Un système de médiation distribué pour l'e-santé et l'épidémiologie

Sébastien Cipière

N° d'ordre : D.U : 2716
EDSPIC : 764

UNIVERSITÉ BLAISE PASCAL – CLERMONT II

École Doctorale
Sciences Pour l'Ingénieur de Clermont-Ferrand

THÈSE

Présentée par

Sébastien Cipière

Ingénieur en informatique

pour obtenir le grade de

Docteur d'Université

Spécialité : Informatique

Un système de médiation distribué pour l'e-santé et l'épidémiologie

---

Soutenue publiquement le 12/07/2016, devant le jury :

Directeur de thèse :	David R.C. Hill, Pr, UBP, LIMOS UMR 6158
Co-Directrice :	Lydia Maigne, McF HDR, UBP, LPC UMR 6533
Rapporteur :	Ignacio Blanquer Espert, Pr, Universitat Politècnica de València
Rapporteur :	Pascal Staccini, PU-PH, Centre Hospitalier Universitaire de Nice

N° d'ordre : D.U : 2716
EDSPIC : 764

UNIVERSITÉ BLAISE PASCAL – CLERMONT II

École Doctorale
Sciences Pour l'Ingénieur de Clermont-Ferrand

THÈSE

Présentée par

Sébastien Cipière

Ingénieur en informatique

pour obtenir le grade de

Docteur d'Université

Spécialité : Informatique

Un système de médiation distribué pour l'e-santé et l'épidémiologie

---

Soutenue publiquement le 12/07/2016, devant le jury :

Directeur de thèse :	David R.C. Hill, Pr, UBP, LIMOS UMR 6158
Co-Directrice :	Lydia Maigne, McF HDR, UBP, LPC UMR 6533
Rapporteur :	Ignacio Blanquer Espert, Pr, Universitat Politècnica de València
Rapporteur :	Pascal Staccini, PU-PH, Centre Hospitalier Universitaire de Nice

Résumé

Réseau Sentinelle pour l'e-santé et l'épidémiologie en Auvergne

Améliorer le suivi des patients grâce à une infrastructure distribuée

À ce jour, les mesures de risque des cancers ou d'efficacité de leur suivi, se font à partir de recueils de données médicales spécifiques initiés par les médecins épidémiologistes. Ces recueils disposent néanmoins de certaines limites : perte d'information, biais de déclaration, absence de données pour un risque non connu, biais de mesure (par exemple pour les données de nature médico-économiques).

Le partage sécurisé de données médicales entre différentes structures médicales publiques et/ou privées est à ce jour en pleine mutation technologique. Les technologies proposées doivent rendre possible un partage électronique et sécurisé de ces données de manière à les rendre disponible à tout instant dans le cadre de l'observation sanitaire à l'évaluation de prises en charge ou de politiques de santé. Pour répondre à ces besoins, l'infrastructure GINSENG se base sur des informations produites dans le cadre des soins, sans nouvelles modalités de recueil, permettant à la fois une vitesse d'accès à l'information et une exhaustivité accrue. Ce recueil se fait par ailleurs avec de meilleures garanties d'anonymat et un chaînage de l'information médicale pour chaque patient. Une autorisation de la CNIL a été octroyée à l'infrastructure informatique du projet ainsi qu'à son utilisation pour le suivi des cancers en octobre 2013.

Un accès facilité et sécurisé aux données médicales

Depuis le portail web e-ginseng.com, les médecins habilités s'authentifient grâce à leur Carte de Professionnel de Santé (CPS). Chaque patient, dont les données médicales sont réparties dans les établissements de santé, est identifié avec son accord, par les attributs suivants : nom, prénom, année et mois de naissance ainsi que son code postal de résidence avant d'être assigné à un numéro d'identification unique et anonyme. La mise à jour des données médicales de chaque patient est réalisée une fois par semaine ; chaque médecin peut alors consulter toutes les informations médicales relatives à chaque patient par une simple connexion au réseau. Ces informations lui apparaissent sous forme d'une arborescence d'évènements médicaux. Par exemple, un médecin chargé du suivi des patients dans le cadre du dépistage organisé pourra accéder directement depuis le portail web aux informations médicales dont il aura besoin pour établir une fiche médicale exhaustive du parcours du patient pour lequel un cancer aurait été détecté ou bien une suspicion de cancer qui se serait avérée négative suite à plusieurs examens médicaux. Un médecin épidémiologiste peut également réaliser des requêtes

statistiques d'envergure sur les données médicales afin de répondre à des questions d'intérêt en santé publique. Pour aller plus loin, les requêtes épidémiologiques lancées sur les données médicales peuvent être couplées à des informations d'utilité publique recueillies sur d'autres bases de données en accès libre sur internet.

