ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

RAPPORT TECHNIQUE
PRÉSENTÉ À
L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

COMME EXIGENCE PARTIELLE
À L'OBTENTION DE LA
MAÎTRISE
EN GÉNIE DES TECHNOLOGIES DE L'INFORMATION

PAR
Liliana ALVARADO MALLMA

ANALYSE DE L'UTILISATION POTENTIELLE DE LA STRUCTURE DE DONNÉES DE L'UNIVERSITÉ BERKELEY (ADAM)

MONTRÉAL, LE 30 NOVEMBRE 2015

PRÉSENTATION DU JURY

CE RAPPORT DE PROJET A ÉTÉ ÉVALUÉ

PAR UN JURY COMPOSÉ DE :

Professeur Alain April, directeur de projet
Département de génie logiciel et TI à l'École de technologie supérieure

Professeur Sègla Kpodjedo, président du jury
Département de génie logiciel et TI à l'École de technologie supérieure

REMERCIEMENTS

Je tiens tout d'abord à offrir mes remerciements à M. Alain April, pour avoir accepté de me diriger dans ce projet. Je le remercie pour son temps, sa patience et son support, ses judicieux conseils m'ont aidé à encadrer le travail et à simplifier la complexité de la recherche.

Je voudrais également souligner le travail de M. David Lauzon, étudiant de doctorat sous la supervision de M. April, pour son support et ses commentaires pertinents qui m'ont aidé à enrichir ce travail.

Je suis également redevable envers Dr Pavel Hamet et M. Michael Philips, du Centre de Recherche du CHUM (CRCHUM), qui m'ont accueilli au sein de l'organisation et qui ont consacré de leur temps afin que je puisse avoir les ressources nécessaires pour effectuer mon travail. Un remerciement spécial à M. François Harvey et à M. François Marois, bio-informaticiens du CRCHUM, qui m'ont grandement aidé à la réalisation de ce document.

Finalement, je remercie ma famille pour son support et ses encouragements tout au long de la réalisation de mes études de maîtrise.

ANALYSE DE L'UTILISATION POTENTIELLE DE LA STRUCTURE DE DONNÉES DE L'UNIVERSITÉ BERKELEY (ADAM)

Liliana ALVARADO MALLMA

RÉSUMÉ

Face aux récents développements technologiques, notamment la technologie ADAM (c.-à-d. le BigData appliqué à la génomique) et les nouvelles plateformes de séquençage du génome humain, la recherche dans le domaine de la médecine personnalisée se trouve aujourd'hui en pleine expansion. Les centres de recherche et laboratoires spécialisés du domaine de la santé utilisent de plus en plus les données génétiques et doivent faire évoluer rapidement leurs infrastructures technologiques afin de faire face aux nouveaux défis de traitements de données massives en temps réel. À Montréal, le Centre de Recherche Hospitalier de l'Université de Montréal (le CRCHUM), en partenariat avec les étudiants de la maîtrise en génie logiciel de l'École de Technologie Supérieure (ÉTS), visent cet objectif. Plusieurs opportunités d'amélioration du flux de travail du séquencement et de l'analyse du génome sont possibles, notamment par l'utilisation de logiciels libres disponibles, par exemple : la technologie ADAM. Cependant, l'adoption de nouvelles technologies doit suivre une méthodologie de travail structurée qui débute par la compréhension du domaine, des problématiques et des besoins des bio-informaticiens. Une mauvaise compréhension des besoins peut entrainer, des dépenses d'efforts sans résultats probants, des développements inutiles et même l'échec du projet d'amélioration. C'est pour cette raison que l'analyse d'affaires, suivant l'approche du « *Business Analysis Body of Knowledge* » (BABOK), devient une étape incontournable dès le début de la réalisation de tout projet de génie logiciel.

Ce rapport technique a été réalisé dans le cadre d'un projet pratique et appliqué, d'une valeur de 6 crédits, à la fin de la maîtrise en génie logiciel et en technologies de l'information. Ce projet a pour but d'analyser la situation actuelle du flot de travail de génotypage du laboratoire de recherche du Dr Hamet au CRCHUM en suivant la méthodologie d'analyse d'affaires structurée du BABOK qui permettra de comprendre le domaine d'affaires et les

problématiques existantes, d'identifier et de documenter les processus d'affaires afin d'évaluer l'utilisation potentielle de la structure de données proposée par la technologie ADAM de l'Université de Berkeley. Cette première recherche exploratoire précède plusieurs travaux de recherche qui visent à concevoir et rendre disponible une plateforme technologique BigData en logiciel libre utile pour les petits centres de recherche en santé au Québec.

ANALYSYS OF THE POTENTIAL USE OF BERKELEY'S UNIVERSITY ADAM DATA STRUCTURE

Liliana ALVARADO MALLMA

ABSTRACT

Faced with the latest technological developments, including the ADAM Technology (i.e. Big Data technologies applied to genomics) and new platforms for sequencing the human genome, research in the field of personalized medicine is evolving rapidly. Research centers and genomics laboratories must adapt their technological infrastructure to meet new challenges. In Montreal, the Hospital Research Center at the University of Montreal (CHUM Research Centre), in partnership with the École de Technologie Supérieure (ÉTS) Software Engineering master program, are investigating the use of ADAM. Several opportunities for the improvement of current health research work flows have been identified using the freely available ADAM technology. However, the adoption of new technologies must follow a structured methodology that begins by understanding the issues and bioinformatics pressing needs. Poor understanding of needs often results in failure to determine key/strategic requirements and may result in loss of efforts, quick fix and worsening the existing situation. It is for this reason that business analysis becomes a crucial step before the design and implementation of any IT project.

This 6 credits applied research report was produced as part of the completion of a Masters in Software Engineering. The aim of this applied research project was to analyze the genomics research workflow of Montreal hospital research center. It followed a structured business analysis methodology, the BABOK, for understanding the genetics research domain, identify and document the existing genetic workflow processes with the objective of replacing the existing data structures by the ADAM technology developed by the university of Berkeley, California. This first exploration, at Dr Pavel Hamet's CHUM research lab, is an important analysis step required for several potential research projects that aim to design and make available open source BigData technologies and platforms for small health research centers in Quebec.

TABLE DES MATIÈRES

Page

INTRODUCTION 1

CHAPITRE 1 REVUE DE LA LITTÉRATURE 5
1.1 L'analyse des exigences 5
1.1.1 Définitions préliminaires 6
1.2 Modèles d'analyse des exigences 6
1.2.1 Selon la perspective de l'analyse d'affaires 6
1.2.2 Selon la perspective de l'analyse de systèmes 9
1.3 Comparaison entre le BABOK et le SWEBOK 12
1.4 Sommaire du chapitre 14

CHAPITRE 2 ÉTUDE DE CAS LABORATOIRE DU Dr HAMET DU CRCHUM 15
2.1 Mise en contexte 15
2.1.1 Introduction à la génomique 15
2.1.1.1 L'étude du génome 16
2.1.1.2 Problématique du domaine 17
2.1.2 La technologie ADAM 17
2.1.3 Le laboratoire du Dr Hamet au CRCHUM 19
2.2 Problématique spécifique du projet 20
2.2.1 Objectifs du projet 20
2.2.2 Portée du projet 21
2.3 Méthodologie 21
2.3.1 Méthodologie utilisée pour l'analyse 21
2.3.2 Méthodologie utilisée pour la modélisation de la base de données 28
2.4 Sommaire du chapitre 31

CHAPITRE 3 PRÉSENTATION DES RÉSULTATS 33
3.1 Présentation des résultats 33
3.1.1 Les résultats de l'analyse d'affaires 33
3.1.2 Les résultats de l'analyse des exigences 36
3.1.2.1 Les exigences du modèle de données 36
3.1.2.2 Les exigences des données 37
3.1.3 Les résultats de la modélisation 41
3.1.3.1 Le modèle relationnel de données 41
3.1.3.2 Correspondance avec le modèle d'ADAM 43
3.2 Analyse des résultats 45
3.2.1 Analyse des résultats de l'élicitation des exigences 45
3.2.2 Analyse des résultats de la modélisation 45
3.2.3 Analyse des résultats de la validation 46
3.3 Revue critique du travail 46
3.4 Travaux futurs 48

CONCLUSION 51

ANNEXE I DOCUMENT DE SPÉCIFICATION DES EXIGENCES D'AFFAIRES ET D'APPLICATION 53

ANNEXE II DESCRIPTION DES PROCESSUS AU CRCHUM 89

ANNEXE III ACTIVITÉS DE LA PRÉPARATION DES DONNÉES 91

ANNEXE IV DICTIONNAIRE DE DONNÉES DE LA BASE DE DONNÉES PROGNOMIX 103

ANNEXE V MODÈLE DE DONNÉES D'ADAM 107

ANNEXE VI LE MODÈLE DE DONNÉES PROPOSÉ 109

LISTE DE RÉFÉRENCES BIBLIOGRAPHIQUES 113

BIBLIOGRAPHIE 115

LISTE DES TABLEAUX

		Page
Tableau 2.1	Énoncé du problème	26
Tableau 2.2	Positionnement de la proposition	26
Tableau 3.1	Les besoins d'affaires de l'étape de préparation des données	34
Tableau 3.2	Les exigences du modèle de données	37
Tableau 3.3	Les exigences des données des études	38
Tableau 3.4	Les exigences des données de génotypage	40

LISTE DES FIGURES

		Page
Figure 1.1	Les domaines de connaissance selon le BABOK	7
Figure 1.2	Les éléments de l'analyse des exigences selon le SWEBOK	10
Figure 2.1	Les étapes de la méthodologie d'analyse d'exigences	22
Figure 2.2	Les étapes de la méthodologie de modélisation de la base de données	29
Figure 3.1	Modèle de données proposé	42
Figure 3.2	Correspondance entre le modèle de données proposé et le modèle ADAM	44

LISTE DES ABRÉVIATIONS, SIGLES ET ACRONYMES

IIBA	International Institute of Business Analysis
IEEE	Institute of electrical and electronics engineers
BABOK	Guide du corpus de connaissances de l'analyse d'affaires
SWEBOK	Software engineering body of knowledge
CRCHUM	Centre de recherche du centre hospitalier de l'Université de Montréal
T2D	Acronyme en anglais du diabète de type 2
GWAS	Acronyme en anglais de « *Genome wide association study* »

INTRODUCTION

La recherche dans le domaine de la génomique appliquée à la santé peut sauver des vies grâce aux avancements en médecine préventive personnalisée. Le diagnostic précoce de maladies à l'aide de l'analyse du génome du patient permet de personnaliser ses traitements. Plus le diagnostic est posé tôt, plus le traitement sera efficace. Actuellement, deux tendances technologiques convergent afin d'aider à accomplir les objectifs de la médecine personnalisée.

En premier lieu, les développements des technologies de séquençage du génome humain permettent d'identifier le génome des individus de plus en plus rapidement et à des coûts de plus en plus bas. L'analyse complète du premier génome humain a pris dix ans et a coûté 3 $ milliards de dollars. Aujourd'hui, il est devenu possible de séquencer le génome complet d'un individu en quelques heures, et ce, pour seulement quelques milliers de dollars. Ces nouvelles technologies produisent une quantité massive de données (c.-à-d. de l'ordre des pétaoctets), et ce, en très peu de temps (Bhardwaj *et coll.,* 2014, p. 1).

En deuxième lieu, les technologies émergentes du domaine du BigData sont en train de révolutionner les méthodes traditionnelles de traitement de l'information. Le BigData est un ensemble de technologies et de logiciels de traitement distribué utilisés pour gérer une quantité massive de données très rapidement. Les volumes de données, du domaine de la génomique, peuvent rapidement représenter des pétaoctets, voire des exaoctets (c.-à-d. un milliard de gigaoctets) (Andreu-Perez *et coll.,* 2015, p. 1). Les structures de données existantes (c.-à-d. les bases de données relationnelles) et les infrastructures matérielles traditionnelles n'ont pas la capacité de traiter une telle quantité de données avec un temps de réponse rapide. C'est à partir de cette problématique que les technologies émergentes du BigData ont été conçues.

Dans le but de répondre plus efficacement aux besoins de la médecine préventive, et en intégrant ces deux tendances technologiques, l'Université de Berkeley a développé la

technologie ADAM. Créé pour régler les problèmes de BigData en génomique, ADAM est un cadriciel de traitement d'une quantité massive d'information et un format de données adapté à la génomique. Avec une structure de données plus appropriée aux patrons de lecture et d'écriture parallèles, des méthodes de traitement de données plus performantes et une architecture distribuée, il est possible de réduire radicalement le temps d'analyse d'une grande quantité de génomes (Massie *et coll*., 2013, p. 3).

Au Centre hospitalier de recherche de l'Université de Montréal (le CRCHUM), l'équipe de chercheurs du laboratoire du Dr Pavel Hamet tente de trouver les causes génétiques des complications du diabète type 2 (acronyme anglais T2D). L'approche technologique utilisée par ses bio-informaticiens s'appuie sur des technologies de traitement de données traditionnelles, c'est-à-dire les bases de données relationnelles, qui sont de moins en moins adaptées à la croissance importante des données génomiques et les besoins d'analyse de ces données en temps réel. Des occasions d'améliorations de ces technologies pourraient être réalisées en utilisant le logiciel libre ADAM.

L'objectif de ce projet de recherche appliquée est d'analyser les problématiques de génotypage, du laboratoire du Dr Hamet, en suivant une méthodologie d'analyse d'affaires structurée qui permettra d'identifier les problématiques existantes, de préciser les exigences pertinentes d'amélioration, et de les documenter afin d'évaluer l'utilisation potentielle de la technologie et la structure de données proposée par ADAM.

Le premier chapitre de ce rapport présente une revue de la littérature du domaine de l'analyse des exigences logicielles. Le deuxième chapitre présentera la méthodologie de travail utilisée pour aborder la problématique d'étude et de compréhension du génotypage du laboratoire du Dr Hamet. Ce chapitre débute par une mise en contexte des technologies à utiliser et de l'organisation du laboratoire suivi d'une présentation de l'objectif du travail de recherche permettant de comprendre ses enjeux. Par la suite, une description de la méthodologie d'analyse utilisée basée sur les concepts abordés dans la revue de la littérature scientifique sera présentée. Le troisième chapitre consistera en une présentation des résultats obtenus lors

de l'expérimentation et d'une revue critique de cette expérience. Pour terminer, une conclusion récapitulera l'ensemble des observations et des travaux effectués.

CHAPITRE 1

REVUE DE LA LITTÉRATURE

En 1994, une étude réalisée par la firme de services-conseils « Standish Group » mettait en évidence que parmi les causes d'échecs des projets de développement de logiciels, 13% correspondaient à une pauvre analyse des exigences (« *The Chaos Report* », 1994). En effet, non seulement les erreurs des exigences étaient fréquentes, elles étaient les plus chères à corriger et pouvaient consommer de 25% à 40% du budget du projet. Des études subséquentes du même groupe ont permis de constater que cette tendance se maintient (Eveleens et Verhoef, 2010). Pour aborder cette problématique, depuis plusieurs années, les chercheurs et professionnels travaillent à établir une méthodologie d'analyse des exigences, permettant de récolter des exigences de qualité. Plusieurs modèles d'analyse existent actuellement. Ce chapitre consiste en une revue de la littérature du domaine de l'analyse des exigences logicielles vue sous deux perspectives : le modèle d'analyse d'affaires et le modèle d'analyse de systèmes. À la fin du chapitre, un résumé des concepts traités sera réalisé pour souligner ceux qui seront utiles à la réalisation du présent document.

1.1 L'analyse des exigences

Un des plus grands défis de l'étape d'analyse de systèmes informatiques est la gestion des exigences. La qualité des exigences récoltées lors de l'analyse joue un rôle important dans la réussite d'un projet informatique. Bien comprendre les exigences demande non seulement des habilités personnelles et communicationnelles, mais aussi l'utilisation d'une méthodologie d'analyse structurée. C'est ce que sera résumé dans cette section.

1.1.1 Définitions préliminaires

Avant de poursuivre, deux définitions proposées par Leffingwell et Widrig (2000) sont utiles à la compréhension :

- « Une **exigence logicielle** est la *capacité dont un utilisateur a besoin pour résoudre un problème ou atteindre un objectif* ». Une exigence cherche à répondre à la question du « *quoi* », c'est-à-dire ce que le logiciel doit faire.
- La **gestion des exigences** est une « *approche systématique pour expliciter, organiser et documenter les exigences d'un système* ». C'est un processus qui permet d'établir et de maintenir l'accord entre le client et l'équipe du projet sur les exigences changeantes du système.

1.2 Modèles d'analyse des exigences

Deux modèles d'analyse des exigences seront étudiés. Ils représentent deux perspectives d'analyse qui se complémentent et qui permettront de mieux atteindre les objectifs de ce projet de recherche appliquée.

1.2.1 Selon la perspective de l'analyse d'affaires

L'analyse d'affaires se concentre sur l'analyse du domaine d'affaires avant de proposer une solution, qui ne sera pas nécessairement technologique. L'« *International Institute of Business Analysis* » (IIBA) définit un corpus de connaissance pour l'analyste d'affaires (c.-à-d. le « *Business Analysis Body of Knowledge* » (BABOK)). Ce guide regroupe les meilleures pratiques du domaine, structurées en sept domaines de connaissance.

La figure 1.1 présente les domaines de connaissances proposées par le BABOK. Les sections prochaines consistent à un résumé des six domaines de connaissances qui couvrent l'analyse des exigences (IIBA 2009).

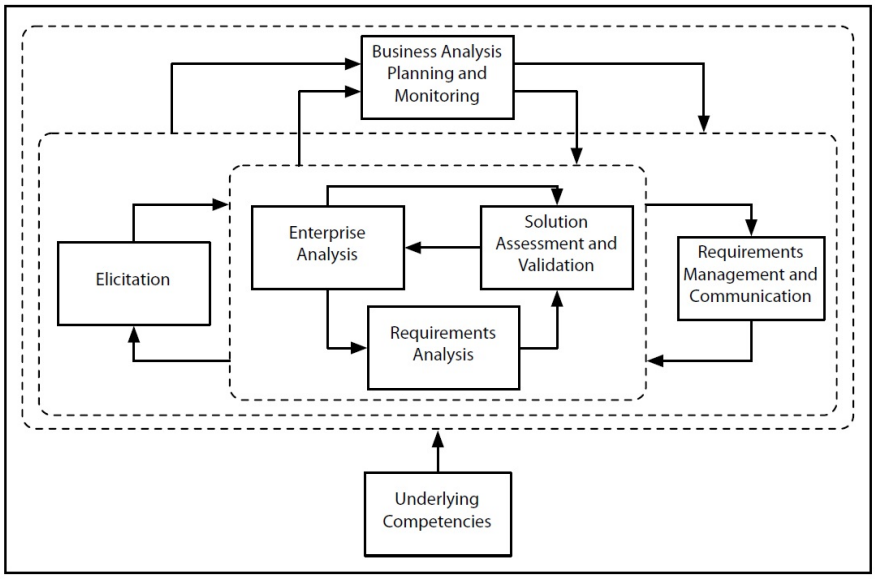


Figure 1.1 Les domaines de connaissance selon le BABOK (IIBA 2009, p. 7)

1) *La planification* suggère les activités nécessaires pour bien planifier l'effort d'analyse. Parmi ces activités, se retrouvent: la planification de la démarche, l'analyse des parties prenantes, la sélection de techniques d'analyse d'affaires, le processus de gestion des exigences et l'évaluation du progrès du travail.

2) *L'élicitation* a pour objectif d'identifier et de comprendre les besoins et préoccupations des parties prenantes par rapport à une problématique existante. Les activités de l'élicitation sont:

   - La préparation: pour assurer les ressources nécessaires pour l'élicitation,
   - L'exécution: consiste à rencontrer les parties prenantes afin de connaître leurs besoins,
   - La documentation des résultats: pour rendre explicite les besoins, et,
   - La confirmation des résultats: pour valider les exigences récoltées.

3) *La gestion et la communication des exigences* décrivent comment gérer les conflits, difficultés et changements lors de l'analyse. L'objectif est d'assurer la compréhension des exigences entre les parties prenantes et l'équipe du projet.

4) *L'analyse de l'entreprise* permet de mieux comprendre l'organisation. Cette analyse permet d'identifier les besoins d'affaires et de définir la portée de la solution qui sera proposée. La définition du besoin d'affaires permet de comprendre l'entreprise, ses objectifs, les problèmes à résoudre et les résultats attendus par les parties prenantes. Les activités de ce domaine de connaissance sont :

   - La détermination des besoins d'affaires : cette activité sert à identifier le pourquoi du changement. Un besoin d'affaires peut être exprimé par le client comme un problème, un souhait ou une préoccupation. Pour valider l'existence réelle d'un tel problème, l'analyste peut s'appuyer sur l'étude des buts et objectifs ainsi que sur les expectatives des parties prenantes. Une des approches suggérées par le BABOK est l'approche de haut en bas (de l'anglais top-down), c'est-à-dire, exprimer le problème en termes généraux et aller de plus en plus en détail jusqu'à trouver les causes réelles.

   - L'évaluation des capacités de l'organisation pour satisfaire les besoins d'affaires : un besoin d'affaires peut ne pas être réalisable si l'entreprise n'a pas les capacités requises pour l'atteindre (argent, équipement, technologie, etc.). Cette activité d'évaluation permet de s'assurer que tous les éléments nécessaires sont ou seront en place pour accomplir un objectif d'affaires.

   - La détermination de la solution la plus faisable : selon les capacités de l'entreprise.

   - La détermination de la portée de la solution : cette activité permet de raffiner et de valider la portée de la solution à proposer. Avec toute cette information, le cas d'utilisation peut être établi.

5) *L'analyse des exigences* permet de définir les exigences selon les besoins des parties prenantes. La formulation de ces exigences permettra de proposer des alternatives de solution réalisables. Les activités de ce domaine de connaissance sont : la priorisation des exigences, l'organisation des exigences, la précision et modélisation des exigences, la détermination des hypothèses et contraintes, la vérification et la validation des exigences.

6) *La validation et l'évaluation de la solution* décrivent les activités d'évaluation des solutions possibles pour trouver la meilleure solution. L'objectif est d'assurer que la solution retenue correspond aux besoins d'affaires.

1.2.2 Selon la perspective de l'analyse de systèmes

L'analyse de systèmes se concentre sur l'étude des différentes manières de mettre en œuvre un besoin logiciel au sein d'un système d'information, selon une perspective du génie logiciel. Le « *Institute of Electrical and Electronics Engineers* » (IEEE) définit un corpus de connaissances (c.-à-d. le guide « *Software Engineering Body of knowledge* » (SWEBOK)). Ce corpus est conformé de dix chapitres dédiés aux domaines de connaissances spécifiques de l'ingénieur(e) logiciel. Le chapitre 2 du SWEBOK traite de l'analyse des exigences en détail. Les éléments abordés dans ce chapitre sont représentés à la figure 1.2. Les prochaines pages font une synthèse des principaux concepts traités dans le SWEBOK (IEEE 2004).

