la CIM (Commission d'Information Médicale) qui examine l'objet de la requête et statue.

Le DIM répond aux requêtes validées en fournissant les données extraites et la liste nominative (si demandée) des patients au responsable médical ayant contresigné la demande. Il est de son ressort de communiquer les données extraites à l'investigateur. Il conserve la liste nominative identifiant les patients, dont il décide de la communication avec les membres de son équipe. Il s'engage à respecter les règles normales de protection de la confidentialité des patients.

L'objectif ici a été de refléter au mieux l'organisation du circuit des requêtes dans l'accès des praticiens aux données à travers l'outil de requête d'i2b2.

i2b2 définit de manière interne plusieurs niveaux de protection aux données (20), inspirés par les règles décrites dans l'HIPAA (Health Insurance Portability and Accountability Act) :

- DATA_PROT : accès à l'intégralité des données individuelles nominatives
- DATA_DEID : accès aux données individuelles, anonymisées, avec champs chiffrés
- DATA_LDS : accès aux données individuelles, anonymisées, sans les champs chiffrés
- DATA_AGG : accès à des données agrégées (compte de patient)
- DATA_OBFSC : accès à des données agrégées masquées (compte approximatif de patients)

et trois types de comptes utilisateur de la machine :

- ADMIN : administration de la machine, gestion des utilisateurs (création, suppression)
- MANAGER : accès aux requêtes des utilisateurs
- USER : utilisateur, avec niveau d'accès conditionné par DATA_*

Les résultats de requête qu'i2b2 peut produire sont de trois types. Soit des données agrégées masqués : la machine donne un effectif approximatif avec un indice de l'incertitude (par exemple, 1345 ± 5 patients, ou <10 lorsque le résultat est petit). La ré-exécution de la même requête donnera un résultat différent à chaque fois (par exemple, 1349 ± 4). Cette méthode permet de prévenir contre la réidentification par raffinement de la requête et croisement de données. Soit des données agrégées, donnant l'effectif exact du groupe de patients sélectionnés. Soit le listing complet des patients (directement identifiant ou non), et l'extraction de données correspondantes.

i2b2 permet de définir des «projets», et d'attribuer des patients et des utilisateurs à chaque projet. Cependant l'interface web de requête n'implémente pas la vérification de la présence d'un patient dans un projet affecté à un utilisateur, tant que les données concernant ce patient sont présentes dans la table **observation_fact**, rendant ainsi toutes les données disponibles à l'utilisateur.

L'architecture d'i2b2 étant modulaire, il est possible de créer plusieurs *cells* d'un même type, et d'affecter ces *cells* à des projets spécifiques. En procédant ainsi, c'est à la connexion qu'un *datamart* est choisi, et plus au moment d'effectuer la requête, rendant impossible l'accès à des données extérieures au projet sélectionné.

Les fonctions de gestion de projets de R2b2 gèrent la création des bases et *cells* pour chaque nouveau projet, et l'import de données vers la base correspondante.

Le choix a été fait ici de bénéficier de la possibilité de créer plusieurs *datamarts* pour séparer les données selon leur périmètre et attribuer des rôles et niveaux d'accès pour chaque utilisateur selon son appartenance à chaque projet.

Des projets et leur *datamart* associés ont été créés pour stocker et gérer les données à plusieurs niveaux de granularité :

- un projet global contenant les données de l'ensemble des patients du CHRU
- un projet par pôle contenant les données produites lors du passage des patients par un ou plusieurs services de ce pôle
- un projet par service contenant les données produites lors du passage des patients par une ou plusieurs unités médicales de ce service

Chaque médecin du CHRU a accès à au moins trois projets : le projet global, le projet de son pôle, et le projet de son service. Un chef de service a accès au projet global, au projet de son pôle, et au projet de son service, sur lequel il a le rôle de MANAGER. Un chef de pôle a accès au projet global, au projet de son pôle et à l'ensemble des projets des services dépendant de son pôle, sur lesquels il a le rôle de MANAGER. Le chef du DIM (et les membres accrédités) a accès à l'ensemble des projets avec le rôle de MANAGER, et le rôle d'ADMIN sur l'instance i2b2.

Chaque utilisateur avec un rôle de MANAGER a un accès DATA_PROT (visualisation complète des données) sur les données et peut gérer les requêtes des utilisateurs dont il est responsable. Au niveau de granularité directement au-dessus, l'utilisateur a un rôle USER, avec un accès DATA_AGG (effectifs) sur les données, ne permettant de ne voir que des données agrégées. À tous les niveaux de granularité supérieurs, l'utilisateur n'a plus qu'un accès DATA_OBFSC (effectifs masqués) sur les données. L'application de ces règles donne la matrice de droits présentée à la **Figure 3**.

		Responsabilité médicale			
		DIM	**Chef de pôle**	**Chef de service**	**Médecin du service**
Niveau d'accès	**CHRU - Tous services**	Données complètes	Données agrégées	Données masquées	Données masquées
	Pôle	Données complètes	Données complètes	Données agrégées	Données masquées
	Service	Données complètes	Données complètes	Données complètes	Données agrégées

Figure 3: Matrice des droits d'accès

Ainsi un investigateur pourra avoir accès aux données agrégées des patients de son service, et toutefois procéder à des requêtes exploratoires sur le reste du pôle ou de l'établissement. Une fois une requête fixée, il pourra s'adresser à son chef de service pour obtenir une extraction de données produites lors du passage dans le service de ces patients, ou au chef de pôle pour obtenir l'extraction des données produites lors du passage des patients dans le pôle, ou au DIM, qui proposera la requête en CIM, pour l'extraction complète de la liste des patients et des données correspondantes.

Les médecins des plateaux techniques possèdent déjà des outils permettant d'explorer l'ensemble des données qu'ils produisent. L'accès aux données cliniques et médico-administratives étant conditionné à l'accord des chefs des services cliniques et au DIM, ils n'auront accès qu'aux données agrégées masquées sur l'ensemble de l'établissement. L'extraction de données est sujette aux mêmes règles que pour un investigateur dans un service clinique souhaitant obtenir une extraction de données de patients, selon le périmètre désiré.

La constitution de l'entrepôt de données a fait l'objet d'un accord par la Correspondante Informatique et Liberté du CHRU de Nancy, dans le cadre d'une Déclaration Normale.

L'accès aux données agrégées (simples comptes d'effectifs) et agrégées masquées (limitant le risque de ré-identification) ne nécessite aucune autorisation ni déclaration particulière, et se fait sans nécessité de l'accord du patient, puisque les données sont entièrement anonymisées.

L'utilisation des données anonymes produites lors du soin, et extraites par le médecin responsable, ne nécessite pas l'accord du patient et rentre dans le cadre de la non-opposition à la collecte des données de soin. Il convient néanmoins d'assurer la présence de la notion

d'opposition à la collecte des données dans le SIH afin de ne pas intégrer les données de ces patients à l'entrepôt.

Selon le choix de fonction de transformation de l'identifiant patient (conditionnant la possibilité d'identification directe), le responsable médical pourrait, en plus de l'extraction de données, obtenir directement l'identification des patients. Le choix pourra être fait par le DIM d'autoriser «automatiquement» l'identification de patients et l'extraction des données par le responsable médical dans son périmètre. Une formation spécifique des chefs de pôles et de services est nécessaire pour communiquer les règles d'utilisation de la plateforme et rappeler les règles de confidentialité. Il reste possible et recommandé, dans le cas d'un accès identifiant aux données, que le responsable médical seul conserve une table de correspondance des identités des patients et fournisse un fichier anonymisé à l'investigateur. La plateforme permettant la traçabilité des requêtes et des extractions de données, il restera possible au DIM de monitorer les accès.

Résultats

Le premier résultat de ce travail est la publication d'un package R, R2b2 (https://github.com/MaximeWack/R2b2), accompagné de son script d'initialisation. Ces outils permettent de configurer une machine i2b2 à partir de l'image de machine virtuelle fournie officiellement, de réinitialiser cette machine, de gérer les projets, les utilisateurs, ajouter des ontologies, importer des données. Ceci sans nécessiter de connaissance particulière de l'architecture i2b2 ou du détail de la représentation des données dans les bases. Ceci ne fait pas non plus usage de modifications de la structure des bases de données i2b2, permettant son utilisation avec des entrepôts préexistants.

Un second résultat est la production des fichiers d'ontologie organisant les données du SIH du CHRU de Nancy, les requêtes d'extraction de données du SIH, et les scripts de transformation des données extraites vers le format accepté par R2b2 pour le chargement dans i2b2.

Enfin, le résultat principal est la mise en œuvre du package et des ressources développées pour déployer un prototype d'entrepôt de données au CHRU de Nancy, intégrant les données et implémentant les règles d'accès aux données décrites dans la méthode.

Les données démographiques, de PMSI (diagnostics et actes), et de biologie ont été extraites pour tous les patients dont l'hospitalisation s'est terminée entre le 1er janvier 2016 et le 24 août 2017, et intégrées à l'entrepôt, totalisant 118'330 patients uniques au cours de 247'580 hospitalisations.

La représentation des données comporte 49818 concepts uniques, dont :

- 39011 appartenant à la CIM 10
- 9471 appartenant à la CCAM
- 1320 pour la représentation des résultats d'examen biologique
- 16 concepts liés aux données démographiques et morphologiques, et relatives au séjour

Totalisant 18'632'466 observations uniques, dont 2'916'787 observations cliniques (diagnostics, actes, mesures morphologiques), et 15'715'679 valeurs biologiques.

Un projet global contenant l'ensemble des données du CHRU, et les projets spécifiques du pôle «Gynécologie-Obstétrique» et de ses services (AMP Clinique, Gynécologie et Cancérologie, Orthogénie, Anténatal, Endocrinologie Maternité, Post-Natal, et Nouveau-Nés) ont été créés, accompagnés des utilisateurs respectifs, à des fins de démonstration.

Discussion

Choix de la solution

Plusieurs solutions d'intégration de donnée et de gestion d'instance i2b2 ont déjà été développées, dont des solutions basées sur des standards, et la littérature est riche d'exemples d'intégration de données hospitalières dans i2b2. (24)

Le choix de développer une nouvelle solution répond à plusieurs considérations :

- l'absence de structuration des données du système hospitalier à ce jour, rendant caduque d'utilisation de solutions basées sur des standards
- l'architecture de support des règles d'accès aux données développées ici, nécessitant une adaptation des méthodes d'insertion de données prenant en fonction la présence de plusieurs *datamarts*
- le désir de fournir une interface utilisant le langage R, très populaire pour l'analyse et la transformation de données, permettant de procéder à la transformation et au chargement des données dans un unique environnement

Représentation des données

La représentation des données, pour ce qui est des données déjà structurées (diagnostics et actes), a été directe, la seule nécessité ayant été de fournir les ontologies au bon format pour accueillir ces concepts dans la plateforme.

Il aurait été souhaitable pour représenter les résultats biologiques d'utiliser une terminologie existante et validée telle que la LOINC (Logical Observation Identifiers Names and Codes) (29), permettant l'interopérabilité et le partage de données. La LOINC propose une structure combinatoire et non pas hiérarchique, exprimant une vaste gamme de concepts, qui aurait nécessité des choix de représentation tant dans les codes de concept à utiliser que pour définir une arborescence répondant au modèle d'ontologie d'i2b2.