Résultats majeurs du projet

L'infrastructure informatique GINSENG est actuellement déployée pour le suivi des cancers en région Auvergne entre les structures de gestion du dépistage organisé du cancer (SGDO) et le cabinet d'anatomie et cytologie pathologiques (ACP) Sipath-Unilabs. Le recours à un hébergeur de données de santé (HADS), nommé Informatique de sécurité (IDS), est également proposé pour le stockage des informations confidentielles des patients. Cette infrastructure permet actuellement de collecter toutes les informations médicales d'intérêt pour le suivi des cancers et l'évaluation des pratiques médicales. Les équipes de bio-statistiques et de santé publique du CHU de Clermont-Ferrand établissent actuellement les analyses épidémiologiques d'intérêt à partir des données collectées par le réseau.

Abstract

The implementation of a grid network to support large-scale epidemiology analysis (based on distributed medical data sources) and medical data sharing require medical data integration and semantic alignment. In this thesis, we present the GINSENG (Global Initiative for Sentinel eHealth Network on Grid) network that federates existing Electronic Health Records through a rich metamodel (FedEHR), a semantic data model (SemEHR) and distributed query toolkits. A query interface based on the VIP platform, and available through the e-ginseng.com web portal helps medical end-users in the design of epidemiological studies and the retrieval of relevant medical data sets.

Remerciements

Merci à tous ceux qui m'ont permis d'arriver jusqu'ici et qui me permettront de continuer encore longtemps à apprendre.

Merci à vous tous ! À vous lecteurs de ce manuscrit, sans qui il n'a aucune valeur.

À Tous les anonymes dont la contribution favorise l'avancée de la recherche.

À David R.C. Hill mon directeur de thèse qui m'a permis de trouver ma place dans la très grande famille que forment ses étudiants.

À ma co-directrice qui m'a supportée au bureau pendant les trois dernières années.

À ma famille qui a toujours était présente et ne m'a jamais laissé manquer de rien.

À tous mes relecteurs à qui j'ai donné bien du travail.

À tous les partenaires du projet GINSENG qui ont su me donner une vision réaliste d'un projet d'envergure.

À tous les administrateurs de systèmes d'information qui ont eu à subir mes cahiers des charges.

Merci à tous les membres de mes deux laboratoires d'accueil LPC et LIMOS pour leurs soutiens, et particulièrement l'équipe PCSV.

À tous ceux que j'ai pu oublier, mais à qui, s'ils se manifestent, je promets de trouver compensation à hauteur de leurs contributions.

Merci au système éducatif Français et à la France en général qui permet à tout citoyen de s'élever.

TABLE DES TABLEAUX

Tableau 1	Liste des différents registres qualifiés par le CNR en janvier 2013	20
Tableau 2	Liste des fichiers constitutifs de SNOMED 3.5 VF	44
Tableau 3	Liste des chapitres du classeur de diagnostics de la classification SNOMED 3.5 VF	44
Tableau 4	Liste des sections du chapitre Maladies de la peau et des tissus sous-cutanés de la nomenclature SNOMED 3.5 VF	45
Tableau 5	Exemples de codage issus de la nomenclature SNOMED 3.5 VF	45
Tableau 6	Tableau des caractéristiques des examens cytologiques de la norme d'échange ACP	47
Tableau 7	Classification ACR et suites recommandées	59
Tableau 8	Économies réalisées grâce à l'utilisation de logiciels Open Source	63
Tableau 9	Fiche technique des machines testées durant la phase de prototypage version 2012	92
Tableau 10	Fiche technique des machines équipant les sites Sipath-Unilabs/ARDOC/ABIDEC	93
Tableau 11	Fiche technique d'un serveur rackable GINSENG version 2015	93
Tableau 12	Liste des paquets installés sur le serveur de virtualisation	94
Tableau 13	Liste des paquets installés sur la machine « serveur de données »	94
Tableau 14	Liste des composants Perl nécessaires aux scripts GINSENG	105
Tableau 15	Table de redirection des ports d'un site GINSENG	109
Tableau 16	Liste et signification des variables de la base métier Sipath-Unilabs	114
Tableau 17	Liste des variables (avec leur description) InVS contenues dans l'export du DOCS	118
Tableau 18	Liste des variables (avec leur description) InVS contenues dans l'export du DOCU	120
Tableau 19	Tableau de correspondance entre les caractères particuliers et leurs substitutions	127
Tableau 20	État d'avancement des différents sous projets constituants le réseau GINSENG	134
Tableau 21	Nombre d'enregistrements par an dans la base Sipath-Unilabs - avril 2015	138
Tableau 22	Nombre d'enregistrements par an dans la base DOCCR ABIDEC - juillet 2015	140
Tableau 23	Nombre d'enregistrements par an dans la base DOCS ABIDEC - juillet 2015	141
Tableau 24	Nombre de patients par tranches d'âge dans l'annuaire ABIDEC- juillet 2015	142
Tableau 25	Nombre d'enregistrements par an dans la base DOCCR ARDOC - juillet 2015	143
Tableau 26	Nombre d'enregistrements par an dans la base DOCS ARDOC - juillet 2015	144
Tableau 27	Nombre de patients par tranches d'âge dans l'annuaire ARDOC - juillet 2015	145
Tableau 28	Nombre d'enregistrements par an dans la base DOCU ABIDEC-ARDOC - juillet 2015	146
Tableau 29	Nombre de patients par tranches d'âge dans l'annuaire ABIDEC-ARDOC - juillet 2015	147
Tableau 30	Temps d'exécution du script d'import en fonction du remplissage de la base	155