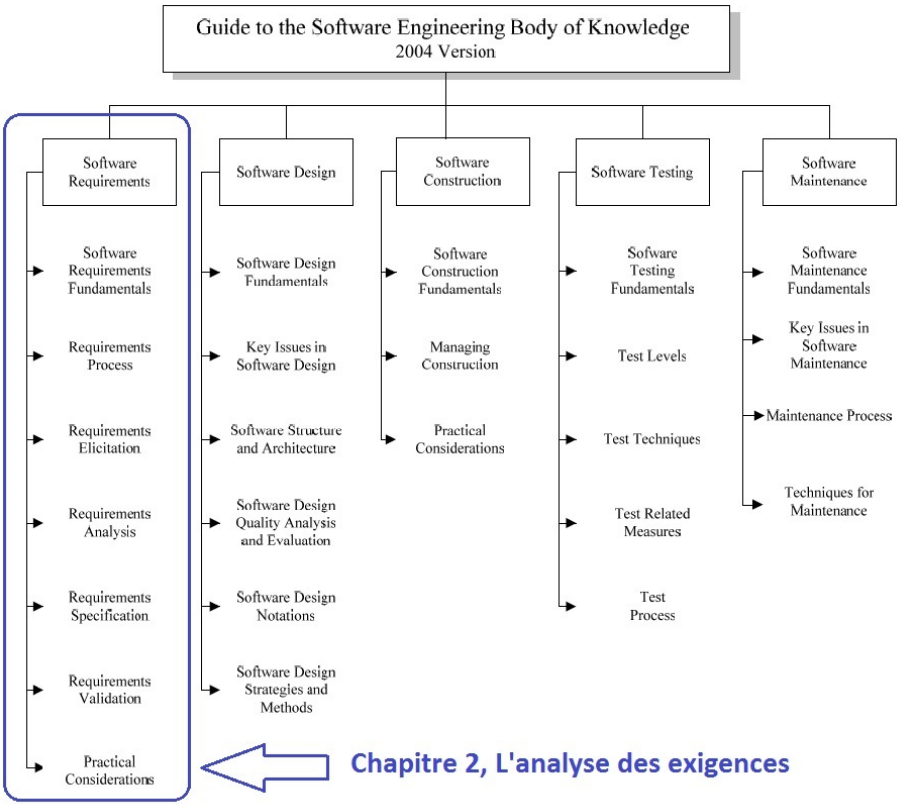


Figure 1.2 Les éléments de l'analyse des exigences selon le SWEBOK (IEEE 2004, p.8)

1) Les fondamentaux des exigences logicielles

Le SWEBOK définit les différents types d'exigence. Une distinction est faite entre les ***exigences fonctionnelles*** aussi appelées capacités, et les ***exigences non fonctionnelles*** aussi appelées contraintes ou exigences de qualité. Dans le premier groupe se retrouvent par exemple, l'entrée de données et les transformations à faire sur ces données, tandis que dans le deuxième, se retrouvent, les exigences de performance du logiciel, la rapidité, la sécurité, la disponibilité, etc. Le SWEBOK stipule aussi que les exigences doivent être

spécifiées clairement sans aucune ambigüité, spécialement dans le cas des exigences non fonctionnelles.

2) Le processus des exigences

   Le SWEBOK stipule que le processus des exigences n'est pas une activité ponctuelle, mais au contraire une activité continue qui s'initie au début du projet et qui est constamment améliorée tout au long du cycle de vie du projet. Ce processus n'est pas rigide, il doit être adapté selon le contexte de l'organisation et du projet. De plus, il s'agit d'un processus multidisciplinaire, l'analyste joue un rôle de médiateur entre les besoins des parties prenantes et l'équipe de développement.

3) L'élicitation des exigences

   L'activité d'élicitation permet de comprendre le problème à résoudre. La communication est très importante dans cette activité. L'analyste doit non seulement comprendre le langage d'affaires des différentes parties prenantes, mais aussi comprendre le langage technique de l'équipe de développement. Pour le SWEBOK, il ne suffit pas de collecter les exigences explicites, exprimées par les utilisateurs, il faut chercher d'autres sources d'exigences. Les exigences peuvent se retrouver dans les objectifs d'affaires, dans les informations sur le domaine, dans l'environnement où le logiciel doit s'exécuter, etc. Différentes techniques d'élicitation sont traitées dans le SWEBOK : les entrevues, la construction de scénarios comme les cas d'utilisation, les prototypes pour clarifier les ambigüités, les rencontres de groupes pour faciliter les échanges d'idées et l'observation. Le choix de la technique dépendra du problème, mais aussi du temps et du budget assigné au projet.

4) L'analyse des exigences

   Cette activité a pour but de résoudre les conflits entre les exigences et d'établir les limites des exigences pour ainsi élaborer les exigences du système. Pour aider à la compréhension du problème, le SWEBOK suggère l'utilisation de modèles conceptuels. Par exemple, le modèle contextuel du logiciel présente graphiquement les interactions

entre le logiciel et son environnement. Ceci aide à l'identification des interfaces. L'analyste peut aussi désigner l'architecture de la solution, cet outil permettra de connecter les exigences avec les composantes logicielles responsables de les satisfaire.

5) La spécification des exigences
   En génie logiciel, cette activité consiste en la création d'un document de spécifications (document Vision). Celui-ci résume les exigences récoltées, et sert non seulement comme un moyen de communication entre les parties impliquées, il permet aussi de maintenir l'accord entre le client et le fournisseur de la solution.

6) La validation des exigences
   Cette étape est nécessaire pour assurer la compréhension des exigences et leurs conformités aux standards de la compagnie et aux attentes du client. Plusieurs moyens de validation sont proposés par le SWEBOK : la revue des exigences, le prototypage, la validation des modèles et les tests d'acceptation.

1.3 Comparaison entre le BABOK et le SWEBOK

Deux corpus de connaissance de l'analyse des exigences ont été présentés. Une comparaison de ces deux modèles a été réalisée par Tu Dang Vuong (2015). Son analyse a servi de référence pour établir les différences. Les points principaux sont résumés ci-dessous:

- **Le domaine d'affaires :** « *le BABOK n'est pas destiné à un domaine spécifique. Le SWEBOK est spécifique au domaine de génie logiciel. La définition des exigences du BABOK est plus élargie que celle du SWEBOK* ». Effectivement, le BABOK classifie les exigences par : exigences d'affaires, exigences des parties prenantes, de la solution et de transition, tandis que le SWEBOK part des exigences du produit ou du processus (c.-à-d. de la solution). Le BABOK propose l'activité d'analyse de l'entreprise pour comprendre les exigences d'affaires, sujet qui n'est pas couvert par le SWEBOK.

- **L'élicitation :** « *Le BABOK se concentre plus sur les activités nécessaires pour bien faire l'élicitation tandis que le SWEBOK se préoccupe plus des sources d'exigences possibles* ». Le BABOK propose des activités spécifiques pour encadrer l'élicitation; tout commence par la préparation de l'élicitation, suivie de l'exécution, la documentation et la confirmation des exigences. Pour chaque activité les intrants et extrants sont bien définis, cela permet à l'analyste de mieux planifier sa démarche. Le SWEBOK, par contre, se concentre sur les différentes sources d'exigences dont il faut tenir compte. Les deux coïncident sur les techniques à utiliser pour l'élicitation.

- **L'analyse :** Les deux modèles couvrent la classification, la priorisation et la modélisation des exigences, mais, à différence du SWEBOK, le BABOK inclut dans cette activité la validation et la documentation des exigences. Pour la modélisation, le BABOK énumère une série de techniques disponibles, et cette information peut être très utile pour l'analyste. Le SWEBOK, par contre, met l'accent, sur l'importance de modéliser les exigences, peu importe la technique. Le SWEBOK considère aussi la conception de l'architecture de la solution et les mesures par points de fonction comme activités qui font partie de l'analyse.

- **La documentation :** « *Le SWEBOK suggère l'utilisation d'un document de spécification des exigences. Pour le BABOK, cette activité est comprise dans l'activité d'analyse des exigences* ». En effet, la documentation du BABOK est incluse dans les activités d'élicitation et d'analyse des exigences, la technique utilisée dans chaque activité déterminera le type de documentation à produire. Le SWEBOK propose l'activité spécification d'exigences qui a pour but de compiler toute l'information collectée lors de l'élicitation et l'analyse dans un seul document. Trois niveaux de détail sont possibles : le document de définition du système, le document de spécifications des exigences du système, et le document de spécification des exigences logiciel (ou document Vision).

- **La validation :** Le BABOK inclut la validation dans l'analyse des exigences, le SWEBOK propose une activité additionnelle de validation des exigences. Les deux modèles font mention aux mêmes techniques de validation.

1.4 Sommaire du chapitre

Deux corpus de connaissances d'analyse des exigences ont été présentés, l'un d'ordre plutôt général (le BABOK), et l'autre, plutôt spécifique aux problèmes du génie logiciel (le SWEBOK). Les deux visent le même objectif. Ils suivent des démarches complémentaires et produisent des exigences avec différents niveaux de détail. L'utilisation de l'un ou l'autre dépendra du niveau de compétence technique et du type de projet. Le BABOK est plus destiné aux analystes d'affaires, qui s'appuient généralement sur des spécialistes techniques pour compléter leurs analyses. Le SWEBOK est un guide destiné aux professionnels du génie logiciel, qui manquent généralement des compétences dans le domaine des affaires. L'écart entre ces deux profils existera toujours, mais il peut être comblé avec de la formation et surtout avec de l'expérience.

Pour la réalisation du projet, une méthodologie fondée sur le SWEBOK sera utilisée, mais elle sera complémentée avec l'activité d'analyse de l'entreprise proposée par le BABOK. De cette façon, il sera possible de s'attarder à la compréhension du domaine de la génomique et de l'organisation avant de déterminer la problématique et les exigences de la solution à proposer. Dans le prochain chapitre, une explication de la méthodologie sera présentée afin de mieux comprendre comment la problématique du CRCHUM sera abordée.

CHAPITRE 2

ÉTUDE DE CAS LABORATOIRE DU Dr HAMET DU CRCHUM

Ce chapitre présente l'étude de cas du laboratoire de recherche en santé du Dr Pavel Hamet du CRCHUM où l'activité d'analyse a été réalisée afin de comprendre les problématiques existantes de manière à pouvoir proposer une solution d'amélioration de gestion de grandes quantités de données. Le chapitre commence par une mise en contexte du domaine de la génomique et de la technologie ADAM, suivie par la présentation de l'étude de cas. Ensuite, la problématique de la recherche sera décrite. Finalement, la méthodologie utilisée sera expliquée en détaillant, pour chaque étape, les constats et observations qui ont été faits lors de l'expérimentation.

2.1 Mise en contexte

2.1.1 Introduction à la génomique

La génomique est l'étude du génome, c'est-à-dire de l'ensemble du matériel génétique de tout être vivant, qu'il s'agisse d'un humain, d'une plante, d'un animal et même d'un virus. Le génome, mot formé à partir des mots « gène » et « chromosome », est ainsi l'ensemble de l'information héréditaire, plus spécifiquement l'ADN (acide désoxyribonucléique). L'ADN est une molécule présente dans chaque cellule de tout organisme vivant, elle est nécessaire pour son développement et fonctionnement. L'ADN est composé de quatre nucléotides de base : A (adénine), G (guanine), T (thymine) et C (cytosine). Ce qui fait que chaque individu soit unique est la façon dont les paires de ces nucléotides sont organisées. Parfois, les enzymes responsables de transformer l'ADN produisent des mutations ou des altérations de l'encodage normal du génome, ce qui peut être l'origine des certaines maladies (Bhardwaj *et coll*., 2014, p. 1).

La génomique est la clé du futur de la médecine personnalisée. Puisque chaque être humain a son génome personnel avec les différentes mutations qui affecteront positivement ou négativement sa santé, en identifiant ces mutations, il est possible de personnaliser le traitement des maladies. Le but de la génomique appliquée à la santé est de comprendre la composante génétique responsable de l'occurrence de certaines maladies chez l'individu. Les chercheurs essaient d'établir une relation entre le risque d'un individu d'avoir une maladie et les variations génétiques qu'il présente.

2.1.1.1 L'étude du génome

Cette étude peut être réalisée en suivant deux approches : le séquençage et le génotypage. L'approche du séquençage inclut la cartographie et l'analyse de l'ADN. Tout d'abord, le séquençage utilise un analyseur génétique pour lire les segments d'ADN d'un individu. Ces segments sont réassemblés afin de former le génome complet. Ensuite, dans la cartographie, les différents gènes sont identifiés et annotés. Finalement, une fois le génome reconstruit et cartographié, différentes transformations s'appliquent afin de réduire la taille des données et d'en faciliter l'analyse. En pratique, les analystes ont besoin d'effectuer plusieurs lectures-assemblage de génomes pour faire des comparaisons statistiques sur les génomes de plusieurs individus (illumina, 2015). Lors de l'utilisation d'une approche du génotypage, seules les variations (c.-à-d. les mutations) d'intérêt sont analysées. Un certain niveau de connaissance des variations génétiques est nécessaire, et les bio-informaticiens possèdent cette expertise. Plusieurs variations peuvent être analysées en même temps en utilisant les « arrays » ou « biochips » de génotypage.

Pendant que le séquençage essaie de déterminer la séquence exacte d'un segment d'ADN pour identifier toutes les variations possibles, le génotypage cherche à découvrir certaines variations sélectionnées d'avance. Comme l'ont peut s'y attendre, le séquençage demande plus de ressources informatiques et prend plus de temps que le génotypage. Mais, quelle que soit l'approche utilisée, la quantité de données à traiter reste énorme (Tikhomirov *et coll*., 2008, p. 1).

2.1.1.2 Problématique du domaine

La problématique sous-jacente à l'analyse de génomes est son grand volume de données. En effet, le génome humain contient 23 paires de chromosomes qui sont à leur tour composés d'entre 20,000 et 25,000 gènes. Le génome au complet est de l'ordre de 10 à 100 gigaoctets, avec les techniques de réduction de la taille de données on peut obtenir des fichiers d'environ 100 mégaoctets. Cette taille dépend du format de données et des techniques de réduction. Pour le format de données (SAM/BAM), un des plus utilisés pour le séquençage, la lecture et l'assemblage de ces données peut prendre trois jours de traitement. Ces transformations constituent la source de véritables goulots d'étranglement des implémentations actuelles. Les technologies à utiliser doivent permettre de traiter efficacement les données en ce qui a trait au temps de réponse de lecture et d'écriture de données, de capacité de stockage, d'accès aux différents formats de données, etc. Pour régler ce problème, les technologies Big Data ont été envisagées (Burriel, 2012).

2.1.2 La technologie ADAM

ADAM est le résultat de l'application de la technologie Big Data à la génomique. Il s'agit d'un ensemble de formats de données et d'algorithmes pour les projets d'infonuagique à grande échelle. ADAM a été conçu à l'aide de logiciels libres. Les formats de données d'ADAM sont implémentés sur la technologie Avro et accessibles avec la technologie Parquet, et le tout est encapsulé par le cadriciel Spark (Massie *et coll*., 2013, p. 1). Cette configuration a les avantages suivants :

- **Spark** est un cadriciel de traitement distribué en mémoire vive qui a comme principale caractéristique la minimisation des accès (c.-à-d. les entrées/sorties) sur disque.

- **Avro** est un logiciel libre de sérialisation de données multiplateforme qui implémente le schéma de données d'ADAM. Dans une implémentation Avro, les données sont enregistrées avec leurs schémas permettant ainsi la portabilité.

- **Parquet** permet l'accès à la base de données d'ADAM. Parquet utilise une structure de stockage par colonnes et permet l'utilisation des schémas de compression et d'encodage de données à haute performance.

L'architecture de données d'ADAM

L'architecture de données d'ADAM repose sur le principe selon lequel les bio-informaticiens doivent se préoccuper des données et non des formats de données comme c'est le cas actuellement au CRCHUM. Quatre composantes sont utilisées dans l'architecture de données ADAM :

1) **Le schéma ADAM** contient les formats de données implémentés sur Avro. La représentation des données d'ADAM est décrite en utilisant le système de sérialisation de données Avro. Le système Avro permet de définir un schéma déterminé pour chaque type de données (soit séquençage, variations génétiques ou génotypes), différents formats de données peuvent être créés. Avec cette caractéristique, Avro élimine la nécessité de développer des librairies dans un langage donné pour formater (c.-à-d. coder et décoder les données entre différents formats); ainsi les incompatibilités entre les librairies sont pratiquement éliminées. Cette portabilité assure que les logiciels conçus ne soient pas confinés à une technologie spécifique. De plus, les fichiers ADAM occupent moins d'espace sur disque (c.-à-d. par exemple 25% de moins pour les fichiers BAM).
2) **Les données matérialisées** implémentées sur Parquet. Les schémas ADAM sont encapsulés à l'aide de Parquet qui utilise un stockage par colonnes.
3) **Les librairies d'accès aux formats de données** implémentées sur Avro et Parquet.
4) **Les librairies de transformation de données** implémentées sur Spark.

(Massie *et coll*., 2013)

Les principaux concepts de la génomique ont été introduits et la problématique du domaine a été présentée. Ensuite, le schéma d'ADAM a été décrit comme une alternative pour répondre à cette problématique. La section suivante présente le cas d'étude où tous les concepts appris seront appliqués.

2.1.3 Le laboratoire du Dr Hamet au CRCHUM

Le CRCHUM est un organisme dédié à la recherche dans le domaine médical. Ces activités de recherche ont pour objectif de tenter de faire des découvertes qui visent à améliorer les traitements de patients et de la population. Au laboratoire du Dr Pavel Hamet du CRCHUM, l'équipe de chercheurs concentre ses efforts de recherches sur les traitements du diabète type 2 (c.-à-d. acronyme en anglais T2D). Les personnes ayant cette maladie ont un risque plus élevé d'avoir des complications cardiaques, rénales ou cardiovasculaires. L'intérêt médical de la recherche est d'identifier, à une étape précoce, les patients à risque élevé qui pourront mieux bénéficier des traitements préventifs. Cette stratégie, étant très couteuse, ne peut être adoptée par le système de santé actuel. Les résultats des recherches auront un grand impact non seulement sur la diminution des coûts du système de santé, mais aussi sur la santé des patients T2D.

Les recherches sur le T2D bénéficient des avancements récents sur le génome humain. En identifiant des biomarqueurs associés au T2D, et en les intégrant avec l'information génétique des patients T2D, il est possible d'améliorer les prédictions des complications vasculaires ce qui permettrait d'atteindre les objectifs du Dr Hamet. Au laboratoire, l'équipe de chercheurs a implémenté son propre flux de travail qui vise à appuyer la découverte de ces biomarqueurs. L'approche utilise des modèles de prédiction génétiques et des outils de visualisation afin d'isoler un traitement personnalisé approprié pour chaque patient. Toutefois, ce flux de travail pourrait probablement être optimisé, les technologies de traitement de données traditionnelles utilisées ne gèrent pas bien le grand volume de données impliquées. Plusieurs opportunités d'améliorations de cette approche peuvent être réalisées en utilisant les logiciels libres disponibles ainsi que les connaissances publiées récemment par l'Université Berkeley (ADAM).

2.2 Problématique spécifique du projet

Lors des premières rencontres avec le Dr Hamet et son équipe de chercheurs, plusieurs préoccupations sur la gestion de données et les outils informatiques utilisés ont été mentionnées. Effectivement, il existe des occasions d'amélioration tout au long du flux de travail. Avant d'effectuer des améliorations, une étude détaillée sera réalisée pour identifier les étapes les plus problématiques et concentrer les efforts d'analyse et d'améliorations futures sur ces étapes.

2.2.1 Objectifs du projet

Ce projet de recherche appliquée a pour but d'étudier le flux de travail d'analyse de variations de génomes utilisé au laboratoire du Dr Hamet. L'analyse cherchera à décrire, en détail, la situation actuelle des processus et des technologies, ce qui permettra d'identifier les problématiques en gestion de données et ainsi d'évaluer l'apport éventuel du modèle de données et de la technologie BigData ADAM.

Cette étude permettra :

- D'identifier les fonctionnalités à améliorer;
- De formaliser les exigences tout au long des processus à améliorer;
- De proposer des améliorations au modèle de données existant;
- D'évaluer l'application potentielle de la plateforme Adam comme solution de rechange.
- De documenter le flux de travail (technologies et processus) existant.

2.2.2 Portée du projet

Le flux de travail utilisé au CRCHUM est composé de plusieurs étapes qui vont de la réception de données sources jusqu'à la découverte de biomarqueurs et les analyses et prédictions finales. Ce projet cherche à comprendre l'ensemble des processus et à identifier les étapes problématiques qui pourraient bénéficier des innovations proposées par ADAM. Cependant, vu la complexité du flux de travail, l'effort de l'analyse se concentrera sur une seule étape. Finalement, une solution d'amélioration sera proposée pour l'étape en étude.

2.3 Méthodologie

2.3.1 Méthodologie utilisée pour l'analyse

La méthodologie d'analyse utilisée est une combinaison des deux modèles d'analyse présentés dans la revue de la littérature. Cette approche s'adapte mieux à la problématique du projet, et ceci, pour deux raisons. Premièrement, il s'agit de comprendre l'organisation, ses processus et ses besoins d'affaires; ainsi le modèle d'analyse d'affaires est mieux adapté pour cette partie initiale de la recherche. Deuxièmement, les exigences collectées lors de l'analyse doivent être détaillées selon les spécifications du génie logiciel, car il s'agit d'un projet technologique, et c'est le modèle d'analyse de systèmes qui propose les activités les plus appropriées pour arriver à cet objectif.

Le travail effectué par l'étudiant de maîtrise Tu Dang Vuong intitulé « *Déterminer les exigences d'affaires et d'application pour les fonctionnalités front-end de Snoobe* » (Vuong, 2015) a été une référence importante pour la définition de la méthodologie d'analyse. Il a utilisé une approche semblable (c.-à-d. comportant deux perspectives d'analyse). Une adaptation à cette méthodologie a été apportée pour l'ajuster aux besoins spécifiques du projet.

La méthodologie d'analyse des besoins d'affaires est typiquement composée de six étapes illustrées à la figure 2.1. Il s'agit d'un processus itératif, séquentiel à la base, mais qui peut se répéter à n'importe quelle étape lorsque de nouvelles informations sont découvertes. Cela permet de s'assurer que toutes les informations nécessaires seront collectées.

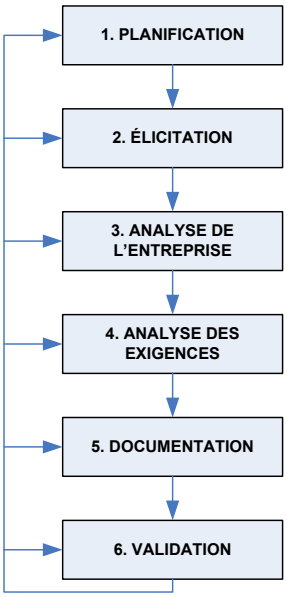


Figure 2.1 Les étapes de la méthodologie d'analyse d'exigences

1) Planification

   Cette étape suit la proposition du modèle d'analyse d'affaires. C'est l'étape de planification de l'effort d'analyse. Les activités à suivre sont :

   La définition de la portée de l'étude

   La portée de l'étude permet de mieux diriger l'analyse, elle déterminera les limites du projet et les activités à suivre. Une analyse préliminaire de l'organisation a permis d'identifier les étapes du flux de travail représenté dans la figure A I-1. Six étapes

conforment le flux de travail du Dr Hamet : le génotypage, la préparation des données, l'imputation, la permutation, l'association GWAS, et la visualisation. Suite à cette analyse et avec l'approbation du client, il a été conclu que l'étape de ***préparation des données***, puisqu'elle à la fois importante et problématique, pourrait mieux bénéficier d'une étude en profondeur. En effet, c'est à cette étape que les données brutes qui seront utilisées dans les étapes subséquentes sont recueillies et formatées. La qualité des données produites par cette étape a un impact sur l'ensemble des processus. La portée de l'étude se limite donc à analyser l'étape de la préparation des données.

L'analyse des parties prenantes

Elle sert à identifier les différentes parties intéressées à la réalisation du projet. Il s'agit non seulement des intervenants du flux de travail, mais aussi des personnes de l'organisation et de l'externe qui influencent l'exécution du projet. Le tableau A I-5 liste les parties prenantes participantes dans l'étape de préparation des données.

La définition des activités à effectuer dans l'analyse

Les activités sont choisies selon la portée du projet. Tel qu'il a été présenté à la figure 2.1, les activités principales à exécuter lors de l'analyse sont : l'élicitation, l'analyse des exigences, l'analyse de l'entreprise, la documentation des exigences et la validation des exigences.

La définition des techniques d'analyse d'affaires

Ce sont les techniques à utiliser pour chaque activité d'analyse. Ces techniques seront détaillées dans les sections suivantes.

2) Élicitation

L'élicitation est l'étape de la récolte des informations ayant pour objectif de mieux comprendre les besoins de l'organisation et des parties prenantes. Les quatre activités proposées par l'IIBA seront incluses dans la méthodologie.