La difficulté d'assignation de codes de concept LOINC est reconnue comme complexe, et l'on trouve de nombreux exemples d'études se concentrant sur l'assignation de codes depuis les concepts internes utilisés par les hôpitaux (30). Diverses approches palliant à la difficulté de produire ces assignations manuellement ont été explorées, comme des techniques d'analyse de texte (34), ou le recours à du travail participatif (35).

Les ressources et le temps limités ainsi que l'absence d'avantage direct à l'interopérabilité dans le cadre d'un entrepôt interne à un seul établissement ont mené à préférer la réalisation d'une organisation *ad hoc*, proche de l'organisation à laquelle les praticiens sont habitués. Un effort de standardisation de la représentation de ces résultats ne saurait également se faire sans la collaboration des différents laboratoires pour systématiser l'attribution de codes de concepts à la source.

Organisation des données et confidentialité

La loi Informatique et Libertés (36) demande que le patient ait donné son accord pour l'accès à des données à caractère personnel, dans le cadre d'une étude précise. Les entrepôts de

données contenant idéalement l'ensemble des données produites dans un établissement, il est impossible de connaître à l'avance les finalités des études qui seront réalisées par son usage. Ce cadre strict rend délicate la mise en place d'un entrepôt de données.

Un groupe de travail conjoint entre le CIMES (Collège d'enseignants d'Informatique Médicale, biomathématiques, méthodes en Épidémiologie et Statistiques), le CUESP (Collège Universitaire des Enseignants en Santé Publique), le CNOM (Conseil National de l'Ordre des Médecins), le CCTIRS (Comité Consultatif sur le Traitement de l'Information en matière de Recherche dans le domaine de la Santé) et la CNIL (Commission Nationale de l'Informatique et des Libertés) a produit en 2014 un guide de bonne pratique sur la réutilisation de données de santé recueillies en milieu hospitalier lors de leur réutilisation à des fins statistiques, (37) émettant des recommandations sur les pratiques à suivre lors de la mise en place d'un entrepôt de données dans le cadre législatif actuel.

Ces recommandations ont été suivies pour l'élaboration des règles d'isolation des données et régissant l'accès à ces données.

Sur ce prototype, la fonction de transformation des identifiants uniques de patients utilisée est la fonction d'identité, qui autoriserait les responsables médicaux non seulement à extraire eux-mêmes les données collectées par l'activité de leur service/pôle, mais également à obtenir une liste identifiante de patients.

Ceci représente le scénario le plus permissif, donnant une autonomie aux services et pôles pour accéder par leurs propres moyens aux données ainsi qu'à l'identification des patients. Il est soumis à la volonté du DIM de considérer chaque *datamart* lié à un service ou un pôle comme un projet validé pour ce périmètre, tel un «mini-entrepôt» ne contenant que les données attachées à ce périmètre. Cette option offrant une plus grande autonomie aux services pourrait bénéficier à la conduite de recherches biomédicales au sein de l'établissement, au détriment d'un niveau de protection optimal de la confidentialité.

L'option fournie par la plateforme d'extraire les données avec une renumérotation du seul échantillon extrait (et en continuant d'avoir accès aux identifiants) permettrait au médecin encadrant du service de garder une liste de correspondance des patients tout en communiquant des données anonymisées aux membres de l'équipe de recherche, ce qui reste compatible avec les recommandations de bonnes pratiques. Il faut s'assurer que les responsables seront formés à cette pratique.

Le remplacement de la fonction de transformation par la fonction de renumérotation représente un scénario plus conservateur. Il permettrait toujours aux services d'extraire eux-mêmes des données contenues dans l'entrepôt, mais en limitant la capacité d'identification des patients et proposant ainsi la meilleure solution envisageable dans ce contexte pour la protection de leur anonymat. L'accès à ces données serait donc toujours conditionné à une communication formelle avec le DIM, en application stricte des recommandations de bonne pratique, en ne donnant jamais un accès «indépendant» à des données identifiantes par les utilisateurs.

La politique mise en place à l'Hôpital Européen Georges Pompidou (38), autre établissement public français utilisant i2b2, consiste à limiter par défaut l'accès aux seules données agrégées (niveau 1). L'accès aux données complètes (anonymisées (niveau 2) ou identifiantes (niveau 3))

nécessite le renseignement d'un protocole et l'approbation du comité d'éthique de l'établissement.

Notre approche est hybride et consiste à donner accès aux responsables médicaux aux données anonymes collectées lors des soins sous leur responsabilité, tout en continuant à demander l'entrée dans le circuit recherche dès lors qu'il est demandé accès aux données extérieures au périmètre ou à l'identification des patients. Ceci permettrait la conduite d'études observationnelles rétrospectives en autonomie vis-à-vis du DIM. Les requêtes et extractions conduites dans ce cadre pourraient toujours être modérées *a posteriori* par le DIM responsable de l'entrepôt, puisque toutes les opérations sont tracées dans la plateforme.

Le livret d'accueil patient ainsi que la charte informatique signée par le personnel sont déjà compatibles avec ces dispositions et ne nécessitent pas de changement, puisqu'ils comportent respectivement la mention de réutilisation des données à des fins de recherche et la possibilité de s'y opposer ; et la définition des règles à observer concernant l'accès et le traitement des données. Une note spécifique mentionnant l'existence d'un entrepôt de données cliniques pourra être ajoutée à des fins de transparence complète envers le patient et le recueil de son opposition à ce traitement en particulier.

Perspectives

La création des outils de manipulation de la plateforme i2b2, d'un modèle de création d'ontologies, et des outils de chargement de données permettront la pérennisation du projet, et la possibilité pour les équipes du DIM de développer l'intégration d'autres sources de données du SIH du CHRU de Nancy, qu'il s'agisse de données possiblement structurées telles que les prescriptions médicamenteuses, les signes «de pancarte», ou de données non structurées (comptes-rendus d'hospitalisation, d'anatomo-pathologie et d'imagerie).

L'accès direct aux bases de données des logiciels du SIH permettra une intégration continue des données des patients, de manière à donner accès aux praticiens aux données les plus récentes, y compris celles déjà disponibles pour les patients en cours d'hospitalisation.

L'accès aux bases de données des logiciels utilisés en interne (tel que GLIMS pour les résultats de dosages biologiques) permettrait l'intégration de l'antériorité des résultats de biologie avant la mise en place du dossier patient informatisé avec DxCare, afin d'enrichir les données présentes dans la plateforme.

Il est attendu une collaboration avec la DSI pour obtenir l'accès direct aux différentes bases, augmenter les capacités de la machine hébergeant la plateforme pour intégrer la totalité des données déjà produites, et le déploiement effectif à tous les pôles et services. Un temps ingénieur dédié est recommandé pour gérer les tâches d'administration routinières de l'entrepôt : administration et maintenance système, allocation de ressources, gestion des utilisateurs et de leurs droits, et gestion des flux d'intégration une fois conçus par le DIM.

Un manuel d'utilisation de l'outil de requête a été produit pour communication aux praticiens lors de la mise en production de la plateforme (**Annexe 2**).

Conclusion

Un package R, R2b2, permettant la manipulation et la gestion d'une instance i2b2 a été produit et publié sous licence libre.

Il permet une manipulation simplifiée de la plate-forme i2b2, et réunit les fonctionnalités majeures (gestion de la plateforme, des utilisateurs, de l'organisation des données, de l'intégration des données, des règles d'accès et de confidentialité) sous une seule interface dans un langage populaire auprès des médecins d'informatique médicale.

Des modèles de représentation des données provenant de divers logiciels du SIH ont été conçus pour décrire les données du PMSI (actes et diagnostics), les données démographiques, et les résultats d'examen biologiques des patients. Les fichiers d'organisation des données propres à l'établissement, ainsi que les outils permettant l'intégration de ces données ont été conçus et livrés au DIM, permettant la mise en production de cette plateforme.

L'organisation de l'accès aux données et de la préservation de l'anonymat des patients a été décidée de manière à fournir une certaine autonomie aux praticiens, en respectant la législation, et en fournissant un mécanisme de traçabilité.

Références

1. Yoo S, Lee KH, Lee HJ, Ha K, Lim C, Chin HJ, et al. Seoul national university bundang hospital's electronic system for total care. Healthc Inform Res [Internet]. 2012 Jun [cited 2016 Apr 8];18(2):145–52. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3402557/

2. Lowe HJ, Ferris TA, Hernandez PM, Weber SC. STRIDE – an integrated standards-based translational research informatics platform. AMIA Annu Symp Proc [Internet]. 2009 [cited 2016 Apr 8];2009:391–5. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2815452/

3. Chelico JD, Wilcox A, Wajngurt D. Architectural design of a data warehouse to support operational and analytical queries across disparate clinical databases. AMIA Annu Symp Proc. 2007;901.

4. Cuggia M, Garcelon N, Campillo-Gimenez B, Bernicot T, Laurent J-F, Garin E, et al. Roogle: An information retrieval engine for clinical data warehouse. Stud Health Technol Inform. 2011;169:584–8.

5. Delamarre D, Bouzille G, Dalleau K, Courtel D, Cuggia M. Semantic integration of medication data into the EHOP clinical data warehouse. Stud Health Technol Inform. 2015;210:702–6.

6. Brammen D, Katzer C, Röhrig R, Weismüller K, Maier M, Hossain H, et al. An integrated data-warehouse-concept for clinical and biological information. Stud Health Technol Inform. 2005;116:9–14.

7. AlHazme RH, Rana AM, De Lucca M. Development and implementation of a clinical and business intelligence system for the florida health data warehouse. Online J Public Health Inform [Internet]. 2014 Oct 16 [cited 2016 Apr 8];6(2). Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4221087/

8. Chute CG, Beck SA, Fisk TB, Mohr DN. The enterprise data trust at mayo clinic: A semantically integrated warehouse of biomedical data. J Am Med Inform Assoc [Internet]. 2010 [cited 2016 Apr 8];17(2):131–5. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3000789/

9. Lyman JA, Scully K, Harrison JH. The development of health care data warehouses to support data mining. Clin Lab Med. 2008 Mar;28(1):55–71, vi.

10. Murphy S, Barnett G, Chueh H. Visual query tool for finding patient cohorts from a clinical data warehouse of the partners HealthCare system. Proc AMIA Symp [Internet]. 2000 [cited 2016 Apr 8];1174. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2243863/

11. Murphy SN, Chueh HC. A security architecture for query tools used to access large biomedical databases. Proc AMIA Symp [Internet]. 2002 [cited 2017 Sep 19];552–6. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2244204/

12. Mendis M, Wattanasin N, Kuttan R, Pan W, Philips L, Hackett K, et al. Integration of hive and cell software in the i2b2 architecture. AMIA Annu Symp Proc. 2007;1048.