La préparation

Cette étape vise à comprendre les différentes sources d'exigences possibles au sein de l'organisation (IEEE 2004). Pour mieux comprendre les sources des exigences, le site Web de l'organisation et quelques documents expliquant le flux de travail du point de vue des chercheurs du domaine ont été consultés. Cependant, étant donné qu'aucune documentation formelle sur la préparation des données n'existait, la documentation a dû être conçue au fur et à mesure avec l'aide des bio-informaticiens. La technique d'élicitation utilisée a été principalement des sessions d'entrevues avec les bio-informaticiens, et à cet effet, un bureau de travail a été assigné au laboratoire du Dr Hamet pour la réalisation du projet.

L'exécution

Le travail d'élicitation s'est effectué par une interaction constante avec les bio-informaticiens. Cette proximité a permis de comprendre graduellement la terminologie complexe utilisée dans le travail quotidien des chercheurs du domaine de la santé.

La documentation

Pour chaque activité d'élicitation, un schéma, un diagramme, un tableau et d'autres documents pertinents ont été préparés. Cette documentation a été présentée pour sa validation. Lors du démarrage du projet de recherche, un compte rendu de la rencontre de démarrage a été réalisé en consignant les attentes du client, ce qui a servi à encadrer la portée de la recherche.

La confirmation des résultats

Pour vérifier si les exigences ont bien été comprises, l'avis et la confirmation de chaque bio-informaticien (c.-à-d. les clients) ont été demandés. Une attention particulière a été portée à la préparation d'une synthèse des découvertes sous la forme d'un document, d'un tableau, ou d'un diagramme autour duquel se déroulaient les discussions portant sur la description de la situation actuelle.

Lors de cette étape d'élicitation, les documents suivants ont été élaborés:

- La description du processus de préparation des données : pour comprendre le sous-processus à l'intérieur de l'étape en étude et les interactions avec les autres étapes du flux de travail au complet (voir figure A I-2),
- Le diagramme des activités de la préparation des données : sous la forme d'un diagramme de processus d'affaires (voir figure A I-3). Ce diagramme présente la séquence d'activités et les responsables de les exécuter,
- La description des activités de la préparation des données : permettant de mieux se situer dans l'utilisation des outils et bases de données existantes (voir l'annexe III),
- La description de l'environnement de travail : incluant tous les intervenants et les composants de l'infrastructure, ainsi que les logiciels utilisés lors de la préparation des données (voir la section 3.3 de l'annexe I et les tableaux A I-9 et A I-10),
- La liste des exigences récoltées : (voir le tableau A I-11).

3) Analyse de l'entreprise

Cette étape proposée par l'IIBA a été incluse, parce qu'elle permet de bien saisir l'organisation et son contexte opérationnel, et sert ainsi de point de départ pour identifier ses besoins d'affaires. Puisque le domaine de la recherche en génomique peut devenir très complexe, pour une analyste qui n'a pas d'expérience dans ce secteur d'affaires, cette étape s'avère très importante. Elle permettra de mieux comprendre les besoins d'affaires et de raffiner la portée des solutions potentielles. Trois éléments ont été étudiés lors de cette étape : les buts et objectifs de l'entreprise (c.-à-d. vue d'ensemble et par rapport à l'étude visée), les problématiques de traitement des données du laboratoire et les résultats attendus par les parties prenantes.

Pour la réalisation de cette tâche, la consultation du plan stratégique du CRCHUM a permis de connaître la vision, la mission et les objectifs de l'organisation à laquelle le laboratoire appartient. Les premières rencontres avec le Dr Hamet, la consultation du site Web du CRCHUM, les documents de recherche publiés par le Dr Hamet ainsi que les diverses entrevues qu'il a accordées aux médias, où il résume les objectifs de sa

recherche, ont aidé à bien comprendre l'organisation et ses objectifs. Cette activité a permis de produire l'énoncé du problème selon la structure suivante (voir tableau A I-3).

Le problème actuel est	Description du problème
Affecte	Les intervenants affectés par la problématique
dont l'impact est	Les conséquences de la problématique
Une bonne solution serait	La solution proposée

Tableau 2.1 Énoncé du problème

Aussi de positionner la solution proposée selon la structure suivante (voir tableau A I-4).

Pour	Le client
Qui	Le besoin ou l'opportunité
La nouvelle structure de données	Définition de la solution
Qui	Le bénéfice principal de la solution
Contrairement à	La situation actuelle
Notre proposition	Les bénéfices apportés par la solution

Tableau 2.2 Positionnement de la proposition

4) Analyse des exigences

Cette étape suit la proposition du modèle d'analyse de systèmes (SWEBOK). En concordance avec les objectifs du projet, deux tâches ont été incluses : la définition des hypothèses et contraintes, et la modélisation des exigences. L'activité de classification des exigences, qui appartient à cette étape, a été omise puisque la solution proposée est une nouvelle structure de données (c.-à-d. ADAM de Berkeley) qui a seulement quatre exigences prioritaires de haut niveau, qui seront développées lors de la réalisation de l'étape de la modélisation (voir section 5.1 de l'annexe I).

Présomptions initiales

Deux présomptions ont été établies pour ce projet. Il s'agit des choix technologiques actuels et potentiels qui influenceront le modèle de données à proposer et à adapter. D'abord, la première présomption est que le CRCHUM continuera à utiliser l'approche du génotypage pour la détection de variations génétiques (l'approche alternative est le séquençage complet du génome). Selon l'approche utilisée, les formats de données changent, ce qui influencerait directement le modèle de données futur. Ensuite, la deuxième présomption est que le modèle d'ADAM, conçu initialement pour aborder la problématique du séquençage complet du génome humain, pourra être réutilisé dans le contexte du CRCHUM. Ces deux présomptions ont été établies après l'analyse initiale des technologies utilisées actuellement au laboratoire du Dr Hamet et la documentation disponible de Berkeley sur la technologie émergente ADAM. Pour valider ces deux présomptions, l'avis du client et d'autres professionnels du génie logiciel ont été demandés. Il s'avère qu'ADAM pourra être adapté pour y inclure le génotypage et aussi les résultats des études GWAS du client.

La modélisation des exigences

Pour modéliser la nouvelle base de données, les exigences de haut niveau doivent être analysées plus en profondeur. Cette analyse est expliquée à la section 2.3.2 de ce chapitre.

5) Documentation

La documentation de l'état actuel est l'étape qui s'occupe de la documentation des processus existants et soulève des exigences futures. Pour cette étape, la réalisation d'un document de spécification des exigences (c.-à-d. un document de vision) suggéré par le processus de génie logiciel sera réalisée. Ce document préconise une abstraction à haut niveau du problème sous étude et de la solution proposée. Le document vision permet de décrire la solution en termes généraux facilement compréhensibles par le client, incluant la description du marché potentiel de la solution, les utilisateurs du système et les

spécifications de haut niveau de la solution proposée. Il sert comme cadre de discussion facilitant l'entente entre les principales parties prenantes du projet de génie logiciel. Le document vision produit suite à l'analyse des processus de génotypage du laboratoire du Dr Hamet est présenté à l'annexe I.

6) Validation

   Deux activités sont nécessaires lors de la réalisation de cette étape : la vérification et la validation des exigences avec le client. Ces activités seront réalisées par le moyen d'entrevues avec les bio-informaticiens. Ce contact direct a permis d'assurer que les exigences découvertes respectaient bien la réalité et les besoins formulés par les utilisateurs.

À cette étape de l'analyse, la compréhension de la problématique de l'étape de la préparation des données est complétée et une proposition d'amélioration de cette problématique (c.-à-d. une nouvelle conception de base de données) ainsi que des exigences de haut niveau ont été formulées. La prochaine étape consiste à détailler et prioriser les exigences collectées, concevoir le modèle de données et à proposer une nouvelle structure de données qui prend en compte de ces exigences.

2.3.2 Méthodologie utilisée pour la modélisation de la base de données

Comme nous l'avons précisé, ce projet de recherche appliquée a pour but de formuler une première version d'une structure de données mieux adaptée aux besoins de traitement de grandes quantités de données actuels et futurs du laboratoire. Ce modèle de données pourra servir de base pour des travaux futurs de développements de logiciels de séquençage et d'analyse du GWAS. Cette section décrit la méthodologie utilisée pour la modélisation de cette base de données. La méthodologie est composée de sept étapes telles qu'illustrées à la figure 2.2.

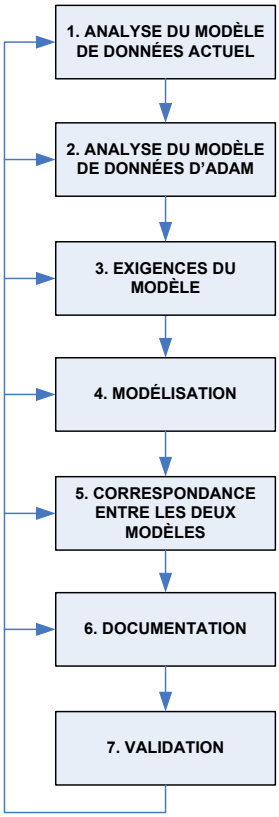


Figure 2.2 Les étapes de la méthodologie de modélisation de la base de données

1) Analyse du modèle de données actuel

La première tâche à réaliser, en vue de la modélisation de la base de données, est l'étude du modèle actuel de données. Cela permettra de comprendre les entités, les relations, la cardinalité, et aussi de vérifier les problèmes encourus actuellement par l'utilisation quotidienne de cette structure de données. L'analyse du modèle actuel de données a permis de produire le modèle de la base de données Prognomix (voir figure A I-4). Cette analyse a été complétée et est accompagnée de la réalisation d'un dictionnaire de données détaillé (voir annexe IV) où chaque entité de données, de cette base de données, a été

analysée. Cette étude a aidé à mieux comprendre le domaine d'affaires ainsi que la nature des données utilisées quotidiennement.

2) Analyse du modèle de données d'ADAM

   Le but du projet est d'évaluer l'utilisation potentielle de la structure de données d'ADAM. Cette étape a pour objectif de comprendre le modèle de données d'ADAM. Pour la réalisation de cette étape, quelques questions qui aident à orienter le travail ont été formulées:

   - Dans quel contexte ou problématique spécifique s'applique le modèle ADAM?
   - Quels problèmes spécifiques seront résolus avec l'adoption de ce modèle?
   - Quelles sont les contraintes et hypothèses pour l'implémentation du modèle d'ADAM,
   - Quels sont les entités d'information, les relations entre ces entités, les types de données utilisés dans le modèle ADAM.

   Les réponses à ces questions seront présentées dans le chapitre suivant intitulé « présentation de résultats ».

3) Exigences du modèle de données

   Cette étape utilise, à l'entrée, les exigences formalisées lors de l'étape d'analyse, et cherche à les détailler jusqu'au niveau des données mêmes. Par exemple, pour détailler l'exigence « *le modèle doit être capable d'enregistrer l'information de plusieurs études T2D* », il faut spécifier les différentes études qui peuvent être acceptées, comprendre leur structure de données et l'information qui sera utilisée au départ, transférée d'autres sources, transformées et produites. Les résultats de cette étape permettront de modéliser la base de données future.

4) Modélisation

La modélisation est l'étape où l'analyste produira un diagramme de classes de la nouvelle structure de données en tenant compte de toutes les exigences soulevées durant l'étude. Au lieu de modéliser directement la base de données future en s'inspirant directement des schémas de données d'ADAM, un modèle relationnel de données amélioré a été produit en tenant compte de toutes les exigences récoltées pour ensuite faire le parallèle avec le modèle d'ADAM. Cette approche permet de ne pas trop tenir compte de l'influence de la structure des données d'ADAM au cas où il ne contiendrait pas la même information.

5) Correspondance entre le modèle de données proposé et celui d'ADAM
   Cette étape a pour but d'identifier les entités du modèle de données proposé par ADAM qui pourraient être utilisées. Cette correspondance permet d'identifier les différences entre le modèle de données du laboratoire et les parties qui pourront éventuellement migrer vers le modèle ADAM telles quelles ou adaptées.

6) Documentation
   Lors de toute l'analyse, une étape de documentation est nécessaire pour décrire le modèle. Dans cette étape, un modèle de classes et une liste de tables accompagnées de leurs descriptions, la provenance des données et la fréquence des mises à jour ont été produits.

7) Validation
   La réalisation du modèle proposé s'est effectuée à l'aide de plusieurs itérations de conception suivie de validation. À chaque itération, une validation du modèle de la part du client (c.-à-d. les bio-informaticiens) et de la part d'un expert ADAM a été effectuée.

2.4 Sommaire du chapitre

Dans ce deuxième chapitre, le travail d'analyse des processus d'affaires effectué au laboratoire du Dr Hamet a été décrit. Débutant par une mise en contexte, une description des objectifs de la recherche a été réalisée suivie d'une description de la méthodologie utilisée

pour l'analyse des exigences. Par la suite l'approche de modélisation de la nouvelle structure de données a été présentée. Deux documents ont été produits : 1) le document Vision; et 2) le modèle de données. Le prochain chapitre abordera l'interprétation des résultats de l'étude.

CHAPITRE 3

PRÉSENTATION DES RÉSULTATS

Le troisième chapitre présente les résultats obtenus à la suite de l'expérimentation effectuée au laboratoire du Dr Hamet. Les résultats sont présentés suivis d'une discussion concernant ce qui reste à étudier pour améliorer et compléter cette étude. Finalement, une revue critique du projet est réalisée, permettant d'identifier les points forts et les points faibles de cette recherche et expérimentation.

3.1 Présentation des résultats

Au début du projet, il a été établi que l'objectif de la recherche était d'identifier une étape problématique du flux de travail de génotypage du laboratoire et, par la suite, d'analyser cette étape en profondeur afin de proposer une solution d'amélioration. La piste d'amélioration envisagée est l'utilisation potentielle de la technologie et du modèle de données du projet ADAM de l'université Berkeley. Pour ce faire, la méthodologie utilisée proposait d'initier le travail par une analyse des processus d'affaires suivie d'une analyse des exigences du processus de génotypage. Finalement, les résultats du projet consistent en un ensemble d'exigences et de documentation permettant la modélisation d'une base de données qui tiendra compte de ces exigences.

3.1.1 Les résultats de l'analyse d'affaires

L'étape d'analyse d'affaires a permis de comprendre l'organisation et son contexte, de détailler le flux de travail de génotypage du laboratoire, et d'identifier l'étape de la préparation des données comme étant une des étapes les plus problématiques par rapport à la gestion de données. En suivant la méthodologie décrite dans le chapitre précédent, la problématique, les besoins d'affaires, et finalement, les exigences de la solution proposée ont été identifiés.

Les besoins d'affaires de l'étape de la préparation des données

Lors de cette étape, tous les besoins d'affaires ont été identifiés. Par contre, trois besoins d'affaires concernent plus particulièrement notre sujet d'intérêt, le problème du modèle de données. Les exigences d'intérêt sont résumées dans le tableau suivant.

Besoin	Situation actuelle
B01 – La structure de l'information utilisée tout au long du pipeline n'est pas flexible.	Une base de données relationnelle avec plusieurs fichiers d'entrée et de sortie. Nécessite des scripts/outils pour traiter chaque type de fichier. Information pas standardisée et/ou manquante.
B02 – Intégrer l'information de plusieurs études T2D.	Une nouvelle BD est créée pour supporter chaque nouvelle étude. Plusieurs scripts d'insertion de données existent.
B03 – Traiter l'information provenant du génotypage pour rendre son accès plus efficace.	Les génotypes sont sauvegardés dans une table sur une seule colonne qui a environ 950,000 caractères. Un script spécial de lecture est nécessaire.

Tableau 3.1 Les besoins d'affaires de l'étape de préparation des données

La formulation d'une solution

La solution proposée est un nouveau modèle de données qui permettra de combler ces besoins d'affaires. À la suite de la revue de la littérature et de l'étude du projet ADAM, quelques questions, afin de valider l'utilité du modèle de données d'ADAM comme solution potentielle, ont été formulées.

1) Dans quel contexte ou problématique spécifique s'applique le modèle ADAM?
   Cette première question est l'une des principales préoccupations du client. En effet, dès les premières lectures des articles publiés sur ADAM, il a été constaté que cette technologie était conçue pour résoudre des problématiques de séquençage du génome (c.-

à-d. l'identification de toutes les variations génétiques possibles du génome humain) tandis qu'au laboratoire ils utilisent du génotypage (c.-à-d. l'analyse des variations génétiques d'intérêt). Malgré cette différence, l'objectif final reste le même, c'est-à-dire d'identifier les variations génétiques qui peuvent conduire à l'apparition de maladies. En ce qui concerne la structure des données, une correspondance entre les entités de chaque modèle (c.-à-d. le laboratoire et ADAM) peut être effectuée.

2) Quels problèmes spécifiques seront résolus avec l'adoption de ce nouveau modèle de données?

Les principaux avantages de cette technologie émergente sont :

- Le format de données d'ADAM assure la portabilité de la solution.
  Avec l'utilisation des schémas « Avro » où la représentation des données est une couche indépendante des données elles-mêmes, il est possible d'intégrer n'importe quel format de données dans un même modèle. Le génotypage, l'imputation et l'association GWAS (c.-à-d. les étapes clés du flux de travail du laboratoire du Dr Hamet) utilisent des outils qui ont tous leurs propres formats de données. Les divers formats de fichiers (c.-à-d. CEL, PED, MAP, SAMPLE, etc.) pourraient s'intégrer directement à un seul modèle, c'est-à-dire ADAM, ce qui éliminerait la dépendance envers les différents outils spécifiques de traitement de données et simplifierait les traitements.

- L'amélioration de la performance des traitements de données.
  Dans le projet ADAM, la couche de représentation de données (c.-à-d. celle gérée par « Avro ») est encapsulée avec la technologie « Parquet », un logiciel libre qui permet le stockage de données par colonnes. Cette caractéristique permet d'atteindre des niveaux de compression de données très élevés, de sauvegarder de l'espace sur disque, ainsi que d'utiliser moins de ressources computationnelles pour analyser et formater les données. Cette combinaison de technologies BigData « Avro-Parquet » permet d'offrir une solution à haute performance de lecture/écriture des données. Ces

technologies seront plus adaptées aux besoins grandissants en traitement de données massives que celles qui sont utilisées actuellement.

- L'adoption d'une technologie plus adaptée à la génomique (BigData ADAM).
  Avec les derniers avancements en génomique, la tendance est de traiter et de produire de grands volumes d'information qui peuvent être plus difficilement traitées avec les technologies de bases de données traditionnelles. L'adoption de la technologie et de la structure de données ADAM permettraient au laboratoire de mettre à jour son infrastructure technologique afin d'introduire ces innovations.

3.1.2 Les résultats de l'analyse des exigences

L'identification des besoins d'affaires, décrits au tableau 3.1 de la section précédente, a permis d'établir les exigences à haut niveau du nouveau modèle de données. Ces exigences sont présentées dans le tableau ci-dessous. Une analyse plus détaillée des profils de données a été réalisée dans le but d'approfondir la compréhension de ces exigences.

3.1.2.1 Les exigences du modèle de données

#	Description	Explication
EX01	S'adapter au modèle de données proposé par ADAM.	Le modèle d'ADAM s'adapte bien à la problématique du séquençage du génome humain. Une correspondance peut se faire pour l'adapter au flux de travail du CRCHUM.
EX02	Intégrer plusieurs sources de données d'études T2D.	Notamment ADVANCE et MONICA. L'étude CARTaGENE pourrait s'utiliser dans le futur.
EX03	Intégrer l'information manquante et standardiser l'information existante.	Voir les exigences de données (section 3.1.2.2).
EX04	Accéder plus efficacement aux données de génotypage.	Les génotypes d'un individu sont sauvegardés dans un long string d'environ 950,000 caractères

		qui en rend la consultation très difficile. S'inspirer du modèle ADAM pour proposer une solution.

Tableau 3.2 Les exigences du modèle de données

3.1.2.2 Les exigences des données

Les bio-informaticiens, du laboratoire du Dr Hamet, ont donné accès à la base de données clinique utilisée actuellement. Une analyse de données a été effectuée sur le modèle de données actuel afin d'identifier les entités, les relations entre ces entités, la cardinalité des relations, les contraintes, etc. Ainsi, une analyse du profil des données a été effectuée pour connaître sa nature. Les exigences des données sont résumées ci-dessous.

Résumé des exigences des données

Les exigences des données ont été classifiées en deux catégories :

1) Exigences des données des études T2D : correspondant aux données démographiques et médicales des patients T2D.

2) Exigences des données du génotypage : correspondant aux annotations et génotypes. C'est cette catégorie qui bénéficierait particulièrement des propositions du modèle ADAM.

1) Les exigences des données des études

#	Exigence	Explication
1	Le modèle doit permettre	Chaque étude contient un ensemble

	d'enregistrer les données de plusieurs études (ex. ADVANCE[1], MONICA[2], CARTaGENE[3]).	d'informations sur les patients T2D (données démographiques et tests médicaux). Les données des études sont chargées dans la base de données par lots selon l'ordre d'arrivée des nouvelles actualisations.

Tableau 3.3 Les exigences des données des études

#	Exigence	Explication
2	Une étude peut contenir plusieurs types de tests (sang, cholestérol, salive, etc.)	Au laboratoire du Dr Hamet ce sont les tests sanguins qui sont utilisés.
3	Les patients font des visites chez le médecin. Chaque visite a un ou plusieurs diagnostics, des prescriptions (médicaments) et des phénotypes (traits observables).	Plusieurs données de référence (métadonnées) sont associées aux visites : - Liste de diagnostics - Liste de médicaments - Types de phénotypes - Types de mesure - Unité de mesure
4	Les diagnostics doivent indiquer la présence ou l'absence d'une maladie.	Il existe une liste de diagnostics de référence (ex. cancer étape I).
5	Les prescriptions doivent indiquer la présence ou l'absence d'un médicament.	Il existe une liste de médicaments de référence.
6	Les phénotypes (ex. Hémoglobine, Hypertension, etc.) comportent généralement des mesures, mais il peut	

[1] « *Action in diabetes and vascular disease* ». L'étude contient une base de données de patients T2D
[2] « *Multinational **MONI**toring of trends and determinants in **CA**rdiovascular disease* ». Étude des facteurs de risque et de la tendance de la population par rapport aux maladies cardiovasculaires
[3] Banque d'échantillons biologiques d'hommes et de femmes québécoises

	y avoir des phénotypes sans mesure (ex. peau rouge). Dans le sens inverse, chaque mesure correspond à un phénotype.	
7	Une mesure correspond à la valeur d'un phénotype (trait observable). Les mesures peuvent avoir différentes unités de mesure (ex. mg/L, années, ug/mg, battement/minute, etc.).	Il existe plusieurs types de mesure de référence.

Tableau 3.3 Les exigences des données des études « suite »

2) Les exigences des données du génotypage

<table>
<tr><th>#</th><th>Exigence</th><th>Explication</th></tr>
<tr><td>1</td><td>Les annotations proviennent des publications d'Affymetrix[4]. Chaque actualisation des annotations doit générer une nouvelle version de l'annotation.</td><td></td></tr>
<tr><td>2</td><td>Une annotation d'Affymetrix contient certaines données de base.</td><td>Annotations<table><tr><td>rs</td><td>SNP unique identifier</td></tr><tr><td>chromosome</td><td>Chromosome number of SNP</td></tr><tr><td>position</td><td>Chromosome physical position</td></tr><tr><td>marshfield</td><td>Chromosome Marshfield position</td></tr><tr><td>allele_a</td><td>Nucléotide A,C,G ou T de l'allèle A</td></tr><tr><td>allele_b</td><td>Nucléotide A,C,G ou T de l'allèle B</td></tr><tr><td>strand</td><td>Strand</td></tr></table></td></tr>
<tr><td>3</td><td>À chaque processus de génotypage enregistrer: la date/heure, la version de l'annotation et la version du génotypage.</td><td></td></tr>
<tr><td>4</td><td>Le génotypage génère une liste de génotypes associés à chaque individu.</td><td>Il y a une correspondance entre les génotypes et les annotations.</td></tr>
<tr><td>5</td><td>Un individu peut être génotypé plusieurs fois. Chaque génotypage doit être enregistré sur différentes versions.</td><td>Les génotypes sont décrits par les deux allèles hérités d'une position déterminée d'un SNP. Affymetrix utilise la notation AA, AB, ou BB pour décrire les paires d'allèles d'un SNP.</td></tr>
</table>

Tableau 3.4 Les exigences des données de génotypage

[4] Lien vers Affymetrix : http://www.affymetrix.com/support/technical/annotationfilesmain.affx)

3.1.3 Les résultats de la modélisation

3.1.3.1 Le modèle relationnel de données

En tenant compte de toutes les exigences, le modèle relationnel de données, présenté à la figure 3.1, a été produit (une description des tables du modèle se trouve à l'annexe VI). Un modèle relationnel de la nouvelle structure des données nécessaires a été représenté afin de bien documenter les besoins et, par la suite, permettre de le comparer au modèle relationnel existant au laboratoire. La correspondance de ce modèle avec le modèle offert actuellement par ADAM se fera ultérieurement lorsqu'il sera validé.

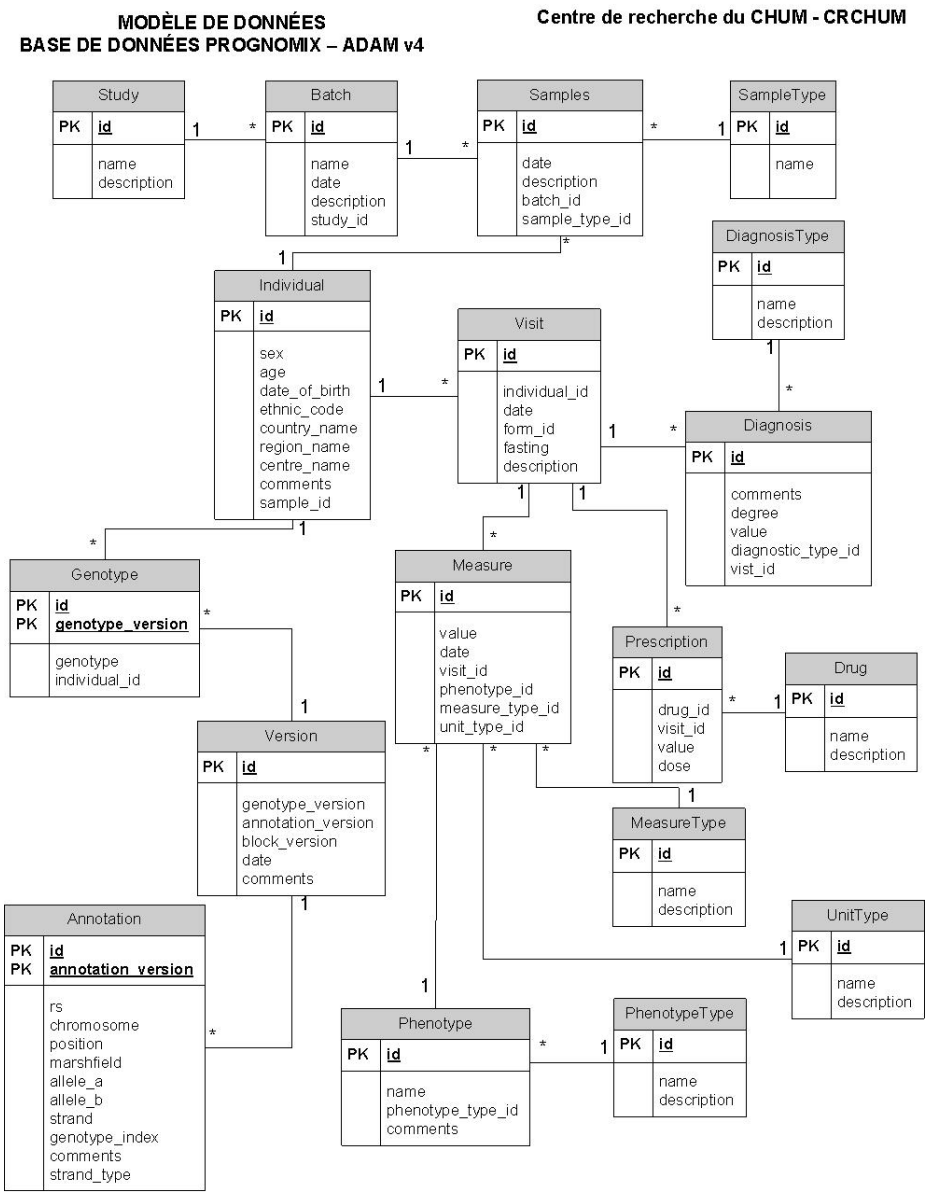


Figure 3.1 Modèle de données proposé

3.1.3.2 Correspondance avec le modèle d'ADAM

Pour faire la correspondance avec ADAM, les entités : « Annotations » et « Génotypes » ont été comparées avec les formats « Genotype » et « Variant » d'ADAM. Des modifications ont été apportées au format actuel du modèle d'ADAM pour tenir compte des besoins particuliers du laboratoire. Cette fois-ci, le format de schémas d'Avro a été utilisé tel que démontré à la figure 3.2. Cette figure est divisée en trois sections : la section A correspond au nouveau modèle (c.-à-d. Prognomix ADAM représenté à la figure 3.1), la section B correspond au modèle ADAM (les modifications à ce modèle sont mises en italique), et la section C correspond aux ajouts à faire au modèle ADAM pour inclure l'information manquante.

Pour intégrer les informations du génotypage, deux champs ont été ajoutés au schéma « Variant » d'ADAM. Il s'agit des informations des allèles A et B. La justification de cet ajout est que l'outil d'analyse de génotype d'Affymetrix utilise la notation allèles A et B (ex. AA/AB/BB) et non la notation standard avec les allèles de base, c'est-à-dire allèles A, C, G et T. De plus, une des présomptions initiales du projet était que l'outil Affymetrix serait toujours utilisé pour le génotypage au laboratoire. C'est donc avec ce codage particulier que l'étape de préparation des données recevra les données du génotypage. Les champs disponibles dans le schéma « Variant » pour les allèles (c.-à-d. « *reference* » et « *alternate* » allèles) ne peuvent pas être utilisés, car ils sont destinés aux allèles de base. Pour être cohérent avec la notation d'allèles AB, le schéma « EnumAB » qui contient les deux types d'allèles mentionnés a été ajouté. Pour transformer les allèles du format AB vers le format standard une correspondance spéciale est nécessaire. La table « Annotation » de la base de données Prognomix ADAM contient cette correspondance. Dans cette table, chaque code SNP (c.-à-d. une variation identifiée par le génotypage) est associé à un allèle A et B. Les entités « Individual », pour le patient, et « Phenotype », pour les traits observables d'un individu, ont été aussi incluses dans le modèle. Les autres tables du modèle proposé dans la figure 3.1 n'ont pas de correspondance avec le modèle ADAM.

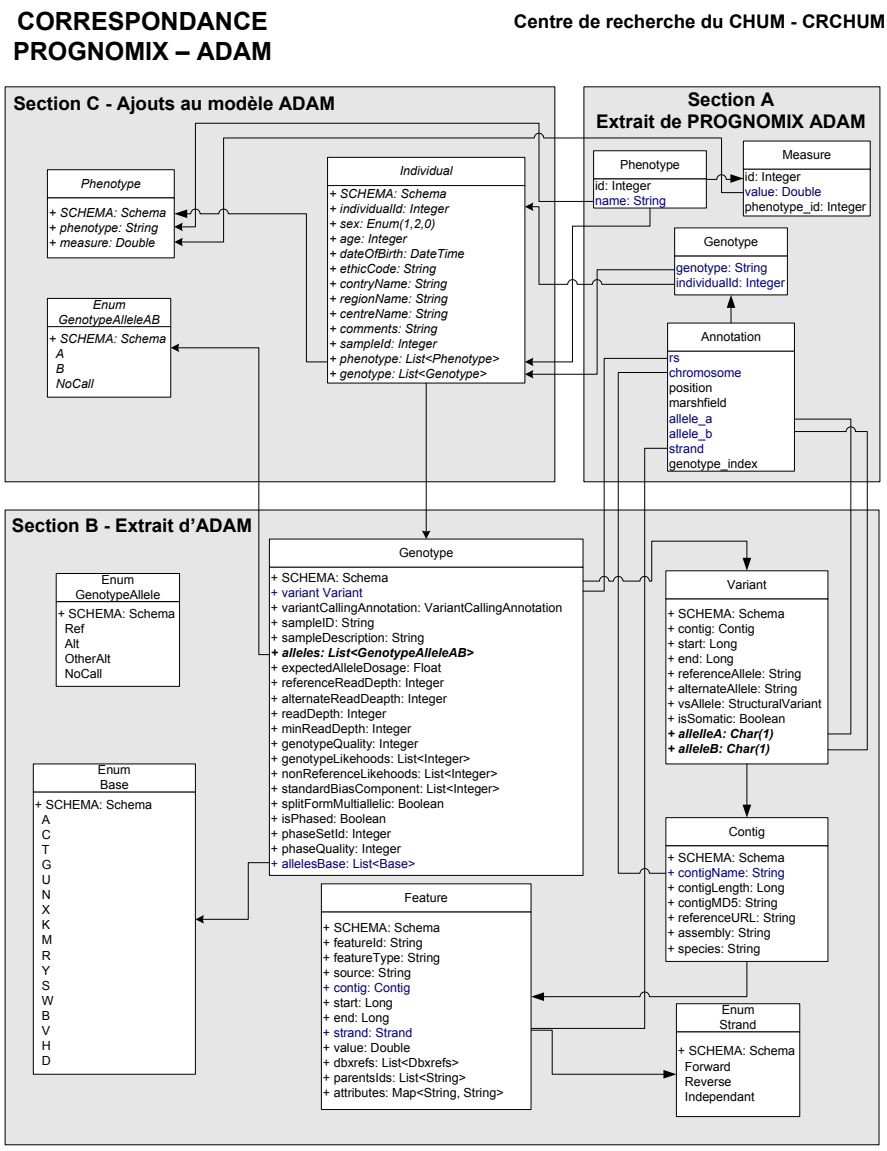


Figure 3.2 Correspondance entre le modèle de données proposé et le modèle ADAM

3.2 Analyse des résultats

L'analyse des résultats observés consiste à examiner les résultats obtenus lors de l'analyse des exigences, de la modélisation de la nouvelle base de données et de la validation du modèle.

3.2.1 Analyse des résultats de l'élicitation des exigences

Les exigences à haut niveau collectées lors de la première analyse (voir le tableau 3.2) ont été détaillées afin d'obtenir les informations nécessaires pour la modélisation. De cette façon, sept exigences ont été recueillies pour les données provenant des études de patients T2D (voir le tableau 3.3) et cinq exigences pour les données provenant du génotypage (voir le tableau 3.4). Puisqu'il s'agit des exigences nécessaires pour le flux de travail du laboratoire du Dr Hamet, toutes ont été classées comme prioritaires. Pour la réalisation de cette étape, la construction d'un dictionnaire de données du modèle actuel a permis d'identifier les informations manquantes ainsi que de comprendre les données, et notamment les types, les attributs et les relations entre les différentes informations.

3.2.2 Analyse des résultats de la modélisation

Suite à la collecte des exigences, une version améliorée du modèle relationnel existant (voir la figure 3.1) a été produite. Ce modèle amélioré n'est pas encore migré au format d'ADAM pour les raisons suivantes : 1) par simplicité; 2) pour faciliter la communication avec les bio-informaticiens; 3) pour faciliter la comparaison avec le modèle de données actuel, et 4) parce qu'il y avait plusieurs exigences de base à mettre en place avant de faire la correspondance avec ADAM. En effet, le modèle de données produit corrige plusieurs erreurs identifiées dans le modèle de données actuel. Par exemple, l'intégration des différentes études T2D, actuellement une nouvelle base de données est créée pour chaque nouvelle étude, l'intégration de tables de références pour les types d'unité de mesure, le type de phénotype et

les types d'échantillons, la relation entre les versions du génotypage, les résultats du génotypage et les annotations.

En ce qui concerne la correspondance entre le nouveau modèle de données et ADAM, il s'agit d'une première ébauche. Cette correspondance permet de trouver les informations qui pourraient se partager entre les deux modèles et les informations manquantes qu'il faudrait ajouter au modèle ADAM. Le but est d'utiliser éventuellement le modèle d'ADAM pour gérer les données produites par le génotypage, mais pour cela, il faut connaître quelles seront les entités et les informations à prendre de chaque modèle.

3.2.3 Analyse des résultats de la validation

La validation a été réalisée directement avec le client. En raison de la complexité des exigences pour modéliser la base de données, quelques itérations d'élicitation supplémentaires ont été réalisées afin d'approfondir la définition de chaque exigence. C'est cette activité qui a entrainé la décomposition de chaque exigence en plusieurs exigences de données. C'est à partir de ce niveau de détail qu'il est possible d'avoir les informations nécessaires pour modéliser la base de données.

3.3 Revue critique du travail

La finalité de ce travail est l'analyse d'une problématique de données du processus de génotypage suivie d'une proposition de solution d'amélioration de la gestion de données. Initialement, une cartographie du flux de travail a été réalisée, suivie d'une étude détaillée d'une étape problématique : la préparation des données. Tout au long des activités de cartographie, la fidélité de la représentation graphique et de la description des processus a été validée avec les bio-informaticiens. Ensuite, l'analyse des besoins et la formulation des exigences de l'étape de la préparation des données ont été achevées avec la production du document Vision. Finalement, un modèle de données en tenant compte de ces exigences, ainsi que la première ébauche d'une correspondance entre le nouveau modèle et celui d'ADAM ont été produits.

Cependant, les découvertes effectuées lors de cette recherche restent théoriques puisque le modèle n'a pas encore été mis en place. De plus, ce projet ne propose pas une solution finale, et peut être amélioré sur différents points. Tout d'abord, les exigences peuvent être validées à nouveau par l'étude approfondie d'autres étapes du flux de travail. Par exemple, l'étape de génotypage est actuellement utilisée comme une « boîte noire » pour les bio-informaticiens du laboratoire. Puisque le processus de génotypage est en dehors de la portée de cette étude, l'analyse du format de données produites par la puce d'Affymetrix n'a pas été réalisée. En fait, pour la réalisation de cette recherche, il a été établi, dès le départ, que les données brutes du génotypage « *allele A et B probe intensities* » produites par Affymetrix étaient déjà converties en génotypes et en annotations. Cette conversion nécessite des logiciels particuliers afin de lire les données et exécuter l'algorithme de génotypage. Plusieurs logiciels gratuits sont disponibles sur le Web. Dans le laboratoire, un logiciel propriétaire est utilisé et son fournisseur n'offre plus de soutien technique. L'étude de cette étape devient donc impérative afin de comprendre et faire évoluer ce logiciel à l'avenir.

Difficultés rencontrées

Pendant la réalisation de ce travail, des difficultés ont dû être surmontées. Ces difficultés sont principalement dues à la structure organisationnelle du CRCHUM et au manque d'expérience, de notre équipe de recherche, en génie logiciel, concernant le domaine de la génomique.

Dès le début du projet, la première difficulté constatée était la compréhension de la terminologie du domaine de la génomique. Le travail d'analyse implique que l'analyste d'affaires maîtrise le domaine à étudier, et qu'il sera alors en mesure de comprendre rapidement les besoins du client et de proposer des solutions innovantes. En effet, la qualité de l'analyse dépendra en grande partie des compétences de l'analyste par rapport à la mise en œuvre des outils d'analyse d'affaires et par rapport au domaine d'affaires. Pour lever ce premier obstacle, il a été décidé d'ajouter, dans la méthodologie de travail, l'étape de l'analyse de l'entreprise proposée par IIBA.

Une autre difficulté rencontrée lors de cette étude de cas a été l'inexistence d'une structure organisationnelle bien définie qui supporte tout le flux de travail du Dr Hamet. Effectivement, Robins *et coll*. (2014 p. 24) définissent la structure organisationnelle comme l'ensemble des règles de répartition de l'autorité, des tâches, du contrôle et de la coordination. Au laboratoire, deux bio-informaticiens ont la responsabilité de gérer le flux de travail au complet. Ils travaillent sous la supervision d'un gestionnaire de projets. Malgré cette organisation, les chercheurs principaux exercent beaucoup d'influence sur la façon de travailler et établissent leurs propres règles. Cette situation est à l'origine d'une organisation informelle parallèle et moins structurée qui a parfois plus d'influence que l'organisation formelle structurée. Conséquemment, les bio-informaticiens du laboratoire se retrouvent dans une situation difficile et n'ont pas le temps d'établir une procédure de travail stable, étant toujours en mode réactif face aux demandes en constantes évolutions.

Une autre difficulté a été la collecte des informations concernant le flux de travail étudié. Parfois les bio-informaticiens ne connaissaient pas le fonctionnement des outils disponibles ou n'avaient pas une compréhension exacte de la source de données utilisée. Cependant, il faut souligner leur grande disponibilité et l'intérêt qu'ils ont démontré à faciliter l'accès à toutes les ressources nécessaires pour la bonne réalisation de ce projet de recherche appliquée. Cette étude leur a permis, parallèlement, de réfléchir sur la façon de travailler et de prendre conscience de lacunes et de problématiques concernant ce flux de travail.

3.4 Travaux futurs

Cette première recherche réalisée au laboratoire du Dr Hamet n'est que le début d'une série d'activités qui visent à améliorer le flux de travail du laboratoire. D'autres travaux suivront afin de compléter cette analyse, d'améliorer la modélisation de données ADAM et de mettre en place un prototype fonctionnel.

Ce qui reste à améliorer :

- Validation des exigences avec l'étude de l'étape de génotypage et notamment les données provenant d'Affymetrix;
- Étude des outils d'analyse utilisés dans d'autres étapes, par exemple l'association GWAS, l'imputation de génotypes et la visualisation pour valider que toutes les informations nécessaires soient comprises dans le modèle de données.
- Validation du modèle de données dans le cas de l'apparition des nouvelles exigences.
- Validation de la correspondance avec le modèle d'ADAM.

Ce qui reste à faire :

- Conception du modèle de données ADAM.
- Transfert des données vers la nouvelle structure.
- Calcul de variations génétiques des individus sur la nouvelle structure.
- Tests de performance du nouveau modèle.

CONCLUSION

Face aux derniers développements technologiques, notamment la technologie ADAM (BigData appliqué à la génomique) et les nouvelles plateformes de séquençage du génome humain, la recherche dans le domaine de la médecine personnalisée se trouve aujourd'hui en pleine expansion. Les centres de recherche spécialisés en génomique doivent adapter leurs infrastructures technologiques pour faire face aux nouveaux défis. L'adoption de ces nouvelles technologies nécessite l'utilisation d'une méthodologie de travail structurée qui permettra d'aboutir à l'objectif final. Une attention spéciale doit être portée à l'étape d'analyse, car elle permet de comprendre les besoins de l'organisation. Une mauvaise compréhension de ces besoins peut entrainer, dans le pire des cas, l'échec du projet.

L'analyse effectuée au CRCHUM a permis de comprendre l'organisation et les processus d'affaires. Ensuite, une étape du flux de travail, la préparation des données, a été identifiée comme la plus problématique en ce qui concerne la gestion des données. Pour approfondir l'étude de cette étape, une méthodologie d'analyse d'exigences fondée sur les notions proposées dans les guides BABOK et SWEBOK a été suivie. Finalement, un document d'analyse de besoins (le document Vision) a été produit ainsi qu'un nouveau modèle de données et une correspondance entre ce modèle et celui d'ADAM.

Une note personnelle concernant ce travail. Ce travail de recherche appliquée m'a permis de constater que l'étape d'analyse des besoins est à la fois incontournable et l'une des activités les plus importantes de tout projet de génie logiciel. Sans l'utilisation d'une méthodologie d'analyse bien structurée, les risques de ne pas repérer les bonnes informations sont toujours présents. Cependant, seule la présence d'un guide méthodologique n'est pas suffisante. L'expertise de l'analyste, l'appui des spécialistes du domaine sont aussi essentiels pour obtenir des résultats probants. Plusieurs compétences, spécifiques à ce métier, et personnelles m'ont été nécessaires pour le bon succès de ce travail. La formation mais aussi l'expérience de travail a grandement contribué à enrichir mes compétences.

ANNEXE I

DOCUMENT DE SPÉCIFICATION DES EXIGENCES D'AFFAIRES ET D'APPLICATION

Analyse de l'utilisation potentielle de la structure de données de l'université Berkeley (Adam)

Version : 0.8.7

DOCUMENT VISION

Réalisé par :
Liliana ALVARADO – ÉTS

Superviseur :
Prof. Alain APRIL – ÉTS
David Lauzon - ÉTS

approuvé le ___/___/___

Collaborateurs (utilisateurs qui expriment leurs exigences)

Collaborateurs
Pavel Hamet CHUM
Michael Phillips CHUM

approuvé le ___/___/___

Historique des révisions

Date	Version	Description	Auteur
16-05-2015	0.1	Construction du document	Liliana Alvarado
19-05-2015	0.1	Ajout de la portée et définition du problème	Liliana Alvarado
23-05-2015	0.1	Ajout de l'annexe A	Liliana Alvarado
29-05-2015	0.2	Ajout de la section des intervenants	Liliana Alvarado
04-06-2015	0.3	Ajout de la section environnement et liste de besoins	Liliana Alvarado
11-06-2015	0.4	Modification de la liste de besoins	Liliana Alvarado
12-06-2015	0.5	Modification de la section opportunité d'affaires.	Liliana Alvarado
12-06-2015	0.6	Commentaires du Dr April concernant le titre, les licences et les besoins	Alain April
14-06-2015	0.7	Révision du format et de la mise en page du document	Liliana Alvarado
16-06-2015	0.8	Modification de la section environnement utilisateur. Correction de la figure A I-2.	Liliana Alvarado
05-07-2015	0.8.1	Revue de l'ensemble du document et premiers commentaires	David Lauzon
07-07-2015	0.8.2	Modification de la figure A I-2. Modification des tableaux A I-1, A I-2 et A I-9. Ajout du tableau A I-10. Ajout de la figure A I-3. Description de la base de données clinique Prognomix.	Liliana Alvarado
15-07-2015	0.8.3	Modification des tableaux A I-1, A I-2, A I-9 et A I-11. Ajout du tableau A I-10. Modification de la figure A I-3. Ajout de la section 4.	Liliana Alvarado
20-08-2015	0.8.4	Ajout de la figure A I-3. Ajout de l'annexe IV	Liliana Alvarado
21-08-2015	0.8.5	Ajout de l'annexe V	Liliana Alvarado
25-08-2015	0.8.6	Ajout de la section 4,5,6,7,8,9 Ajout de l'annexe VI	Liliana Alvarado
19-09-2015	0.8.7	Vérifications finales	Liliana Alvarado

Table des matières

Page

1. Introduction 54
1.1 Objectif 55
1.2 Portée 55
1.3 Définitions, acronymes et abréviations 57
1.4 Références 58
2. Positionnement 58
2.1 Opportunités d'affaires 58
2.2 Énoncé du problème 59
2.3 Positionnement du produit 60
3. Descriptions des intervenants et des utilisateurs 61
3.1 Résumé des intervenants 61
3.2 Résumé des utilisateurs 63
3.3 Environnement utilisateur 64
3.4 Principaux besoins des intervenants et utilisateurs 73
4. Vue d'ensemble du produit 74
4.1 Perspective du produit 74
4.2 Principaux avantages 74
4.3 Hypothèses et dépendances 75
4.4 Licences et installation 75
5. Caractéristiques du produit 76
5.1 CAR01 - Nouveau modèle de la base de données clinique 76
6. Contraintes 76
7. Gammes de qualité 77
8. Attributs des caractéristiques 77
9. Autres exigences du produit 78
9.1 Standards applicables 78
9.2 Exigences du système 78
9.3 Exigences de performance 78
9.4 Exigences environnementales 79
10. Exigences de documentation 79
10.1 Manuel de l'utilisateur 79
10.2 Aide en-ligne 79
10.3 Guides d'installation, de configuration, et fichier à lire 79
A Attributs des caractéristiques 80

Liste des tableaux

Page

Tableau A I-1 Liste de définitions d'acronymes et d'abréviations 57

Tableau A I-2 Liste de références 58

Tableau A I-3 Énoncé du problème 59

Tableau A I-4 Positionnement du produit 60

Tableau A I-5 Résumé des intervenants 61

Tableau A I-6 Résumé des utilisateurs 63

Tableau A I-7 Liste de tables de la base de données Prognomix 67

Tableau A I-8 Liste des principaux fichiers utilisés dans la préparation des données 69

Tableau A I-9 Liste des composantes matérielles 70

Tableau A I-10 Liste des composantes logicielles 71

Tableau A I-11 Résumé des besoins des intervenants et utilisateurs 73

Tableau A I-12 Liste des avantages du produit 74

Tableau A I-13 Liste des attributs des caractéristiques du produit 77

Liste des figures

Page

Figure A I-1 Flux de travail du laboratoire du Dr P. Hamet……56

Figure A I-2 Processus de préparation des données.....64

Figure A I-3 Les activités de la préparation des données.....65

Figure A I-4 Modèle de données – Base de données Prognomix.....68

1. Introduction

Le centre de recherche du centre hospitalier de l'université de Montréal (CRCHUM) est un organisme dédié à la recherche dans le domaine médicale. Ses activités de recherche ont pour objectif de faire des découvertes majeures pour les patients et la population. Le Dr Pavel Hamet, du CRCHUM, et son équipe de chercheurs, concentrent leurs efforts de recherches sur les traitements du diabète type 2 (c.-à-d. acronyme en anglais T2D). Les personnes ayant cette maladie ont un risque plus élevé d'avoir des complications cardiaques, rénales et cardiovasculaires. L'intérêt médical de la recherche est d'identifier, à une étape précoce, les patients à risque élevé qui pourront mieux bénéficier de traitements préventifs. Cette stratégie, étant très couteuse, ne peut pas être adoptée par le système de santé actuel.