13. Murphy SN, Mendis M, Hackett K, Kuttan R, Pan W, Phillips LC, et al. Architecture of the open-source clinical research chart from informatics for integrating biology and the bedside. AMIA Annu Symp Proc [Internet]. 2007 [cited 2016 Apr 8];2007:548–52. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2655844/

14. Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc [Internet]. 2010 [cited 2016 Apr 8];17(2):124–30. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3000779/

15. i2b2. Informatics for integrating biology & the bedside [Internet]. 2017 [cited 2017 Sep 19]. Available from: https://www.i2b2.org/work/i2b2_installations.html

16. i2b2. Sites [Internet]. 2017 [cited 2017 Sep 19]. Available from: http://www.healthmap.org/i2b2/

17. Team RC. R: A language and environment for statistical computing. 2017; Available from: https://www.R-project.org

18. CSP. Code de la santé publique - article l1110-12. 2017.

19. CSP. Code de la santé publique - article l1110-4. 2017.

20. Murphy SN, Gainer V, Mendis M, Churchill S, Kohane I. Strategies for maintaining patient privacy in i2b2. J Am Med Inform Assoc [Internet]. 2011 Dec [cited 2016 Apr 8];18(Suppl 1):i103–8. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3241166/

21. CSP. Code de la santé publique - article l1121-1. 2017.

22. LOI n° 2012-300 du 5 mars 2012 relative aux recherches impliquant la personne humaine. Mar, 2012.

23. CSP. Code de la santé publique - article l1121-4. 2017.

24. Zapletal E, Rodon N, Grabar N, Degoulet P. Methodology of integration of a clinical data warehouse with a clinical information system: The HEGP case. Stud Health Technol Inform. 2010;160(Pt 1):193–7.

25. Abend A, Housman D, Johnson B. Integrating clinical data into the i2b2 repository. Summit on Translat Bioinforma [Internet]. 2009 Mar 1 [cited 2016 Apr 8];2009:1–5. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3041580/

26. Takai-Igarashi T, Akasaka R, Suzuki K, Furukawa T, Yoshida M, Inoue K, et al. On experiences of i2b2 (informatics for integrating biology and the bedside) database with japanese clinical patients' data. Bioinformation [Internet]. 2011 Mar 26 [cited 2016 Apr 8];6(2):86–90. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3082863/

27. Murphy SN, Mendis ME, Berkowitz DA, Kohane I, Chueh HC. Integration of clinical and genetic data in the i2b2 architecture. AMIA Annu Symp Proc [Internet]. 2006 [cited 2016 Apr 8];2006:1040. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1839291/

28. Majeed RW, Röhrig R. Automated realtime data import for the i2b2 clinical data warehouse: Introducing the HL7 ETL cell. Stud Health Technol Inform. 2012;180:270–4.

29. Forrey AW, McDonald CJ, DeMoor G, Huff SM, Leavelle D, Leland D, et al. Logical observation identifier names and codes (LOINC) database: A public use set of codes and names for electronic reporting of clinical laboratory test results. Clin Chem. 1996 Jan;42(1):81–90.

30. Zunner C, Ganslandt T, Prokosch H-U, Bürkle T. A reference architecture for semantic interoperability and its practical application. Stud Health Technol Inform. 2014;198:40–6.

31. Wu J, Finnell JT, Vreeman DJ. Evaluating congruence between laboratory LOINC value sets for quality measures, public health reporting, and mapping common tests. AMIA Annu Symp Proc. 2013;2013:1525–32.

32. Lin M-C, Vreeman DJ, Huff SM. Investigating the semantic interoperability of laboratory data exchanged using LOINC codes in three large institutions. AMIA Annu Symp Proc. 2011;2011:805–14.

33. Kopanitsa G. Mapping russian laboratory terms to LOINC. Stud Health Technol Inform. 2015;210:379–83.

34. Lee L-H, Groß A, Hartung M, Liou D-M, Rahm E. A multi-part matching strategy for mapping LOINC with laboratory terminologies. J Am Med Inform Assoc. 2014 Oct;21(5):792–800.

35. Vreeman DJ, Hook J, Dixon BE. Learning from the crowd while mapping to LOINC. J Am Med Inform Assoc. 2015 Nov;22(6):1205–11.

36. CNIL. Délibération de la commission nationale de l'Informatique et des libertés. Jan 19, 2017.

37. CIMES, CUESP, CNIM, CCTIRS, CNIL. Guide de bonnes pratiques permettant d'assurer la confidentialité des données de santé recueillies en milieu hospitalier lors de leur réutilisation à des fins de traitement statistique. 2013; Available from: http://www.departement-information-medicale.com/wp-content/uploads/2014/01/Guide_Voct_2013_diffusionEMOIS.pdf

38. Jannot A-S, Zapletal E, Avillach P, Mamzer M-F, Burgun A, Degoulet P. The georges pompidou university hospital clinical data warehouse: A 8-years follow-up experience. International Journal of Medical Informatics [Internet]. 2017 Jun 1 [cited 2017 Sep 8];102:21–8. Available from: http://www.sciencedirect.com/science/article/pii/S1386505617300370

Conclusion

La réalisation de ce travail pilote a permis l'installation d'un prototype d'entrepôt de données cliniques pour la recherche au CHRU de Nancy.

Cet entrepôt intègre des données de codage PMSI (actes et diagnostics médicaux), les données démographiques, et les résultats d'analyses biologiques, depuis le début de l'année 2016.

L'architecture développée permet la séparation des données en fonction de leur responsabilité médicale, permettant un retour direct aux praticiens des données produites lors du soin par leur service, tout en protégeant la vie privée des patients et leur anonymat à l'aide de règles d'accès empêchant la réidentification.

Personnellement, ce travail a permis d'explorer en détail le fonctionnement de la plateforme i2b2, son architecture particulière et ses difficultés intrinsèques. S'agissant d'un «produit» non commercial et développé majoritairement dans un milieu académique par une succession de développeurs non professionnels, il n'est pas exempt de bugs (par exemple dans l'interface d'administration web), et autres artefacts d'un développement incrémentiel (nécessité de changer plusieurs champs dans diverses bases de données pour obtenir un unique résultat ; fonctionnalités décrites dans les documents de design, avec existence de l'architecture sous-jacente en base de donnée, mais sans implémentation dans l'interface de requête).

La création du package et de la documentation l'accompagnant permettront la poursuite du travail sur l'entrepôt, en servant d'interface d'abstraction masquant les particularités de fonctionnement interne d'i2b2. Les outils fournis dans le package permettront de futurs travaux d'intégration des autres sources de données présentes dans le SIH.

Le fait d'avoir développé le package à l'aide du langage R, et respectant les paradigmes modernes du langage, rend le niveau de connaissances permettant la manipulation de la plateforme plus accessible, en particulier pour les futurs internes du DIM qui apprennent à manipuler ce langage.

Il est attendu que la plateforme soit pérennisée et mise en production, et ce suivant plusieurs recommandations :

- mise à disposition d'un matériel permettant l'intégration d'une plus grande quantité de données, notamment historiques
- ouverture des accès aux bases de données des applications métier, permettant l'intégration continue des données
- assignation d'au minimum un mi-temps ingénieur pour assurer la maintenance technique de l'entrepôt
- gestion fine des accès par les médecins du CHRU

Les perspectives d'utilisation future de l'entrepôt sont les suivantes :

- projets pouvant être confiés aux internes d'intégration de nouvelles sources de données
- possibilité de recherche propre à l'entrepôt (traitement automatisé du langage pour l'extraction de concepts depuis les comptes rendus en texte libre, outils d'aide à la décision, automatisation de règles d'assurance qualité, etc.)

- retour d'expérience de structuration des données pour façonner, organiser et harmoniser la représentation des données dans les diverses composantes du SIH
- utilisation de l'entrepôt pour guider l'organisation de l'offre de soins

Annexe 1 : Guide administrateur i2b2

Obtention de la machine virtuelle

Télécharger la machine virtuelle i2b2 mise à disposition sur le site officiel i2b2.org, et l'installer avec VMware.

L'image est prévue pour la version personnelle de VMware, mais est transformable vers une image VMware server.

Toutes les opérations suivantes sont à mener avec l'utilisateur **root**. Un utilisateur administrateur sera créé lors des étapes suivantes. Une fois un compte administrateur i2b2 créé, toutes les opérations en-dehors de la gestion de projets peuvent être conduites avec cet utilisateur.

Initialisation de la machine virtuelle et installation de R2b2

Après avoir installé la machine et configuré le réseau et le proxy si nécessaire, télécharger le script d'initialisation de la machine avec la commande suivante :

```
curl https://raw.githubusercontent.com/MaximeWack/R2b2/master/prepare.sh
```

Puis l'exécuter.

Ce script va :

- mettre à jour le système
- installer le dépôt EPEL (Extra Packages for Enterprise Linux) qui contient notamment la dernière version de R
- installer sudo
- installer git
- installer R
- installer les dépendances système pour les packages R requis
- installer les packages R utiles à R2b2
- installer R2b2 depuis github

Installation des fichiers spécifiques au CHRU

Cloner le dépôt contenant les fichiers spécifiques du CHRU :

```
git clone https://github.com/MaximeWack/R2b2_CHRUN
```

Celui-ci contient :

- un fichier `fresh_install.R` qui est un script R contenant une fonction d'initialisation, et les différentes fonctions d'intégration de données spécifiques au CHRU.
- un répertoire `requetes` contenant les fichiers .wid de requête BO pour extraire les données du SIH

- un répertoire `onto` contenant les fichiers d'ontologie pour les types de données extraites du SIH

Initialisation de l'instance i2b2

Dans une console R, sourcer le fichier `fresh_install.R` pour charger le package R2b2 et les fonctions.

La fonction `fresh_install` prend pour arguments un nom de compte et un mot de passe pour l'administrateur i2b2, ainsi qu'un identifiant et nom complet pour le domaine.

Par exemple : `fresh_install("i2b2admin", "mot_de_passe", "CHRUN", "CHRU de Nancy")`

Cette fonction encapsule d'autres fonctions de R2b2 permettant de manipuler la machine :

- set_permissions() : configure les permissions pour les répertoires contenant le client web et l'applicatif i2b2, permettant à l'administrateur i2b2 de les modifier sans besoin de permission **root**.
- create_admin(admin, pass) : créé un compte système et postgreSQL pour l'administrateur i2b2, avec le mot de passe fourni, et modifier le mot de passe de l'administrateur postgreSQL par défaut
- clear_webclient() : supprime les identifiants demo par défaut de l'interface web
- clear_default_*data() : supprime les données de démonstration présentes dans la machine téléchargée
- set_domain(domain_id, domain_name) : configure le domaine de la machine, dans la base de données et dans le client web
- add_users("i2b2", "demouser", data.frame(...)) : utilise le compte administrateur i2b2 par défaut pour ajouter le nouvel administrateur (même nom et mot de passe que pour les comptes système et postgreSQL)
- service("pg", "restart") : redémarre le moteur de base de données pour éviter le verrouillage de la base
- add_project("CHRU", "CHRU - Tous services") : créé le projet global pour l'ensemble de l'hôpital
- delete_users(c("i2b2", "demo")) : suppression des comptes présents pour démonstration
- add_ont("Diagnostics", "CIM")… : création des ontologies avec leurs schémas respectifs
- populate_ont(readr::read_csv("ont/cim.ont"), readr::read_csv("onto/cim.modi"), "CIM")… : remplissage des ontologies à partir des fichiers fournis
- add_ontologies("CHRU") : remplissage des tables de concepts/providers pour le projet global
- service("jboss", "restart")

Après cette étape la machine est configurée avec un domaine, un projet principal, et un administrateur avec un compte système (normal), postgreSQL (admin), et i2b2 (admin).