Les avancements récents des recherches sur le génome humain représentent un bon potentiel pour supporter les objectifs de la recherche. En identifiant des biomarqueurs spécifiques au T2D, et, en les intégrant avec l'information génétique des patients ayant cette maladie, il est possible d'améliorer les prédictions des complications vasculaires ce qui permettrait d'implémenter des programmes de prévention à une étape précoce à fin d'améliorer la santé de ces patients tout en réduisant les coûts des traitements actuels.

Avec l'aide de l'analyse des données génétiques, il pourrait être possible d'individualiser les traitements et d'offrir des diagnostics de prévention des maladies. Pour ce faire, l'équipe du Dr Hamet a implémenté son propre pipeline (comme l'ont fait des nombreuses autres équipes dans le monde), qui vise à appuyer la découverte de biomarqueurs liés aux complications du T2D. Cette approche, souvent retrouvée actuellement, utilise des modèles de prédiction génétiques et des outils de visualisation afin d'isoler un traitement personnalisé approprié pour chaque patient. Cependant plusieurs opportunités d'améliorations peuvent être réalisées en utilisant les logiciels libres disponibles ainsi que les connaissances dernièrement publiées par l'université Berkeley. Afin d'évaluer comment utiliser ces innovations, les activités existantes (c.-à-d. la situation actuelle), doivent être décrites en détail, ce qui est l'objectif du présent document.

1.1 Objectif

Le but de ce document est de décrire la situation actuelle des logiciels d'appui à la recherche et de préciser les besoins et exigences de haut niveau. Lors de cette étape, il est important de décrire, en détail, les fonctionnalités actuelles du processus d'analyse de variations de génomes utilisées par le Dr Hamet. Cette analyse permettra :

- D'identifier les fonctionnalités à améliorer;
- De formaliser les exigences fonctionnelles tout au long des processus à améliorer;
- De proposer des alternatives de solution à partir de la plateforme Adam de Berkeley.

1.2 Portée

Le processus actuel d'analyse de variations de génomes permet la découverte de biomarqueurs liés au T2D. La figure A I-1 présente ce flux de travail (de l'anglais «workflow»).

- La partie bleue illustre les processus de génotypage, la préparation des données et l'imputation, activités déjà fonctionnelles, qui s'effectuent au sein du laboratoire.
- La partie verte illustre le processus de permutation et de visualisation de données qui utilisent l'information produite lors du flux de travail. Au moment de cette étude, la partie visualisation est en étape de prototypage (voir le document de vision GOAT v1.0).
- Entre ces deux processus se trouve le processus d'association GWAS qui s'exécute en utilisant l'infrastructure fourni par l'organisme externe, Calcul Québec.

Deux sources de données sont actuellement utilisées : 1) les diverses études concernant le T2D seront utilisés à l'entrée de l'activité de génotypage, et 2) la base de données de références « *1000 Génomes* » sera utilisée à l'entrée de l'activité d 'imputation. Le génotypage et l'imputation produisent les données nécessaires à l'entrée du processus d'association GWAS. Une description détaillée du flux de travail actuel est disponible dans l'annexe II.

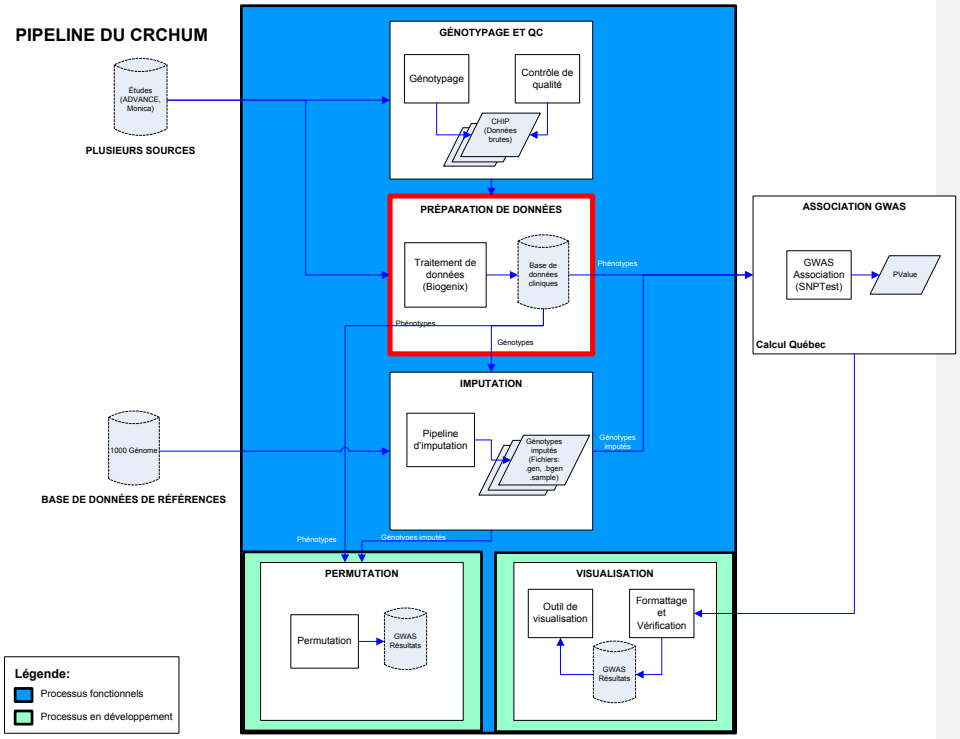


Figure A I-1 Diagramme du flux de travail du laboratoire du Dr Hamet

La portée de ce document de vision se limite au processus de *préparation des données* qui se trouve dans le rectangle rouge à la figure A I-1. La préparation des données est une étape qui permet de formater les données brutes obtenues du génotypage afin qu'elles puissent être utilisées par les processus de permutation, d'imputation et d'association GWAS. Cette étape utilise actuellement un logiciel développé par l'entreprise Biogenix, un partenaire de l'ÉTS, et une base de données clinique utilisant la technologie de logiciel libre PostgreSQL. Cette base de données a été modélisée par l'équipe du CHUM. Une problématique récurrente rapportée est que la structure de cette base de données requiert des traitements manuels à chaque fois qu'une étude GWAS est produite.

1.3 Définitions, acronymes et abréviations

Terminologie	Définition
Génome	L'ensemble d'information génétique d'un organisme.
Biomarqueur	Caractéristique biologique mesurable liée à un processus normal ou non.
Génotype	Information génétique responsable d'un trait physique.
Phénotype	L'ensemble de caractères observables d'un individu. Le phénotype est déterminé par le génotype.
ADVANCE	Acronyme en anglais de: "*Action in diabetes and vascular disease*". Base de données d'études cliniques. L'étude ADVANCE contient une grande base de données des tests médicaux obtenus sur des patients T2D.
MONICA	Acronyme en anglais de: "Multinational **MONI**toring of trends and determinants in **CA**rdiovascular disease". Étude sur les facteurs de risque et sur la tendance de la population par rapport aux maladies cardiovasculaires.
CARTaGENE	Étude qui contient une banque d'échantillons biologiques d'hommes et de femmes québécoises. L'étude a pour but d'identifier les facteurs génétiques et environnementaux causant les maladies chroniques communes qui affectent la population québécoise.
1000 Genome	Base de données de référence qui contient un catalogue de variations génétiques des humaines.
Calcul Québec	Regroupement d'universités québécoises qui offre des ressources matérielles et humaines au service de la recherche et de l'innovation.
GWAS	Acronyme en anglais de: "*Genome wide association study*". Cette étude permet la détermination de la contribution génomique sur une variété de maladies.
P-Value	Résultat de l'association GWAS, mesure de corrélation entre un marqueur génétique et la maladie ou phénotype d'intérêt.
SNP	Acronyme en anglais de: "*Single Nucleotide Polymorphism*". Les SNP's correspondent aux variations mineures du génome au sein d'une population. Ces variations peuvent être à l'origine de maladies génétiques ou de prédispositions à des maladies. Un SNP est une petite fraction du génome humain qui est composé d'environ 3 milliards de nucléotides.
ADAM	ADAM est un ensemble de formats des fichiers, des librairies et des outils client pour traiter, d'une façon efficace, les données génomiques. Développé par l'université de Berkeley, ADAM propose un flux de travail pour traiter des grandes quantités de données. Cette technologie se caractérise pour être hautement performante et facile pour la mise en échelle.

Tableau A I-1 Liste de définitions d'acronymes et d'abréviations

1.4 Références

Nom de la référence	Lien
Affymetrix	http://www.affymetrix.com/
Calcul Québec	http://www.calculquebec.ca/en/
1000 Genome	http://www.1000genomes.org/
ADVANCE	http://www.ncbi.nlm.nih.gov/pubmed/16075030
MONICA	http://www.thl.fi/monica/
CARTAGENE	http://cartagene.qc.ca/
SNPTest	https://mathgen.stats.ox.ac.uk/genetics_software/snptest/snptest.html
Impute2, Shapeit	http://mathgen.stats.ox.ac.uk/impute/pre-phasing.with.SHAPEIT_IMPUTE2.html
PLINK	http://pngu.mgh.harvard.edu/~purcell/plink/
ADAM	http://bdgenomics.org/

Tableau A I-2 Liste de références

2. Positionnement

2.1 Opportunités d'amélioration

L'étude de l'étape de préparation des données permettra aux chercheurs du CRCHUM d'exprimer leurs problématiques et besoins/exigences afin que les ingénieurs logiciel de l'ÉTS:

- Proposent des modifications à la conception actuelle du flux de travail, de manière à rendre plus efficace le travail en fournissant, dès le début, l'information appropriée dans les délais requis;
- Conçoivent une solution qui permet aux chercheurs de concentrer leurs efforts sur la recherche au lieu que sur des tâches informatiques;
- Documentent la situation actuelle d'une manière détaillée et identifient les besoins d'affaires et les différents processus du pipeline;

- Implémentent une structure de données, fondée sur la proposition de l'Université Berkeley Adam, qui supporte mieux les besoins d'affaires actuels et futurs du laboratoire.

2.2 Énoncé du problème

Le problème actuel du flux de travail du Dr Hamet est	• Une structure de données inflexible qui rend difficile l'exécution de requêtes changeantes rapidement. • Le besoin d'exécuter, répétitivement, des scripts pour formater les données à traiter. • La présence de données redondantes ou manquantes. • La présence de données non standardisées. • Le manque d'intégration/utilisation d'information provenant des plusieurs études disponibles. • La présence de plusieurs structures de données : fichiers et bases de données qui ne sont pas intégrées entre elles. • Le manque de sécurité d'accès aux données. • Aucun processus de traçabilité en place.
affecte	• La productivité des bio-informaticiens qui sont souvent en mode réactif et obligés à exécuter des tâches informatiques au lieu de se concentrer à appuyer la recherche. • La productivité des chercheurs qui ont besoin des bio-informaticiens pour extraire l'information nécessaire à la recherche.
dont l'impact est	• La surcharge du travail des bio-informaticiens. • Le travail de chercheurs est moins efficace. • Le risque que le résultat d'une découverte ne soit pas reproductibles et conséquemment qu'il puisse être approuvé par les autorités gouvernementales (c.-à-d. passer l'étape de validation indépendante nécessitant la traçabilité/provenance).
Une bonne solution serait	• La définition d'une structure de données, inspirée de Berkeley Adam, qui supporte mieux les besoins présents et futurs du CRCHUM.

Tableau A I-3 Énoncé du problème

2.3 Positionnement du produit

Pour	• Les bio-informaticiens qui sont les responsables d'exécuter toutes les étapes du flux de travail et fournissent les informations à toute l'équipe de recherche. • Les chercheurs qui utilisent les données stockées dans la base de données.
Qui	Ont besoin d': • Une structure de données plus flexible et facile à gérer. • Avoir une source uniforme d'information pour mieux répondre aux demandes d'information à l'interne. • Augmenter significativement l'efficacité du pipeline. • Inclure une fonctionnalité de provenance permettant facilement la répétition des expériences.
La nouvelle structure de données	est une structure d'information normalisée, de partage d'information, qui permet l'élasticité (c.-à-d. le BigData) des requêtes peu importe le type et la quantité des données futures à traiter.
Qui	Fourni de l'information aux différentes étapes du flux de travail et qui supporte les besoins des utilisateurs et des parties prenantes.
Contrairement à	la structure de données actuelle qui n'est pas uniformisée et est peu flexible.
Notre proposition	• Uniformise l'information qui sera utilisée par le flux de travail dans une structure de données moderne, normalisée et flexible, plus facile à entretenir et qui peut être traitée par les technologies modernes du BigData. • Intègre l'information nécessaire pour supporter le flux de travail existant du Dr Hamet. • Intègre les besoins d'affaires présentés lors de rencontres de collecte d'information des différentes intervenants. • Enregistrera les informations pertinentes pour pouvoir effectuer la provenance des expériences. • Sécurisera l'information la plus sensible selon les exigences du Dr Hamet.

Tableau A I-4 Positionnement de la proposition

3. Descriptions des intervenants et des utilisateurs

3.1 Résumé des intervenants

Nom	Description	Responsabilités
Dr Pavel Hamet	Directeur, chercheur principal	• Promouvoir les activités de recherche. • S'assurer de l'obtention de fonds pour le développement des projets de recherche et de l'optimisation du pipeline. • Approuver les projets au sein du CRCHUM. • Proposer des améliorations tout au long du pipeline. • Effectuer des requêtes et des extractions de données pour des analyses spécifiques. • Préparer des conférences et des publications sur les sujets de recherche.
Dre. Johanne Tremblay	Directeur, chercheur	• Approuver, conjointement avec le chercheur principal, les projets du CRCHUM. • Proposer des améliorations tout au long du pipeline. • Effectuer des requêtes et des extractions de données pour des analyses spécifiques. • Préparer des conférences et des publications sur les sujets de recherche.
Michael Philips	Directeur	• Proposer des améliorations tout au long du pipeline. • Superviser le développement de projets au CRCHUM. • Chercher des partenaires pour assurer la veille technologique au sein du CRCHUM.

Tableau A I-5 Résumé des intervenants

Nom	Description	• Responsabilités

François Harvey François Marois Gilles Godefroid	Bio-informaticien	• Effectuer le contrôle de qualité des données tout au long du pipeline. • Administrer les différentes bases de données utilisées par le pipeline. • Administrer les applications utilisées pour l'opération du pipeline. • Exécuter des analyses spécifiques à la recherche. • Supporter les besoins d'information de l'équipe de recherche.
Liliana Alvarado	Étudiante en maîtrise des TI's - ÉTS	• Responsable de l'étude et de la rédaction du document de vision.
Dr Alain April	Superviseur de projets (ÉTS)	• Assurer la bonne application des techniques et méthodes informatiques pour la réalisation des projets effectués par les étudiants au CRCHUM. • Assurer la mise en place de la technologie adéquate selon les besoins du CRCHUM. • Conseiller l'équipe du CRCHUM sur les technologies disponibles.
David Lauzon	Étudiant de Phd ÉTS	• Commenter le document de vision • Ajouter les questions relatives à la correspondance avec la structure de données Adam

Tableau A I-5 Résumé des intervenants « suite »

3.2 Résumé des utilisateurs

Nom	Description	Responsabilités	Intervenant
Bio-informaticien	Professionnel qui utilise des outils informatiques spécialisés pour effectuer des recherches et des analyses dans le domaine des sciences biologiques ou médicales.	• Mettre à jour la base de données à partir des données provenant des études. • Programmer les scripts nécessaires pour le transfert de données aux différentes étapes du pipeline. • Programmer des requêtes spécifiques pour l'équipe de chercheurs. • Effectuer le contrôle de qualité des données tout au long du pipeline. • Faire des analyses sur les données existantes. • Administrer les bases de données et les applications utilisées tout au long du pipeline.	Michael Philips Dr Pavel Hamet Dre. Johanne Tremblay

Tableau A I-6 Résumé des utilisateurs

3.3 Environnement utilisateur

Cette section du document de vision, présente les caractéristiques de l'environnement de travail sur lequel s'exécute l'étape de « Préparation des données » (voir figure A I-2).

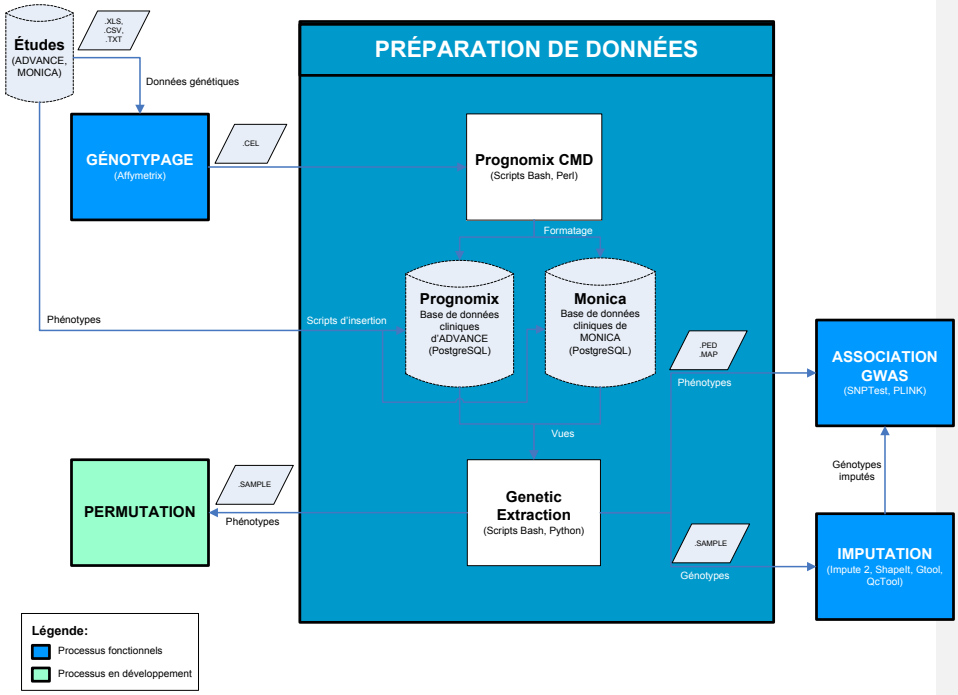


Figure A I-2 Processus de préparation des données

La séquence d'activités exécutées lors de la préparation des données est représentée dans la figure A I-3 (une description détaillé des activités se trouve dans l'annexe III).

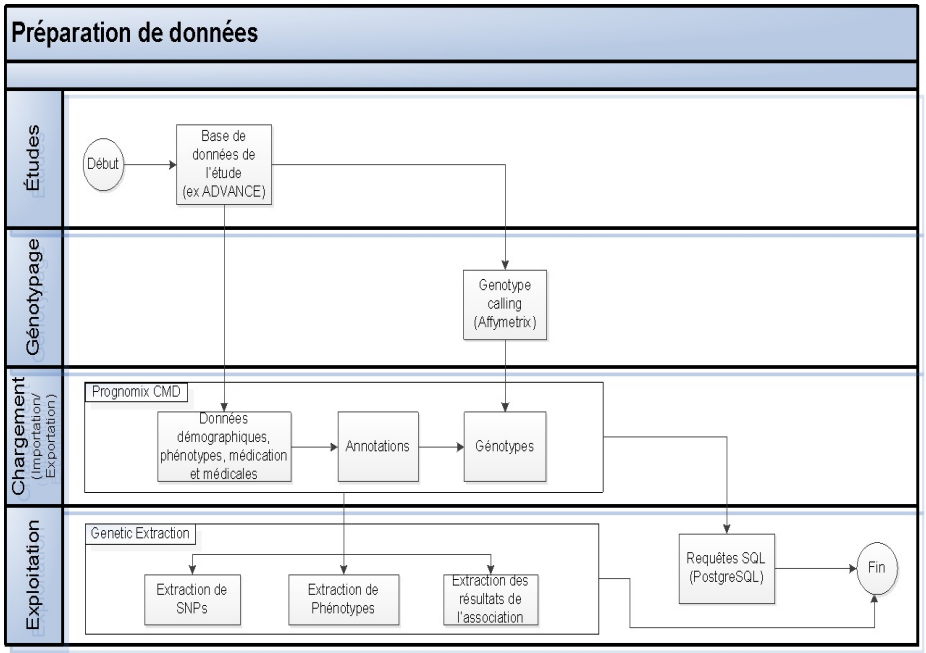


Figure A I-3 Les activités de la préparation des données

Intervenants

Les seuls intervenants, lors de cette étape sont les bio-informaticiens, qui connaissent le fonctionnement du flux de travail actuel et qui sont les responsables de son exécution. Ils exécutent des requêtes sur la base de données pour effectuer leurs analyses.

Sources de données

Les données d'entrée proviennent de l'externe. Il s'agit des études sur le T2D et les maladies cardiovasculaires. Actuellement s'utilisent les études ADVANCE et MONICA. Ces deux bases de données sont disponibles en ligne pour les chercheurs participants à ces études. Les données sont reçues en format .xls, .txt ou .csv.

Note: D'autres études seront utilisées dans le futur, c'est le cas de CARTaGENE, base de données sur la population québécoise, disponible pour les membres de la communauté scientifique et académique.

Base de données clinique

Contient les données des études et du résultat du génotypage d'Affymetrix. Fourni l'information aux prochaines étapes du flux de travail du CRCHUM soit: l'imputation, la permutation et l'association GWAS. Actuellement, chaque étude génère une nouvelle base de données qui suit le même modèle; cette façon de travailler constitue une des principales problématiques au niveau de l'étape de préparation des données. Le tableau A I-7, ci-dessous, décrit la liste de tables utilisées par la base de données clinique généré pour l'étude ADVANCE et la figure A I-4 présente le modèle actuel de la base de données Prognomix (un dictionnaire de données détaillé se trouve dans l'annexe IV).