L'administration de la machine et de la plateforme sont ainsi décorrélées. Il reste cependant toujours possible de donner des droits «sudo» à l'administrateur i2b2 en l'ajoutant au groupe `wheel` et en modifiant le fichier `/etc/sudoers` pour donner les droits sudo au groupe wheel.

Les autres choix de configuration sont à la discrétion de l'administrateur système de la machine (configuration matérielle, sauvegardes, configuration réseau, etc.). Deux éléments de configuration spécifiques à i2b2 restent à connaître, qui ne sont pas choisis automatiquement par le script d'initialisation :

- le client web possède son propre «proxy» permettant de limiter les accès. Le fichier `/var/www/html/webclient/index.php` contient une variable `$WHITELIST` contenant la liste des domaines ayant accès à l'interface web. Ajouter une entrée "i2b2.chu-nancy.fr" permet d'assurer l'accès avec le nom attribué sur le réseau interne du CHRU. Ajouter une entrée "http://" permet d'enlever toute restriction par ce «proxy». Les règles d'accès réseau du CHRU sont déjà suffisantes pour limiter l'accès à i2b2 depuis l'extérieur.
- la configuration postgreSQL de base est suffisante pour l'administration depuis la machine locale. Pour ouvrir la possibilité d'effectuer certaines tâches d'administration i2b2 depuis une autre machine sur le réseau, quelques modifications sont nécessaires :
  - Le fichier `/var/lib/pgsql/9.1/data/postgresql.conf` contient une option "listen_addresses", qui peut être renseignée avec un bloc CIDR pour spécifier le sous-réseau ayant accès à distance à l'instance postgreSQL
  - Le fichier `/var/lib/pgsql/9.1/data/pg_hba.conf` permet de spécifier les types d'accès autorisés pour chaque utilisateur. Par défaut, en local, tous les utilisateurs ont accès en mode `peer` (soit avec la session système). Il possible d'ajouter un accès distant avec une ligne du type : `host all user subnet md5`, autorisant l'accès à l'utilisateur `user` depuis le sous-réseau `subnet` (par exemple 192.168.30.1/24 pour le réseau interne du CHRU), en mode `md5` (soit à l'aide d'un mot de passe spécifique à postgreSQL)

Les fichiers de configuration postgreSQL peuvent être accédés et modifiés uniquement par l'utilisateur `postgres`. Avec un utilisateur administrateur système, il est possible de se logger en tant que `postgres` avec la commande suivante : `sudo su postgres`

Gestion de projets

Les fonctions d'ajout/suppression de projets doivent être exécutées en local sur la machine, ayant un effet sur des fichiers de configuration locaux et entraînant un redémarrage de certains services (la gestion des services dépend du pouvoir d'administration sur la machine).

Ajouter un projet

Exécuter la fonction `add_project(id, label)` (où `id` est un identifiant unique, et `label` un texte descriptif, par exemple "1400", "Cardiologie Médicale ILM".

Cette fonction créé un datamart pour le nouveau projet en créant la base de données associée, les entrées dans la base de données i2b2 référençant le nouveau projet et le chemin d'accès, et

l'entrée dans le fichier de configuration pour lancer une instance de la *cell* CRC pour ce datamart. L'instance i2b2 est redémarrée après ajout du nouveau projet.

Il peut être nécessaire de redémarrer le moteur postgreSQL pour libérer d'éventuels verrous sur la base de données avant de créer un nouveau projet (`service("pg", "restart")`).

La fonction `add_ontologies(id)` permet d'ajouter les concepts des différentes ontologies au datamart du nouveau projet.

La fonction `add_ontologies` devra être mise à jour si de nouveaux types de données sont intégrées.

Lister les projets existants

La fonction `list_projects` permet d'obtenir la liste des projets existants.

Supprimer un projet

La fonction `delete_project(id)` permet de supprimer un projet.

Gestion des utilisateurs

Lister les utilisateurs

La fonction `list_users()` renvoie la liste des utilisateurs avec leur identifiant, nom complet et leur adresse email.

Créer un utilisateur

La fonction `add_user(admin, pass, id, name, email, password)` permet d'ajouter un utilisateur à la machine.

Les paramètres sont :

- admin : nom du compte administrateur i2b2
- pass : mot du passe du compte administrateur i2b2
- id : identifiant unique de l'utilisateur
- name : nom complet de l'utilisateur
- email : adresse email de l'utilisateur
- password : mot de passe pour le nouvel utilisateur

Supprimer un ou plusieurs utilisateurs

La fonction `delete_users(users)` permet de supprimer un ou plusieurs utilisateurs. `users` est un vecteur d'identifiants uniques d'utilisateurs.

Lister les autorisations pour un utilisateur

La fonction `list_user_roles(user)` permet de lister les autorisations pour un utilisateur. `user` est l'identifiant unique de l'utilisateur.

Gérer les autorisations des utilisateurs

La fonction `add_user_roles(admin, pass, id, project, roles)` permet d'ajouter des rôles à un utilisateur.

Les paramètres sont :

- admin : nom du compte administrateur i2b2
- pass : mot de passe du compte administrateur i2b2
- id : identifiant unique de l'utilisateur à gérer
- project : projet pour lequel donner des autorisations
- roles : un vecteur de rôles à attribuer (par exemple c("USER", "DATA_AGG"))

Ajout d'utilisateurs en masse

La fonction `add_users(admin, pass, users)` permet d'ajouter des utilisateurs avec leurs autorisations respectives en masse.

Les paramètres sont :

- admin : nom du compte administrateur i2b2
- pass : mot du passe du compte administrateur i2b2
- users : une dataframe avec les colonnes suivantes :
  - id : identifiant unique de l'utilisateur
  - name : nom complet de l'utilisateur
  - email : adresse email de l'utilisateur
  - password : mot de passe pour le nouvel utilisateur
  - project : projet pour lequel donner des autorisations
  - roles : un de :
    - ADMIN : donne tous les rôles à l'utilisateur (admin + manager + user + accès à toutes les données)
    - MANAGER : donne le rôle de manager et accès à toutes les données
    - USER : donne uniquement le rôle d'utilisateur avec accès aux données masquées (DATA_OBFSC)
    - autre DATA_*, donnant le rôle d'utilisateur et les accès aux données jusqu'à celui spécifié (DATA_OBFSC < DATA_AGG < DATA_LDS < DATA_DEID < DATA_PROT). Dans cette installation, les rôles au-dessus de DATA_LDS sont confondus. On utilisera donc soit DATA_OBFSC pour l'accès masqué aux données agrégées, DATA_AGG pour les données agrégées et DATA_PROT pour les données complètes.

Gestion des projets multiples

La fonction `accounts_obgyn` donne exemple de séquence d'appels de fonctions R2b2 permettant de créer les utilisateurs et attribuer les rôles correspondants aux règles décrites dans la thèse, pour un pôle clinique.

Gestion des ontologies

Lister les ontologies existantes

Les fonctions `list_schemes()` et `list_ont()` permettent respectivement de lister tous les schémas existants, et toutes les ontologies avec les paramètres associés.

Consulter une ontologies

La fonction `get_ont(ont)` où `ont` est le nom de la table contenant l'ontologie à consulter (tel qu'apparaît dans la colonne `c_table_cd` lors d'un appel à `list_ont`) renvoie une dataframe contenant toute l'ontologie sélectionnée.

Ajouter une ontologie

La fonction `add_ont(name, scheme)` permet de créer un emplacement pour une ontologie, en donnant son nom et son schema (par exemple, "CIM", "Diagnostics")

Renseigner une ontologies

La fonction `populate_ont(ont, modi, scheme)` sert à renseigner le contenu d'une ontologie. `ont` est une dataframe contenant au minimum une colonne `c_fullname` contenant le chemin complet des «feuilles» de l'ontologie. `modi` est un argument optionnel (défaut = NULL), une dataframe des modificateurs pour cette ontologie. `scheme` est le schéma de l'ontologie à renseigner avec ces valeurs.

Ceci ne fait que renseigner l'ontologie dans la *cell* metadata. Il convient de renseigner les concepts correspondants dans chaque *cell* demodata.
Ceci peut être fait à l'aide de fonctions telles que `populate_concept` et `populate_provider`. (voir le corps de la fonction `add_ontologies` pour un exemple d'usage)

Supprimer une ontologie

La fonction `delete_ont(scheme)` permet de supprimer l'ontologie correspondant au `scheme` dans la *cell* metadata.

Intégration de données

Extraction des bases du SIH

Les requêtes au format BO WebRich Client (fichiers .wid) sont fournis pour extraire les données du SIH.
Chaque requête est paramétrée avec une invite pour sélectionner les dates d'intérêt.

- pims : extraction des données concernant les patients et les hospitalisations
- pims_diags : extraction des diagnostics PMSI
- pims_actes : extraction des actes PMSI
- dxcare_mensurations : extraction des données morphométriques à l'entrée en hospitalisation

- dxcare_bio : extraction des données de résultat de laboratoire

Les fichiers sont à enregistrer au format **csv** avec une virgule (",") comme séparateur. Il est nécessaire de copier ces fichiers vers la machine i2b2 pour permettre leur lecture par le package. Utiliser un outil comme `filezilla` ou `scp` pour envoyer les fichiers sur la machine.

Lecture et transformation

Des fonctions spécifique pour chacun de ces fichiers extraits sont fournies :

- read_patients : pour lire le fichier extrait de la requête **pims**
- read_diagnostics : pour lire le fichier extrait de la requête **pims_diags**
- read_actes : pour lire le fichier extrait de la requête **pims_actes**
- read_mensurations : pour lire le fichier extrait de la requête **dxcare_mensurations**
- read_bios : pour lire le fichier extrait de la requête **dxcare_bio**

Elles renvoient chacune un objet conforme aux formats i2b2 acceptés par R2b2.

Chargement des données

Les fonctions `import_patients_visits`, `import_mensurations` et `import_bios` permettent de charger les données de patients et hospitalisations, des données morphométriques et des résultats biologiques.

Elles font appel aux fonctions R2b2 sous-jacentes que sont `add_patients_demodata` (qui permet d'ajout des patients à la base de données), `add_encounters` (permettant d'ajouter des visites), et `add_observations` (qui permet d'ajouter des observations pour les patients et hospitalisations existantes).

La fonction `pop_main(main, patients, diagnostics, actes, mensurations, bios)` rassemble toutes les fonctions précédentes en une seule simple à utiliser pour charger les données de l'établissement entier.

`main` est le nom du projet principal (dans l'installation actuelle, "CHRU"), les autres arguments sont les chemins d'accès aux divers fichiers extraits à l'aide des requêtes pré-citées.

La fonction `pop_projects(top_project, projects, files…)` permet le chargement de données par pôle et service. `top_project` est l'identifiant du projet pour le pôle et `projects` est un vecteur d'identifiants de projets pour les services. Pour que ceci fonctionne il est nécessaire que les projets soient nommés avec leur code d'UM.

Annexe 2 : Guide utilisateur i2b2

Connexion

Aller sur https://i2b2.chu-nancy.fr.

Un écran d'accueil vous permet de rentrer votre nom d'utilisateur et votre mot de passe.

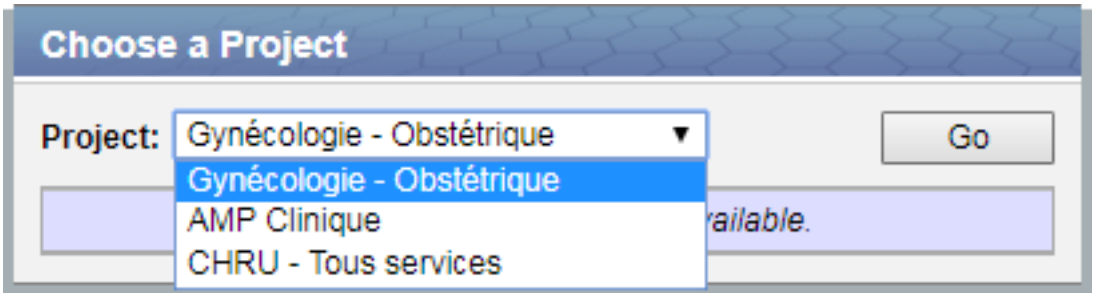
Votre nom d'utilisateur est votre «u», et lors de la première connexion c'est aussi votre mot de passe.

Un seul domaine est accessible, il s'agit des données du CHRU de Nancy.

Cliquez sur `Login` pour vous connecter.

Choix du «projet»

L'écran suivant présente un choix de projet.



Un projet est un périmètre de requête, et il existe des projets pour :

- le CHRU entier
- chacun des pôles
- chacun des services

Par défault vous avez accès à votre service, pôle, et au CHRU.

En sélectionnant votre **service,** vous aurez accès à l'outil de requête, vous permettant d'explorer des effectifs de patients, limités au périmètre de votre service.

En sélection votre **pôle** ou le **CHRU entier,** vous aurez accès à l'outil de requête sur ces périmètres, mais avec des résultats approximatifs (ceci est fait pour prévenir la réidentification de patients avec des requêtes successives).

Si vous êtes **chef de service**, vous avez accès aux requêtes des membres de votre service, ainsi qu'aux outils d'analyses de données et l'accès aux données complètes concernant les passages des patients par votre service.

De la même manière si vous êtes **chef de pôle**, vous avez accès aux requêtes des membres de votre pôle, ainsi qu'aux outils d'analyses de données et l'accès aux données complètes concernant les passages des patients par votre pôle.

En tant qu'utilisateur, sélectionnez le périmètre correspondant au plus près à votre requête, afin de faciliter l'obtention ultérieure de données complètes en passant par votre supérieur direct. Si vous choisissez de faire une requête sur l'ensemble du CHRU par exemple, il vous faudra vous adresser au DIM pour obtenir une extraction de données. Si vous choisissez une requête sur votre service uniquement, votre chef de service pourra directement extraire les données.

Interface

L'interface de requête d'i2b2 se compose de plusieurs panneaux.

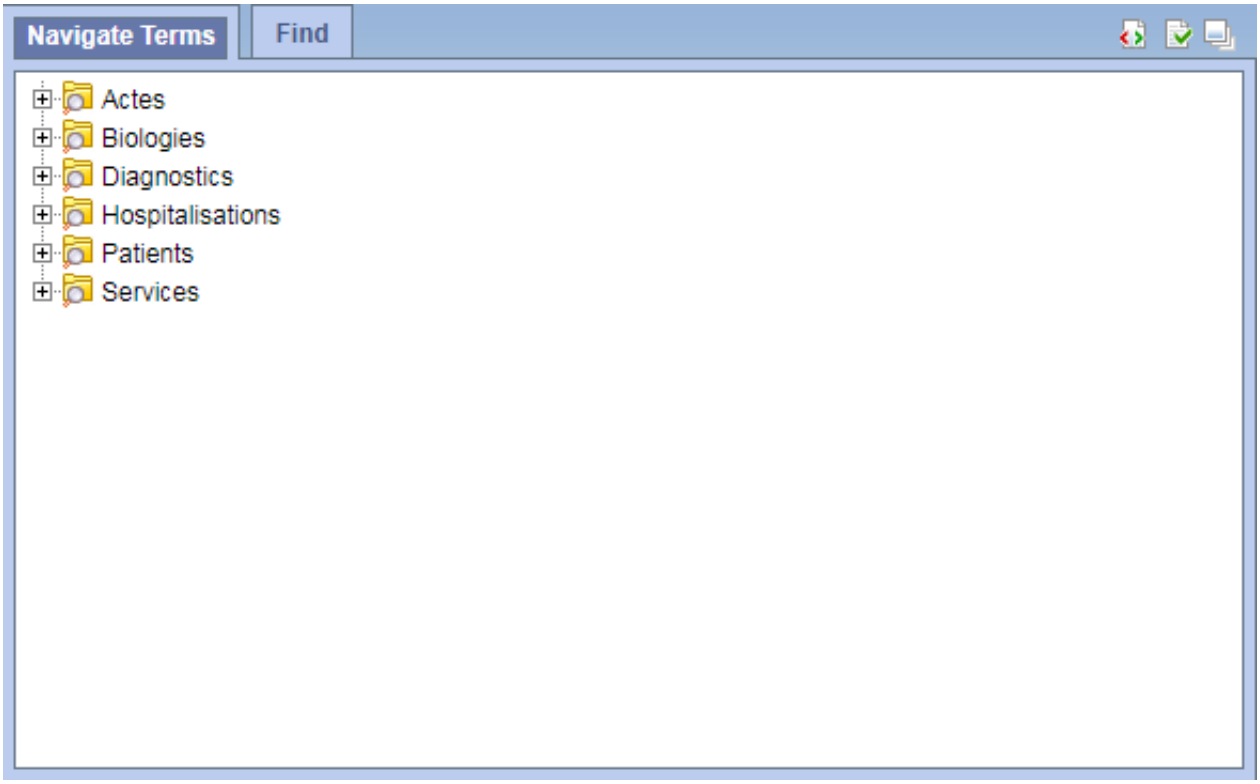
Barre d'outils

En haut de l'interface, la barre d'outils rappelle sur quel projet vous êtes actuellement connecté, votre nom d'utilisateur, et présente des boutons de menu.

Lors de la première connexion il est recommandé de cliquer sur `Change Password` pour changer votre mot de passe vers un plus sécurisé.

Le bouton `Analysis Tools` propose une liste d'outils d'analyse pour les utilisateurs ayant accès aux données complètes (descriptif démographique, frise temporelle des évènements, outils d'export, etc.).

Concepts (*Navigate Terms*)



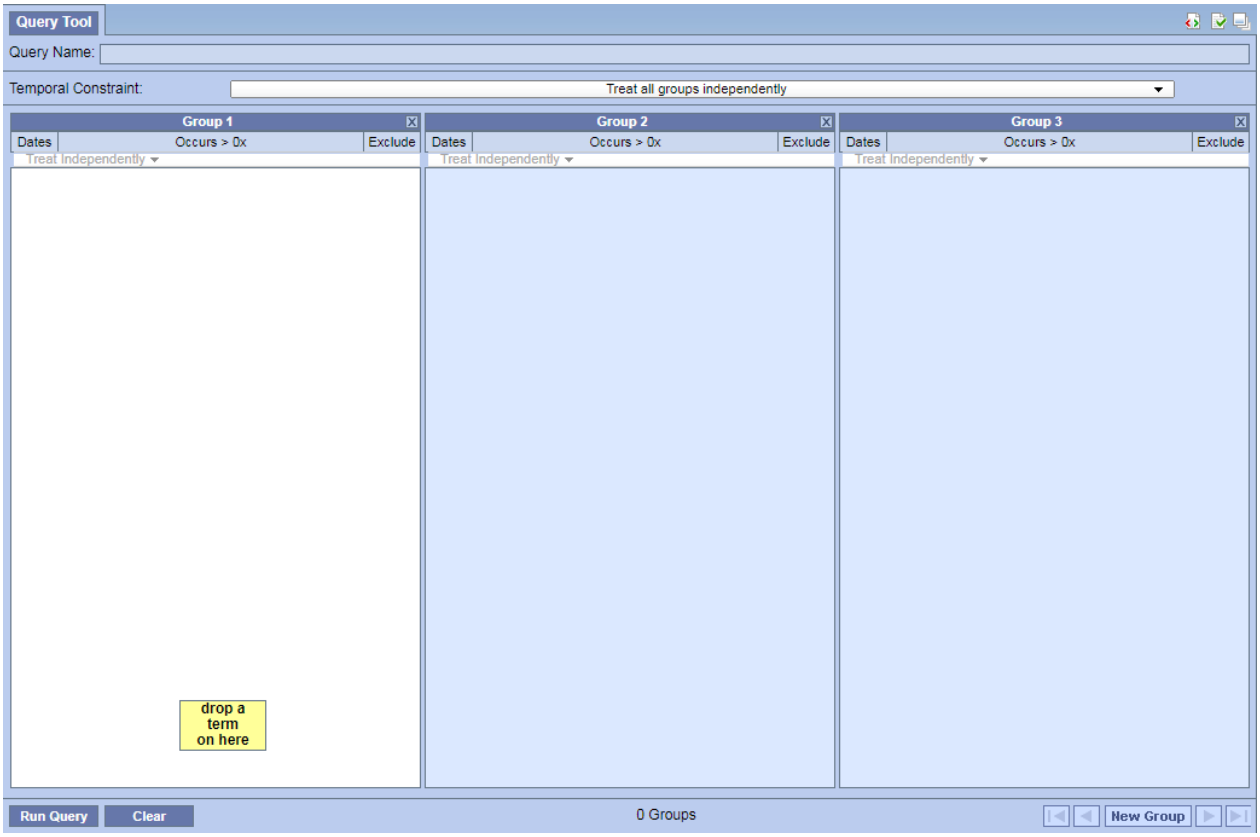
En haut à gauche, l'explorateur de concepts permet de choisir les critères sur lesquels sélectionner une population, et les données à extraire.

Les concepts sont organisés hiérarchiquement, éventuellement selon des terminologies définies, comme la CCAM pour les actes médicaux et la CIM10 pour les diagnostics.

Il est possible de sélectionner des concepts précis, ou des catégories entières, en explorant les terminologies.

L'onglet `Find` permet de rechercher des concepts à partir de mots-clés ou de leur code.

Requête (*Query Tool*)



Le panneau principal de l'interface i2b2, qui permet de construire une requête permettant d'identifier des patients.