- Nom de la base de données : Prognomix
- Moteur de Base de données : PostgreSQL
- Outil de gestion de données : PgAdmin

Nom de la table	Description	Provenance	Fréquence de mise à jour
batch	Liste de processus de génotypage d'Affymetrix	Interne	À chaque processus de génotypage
batches	Personnes analysées dans le génotypage	Interne	À chaque processus de génotypage
checked_phenotype	Phenotypes analysés pour le GWAS	Interne	À chaque processus GWAS
diagnosis	Liste de diagnostics	ADVANCE	Lors des actualisations d'ADVANCE
drug	Liste de médicaments	ADVANCE	Lors des actualisations d'ADVANCE
measure	Mesures associées aux phénotypes d'une personne	ADVANCE	Lors des actualisations d'ADVANCE
medical	Diagnostic de chaque visite	ADVANCE	Lors des actualisations d'ADVANCE
medication	Médicaments prescrits à une personne	ADVANCE	Lors des actualisations d'ADVANCE
pca	Résultat de l'analyse de composantes principales.	Interne	À chaque processus GWAS
person	Liste de personnes provenant d'une étude (ex ADVANCE)	ADVANCE	Lors des actualisations d'ADVANCE
phenotype	L'ensemble des caractères observables d'un individu	ADVANCE	Lors des actualisations d'ADVANCE
phenotype_views	Liste de vues des phénotypes d'intérêt utilisés pour le GWAS	Interne	À chaque processus GWAS
pseudo_autosomal*			
samples	Liste des échantillons provenant d'une étude	Interne	Lors des actualisations d'ADVANCE
snp_annotation	Table de référence de marqueurs génétiques	Affymetrix	Presque statique (Change selon les nouvelles versions dans la littérature scientifique)
snp_genotype	Liste de génotypes (ADN) des personnes	Affymetrix	À chaque processus de Génotypage
version	Liste de processus de génotypage d'Affymetrix	Internet	À chaque processus de Génotypage
visit	Liste des visites au médecin d'une personne	ADVANCE	Lors des actualisations d'ADVANCE
Total de tables: 18			
(*) La table n'est pas utilisée			

Tableau A I-7 Liste de tables de la base de données Prognomix (Étude ADVANCE)

MODÈLE DE DONNÉES
BASE DE DONNÉES PROGNOMIX

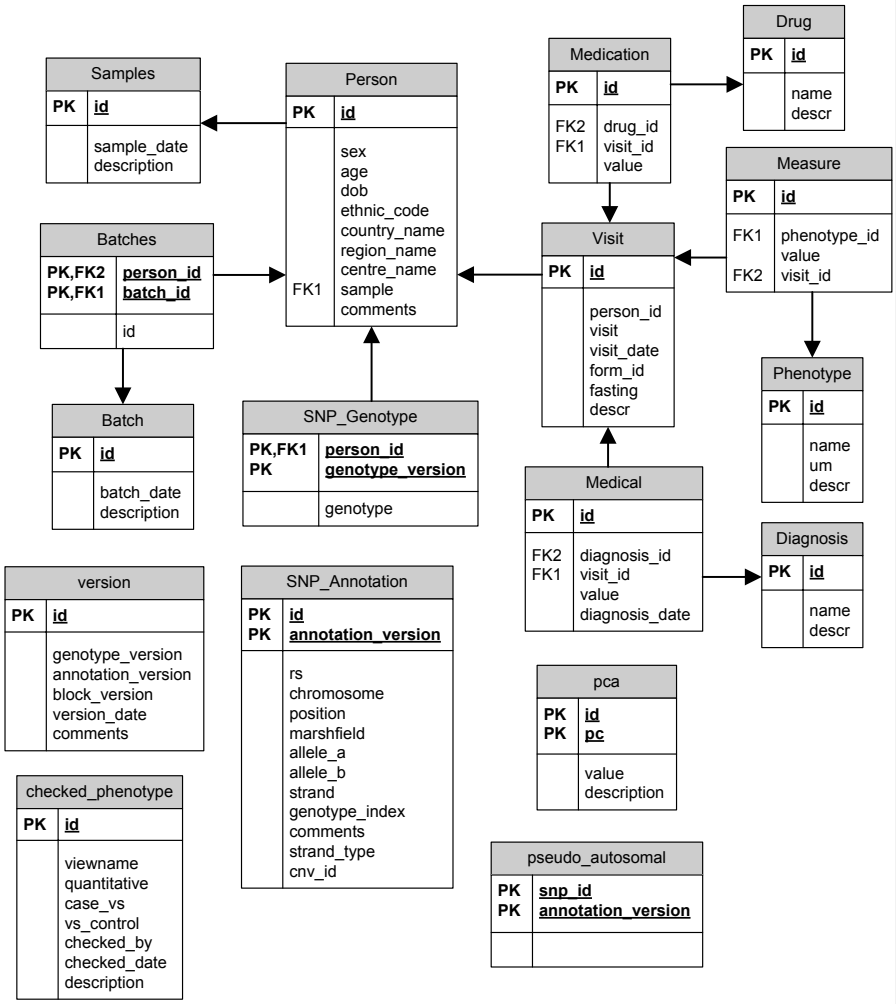


Figure A I-4 Modèle de données actuel – Base de données Prognomix

Fichiers

Description des principaux fichiers utilisés dans l'étape de préparation des données.

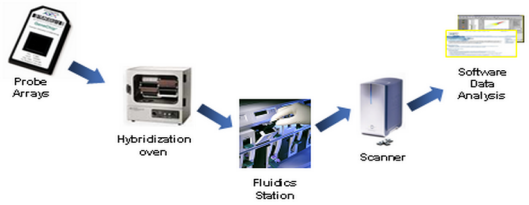
Type de fichier	Description	Utilisation
.CEL	Format de fichier généré par le génotypage d'Affymetrix. Contient les données brutes (de l'anglais « raw data ») du génotypage d'Affymetrix. http://media.affymetrix.com/support/developer/powertools/changelog/gcos-agcc/cel.html	Génotypage
.SAMPLE	Format de fichier produit par l'outil Genetic Extraction, utilisé par l'imputation et la permutation. http://www.stats.ox.ac.uk/~marchini/software/gwas/file_format.html#Sample_File_Format_	Imputation, Permutation
.PED	Format de fichier produit par l'outil Genetic Extraction et format par défaut utilisé par PLINK (association GWAS). Décrive l'information génétique des individus. http://www.shapeit.fr/pages/m02_formats/pedmap.html	Association GWAS
.MAP	Format de fichier produit par l'outil Genetic Extraction, utilisé par SNPTest (association GWAS). Contient la description des SNP, chaque ligne correspond à un SNP. http://www.shapeit.fr/pages/m02_formats/pedmap.html	Association GWAS

Tableau A I-8 Liste des principaux fichiers utilisés lors de la préparation des données

Composantes matérielles et logicielles

Les tableaux A I-9 et A I-10- décrivent les composantes de la plateforme informatique en place au CRCHUM.

A. Composantes matérielles

<table>
<tr><th>Type de composante</th><th>Description</th><th>Caractéristiques</th><th>Emplacement</th></tr>
<tr><td>Équipement pour le génotypage</td><td>Le traitement de biochips et le génotypage utilisent la technologie Affymetrix. Un équipement spécial est utilisé pour cette technologie.</td><td>• Système GeneChip®: un scanneur 3000 7G Plus, trois « fluidic stations » et un « hybridization oven »).
• Système GeneTitan® avec un robot Biomek® FXP qui permet de faire des expériences à haut débit.
 

Source : http://core.nhri.org.tw/encore/Labindex!List.action?lab_id=macore</td><td></td></tr>
<tr><td>Serveurs</td><td>Il existe trois serveurs Linux en production :
• Ancien : BLADE Server
• Nouveau : Rack Server

L’utilisation de chaque serveur n’est pas encore déterminée.</td><td>
<table>
<tr><th colspan="3">Ancien : BLADE Server (14 blades)</th><th colspan="2">Nouveau Serveur : 2 Rack Servers</th></tr>
<tr><th>Blade</th><th>Node</th><th>Caractéristique</th><th>2X Rack</th><th>2X Rack</th></tr>
<tr><td>TSM-NODE</td><td>1</td><td>Intel (R) Xeon (R) E5420/2.50 GHZ</td><td>UCS240 M3 rack servers</td><td>E2700</td></tr>
<tr><td>ADMIN-NODE</td><td>2</td><td>Intel (R) Xeon (R) E5420/2.50 GHZ</td><td>128 GB RAM</td><td>240 TB of data</td></tr>
<tr><td>MGMT-NODE</td><td>3</td><td>Intel (R) Xeon (R) E5420/2.50 GHZ</td><td>2 Intel Xeon E%-2660v2</td><td></td></tr>
<tr><td>DWE-NODE</td><td>4</td><td>Intel (R) Xeon (R) E5420/2.50 GHZ</td><td>300 GB internal HD</td><td></td></tr>
<tr><td>DATA-NODE</td><td>5</td><td>Intel (R) Xeon (R) E5420/2.50 GHZ</td><td></td><td></td></tr>
<tr><td>N156-NODE</td><td>6</td><td>Intel (R) Xeon (R) 5140/2.33 GHZ</td><td></td><td></td></tr>
<tr><td>SPARE-NODE</td><td>7</td><td>Intel (R) Xeon (R) E5420/2.50 GHZ</td><td></td><td></td></tr>
<tr><td>N155-NODE</td><td>8</td><td>Intel (R) Xeon (R) 5140/2.33 GHZ</td><td></td><td></td></tr>
<tr><td>N131-NODE</td><td>9</td><td>Intel (R) Xeon (TM) /3.20 GHZ</td><td></td><td></td></tr>
<tr><td>N136-NODE</td><td>10</td><td>Intel (R) Xeon (TM) /3.20 GHZ</td><td></td><td></td></tr>
<tr><td>N093-NODE</td><td>11</td><td>Intel (R) Xeon (TM) /3.20 GHZ</td><td></td><td></td></tr>
<tr><td>N139-NODE</td><td>12</td><td>Intel (R) Xeon (TM) /3.20 GHZ</td><td></td><td></td></tr>
<tr><td>N207-NODE</td><td>13</td><td>Intel (R) Xeon (TM) /3.20 GHZ</td><td></td><td></td></tr>
<tr><td>N133-NODE</td><td>14</td><td>Intel (R) Xeon (TM) /3.20 GHZ</td><td></td><td></td></tr>
</table>
</td><td>Les serveurs se trouvent physiquement au 2e étage du bâtiment du CRCHUM. Les bio-informaticiens y ont accès.</td></tr>
</table>

Tableau A I-9 Liste des composantes matérielles

B. Composantes logicielles

Type de composante	Nom	Version	Description	Référence	Appartenance
Système d'exploitation	Open SUSE	v11.1, v13.2,	2 Serveurs Linux : Blade Server (14 blades)	https://www.opensuse.org/en/	SUSE
	SUSE Linux Entreprise	v11.1, v11.3		https://www.suse.com/products/server/	SUSE
	Redhat Entreprise Linux	v7	1 Nouveau serveur Linux : 2 Rack servers.	http://www.redhat.com/en	Redhat
	Windows XP		Nécessaire pour exécuter les scripts Prognomix CMD		Microsoft
Base de données	PostgreSQL	v8.4	Moteur de base de données.	http://www.postgresql.org/	Logiciel libre
	PGAdmin	v1.12.1	Outil pour administrer une base de données PostgreSQL, développer et exécuter des requêtes et exécuter SQL.	http://www.pgadmin.org/	Logiciel libre
Outil de travail	Prognomix CMD	Non disponible	Application qui permet d'exécuter les scripts pour traiter les fichiers provenant du génotypage et les études pour les insérer les données dans la base de données.	http://prognomix.com/en/index.php	Prognomix
	Genetic extraction	Non disponible	Application qui contient des scripts Python pour extraire les données génotypiques, phénotypiques, etc. (provenant du génotypage) de la base de données clinique et produire les fichiers nécessaires pour les étapes de permutation, imputation et association GWAS.	http://prognomix.com/en/index.php	Prognomix
	Python	v2.5	Interpréteur, utilisé pour l'exécution des scripts de l'outil Genetic extraction.	https://www.python.org/	Logiciel libre

Tableau A I-10 Liste des composantes logicielles

Type de composante	Nom	Version	Description	Référence	Appartenance
Interfaces	Affymetrix Power Tools	v1.10.2	Exécute les génotypage à partir des données des études. Génère un fichier .CEL qui sera traité par l'outil Prognomix CMD. Il est à noter qu'une nouvelle version sera mise en place : Affymetrix GeneTitan AXIOM Custom Built vers la fin de l'année 2015.	http://www.affymetrix.com/estore/partners_programs/programs/developer/tools/powertools.affx	Logiciel libre
	SNPTest	v2.4.1	Exécute l'association GWAS, utilise les fichiers produits par l'outil « Genetic Extraction ».	https://mathgen.stats.ox.ac.uk/genetics_software/snptest/snptest.html	Logiciel libre
	Impute2	v2	Exécute l'imputation de génotypes. Reçoit le fichier .SAMPLE avec les génotypes extraits de la base de données cliniques.	http://mathgen.stats.ox.ac.uk/impute/impute_v2.html	Logiciel libre
	ShapeIt	v2.778	Exécute l'imputation de génotypes. Reçoit le fichier .SAMPLE avec les génotypes extraits de la base de données cliniques.	https://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html	Logiciel libre

Tableau A I-10 Liste des composantes logicielles « suite »

3.4 Principaux besoins des intervenants et utilisateurs

Besoin	Priorité	Préoccupations	Solution actuelle	Solution Proposée
B01 – La structure de l'information utilisée tout au long du pipeline n'est pas flexible.	Haute	La structure actuelle exige des traitements manuels pour pouvoir fournir les informations aux différents processus du flux de travail.	Une base de données relationnelle avec plusieurs fichiers d'entrée et de sortie. Nécessite des scripts/outils pour traiter chaque type de fichier.	Vérifier la structure actuelle de données et proposer des améliorations (selon le modèle ADAM) pour uniformiser l'information, intégrer l'information manquante et éliminer les redondances.
B02 – Intégrer l'information de plusieurs études T2D.	Haute	La mise à jour de la BD avec l'information des études est une tâche manuelle et peut entrainer des erreurs. Seulement deux sources de données sont supportées.	Une nouvelle BD est créée pour supporter chaque nouvelle étude. Plusieurs scripts d'insertion de données existent.	Standardiser l'information qui sera importée des études. Intégrer l'information des diverses études dans la nouvelle structure de données.
B03 – Traiter l'information provenant du génotypage pour rendre son accès plus efficace.	Haute	Les génotypes sont importés dans un format non relationnel (long string). Il est impossible de faire des requêtes sur les génotypes.	Les génotypes sont sauvegardés dans une table sur une seule colonne qui a environ 950,000 caractères. Un script spécial de lecture est nécessaire.	Adapter la structure de la table contenant les génotypes et l'intégrer au modèle d'ADAM.
B04 – Actualiser la technologie utilisée pour le traitement de données pour rendre l'accès plus efficace.	Haute	Les outils utilisés pour l'importation de données et pour consulter les génotypes sont désuets. Aucun support du fournisseur. Documentation limitée et disperse.	Les bio-informaticiens utilisent souvent des procédures manuelles pour le transfert de l'information. Seul l'importation et la consultation de génotypes sont effectuées avec l'aide de ces outils.	Évaluer la possibilité d'utiliser la technologie proposée par ADAM.
B06 – Sécuriser l'accès aux données sensibles.	Haute	Les données de personnes participantes aux études ne sont pas actuellement sécurisées.	Seulement des mesures de contrôle d'accès à la BD sont en place. Les données sensibles ne sont pas sécurisées.	Identifier les données sensibles à sécuriser et proposer un mécanisme de sécurité.
B07 – Intégrer des mécanismes de traçabilité.	Haute	Les expériences ne peuvent pas être répétées ce qui risque d'invalider les découvertes.	Aucun mécanisme de traçabilité/provenance en place.	Proposer un processus de provenance.
B08 – Intégrer les résultats de l'association GWAS dans une seule base de données.	Moyenne	Il est impossible de faire des requêtes pour associer les résultats GWAS avec les phénotypes et les génotypes.	Les résultats du GWAS sont sauvegardés dans une base de données à part.	Vérifier la structure de la base de données GWAS pour proposer son intégration dans la nouvelle structure de données.

Tableau A I-11 Résumé des priorités des besoins exprimés par les intervenants/utilisateurs

4. Vue d'ensemble du produit

Cette section présente une vue d'ensemble des fonctionnalités de la base de données ainsi que ses interactions avec d'autres applications.

4.1 Perspective du produit

Le produit est une nouvelle base de données qui suivra le plus possible le schéma de fichiers proposé par ADAM ce qui permettra au CRCHUM de comprendre et de documenter leurs besoins d'information actuels et futurs ainsi que d'adopter une structure d'information qui supportera des technologies plus performantes.

4.2 Principaux avantages

Bénéfices pour le client	Caractéristiques correspondantes
B01 – La structure de l'information utilisée tout au long du pipeline n'est pas flexible.	CAR01 – Nouveau modèle de la base de données clinique suivant le schéma ADAM qui intègre les informations de plusieurs études et rendre l'accès aux génotypes plus flexible.
B02 – Intégrer l'information de plusieurs études T2D.	
B03 – Traiter l'information provenant du génotypage pour rendre son accès plus efficace.	

Tableau A I-12 Liste des avantages du produit

4.3 Hypothèses et dépendances

Pour concevoir cette base de données, certaines hypothèses doivent être établies.

H1 : Utilisation de la technologie d'Affymetrix pour le génotypage.
Le CRCHUM a investi dans l'adoption de cette technologie, il est donc nécessaire de tenir compte de cet aspect lors de la modélisation de la base de données pour respecter le format de données utilisé par Affymetrix.

H2 : Il est possible d'utiliser la technologie ADAM pour implémenter des pipelines de génotypage.
La technologie ADAM est conçue principalement pour les cas du séquençage. Au laboratoire s'utilise le génotypage. Ces deux approches de l'analyse de génomes visent au même objectif final, la découverte de variations génétiques. Le transfert du modèle d'ADAM pour l'utiliser dans les cas du génotypage doit être possible.

4.4 Licences et installation

Dans ce document, l'équipe de l'ÉTS ont traduit les besoins des utilisateurs / enrichis (exprimée par les utilisateurs) en détail afin de capturer les exigences générales de l'adaptation du logiciel Adam, à plus long terme, pour rencontrer les exigences spécifiques de l'équipe de recherche du Dr Pavel Hamet. Le document se réfère à de l'information publique sur les projets open-source suivants: MetaboAnalyst, LocusZoom, SNPTest, Navigateur GWAS Schéma, HapGen, Biopython, IGV et de nombreux autres projets open source. Avant qu'Adam puisse utiliser l'une des informations de ces sources, leurs licences individuelles devront être étudiées. Seulement open-source et du matériel disponible publiquement seront intégrés à la chèvre.

GOAT Software License: Le logiciel de GOAT résultant sera autorisé: General Public License: versions v3 ou tout ultérieures Gpl

Installation: L'installation du logiciel sera faite par CRCHUM spécialistes de la bioinformatique au Dr Hamet CRCHUM Lab.
Licence GOAT Vision Document: Le document de vision de GOAT résultant est autorisé: Creative commons - Attribution-ShareAlike 4.0 licence internationale

5. Caractéristiques du produit

5.1. CAR01 – Nouveau modèle de la base de données clinique

Le nouveau modèle doit tenir compte des aspects suivants :

- **S'adapter au modèle de données proposé par ADAM**
  (Voir le modèle de données ADAM dans l'annexe IV)
- **Supporter plusieurs sources de données provenant de différentes études**
  Actuellement deux sources de données sont utilisées : ADVANCE et MONICA. L'étude CARTaGENE pourrait s'utiliser dans le futur.
- **Intégrer l'information manquante et standardiser l'information existante**
  Une analyse plus détaillé des activités de la préparation des données et du modèle actuelle de données a été réalisée pour identifier les problématiques (voir les annexes III et IV).
- **Accéder plus efficacement aux données de génotypage**
  L'information du génotype est sauvegardé sur une colonne qui contient un long string (950,000) caractères ce qui rends la consultation très difficile. S'inspirer du modèle d'ADAM pour proposer une solution.

6. Contraintes

Le modèle actuel de données est implémenté dans PostgreSQL sur Linux, il suit le modèle relationnel. La proposition d'ADAM suit un modèle non relationnel, aucune architecture n'est en place pour supporter le modèle d'ADAM. Par conséquence, le nouveau modèle de base de données doit être initialement implémenté sur PostgreSQL ce qui facilitera la

migration de données et les tests. Ce modèle pourra être d'abord validé sur PostgreSQL pour ensuite être migré vers ADAM.

7. Gammes de qualité

CN01 – Intégralité de données

La structure de données proposée doit tenir compte de toutes les entités d'information participantes dans l'étape de la préparation des données.

CN02 – Consistance entre les différentes sources de données d'entrée

Peu importe la source de données (les études), l'information enregistrée dans la base de données doit être consistante et complète.

CN03 – Accessibilité

L'structure de données doit permettre d'exécuter les recherches facilement. Chaque entité doit être identifiée d'une façon claire et simple en respectant le modèle d'affaires.

CN04 – Contrôle des actualisations

Le modèle doit permettre de contrôler les différentes actualisations de données (ex chargements de données, versions du génotypage et d'annotation, etc.)

8. Attributs des caractéristiques

Cette section résume les caractéristiques selon les définitions données en annexe A.

Caractéristiques	État	Bénéfice	Effort	Risque	Stabilité	Priorité
CAR01	Proposé	Élevé	Moyen	Faible	Élevé	Important

Tableau A I-13 Liste des attributs des caractéristiques du produit

9. Autres exigences du produit

9.1 Standards applicables

Les représentants du CRCHUM n'ont pas stipulé l'utilisation d'aucun standard pour la modélisation de la base de données.

9.2 Exigences du système

- **Sécurité :** Le transfert de données de sources et du génotypage doit être fait de façon sécuritaire pour garantir la consistance et la confidentialité des données.

- **Accessibilité :** L'utilisation de la base de données est plutôt à l'interne. L'accès est destiné à l'équipe de chercheurs et les bio-informaticiens.

- **Portabilité :** Le système d'exploitation utilisé au CRCHUM est Linux sur différents versions (voir les composantes technologiques, tableau A I-10). La base de données doit être implémentée sur cette configuration.

9.3 Exigences de performance

L'extraction/consultation de génotypes est une de principales activités lors de la recherche, cette opération demande beaucoup de temps de réponse. L'ADN d'une personne est représentent pour environ 950,000 caractères, et il existe environ 5,100 personnes dans l'étude ADVANCE ce qui fait un total de 950000 x 5100 registres dans la table de génotypes. La nouvelle structure de données doit faciliter l'accès aux génotypes et l'architecture sur laquelle sera implémentée doit assurer un temps de réponse adéquat.

9.4 Exigences environnementales

Cette section n’est pas applicable pour ce projet.

10. Exigences de documentation

10.1 Manuel de l’utilisateur

Un modèle de données ainsi qu’un dictionnaire de données détaillant les tables et les caractéristiques des champs doivent être fournis.

10.2 Aide en-ligne

Cette section n’est pas applicable pour ce projet.

10.3 Guides d’installation, de configuration, et fichier à lire

Cette section n’est pas applicable pour ce projet.

Annexes A

Attributs des caractéristiques

État	Proposé	Cet état indique que la caractéristique a été proposée aux différents intervenants.
	Approuvé	Cet état indique que la caractéristique a été approuvée par les différents intervenants.
	Incorporé	Cet état indique que la caractéristique a été incorporée dans une itération précédente.
Bénéfice	Faible	Représente une caractéristique pour laquelle le client a peu d'intérêt.
	Moyen	Une caractéristique qui n'est pas essentielle pour le produit, mais le client a un intérêt.
	Élevé	Une caractéristique qui est primordiale pour le client.
Effort	Faible	Nécessite moins de 20 heures.
	Moyen	Nécessite entre 20 et 40 heures.
	Élevé	Nécessite plus de 40 heures.
Risque	Faible	La caractéristique est facilement réalisable dans le temps demandé.
	Moyen	La caractéristique est possible dans le temps demandé.
	Élevé	La caractéristique probable avec des délais dans l'échéancier.
Stabilité	Faible	Caractéristique qui va changer due à l'incertitude du client et au manque de compréhension de l'équipe de développement.
	Moyen	Caractéristique qui peut changer, mais qui est quand même bien comprise et utile pour le client.
	Élevé	Caractéristique qui ne va pas changer pour l'utilisateur et qui est bien comprise par l'équipe de développement.
Priorité	Critique	La caractéristique est essentielle pour le client. Si la caractéristique n'est pas implémentée dans le système, les attentes ne seront pas respectées.

	Important	La caractéristique est importante pour la satisfaction du client. Par contre, son absence ne causera pas de délais dans le développement
	Utile	La caractéristique est intéressante pour le client, mais si elle n'est pas incluse, il n'y aura pas d'impact majeur.