Pour créer une requête il suffit de cliquer-déposer des concepts depuis le navigateur de concepts vers un des groupes. Il est possible de sélectionner des catégories entières en une fois.
Lorsque le concept concerne une valeur numérique, il est possible soit de sélectionner tous les patients ayant eu une mesure de ce paramètre, soit de spécifier des conditions sur la valeur.

Les concepts d'un même groupe sont traités entre eux avec "OU" (par exemple patient en réanimation médicale OU en soins continus), les concepts entre différents groupes avec "ET" (par exemple, sexe masculin ET Âge > 50 ans). Il est possible d'ajouter autant de groupes que possible pour affiner la recherche. Il est également possible de spécifier des périodes temporelles (bouton `Dates`), le fait qu'un évènement ait eu lieu à plusieurs reprises (`Occurs`), ou de transformer un critère d'inclusion en critère d'exclusion (`Exclude`).

Le type de requête le plus basique (par défaut et choisi avec le menu `Temporal Constraint:` "Treat all groups independently") permet de sélectionner tous les patients ayant présenté au moins une fois chacun des critères d'inclusion, que ce soit lors de la même visite ou non.

Une autre possibilité est de faire une requête de type "Selected groups occur in the same financial encounter", qui permet de sélectionner tous les patients ayant présenté *l'ensemble* des critères lors de la même visite. Il reste possible à l'aide du choix "Occurs in Same Encounter" ou "Treat Independently" de choisir quels groupes constituent les critères simultanés.

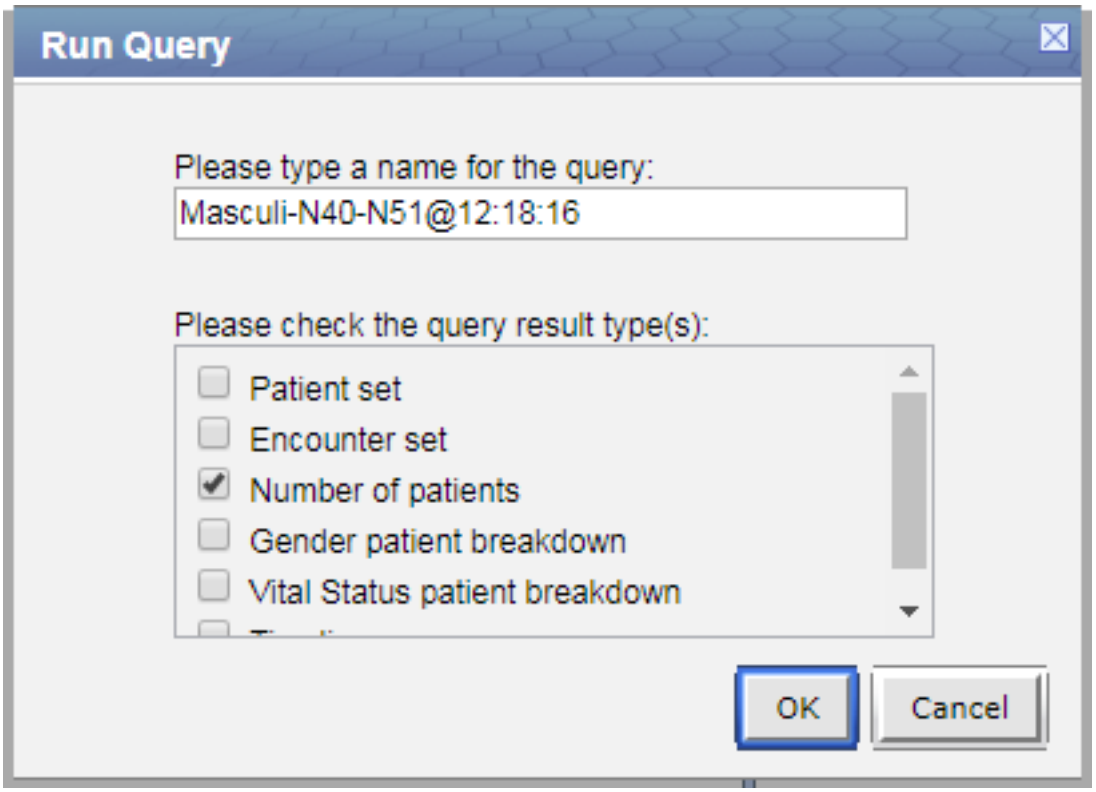
Enfin, le type de requête le plus complexe, "Define sequence of Events", qui fait apparaître un nouveau menu. Ce type de requête permet de définir une population «de base» ("Population in

which events occur"), et de rechercher des séquences d'évènements. Chaque «évènement» fait l'objet d'une requête ("Event n", il est possible de chercher un nombre arbitraire d'évènements). Le panneau "Define order of events" permet de définir les relations temporelles de la séquence d'évènements (par exemple : l'évènement 1 s'est terminé entre 7 jours et 30 jours avant l'évènement 2).

Il est toujours possible de rajouter des groupes de critères de sélection avec le bouton `New Group` situé en bas à droite de l'interface de requête, ainsi que de réinitialiser les champs avec le bouton `Clear`.

Résultats

Une fois les critères de recherche définis, il suffit de cliquer sur `Run Query`.



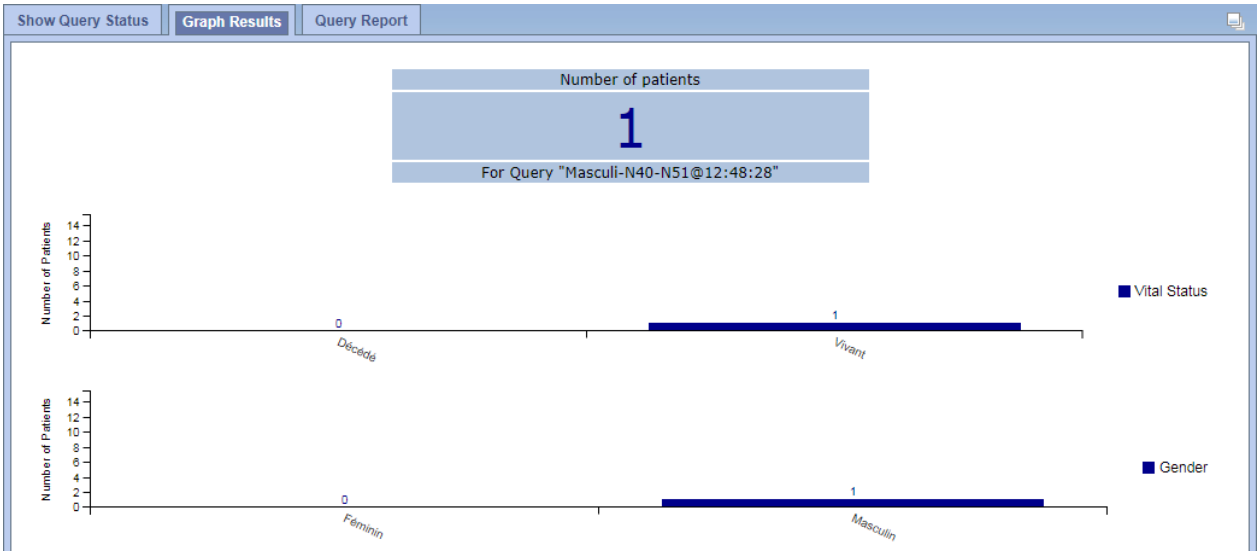
Par défaut, seul le nombre de patients correspondant aux critères sera calculé. Si l'utilisateur n'a pas accès aux données complètes, cocher la case demandant la liste des patients n'aura dans tous les cas aucun effet.

Il est possible de nommer la requête pour la retrouver plus facilement, sinon elle prend par défaut un nom composé des différents critères, avec l'heure et la date à laquelle elle a été effectuée.

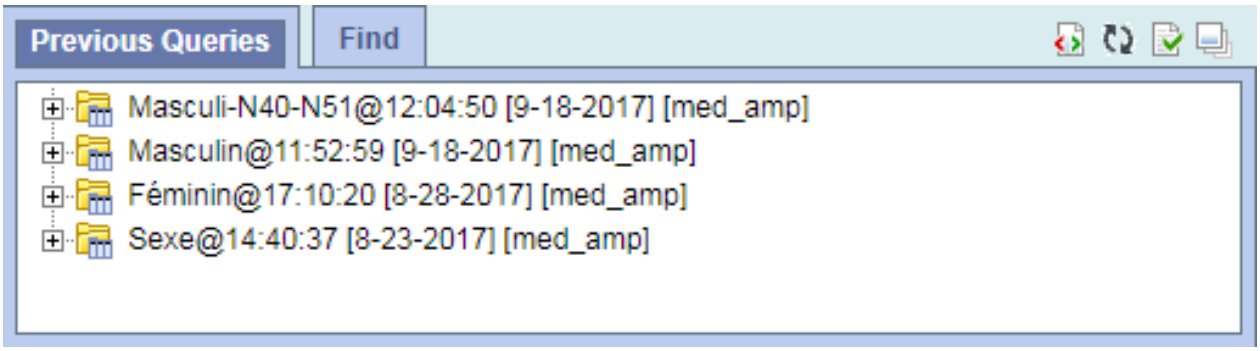
En plus du compte de patients, il est possible de demander l'extraction de la liste des patients et des hospitalisations, pour les utilisateurs ayant accès aux données complètes.

Il est également possible d'obtenir quelques statistiques descriptives de la démographie de l'échantillon : répartition par sexe et par statut vital, et ligne temporelle montrant les occurrences des différents critères de sélection pour chaque patient.

Une fois la requête lancée, et après un petit temps de calcul, les résultats sont affichés dans le panneau `Show Query Status` et ses différents onglets.



Historique



Toutes les requêtes passées sont gardées en historique, et une simple glisser-déposer d'une requête passée vers l'outil de construction de requête permet de ré-exécuter cette requête, éventuellement en ajoutant des critères.

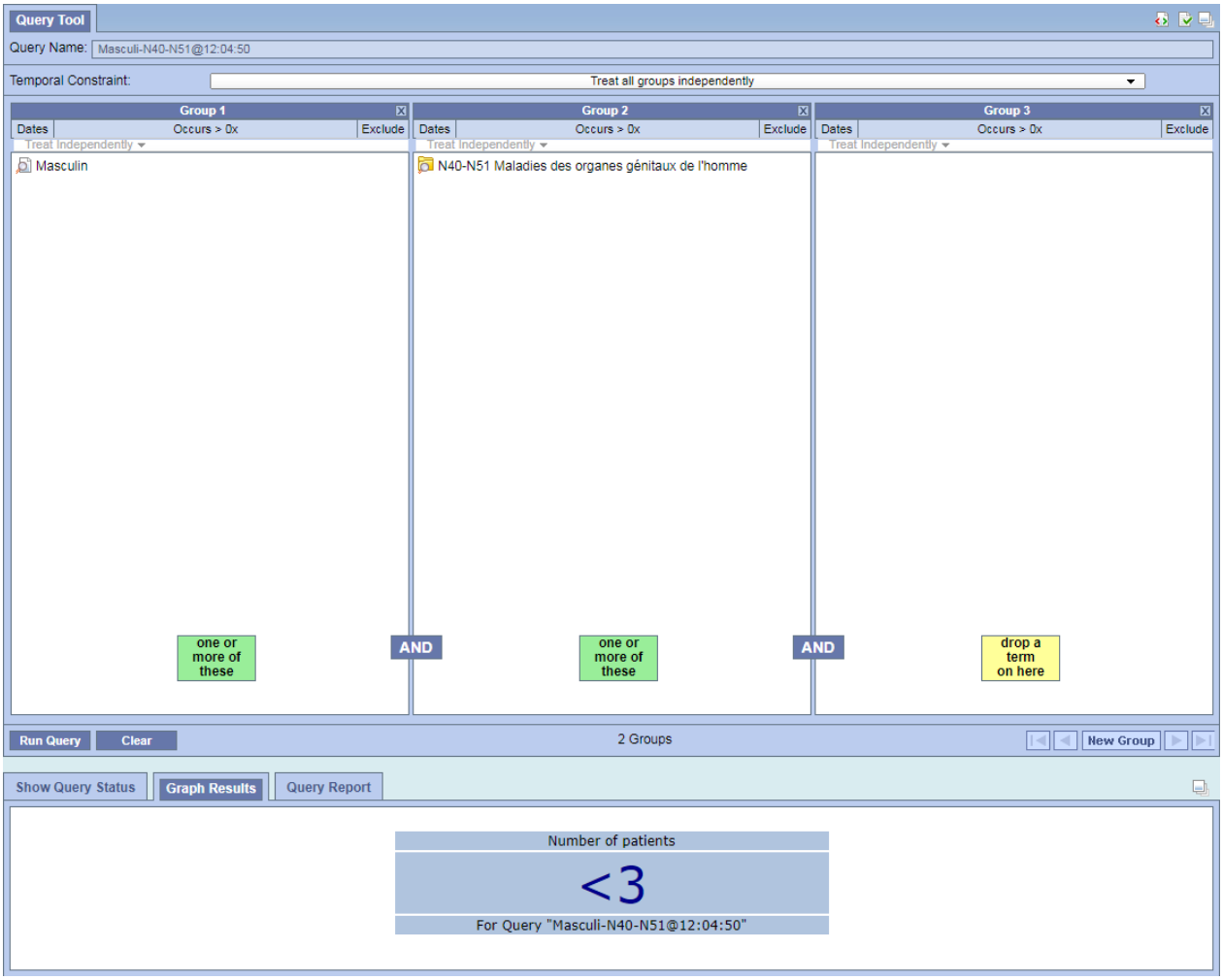
Il est possible de chercher parmi les requêtes passées (onglet `Find`), et les gérer (renommer ou supprimer).

Il est également possible de garder certaines requêtes plus spécifiquement en les déplaçant de la fenêtre `Previous Queries` vers la fenêtre `Workplace`. Les requêtes placées dans la `Workplace` peuvent être annotées avec un texte descriptif plus détaillé, et partagées avec le «manager» du projet.

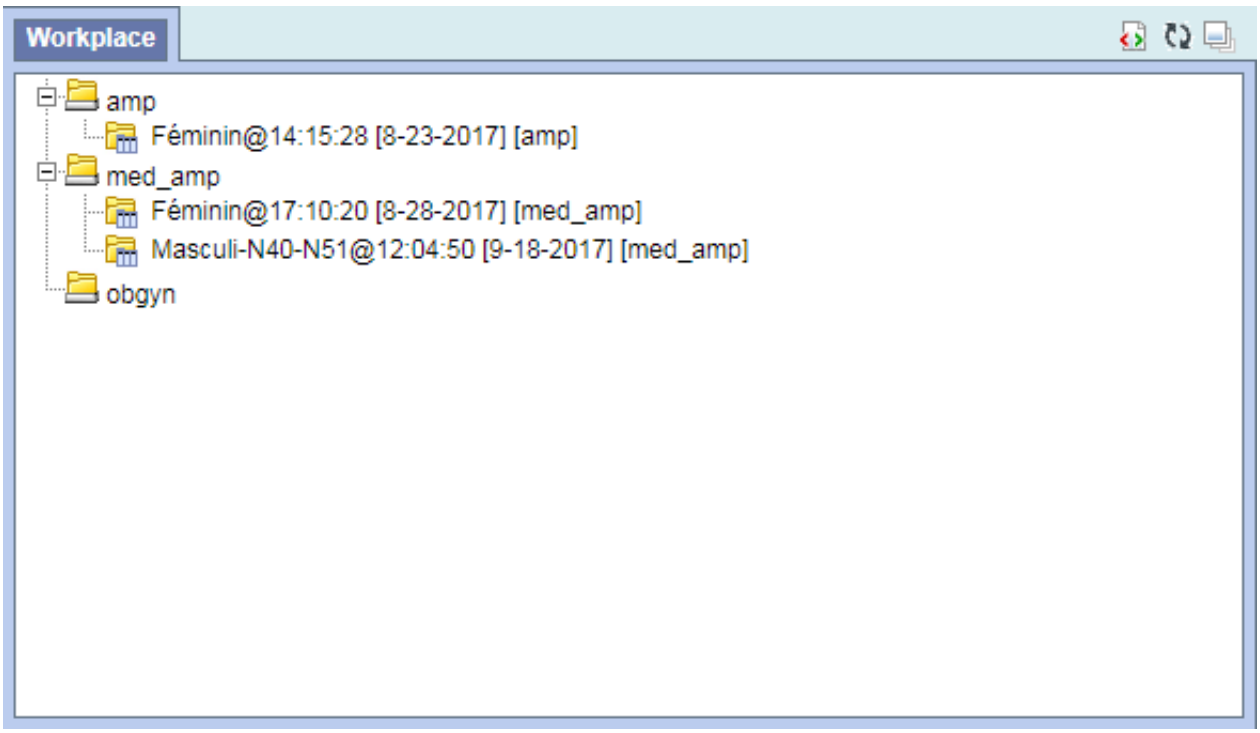
Workplace

Lorsqu'une requête est faite sur un périmètre sur lequel l'utilisateur n'a qu'un accès masqué aux données, le résultat de la requête montrera un compte approximatif.

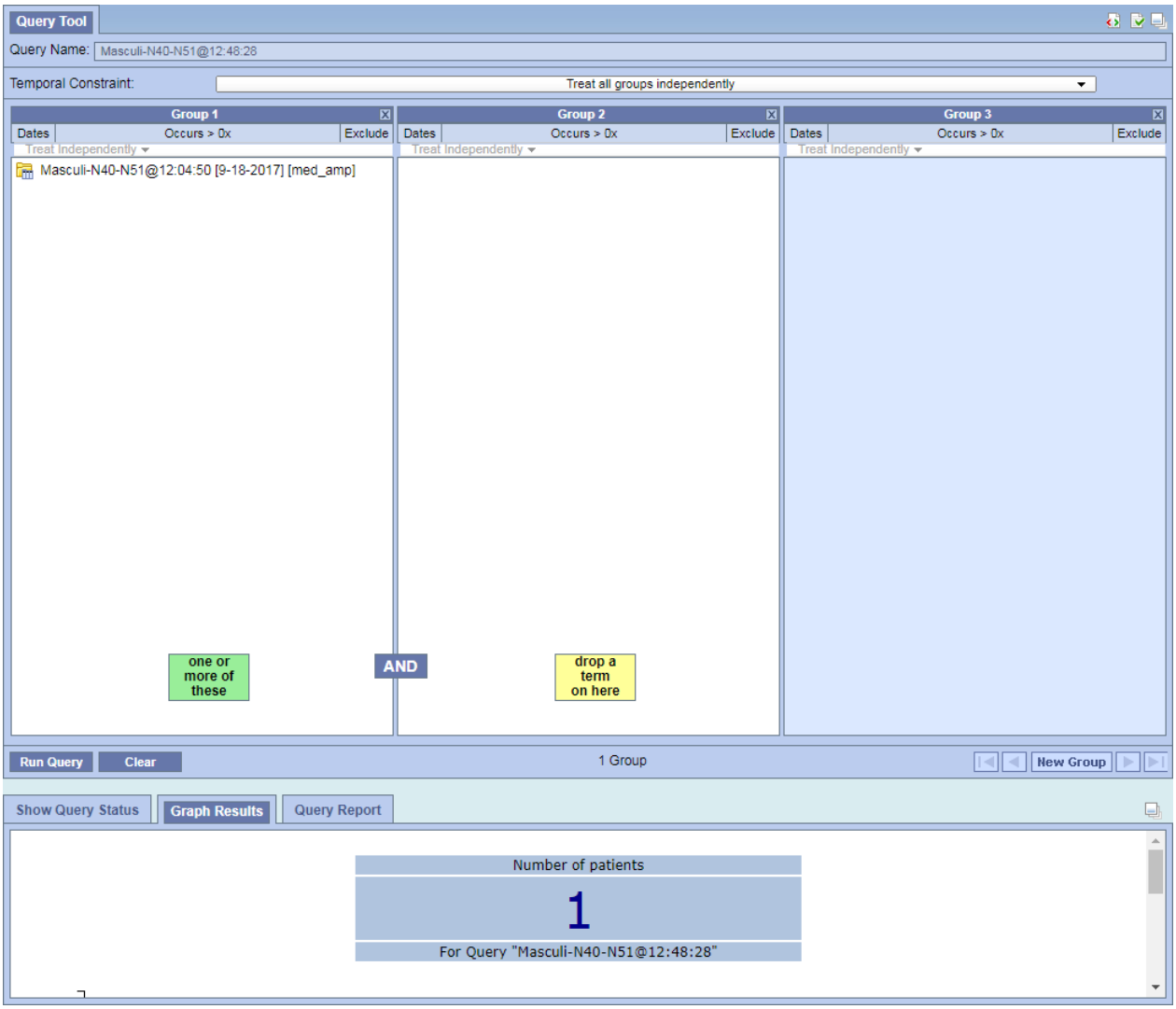
Par exemple, une requête s'intéressant aux patients masculins ayant présenté une maladie des organes génitaux, dans le pôle Gynécologie-Obstétrique, donnera le résultat suivant :



Si l'utilisateur ayant produit cette requête l'enregistre dans son `Workplace` (par exemple l'utilisateur med_amp, représentant un praticien du service d'AMP Clinique), le «manager» du projet Gynécologie-Obstétrique (soit le chef de pôle dans ce cas) pourra avoir accès à cette requête.