ANNEXE II

DESCRIPTION DES PROCESSUS AU CRCHUM

Flux du travail du CRCHUM

Le pipeline commence par la *collecte de données* provenant des diverses études sur le diabète type 2 (T2D). Au CRCHUM s'utilise l'étude ADVANCE et présentement l'étude MONICA vient de s'ajouter au flux de travail. Une fois les données collectées à partir des études, une partie de ces données (les phénotypes) sont insérés dans la base de données clinique et, une autre partie doit être traitée par la plateforme Affymetrix qui effectue le génotypage proprement dit et le contrôle de qualité.

- Le *génotypage* permet de déterminer l'identité d'une variation génétique (génotype), à une position spécifique sur tout ou une partie du génome pour un individu ou un groupe d'individus donné. Un fichier .CEL qui contient l'information génétique d'une personne sera produit par Affymetrix. Le génotypage produit donc plusieurs fichiers .CEL pour chaque personne compris dans l'étude ADVANCE.
- Le *contrôle de qualité* est un processus aussi effectué avec les outils d'Affymetrix.
- La *préparation des données* est une étape qui permet de formater les données brutes obtenus du génotypage pour qu'elles puissent être utilisées par les processus de permutation, d'imputation et d'association GWAS. Cette étape utilise l'outil Biogenix. Le fichier .CEL est traité avec des scripts Python (fournis par Biogenix) pour finalement insérer les génotypes dans la base de données clinique.
- *L'imputation* est le processus de prédiction de génotypes qui n'ont pas été détectés par le génotypage. L'imputation a son propre pipeline, ce processus produit les génotypes imputés qui seront utilisés par les étapes postérieures.
- *L'étude d'association de génomes* (en anglais GWAS association) permet la détermination de la contribution génomique sur une variété de maladies. Une étude GWAS se base sur l'analyse de nombreuses variations génétiques chez de nombreux individus, afin d'étudier leurs corrélations avec des traits phénotypiques. Due à la grande

quantité de données à traiter, le CRCHUM utilise la plateforme fourni par Calcul Québec. Les génotypes imputés et les phénotypes sont la source principale pour le GWAS, cette étude s'effectue à l'aide de l'outil SNIPTest qui produira les « P-VALUE » qui sont des mesures de corrélation entre les marqueurs génétiques (les génotypes) et la maladie.

- La *permutation* est un processus qui effectue des GWAS aléatoires pour valider les résultats. Il utilise sa propre base de données.
- La *visualisation* utilise le résultat du GWAS pour présenter l'information sur des différents formats pour que les chercheurs puissent effectuer leurs analyses plus avancées. Une base de données a été construite spécifiquement pour ce processus.

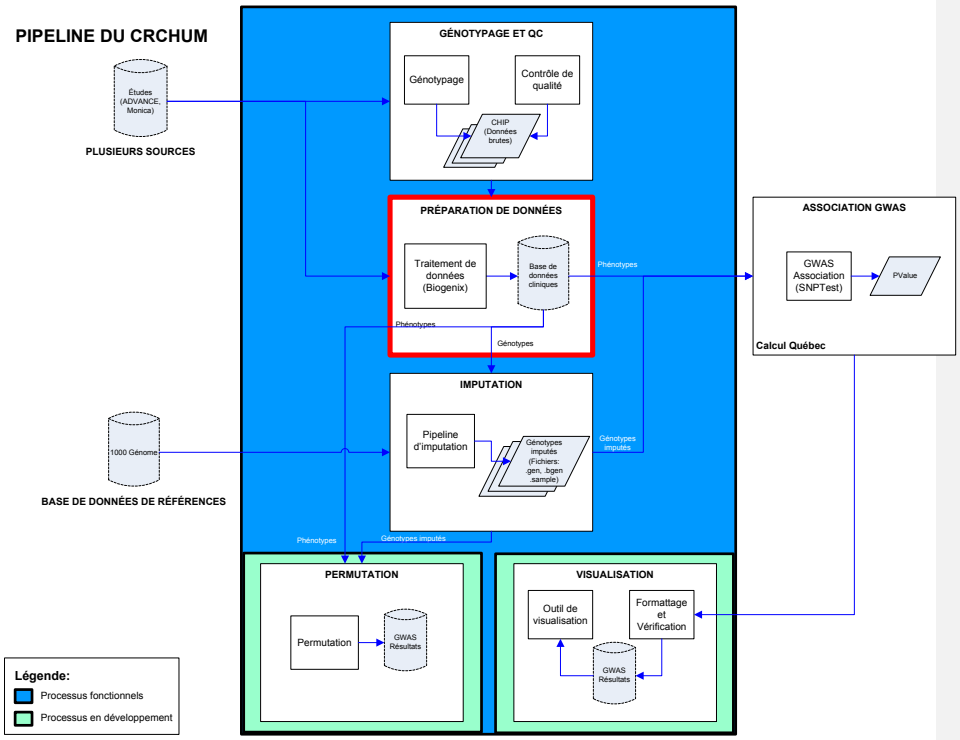


Figure A II-1 Diagramme des processus au CRCHUM

ANNEXE III

ACTIVITÉS DE LA PRÉPARATION DES DONNÉES

1. Collection de données des études

Structure de données d'ADVANCE (Démographique, phénotypes, médicale)

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Informat	Label
96	AC_RATIO	Num	8			AC ratio (µg/mg)
128	AC_RATIO_AVAIL	Num	8	YNF.	YESNOF.	Albumin:creatinine ratio measured within last 3 months and result available
114	AF	Num	8	YNF.	YESNOF.	Atrial fibrillation (current or previous)
150	AGE	Num	8			Age (years)
34	ALPHA_GLUCO	Num	8	YNF.	YESNOF.	Alpha-glucosidase inhibitor
89	ALT	Num	8			ALT (IU/l)
118	AMPUTATION	Num	8	YNF.	YESNOF.	Amputation secondary to vascular disease
5	AMPUTATION_L	Char	1	$NAYNF.		Lower extremity amputation (of one toe or more)
6	AMPUTATION_R	Char	1	$NAYNF.		RIGHT Lower extremity amputation (of one toe or more)
112	ANGINA	Num	8	YNF.	YESNOF.	Hospital admission for unstable angina
13	ANKLE_L	Char	1	$NAYNF.		Ankle reflex present
14	ANKLE_R	Char	1	$NAYNF.		RIGHT Ankle reflex present
83	AVG_CIG	Num	8			Average number of cigarettes smoked per day
142	BACK_PERIN	Num	8	YNF.		Background Perindopril dispensed at rand
17	BACK_RETINO_L	Char	1	$NAYNF.		Background retinopathy evident on fundoscopy
18	BACK_RETINO_R	Char	1	$NAYNF.		RIGHT Background retinopathy evident on fundoscopy
122	BLINDNESS	Num	8	YNF.	YESNOF.	Blindness in either eye thought to be due to diabetes
129	BLOOD_803	Num	8	YNF.	YESNOF.	Blood samples taken for creatinine, sodium, potassium, ALT and HbA1c assays
130	BLOOD_805	Num	8	YNF.	YESNOF.	Blood sample taken for lipids and glucose
131	BLOOD_CENTRAL	Num	8	YNF.	YESNOF.	Blood sample taken for long-term central storage
151	BMI	Num	8	4.1		Body Mass Index
116	CABG	Num	8	YNF.	YESNOF.	Coronary artery bypass graft or percutaneous transluminal coronary angioplasty
25	CATARACT_L	Char	1	$NAYNF.		Cataract
26	CATARACT_R	Char	1	$NAYNF.		RIGHT Cataract
143	CENTRE_ID	Num	8			Centre ID
146	CENTRE_NAME	Char	100			Centre name
100	CONSENT	Char	1	$CONSF.		Signed informed consent to study participation obtained
147	COUNTRY_NAME	Char	64			Country name
88	CREATININE	Num	8			Creatinine (umol/L)
154	CREAT_CLEARANCE	Num	8	5.1		Creatinine clearance
43	CURR_ACE_OTHER	Num	8	YNF.	YESNOF.	Other angiotensin converting enzyme inhibitor
44	CURR_ANGIO_II	Num	8	YNF.	YESNOF.	Angiotensin II receptor antagonist
47	CURR_ANTIHYP	Num	8	YNF.	YESNOF.	Other antihypertensive agent
51	CURR_ANTIPLATELET_OTHER	Num	8	YNF.	YESNOF.	Other anti-platelet agent
50	CURR_ASPIRIN	Num	8	YNF.	YESNOF.	Aspirin
70	CURR_BEER	Num	8			Average standard drinks of beer consumed/week

```
 45 CURR_BETA          Num    8 YNF.     YESNOF.  Beta-blocker
 46 CURR_CALCIUM       Num    8 YNF.     YESNOF.  Calcium antagonist
 49 CURR_CHOL_OTHER    Num    8 YNF.     YESNOF.  Other cholesterol lowering agent
124 CURR_CIG           Num    8 YNF.     YESNOF.  Current cigarette smoker
 41 CURR_DIURETIC_     Num    8 YNF.     YESNOF.  Other diuretic
    OTHER
 29 CURR_DRINKER       Num    8 YNF.     YESNOF.  Currently drink alcohol once a week or more
 48 CURR_HMG_COA       Num    8 YNF.     YESNOF.  HMG Co A reductase inhibitor (statin)
 54 CURR_HRT           Num    8 YNF.     YESNOF.  Hormone replacement therapy
 53 CURR_NITRATES      Num    8 YNF.     YESNOF.  Nitrates
 52 CURR_ORAL_         Num    8 YNF.     YESNOF.  Oral anticoagulant
    ANTICOAG
 42 CURR_PERIN         Num    8 YNF.     YESNOF.  Perindopril
 69 CURR_SPIRITS        Num     8                        Average standard drinks of spirits
consumed/week
 40 CURR_THIAZIDE      Num    8 YNF.     YESNOF.  Thiazide (or thiazide-like) diuretic
 68 CURR_WINE            Num     8                          Average standard drinks of wine
consumed/week
 95 DBP                Num    8                   Diastolic blood pressure
 56 DEMENTIA           Num    8 YNF.     YESNOF.  Does this person have dementia
 60 DIABETES_CHILDREN  Num    8                   Number of children with diabetes
 62 DIABETES_PARENTS   Num    8                   Number of parents with diabetes
 61 DIABETES_SIBLINGS  Num    8                   Number of brothers/sisters with diabetes
155 DIAB_DURATION      Num    8                   Diabetes duration
126 DIAGNOSIS_AGE      Num    8                   Age at first diagnosis of Type 2 diabetes
                                                  (not gestational diabetes)
 37 DIETICIAN          Num    8 YNF.     YESNOF.  Dietician referral
133 DISP_6WK           Num    8 YNF.     YESNOF.  6 week supply of perindopril-indapamide
                                                  fixed dose combination dispensed
149 DOB                Num    8 DDMMYY8. DATE10.  Date of birth
  7 DORSALIS_PEDIS_L   Char   1 $NAYNF.          Dorsalis pedis pulse present
  8 DORSALIS_PEDIS_R   Char   1 $NAYNF.          RIGHT Dorsalis pedis pulse present
 99 ECG_AF             Num    8 YNF.     YESNOF.  ECG results for Atrial fibrillation
132 ECG_PERFORMED      Num    8 YNF.     YESNOF.  ECG performed within last 3 months and copy
available
141 EDUCATION          Num    8                   Age at completion of highest level of formal
education
 27 ERECTILE           Char   1 $NAYNF.          Does this person have erectile dysfunction
136 ETHNIC_CODE        Num    8 ETHNICF.         Ethnic origin
123 EVER_SMOKED        Num    8 YNF.     YESNOF.  Ever smoked cigarettes regularly (i.e.
                                                  on most days for at least a year)
101 EXCL_ACE_CONTRA    Num    8 YNF.     YESNOF.  A definite contra-indication to
                                                  treatment with an ACE inhibitor
102 EXCL_ACE_          Num    8 YNF.     YESNOF.  A definite indication for an ACE inhibitor
    INDICATION                                    other than perindopril 2 or 4 mg daily
104 EXCL_GLICLAZIDE    Num    8 YNF.     YESNOF.  A definite and specific indication for, or
                                                  contra-indication    to,    treatment    with
gliclazide
105 EXCL_HBA1C          Num     8 YNF.      YESNOF.  A definite indication for, or contra-
indication
                                                  to, an HbA1c target of 6.5% or less
106 EXCL_INSULIN       Num    8 YNF.     YESNOF.  Current requirement for long-term therapy
                                                  with full-dose or bed-time insulin
108 EXCL_OTHER         Num    8 YNF.     YESNOF.  Other reason why the participant should be
excluded
103 EXCL_THIAZIDE       Num     8 YNF.      YESNOF.  A definite indication for, or contra-
indication
                                                  to, treatment with a thiazide-like diuretic
107 EXCL_TRIAL         Num    8 YNF.     YESNOF.  Participation in another trial (current
                                                  or within the last month)
 38 EXERCISE           Num    8 YNF.     YESNOF.  Weight control or exercise program
65 EXERCISE_MILD      Num    8                   Mild exercise for more than 15 minutes
 64 EXERCISE_MODR      Num    8                   Moderate exercise for more than 15 minutes
 63 EXERCISE_VIG       Num    8                   Vigorous exercise for more than 15 minutes
 57 FASTING            Num    8 YNF.     YESNOF.  Is patient fasting
135 FORM_Z              Num     8 YNF.      YESNOF.  Contact details form (Z) completed and
stored at LCC
 30 GLICLAZIDE         Num    8 YNF.     YESNOF.  Gliclazide MR
 59 GLINIDE            Num    8 YNF.     YESNOF.  Glinide
```

```
 85 GLUCOSE           Num    8                    Blood glucose (mmol/L)
 67 GUMS              Num    8                    Number of days gums bled in the last year
 84 HBA1C             Num    8                    Haemoglobin A1c (%)
 93 HDL_CHOL          Num    8                    HDL cholesterol (mmol/l)
140 HEIGHT            Num    8                    Height (cm)
 28 HELP              Num    8 YNF.     YESNOF.   Requirement for regular help with everyday
                                                  activities in the last two weeks
113 HF                Num    8 YNF.     YESNOF.   Hospital admission for heart failure
 81 HIP               Num    8                    Hip circumference (cm)
144 HISTORY_MACRO     Num    8 YNF.               Strata: History of macro
145 HISTORY_MICRO     Num    8 YNF.               Strata: History of micro
 39 HOME_BLOOD        Num    8 YNF.     YESNOF.   Home blood glucose monitoring
 72 HR                Num    8                    Heart rate (bpm)
115 HYPER             Num    8 YNF.     YESNOF.   Currently treated hypertension
 36 INSULIN_BED       Num    8 YNF.     YESNOF.   Long-term bedtime insulin
 35 INSULIN_DAY       Num    8 YNF.     YESNOF.   Long-term daytime insulin
 15 KNEE_L            Char   1 $NAYNF.            Knee reflex present
 16 KNEE_R            Char   1 $NAYNF.            RIGHT Knee reflex present
 21 LASER_L           Char   1 $NAYNF.            Previous laser therapy evident on fundoscopy
 22 LASER_R           Char   1 $NAYNF.            RIGHT Previous laser therapy evident on
fundoscopy
 82 LAST_SMOKED_AGE   Num    8                    Age in years last smoked regularly
 91 LDL_CHOL          Num    8                    LDL cholesterol (mmol/l)
 11 LIGHT_TOUCH_L     Char   1 $NAYNF.            Light touch sensation present below the knee
 12 LIGHT_TOUCH_R     Char   1 $NAYNF.            RIGHT Light touch sensation present below
the knee
 98 LV_HYP            Num    8 YNF.     YESNOF.   Left ventricular hypertrophy
 23 MACULAR_L         Char   1 $NAYNF.            Macular oedema
121 MACULAR_OEDEMA    Num    8 YNF.     YESNOF.   Macular oedema
 24 MACULAR_R         Char   1 $NAYNF.            RIGHT Macular oedema
 32 METFORMIN         Num    8 YNF.     YESNOF.   Metformin
110 MI                Num    8 YNF.     YESNOF.   Myocardial infarction
 77 MMSE              Num    8                    Mini Mental State Examination Score (/30)
  1 NEW_ID            Num    8                    blinded patient ID
 55 OPTHALMOLOGIST    Num    8 YNF.     YESNOF.   Are the reported eye findings based
primarily
                                                  on an examination by an ophthalmologist
139 PAST_BEER         Num    8                    Average standard drinks of beer
consumed/week
                                                  in year before diagnosis
127 PAST_DRINKER      Num    8 YNF.     YESNOF.   Drank alcohol regularly before diagnosis of
                                                  diabetes (most weeks for at least a year)
138 PAST_SPIRITS      Num    8                    Average standard drinks of spirits
consumed/week
                                                  in year before diagnosis
137 PAST_WINE         Num    8                    Average standard drinks of wine
consumed/week
                                                  in year before diagnosis
 58 PERIN_BACK_DISP   Num    8 YNF.     YESNOF.   Background perindopril dispensed (if
required)
 79 PERIN_BACK_       Num    8 PIDOSEF.           Background perindopril dose: 2mg daily or
4mg daily
    DOSE_CODE
 78 PERIN_BACK_NUM    Num    8                    Number of Background perindopril packs
dispensed
125 PIPE_CIGAR        Num    8 YNF.     YESNOF.   Current pipe or cigar smoker
  9 POST_TIBIAL_L     Char   1 $NAYNF.            Posterior tibial pulse present
 10 POST_TIBIAL_R     Char   1 $NAYNF.            RIGHT Posterior tibial pulse present
 87 POTASS            Num    8                    Potassium (mmol/l)
120 PROLIF_RETINO     Num    8 YNF.     YESNOF.   Proliferative retinopathy
 19 PROLIF_RETINO_L   Char   1 $NAYNF.            Proliferative retinopathy evident on
fundoscopy
 20 PROLIF_RETINO_R   Char   1 $NAYNF.            RIGHT Proliferative retinopathy evident on
fundoscopy
134 QOL_PROVIDED      Num    8 YNF.     YESNOF.   Quality of life questionnaire provided to
participant
 97 Q_WAVES           Num    8 YNF.     YESNOF.   Q-waves diagnostic of previous myocardial
infarction
148 REGION_NAME       Char  64                    Region name
```

```
119 RETINAL           Num    8 YNF.     YESNOF.  Retinal photocoagulation therapy
117 REVASC            Num    8 YNF.     YESNOF.  Peripheral revascularisation (angioplasty or
surgery)
 94 SBP               Num    8                   Systolic blood pressure
153 SEX               Num    8 SEXF.             Sex
 86 SODIUM            Num    8                   Sodium (mmol/L)
109 STROKE            Num    8 YNF.     YESNOF.  Stroke
  2 STUDY_FORM_ID     Num    8                   Study form ID (visit number)
 31 SULPHO            Num    8 YNF.     YESNOF.  Sulphonylurea other than gliclazide MR
 66 TEETH             Num    8                   Number of natural teeth
 33 THIAZOL           Num    8 YNF.     YESNOF.  Thiazolidinedione
111 TIA                Num    8 YNF.     YESNOF.  Hospital admission for transient ischaemic
attack
 92 TOTAL_CHOL        Num    8                   Total cholesterol (mmol/L)
 90 TRIGLYC           Num    8                   Triglycerides (mmol/l)
  3 ULCER_L           Char   1 $NAYNF.          Chronic ulceration on leg/foot
  4 ULCER_R           Char   1 $NAYNF.          RIGHT Chronic ulceration on leg/foot
 73 VISUAL_1_L        Num    8                   Visual acuity (corrected or through pinhole)
 75 VISUAL_1_R         Num    8                      RIGHT Visual acuity (corrected or through
pinhole)
 74 VISUAL_2_L        Num    8                   Visual acuity (corrected or through pinhole)
 76 VISUAL_2_R         Num    8                      RIGHT Visual acuity (corrected or through
pinhole)
 80 WAIST             Num    8                   Waist circumference (cm)
152 WAISTHIP_RATIO    Num    8 4.2               Waist/hip ratio (cm/cm)
 71 WEIGHT            Num    8                   Weight (kg)
```

2. Génotypage

Pour le génotypage s'utilise les outils d'Affymetrix (Affymetrix Power Tools), ensemble de programmes qui implémentent des algorithmes pour travailler et analyser les GeneChip arrays.
(http://www.affymetrix.com/estore/partners_programs/programs/developer/tools/powertools.affx)

L'outil d'Affymetrix produit un fichier .CEL qui contient le résultat du génotypage. La structure du fichier est disponible sur le lien ci-dessous :
http://media.affymetrix.com/support/developer/powertools/changelog/gcos-agcc/cel.html

3. Chargement de données

Le chargement de données permet l'importation des données des études et du résultat du génotypage d'Affymetrix vers la base de données Prognomix. Les activités de chargement de données utilisent l'outil **Prognomix CMD**. Les principales commandes pour exécuter les trois activités le plus importantes (l'importation de phénotypes, l'importation et l'actualisation des annotations et l'importation de génotypes) sont résumées ci-dessous.
Note : Les informations ont été prises du manuel d'utilisation de PrognomixCmd [1].

PrognomixCMD est un exécutable qui peut être lancé à partir de la ligne de commandes du système d'exploitation. Pour lancer l'application il suffit d'exécuter :

```
PrognomixCmd <command> <command_parameters>
```

Une commande ou une séquence de commandes peuvent être sauvegardé dans un fichier d'extension .cms en utilisant un éditeur de texte. Le fichier doit contenir une ligne par commande <command> <command_parameters>. Ensuite, pour exécuter les commandes utiliser :

```
PrognomixCmd --c <command_file_name>
```

3.1.Importation de phénotypes

Importation des données démographiques à partir de fichiers Excel provenant des études.

PhenoImport: Permet d'importer les informations dans les tables person, visit, medical, diagnosis, medication, drug, phenotype et measure.

Syntaxe

```
PhenoImport -i <input_file> -s <sample> [--nodb -o <output_file>]
            [-usr <user_name> -pwd <password>]
```

Paramètres

`-i <input_file>`	Nom du fichier d'entrée
`-s <sample>`	Numéro de données de tests « data sample number »
`--nodb`	Indicateur pour désactiver l'insertion de données
`-o <output_file>`	Nom du fichier SQL pour l'insertion de données (sans extension) (optionnel)
`--compare`	Indicateur pour activer la comparaison entre les nouvelles valeurs et celles déjà existantes pour la même personne et la même visite (optionnel)
`-usr <user_name>`	Nom de l'utilisateur pour la connexion à la base de données (optionnel)
`-pwd <password>`	Mot de passe pour la connexion à la base de données (optionnel)

Exemples

```
PrognomixCmd PhenoImport -i C:\pheno.txt -s 2 -o C:\pheno.sql -nodb
PrognomixCmd PhenoImport -i c:\temp\pheno.txt -s 1 -usr john_doe -pwd
******
```

3.2.Importation et actualisation des annotations

Les fichiers d'annotation d'Affymetrix sont une référence pour comprendre l'information qui se trouve dans les GeneChip arrays. Ces fichiers sont actualisés trimestriellement par Affymetrix (http://www.affymetrix.com/support/technical/annotationfilesmain.affx).

PrognomixCMD accède aux fichiers d'annotation d'Affymetrix pour importer et mettre à jour les annotations dans la base de données Prognomix.

Importation des annotations

SnpAnnotationImport: Importe les fichiers d'annotation d'Affymetrix dans la table snp_annotation. Une importation génère une nouvelle version de l'annotation.

Syntaxe

```
PrognomixCmd SnpAnnotationImport -i <input_file> -av <anno_version>
            -chp <500k | snp5 | snp6> [-usr <user_name> -pwd <password>]
```

Paramètres

-i <input_file>	Nom du fichier d'entrée
-av <annotation_version>	Nouvelle version de l'annotation
-chp <chip_type>	Type de chip type (500k, snp5 et snp6)
-usr <user_name>	Nom de l'utilisateur pour la connexion à la base de données
-pwd <password>	Mot de passe pour la connexion à la base de données

Exemple

```
PrognomixCmd SnpAnnotationImport -i c:\temp\annotation.csv -av 1 -chp 500k
PrognomixCmd SnpAnnotationImport -i c:\temp\annotation.csv -av 2
            -usr john_doe -pwd ****** -chp snp5
```

Actualisation des annotations

Dans la base de données Prognomix, les annotations sont liées aux génotypes par un index. Cet index représente la position dans la chaine de caractères du génotype où se trouve l'annotation. Chaque marqueur (SNP) de la table snp_annotation peut être identifié dans le string de la table snp_genotype par l'index.