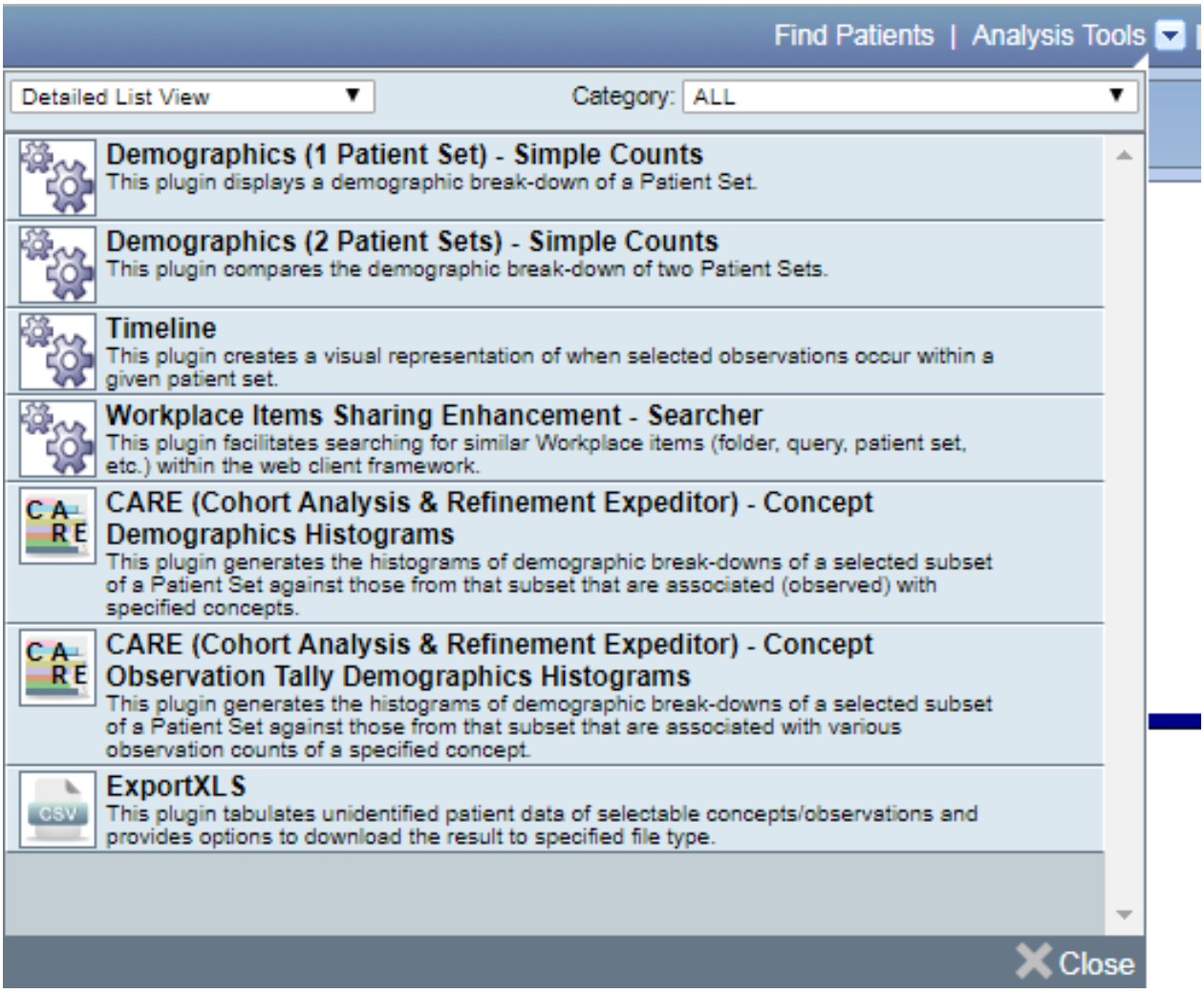


L'exécution de cette requête par le chef de pôle donnera un résultat de compte de patient exact, et permettra l'extraction de donnéresult.png



Analyses

Les utilisateurs ayant accès aux données complètes ont également accès aux outils d'analyse fournis par i2b2.



Les outils disponibles sont les suivants :

Demographics 1

Descriptif sommaire démographique de l'échantillon sélectionné : répartition par âge, sexe, statut vital.

Demographics 2

Descriptif sommaire démographique de deux échantillons provenant de deux requêtes différentes.

Timeline

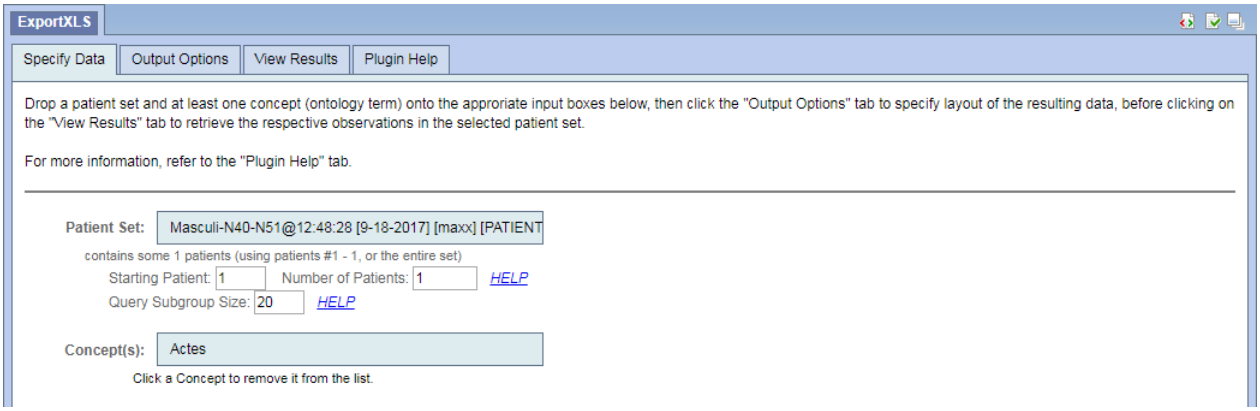
Frise chronologique de survenue de concepts sélectionnés pour une cohorte de patients.

Concept Demographics Histograms

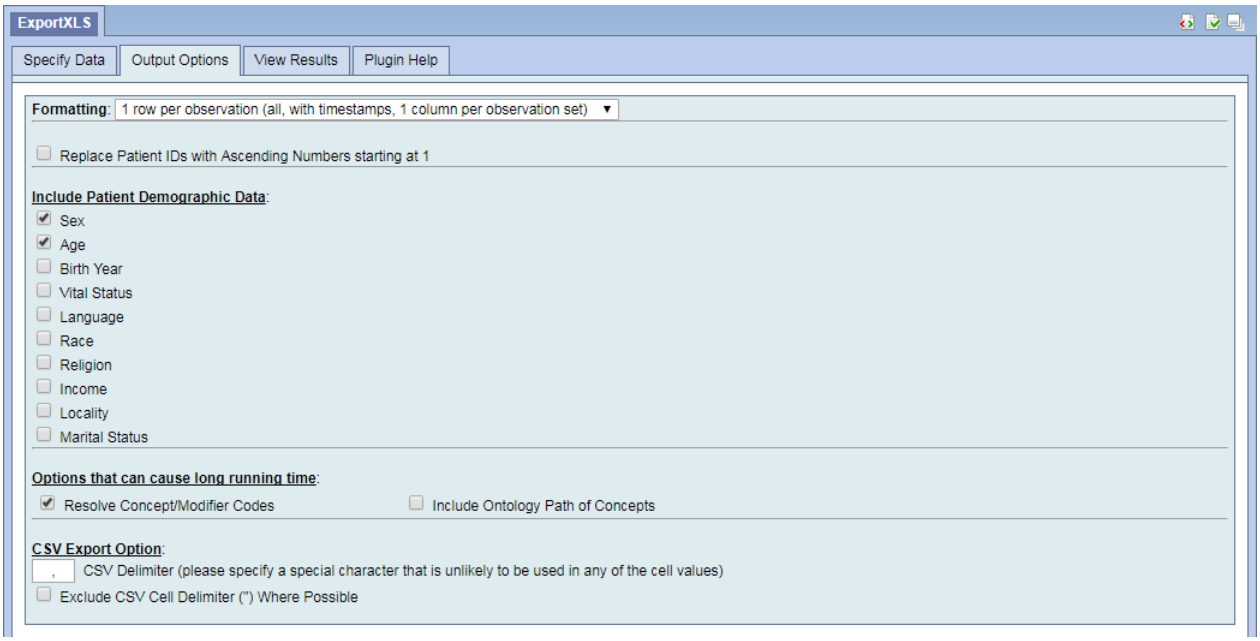
Descriptif sommaire de concepts sélectionnés pour une cohorte de patients.

ExportXLS

Outil d'export de données.



Le premier panneau sert à spécifier la cohorte d'intérêt, le nombre de patients dont on souhaite obtenir les données, et les données à extraire.



Le second panenau permet de régler les options d'extraction (présentation des données, format du fichier, ajout de données démographiques, etc.)

Il est recommandé de choisir le format affiché dans la capture d'écran pour obtenir chacun des concepts pour chacun des patients, avec la date exacte pour chaque occurrence. La case "Resolve Concept/Modifier Codes" permet d'afficher l'intitulé correspondant à chaque code de concept.

VU

NANCY, le **28 août 2017**
Le Président de Thèse

Professeur Nicolas JAY

NANCY, le **07 septembre 2017**
Pour le Doyen de la Faculté de Médecine
Le Vice-Doyen,

Professeur Marc DEBOUVERIE

AUTORISE À SOUTENIR ET À IMPRIMER LA THÈSE/ 10 002

NANCY, le **21 septembre 2017**

LE PRÉSIDENT DE L'UNIVERSITÉ DE LORRAINE,

Pierre MUTZENHARDT

RÉSUMÉ DE LA THÈSE

L'apparition des dossiers patient informatisés et la collection croissante de données liées aux soins a permis l'émergence d'entrepôts de données pouvant être utilisés en recherche clinique pour identifier et retrouver les données liées à un groupe de patients.
La plateforme i2b2 choisie ici est à la fois un entrepôt et un outil de requêtes, disponible sous licence open source.
L'objectif était l'étude de la faisabilité du déploiement de la plateforme i2b2 au CHRU de Nancy, explorant la représentation et l'intégration des données du SIH, et les règles de gouvernance.
Un package R (R2b2) a été développé spécifiquement, permettant de gérer l'entrepôt, manipuler et intégrer les données.
Les sources de données existantes dans le SIH ont été étudiées pour leur intégration, et un modèle de représentation des données a été choisi (ou conçu lorsque cela était nécessaire) pour organiser les concepts.
Un dispositif a été mis en place pour protéger l'accès aux données, en miroir de la responsabilité médicale au CHRU et respectant les règles de collection et d'accès aux données de soins, permettant un retour direct aux praticiens.
Un prototype fonctionnel est fourni, intégrant les données de 118330 patients lors de 247580 hospitalisations depuis début 2016, totalisant plus de 18 millions d'observations.
Le travail conduit autour de ce prototype a permis de dégager plusieurs axes de conduite pour la mise en production du projet, et les règles régissant les accès.

TITRE EN ANGLAIS

Deployment of a Clinical Data Warehouse in Nancy teaching hospital

THÈSE : MÉDECINE SPÉCIALISÉE – ANNÉE 2017

MOTS CLES :

Entrepôt de données, représentation des connaissances, i2b2, gouvernance

INTITULÉ ET ADRESSE :

UNIVERSITÉ DE LORRAINE
Faculté de Médecine de Nancy
9, avenue de la Forêt de Haye
54505 VANDOEUVRE LES NANCY Cedex