SnpAnnotationUpdate: Actualise le champ genotype_index de la table snp_annotation pour une version d'annotation spécifique. Cette opération doit être effectuée à chaque fois qu'une nouvelle version de l'annotation est créée dans la base de données.

Syntaxe

```
PrognomixCmd SnpAnnotationUpdate -av <annotation_version>
            [-usr <user_name> -pwd <password>]
```

Paramètres

-av <annotation_version>	Version de l'annotation
-usr <user_name>	Nom de l'utilisateur pour la connexion à la base de données
-pwd <password>	Mot de passe pour la connexion à la base de données

Exemple

```
PrognomixCmd SnpAnnotationUpdate -av 1
PrognomixCmd SnpAnnotationUpdate -av 2 -usr john_doe -pwd ******
```

3.3.Importation de génotypes

SnpImport: Permet l'importation des génotypes vers la table snp_genotype de la base de données Prognomix.

Syntaxe

```
PrognomixCmd SnpImport -i <input_file> | -f <input_folder> -v <version>
            -b <batch> [-usr <user_name> -pwd <password>
```

Paramètres

-i <input_file(s)>	Liste des fichiers d'entrée (séparés par comma)
-f <input_folder>	Répertoire contenant les fichiers d'entrée (Il ne doit pas contenir d'autres fichiers)
-b <batch>	Numéro de batch pour le génotypage (ce numéro doit exister dans la table batch).
-v <version>	Id version (doit exister dans la table version)
-usr <user_name>	Nom de l'utilisateur pour la connexion à la base de données
-pwd <password>	Mot de passe pour la connexion à la base de données

Note: Les options -i et -f ne peuvent pas s'utiliser ensemble, mais une des deux doit être au moins présente.

Exemple

```
PrognomixCmd SnpImport -f c:\temp -b 2 -v 1
PrognomixCmd SnpImport -i c:\temp\p12345.txt,c:\temp\p88299.txt -b 3 -v 2
            -usr john_doe -pwd ******
```

4. Exploitation de données

4.1. Extraction de phénotype, génotypes et annotations

L'outil **Genetic extraction** permet l'extraction de génotypes, phénotypes et les annotations à partir de la base de données Prognomix. Les résultats peuvent être produits sur plusieurs formats (ex Plink, Haploview, CSV, .PED, .SAMPLE, etc.). Cet outil utilise l'interpréteur Python, les bio-informaticiens ont accès au code source.

dump_genetic: permet l'extraction des SNP (génotypes) et phénotypes en combinant plusieurs paramètres.

Paramètre	**Description**	**Exemple**
`--snp-index`	liste des index du string génotype	1015,200,10002000
`--snp-rs`	liste des rs id	rs7930823,rs11246002
`--snp-affy`	liste d'Affymetrix id	2015080,2077094
`--snp-chrom`	liste de chromosomes	2,38,19,X
`--snp-bp`	lise de chromosomes et de positions des paires de base	2:6002:900,X:150,X:450
`--snp-cm`	liste de chromosomes et de positions Marshfield	2:124.9722:187.111
`--snp-random`	choisi *N* SNPs autosomal aléatoirement	1000
`--snp-60k`	Affymetrix 60k less-performing SNPs	
`--snp-5.0`	SNPs sur Affymetrix 5.0 chip	

--snp-6.0	SNPs sur Affymetrix 6.0 chip	

 # Example #1: Ettore Markov model on HDL candidate SNPs

```
# Category 501 contains 1224 valid caucasians as of Dec 2007
# Version 4 contains their genotypes called together
python /usr/local/bin/dump_genetic.pyc \
--version-id 4 \
--cats 501 \
--outtypes csv-col2 \
--outfile hdl \
--only-geno \
--snp-rs rs5370,rs10503669,rs3890182,rs4775041,rs261332,...
```

 # Example #2: Ettore Markov model on autosomal SNPs

```
# Requires about 765 MB RSS for 1224 subjects, 500k SNPs
# We also dump the disease files
python /usr/local/bin/dump_genetic.pyc \
--version-id 4 \
--cats 501 \
--outtypes csv-col2 \
--outfile markov \
--only-geno \
--snp-chrom 1-22 \
--phenotypes view_dph_albuminuria_baseline,...
```

 # Example #3:

```
python /share/repo/Python/prognomix_python/bin/dump_genetic.py \
--version-id 3 \
--outtypes plink \
--snp-chrom 1-22,X,Y,MT \
--pidfile id_list.txt \
--outfile test
```

Le fichier id_list.txt contient les ids (1 par ligne) des individus que l'on veut extraire. Il est possible de combiner plusieurs paramètres dans le fichier d'entrée par exemple :

```
--snp-index 10-15,200,1000-2000
--snp-rs rs7930823,rs11246002
--snp-affy 2015080,2077094
```

Références

[1] PrognomixCmd User Guide, Prognomix, 2008.

[2] Genetic Extraction User Manual, Prognomix, 2008.

ANNEXE IV

DICTIONNAIRE DE DONNÉES DE LA BASE DE DONNÉES PROGNOMIX

Table / Column	Constraint	Description	Data type	Size	Scope / Example	Cardinality
medication (210,523 rows)						
id	PK	Unique medication identifier	integer	10		210523
drug_id	FK1, UNQ	Drug identifier	integer	10		23
visit_id	FK2, UNQ	Visit identifier	integer	10		51125
value	CHK(0,1), UNQ	Take medication indicator	integer	10	0=The patient does not take the medication, 1=No medication	
pca (1,389 rows)						
id	PK, UNQ	Unique PCA identifier	integer	10	Principal component analysis	3496
pc	PK, UNQ	PCA component number	integer	10	[1,2,3,4,5,6]	6
value		PCA value	double	17,17	ex. -0.5368, 0.6911	
description		PCA description	varchar	2147483647	"pc1 on 1133 East-European in ADVANCE"	
person (5,157 rows)						
id	PK	Unique person identifier	integer	10		5157
	CHK (>0)					
sex	CHK("m","f")	Sex of person	char(1)	1	"m", "f"	2
age		Age of person	integer	10		
dob		Date of birth	date	13	"YYYY-MM-DD"	
ethnic_code		Ethnic code	integer	10	From 1 to 13	13
country_name		Name of country	varchar	64		18
region_name		Name of region	varchar	64	"Canada","Europe - Continental", "Europe - Northern"	4
centre_name		Name of collaborating center	varchar	1024	"Centre de recherche clinique de Laval"	153
sample	FK1	Sample identifier	integer	10	From 1 to 8	8
comments		Comments	text	2147483647		
phenotype (82 rows)						
id	PK	Unique phenotype identifier	integer	10		82
name		Phenotye name	varchar	128	"sodium", "diab duration"	82
um		Measure unit	varchar	32	"years", "ug/mg","umol/L","beats /minute"	24
descr		Phenotype description	varchar	1024	"time since diabetes was diagnosed"	
phenotype_views (248 rows)						
name	PK	View name	text	2147483647		248
qualitative			bool	1	"f", "t"	

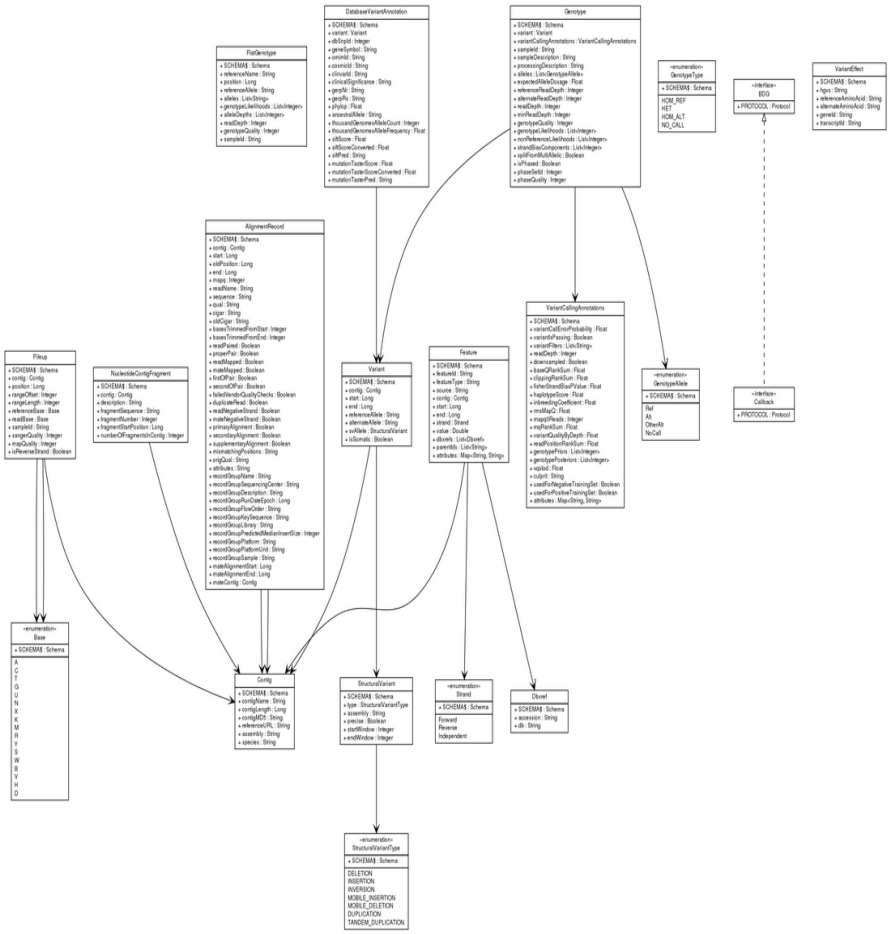
Table / Column	Constraint	Description	Data type	Size	Scope / Example	Cardinality
pseudo_autosomal (840 rows)						
snp_id	PK, CHK(>0)	Unique snp identifier	integer	10	Min(4207749), Max (8574011)	422
annotation_version	PK, CHK (>0)	Annotation version	integer	10	1 or 2	2
samples (9 registres)						
id	PK	Unique sample identifier	integer	10		9
sample_date		Date of sample	date	13	"YYYY-MM-DD"	
description		Description of sample	text	2147483647	"First 200 extreme status caucasians subjects"	
snp_anotation (1,863,892 rows)						
id	PK, CHK(>0)	Unique snp annotation id	integer	10		1863892
annotation_version	PK, UNQ, CHK(>0)	Annotation version	integer	10	1 or 2	2
rs	IDX1	SNP unique identifier	varchar	32	"rs10000012"	932558
chromosome		Chromosome number of SNP	varchar	32	From 1 to 22 and "MT", "X", "Y"	25
position	CHK(>=0)	Chromosome physical position	bigint	19		1850490
marshfield	CHK(>=0)	Chromosome Marshfield position	double precision	17,17	ex. 0.00160588407589416	1099813
allele_a	CHK("A","C","G","T")	Allelle A	char(1)	1	"A","C","G","T"	4
allele_b	CHK("A","C","G","T")	Allele B	char(1)	1	"A","C","G","T"	4
strand	CHK("+"-")	Strand	char(1)	1	"+", "-"	2
genotype_index	IDX2, UNQ, CHK(>=1, <=931946)	SNP genotype string indices	integer	10	From 1 to 931946	931946
comments		Comments	text	2147483647		
strand_type	CHK("s", "r")	Type of strand	char(1)	1	"r", "s"	2
cnv_id	CHK(>0)	Copy number variations identifier	int4	10		15254

Table / Column	Constraint	Description	Data type	Size	Scope / Example	Cardinality
snp_genotype (7,074 rows)						
person_id	PK, FK	Unique snp_genotype identifier	integer	10		3537
genotype_version	PK, CHK(>0)	Gernotype version	integer	10	1 or 2	2
genotype	CHK(length<=931946)	Genotype from Affymetrix	text	21474 83647	Long string containing genotype calling. Each character represent a combination of Allele_A and Allele B, and it has the following format: "1" = AA "2" = BB "3" = AB "4" = 00 "5" = A0 "6" = B0 "X" = 00 0 means allele not found	931946
version (4 rows)						
id	PK	Unique version identifier	serial	10	4	4
genotype_version	CHK(>0)	Version of genotype	integer	10	1 or 2	2
annotation_version	CHK(>0)	Version of annotation	integer	10	1 or 2	2
block_version	CHK(>0)	Version of block	integer	10	1	1
version_date		Date of version	date	13	"YYYY-MM-DD"	
comments			text	21474 83647		
visit (100,157 rows)						
id	PK	Unique visit identifier	integer	10		100157
person_id	FK, UNQ	Person identifier	integer	10		5157
visit	UNQ	Visit number	integer	10	From 0 to 99 (1 = First visit, 99 = Last visit)	31
visit_date		Date of visit	date	13	"YYYY-MM-DD"	
form_id		Study form id	integer	10		200
fasting		Patient fasting indicator	integer	10	0 or 1	
descr		Description	text	21474 83647		

Legend
PK = Primary key constraint
FK = Foreign key constraint
UNQ = Unique constraint
CHQ = Check constraint

ANNEXE V

MODÈLE DE DONNÉES D'ADAM



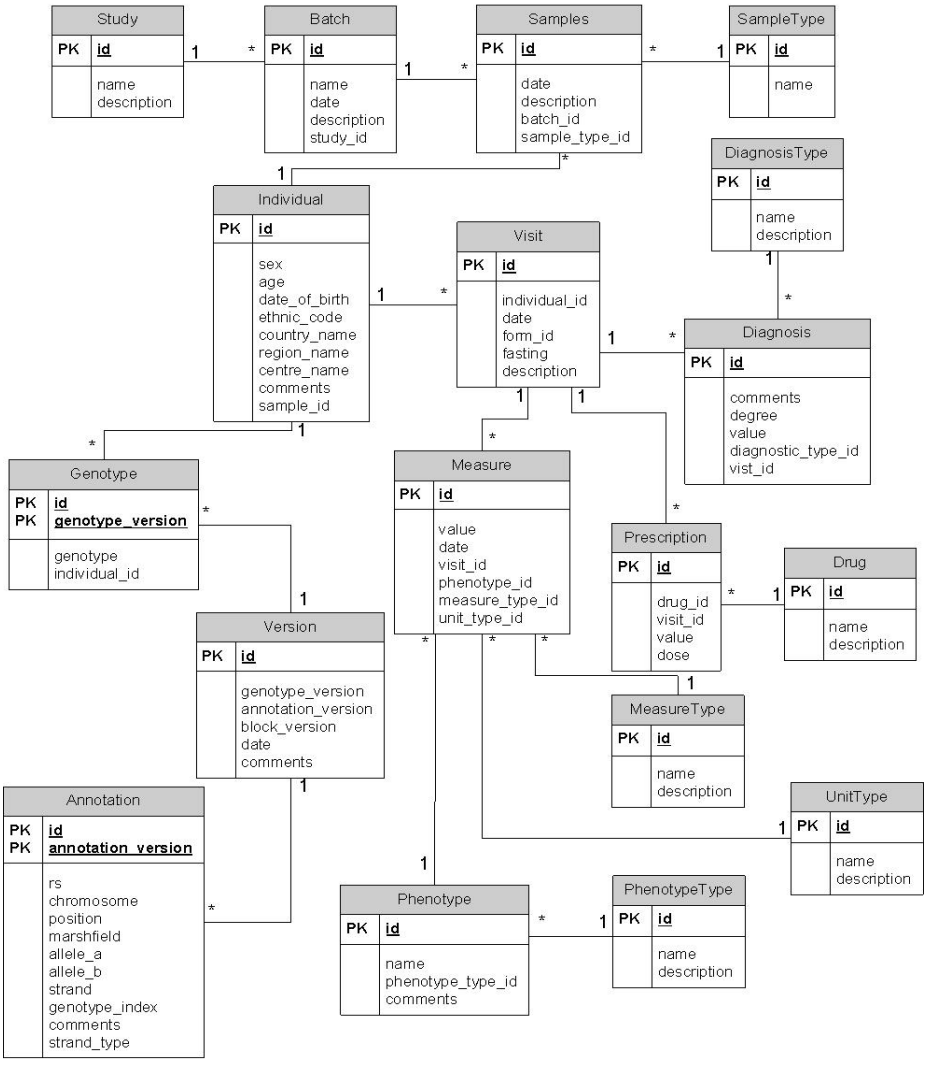
Source : http://genomic.davidlauzon.net/adam/bdg-formats/

ANNEXE VI

LE MODÈLE DE DONNÉES PROPOSÉ

MODÈLE DE DONNÉES
BASE DE DONNÉES PROGNOMIX – ADAM v4

Centre de recherche du CHUM - CRCHUM

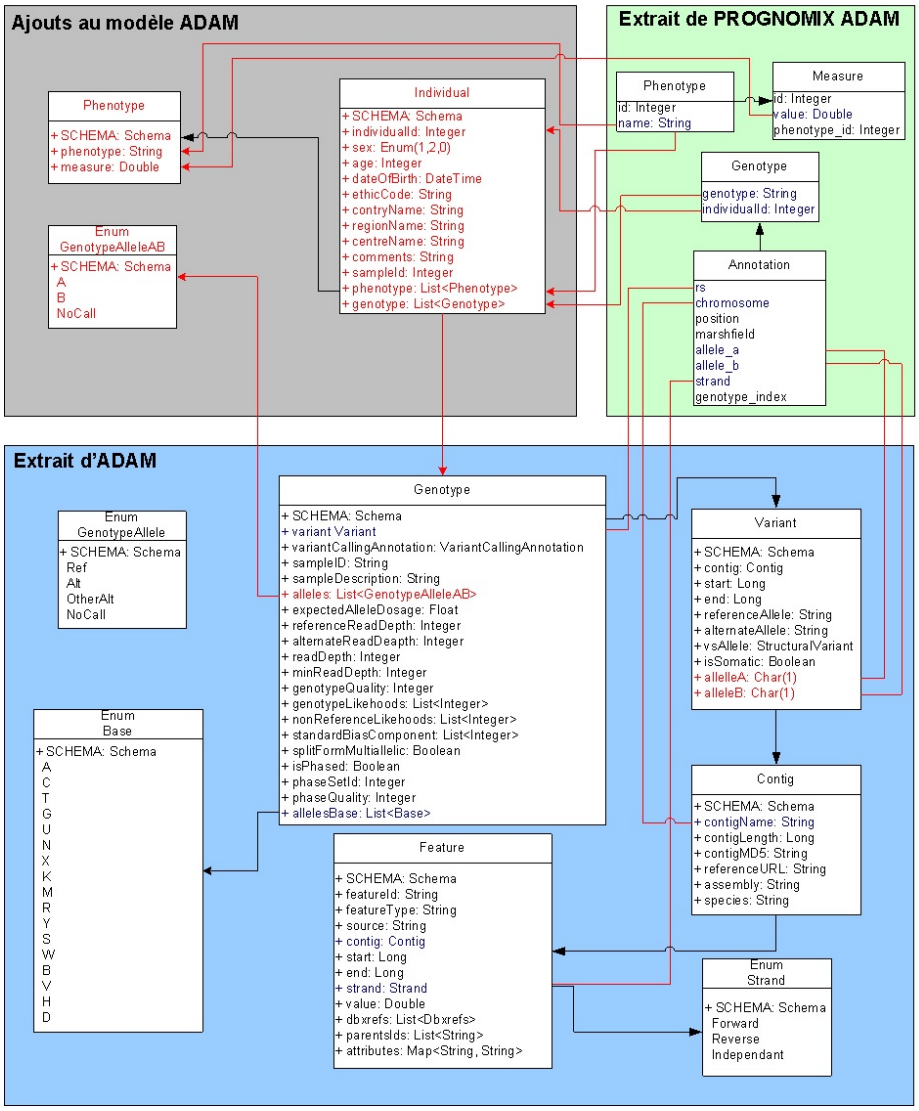


Liste de tables Prognomix ADAM

Nom de la table	Description	Provenance	Fréquence de mise à jour
Study	Liste d'études des patients T2D	Interne	À chaque chargement d'une nouvelle étude
Batch	Liste de processus de chargement de données des études	Interne	À chaque chargement des actualisations des données de l'étude
Samples	Liste des échantillons provenant de l'étude	L'étude	À chaque chargement des actualisations des données de l'étude
SampleType	Les différents types d'échantillon	Interne	Table de référence
Individual	Liste de personnes provenant de l'étude	L'étude	À chaque chargement des actualisations des données de l'étude
Visit	Liste de visites de la personne au médecin	L'étude	À chaque chargement des actualisations des données de l'étude
Diagnosis	Liste de diagnostics associés aux patients par visite	L'étude	À chaque chargement des actualisations des données de l'étude
Measure	Liste de mesures associées aux phénotypes d'une personne par visite	L'étude	À chaque chargement des actualisations des données de l'étude
MeasureType	Liste de types de mesure	Interne	Table de référence
UnitType	Liste de types d'unités	Interne	Table de référence
Prescription	Liste de médicaments prescrits aux patients par visite	L'étude	À chaque chargement des actualisations des données de l'étude
Drug	Liste de médicaments	Interne	Table de référence
Phenotype	Liste de phénotypes	Interne/L'étude	Cette table de référence peut se mettre avec les données des études
PhenotypeType	Liste de types de phénotypes	Interne	Table de référence
Version	Liste de processus de génotypage d'Affymetrix	Interne	À chaque processus de Génotypage
Genotype	Liste de génotype d'une personne	Affymetrix	À chaque processus de Génotypage
Annotation	Liste de marqueurs génétiques (SNP)	Affymetrix	Presque statique (Change selon les nouvelles versions dans la littérature scientifique)
Total de tables: 17			

MAPPING PROGNOMIX – ADAM

Centre de recherche du CHUM - CRCHUM



LISTE DE RÉFÉRENCES BIBLIOGRAPHIQUES

Javier Andreu-Perez, Carmen C. Y. Poon, Robert D. Merrifield, Stephen T. C. Wong, and Guang-Zhong Yang, Big Data for Health, IEEE Journal of biomedical and health informatics, Vol. 19, No. 4, July 2015

Ruchie Bhardwaj. Adhiraaj Sethi, Raghunath Nambiar, 2014, Big Data in Genomics: An Overview, IEEE International Conference on Big Data.

Matt Massie, Frank Nothaft, Christopher Hartl, Christos Kozanitis, André Schumacher, Anthony D. Joseph, David A. Patterson, December 2013, ADAM: Genomics Formats and Processing Patterns for Cloud Scale Computing, Electrical Engineering and Computer Sciences, University of California at Berkeley.

The CHAOS Report, The Standish Group International, 1994, 52 p. [En ligne] https://www.standishgroup.com/sample_research_files/chaos_report_1994.pdf (Consulté le 27 octobre 2015)

The Rise and Fall of the Chaos Report Figures, IEEE Software, 2010. [En ligne] http://www.cs.vu.nl/~x/the_rise_and_fall_of_the_chaos_report_figures.pdf (Consulté le 27 octobre 2015)

Tu Dang VUONG, 2015, Déterminer les exigences d'affaires et d'application pour les fonctionnalités front-end de Snoobe, Montréal, Qc. [En ligne] http://publicationslist.org/data/a.april/ref-491/VUONG_rapport2.1-2.pdf (Consulté le 1 août 2015)

illumina, An Introduction to Next-Generation Sequencing Technology [En ligne] http://www.illumina.com/technology/next-generation-sequencing.html, (Consulté le 23 octobre 2015)

Anna Tikhomirov, Anuar Konkashbaev, Dan L. Nicolae, 2008, On single-array genotype calling algorithms, International Conference on BioMedical Engineering and Informatics.

Unknown
Field Code Changed

Verónica Burriel Coll, May 2012, Design and Development of a Genomic Information System to Manage Breast Cancer Data, Research Challenges in Information Science (RCIS), Sixth International Conference.

BIBLIOGRAPHIE

Leffingwell Dean, Widrig Don, 2003, Managing Software Requirements : A Use Case Approach, 2nd edition, Boston : Addison-Wesley.

International Institute of Business Analysis, 2009. Business Analysis Body of Knowledge, 2e édition. Canada, 286 p.

IEEE, 2004. Software Engineering Body of Knowledge 3e édition. États-Unis: IEEE Computer Society, 204 p.

Stephen Robbins, Mary Coulter, 2014, Management 7e, Pearson