TABLE DES ILLUSTRATIONS

Figure 1 Les 3 grandes entités de la santé en France avec le parcours de soins 8
Figure 2 Frise chronologique de la mise en œuvre de l'e-santé en France 9
Figure 3 CPS version 3 distribuée par l'ASIP santé 12
Figure 4 Boutons de la charte graphique DMP, source ASIP Santé 14
Figure 5 Distribution sur le territoire français des registres qualifiés par le CNR 22
Figure 6 Nouveaux cas de cancer en Loire-Atlantique (moyenne annuelle 2007-2009) 23
Figure 7 Carte mise à disposition par le réseau Sentinelles 26
Figure 8 Proportion de ménages européens ayant accès à Internet en 2010 (source : Eurostat) 28
Figure 9 Exemple d'accès aux données mises à disposition par le CDC 31
Figure 10 Diagramme de séquence UML d'une consultation médicale 38
Figure 11 Diagramme de séquence UML d'une étude de santé publique 39
Figure 12 Alignement direct des champs de deux sites 40
Figure 13 Alignement sur un standard ou une norme 40
Figure 14 Alignement des champs de données de 3 sites en direct 40
Figure 15 Alignement des données en utilisant une norme commune 41
Figure 16 Représentation de la structure d'un code ADICAP 46
Figure 17 Représentation d'un code ADICAP codant pour un Frottis Cervical 46
Figure 18 Déroulement d'une invitation pour le dépistage du cancer colorectal 57
Figure 19 Mammographie seins Gauche et Droit (source : SFR) 60
Figure 20 Logos des partenaires privés du projet GINSENG 64
Figure 21 Logos des partenaires institutionnels et associatifs 65
Figure 22 Visuel de l'accord patient recto RSCA 69
Figure 23 Accord patient RSCA 70
Figure 24 Accord patient RSPA 71
Figure 25 Maquette de l'interface d'aide à la création de requête pour GINSENG 83
Figure 26 Exemple d'une vue de l'interface Zeus pour l'ARDOC 85
Figure 27 Architecture informatique de GINSENG – vue globale simplifiée RSCA 90
Figure 28 Architecture informatique GINSENG – ABIDEC/ARDOC- vue composants réseaux 91
Figure 29 Volume de fichiers patients (en attente/traité/produit) par site intégré dans GINSENG 92
Figure 30 Représentation d'un serveur de virtualisation GINSENG 97
Figure 31 Capture d'écran d'un tableau de bord Icinga2 102
Figure 32 Capture d'écran d'une interface de gestion de ticket Flyspray 104
Figure 33 Illustration de la structure du stockage de la documentation 104

Figure 34 Vue du réseau « interne » GINSENG, exemple donné pour le site Sipath-Unilabs 108
Figure 35 Exemple réduit à un patient fictif d'un export hebdomadaire Sipath-Unilabs 111
Figure 36 Représentation graphique de l'exemple de la Figure 35 (XML viewer) 112
Figure 37 Représentation d'une partie des bases de données MySQL de l'infrastructure GINSENG 115
Figure 38 Requête SQL de la création de la table stockant les données originales du DOCCR 121
Figure 39 Processus d'identification d'un patient lors de l'importation des données 123
Figure 40 Exemple de résultats pour la validation des algorithmes d'identification 125
Figure 41 Représentation graphique la mesure de distance entre les patients 126
Figure 42 Noyau de l'ontologie SemEHR 130
Figure 43 Répartition du genre des patients dans la base Sipath-Unilabs avril 2015 136
Figure 44 Pyramide des âges des patients dans la base Sipath-Unilabs avril 2015 137
Figure 45 Répartition homme/femme pour le DOCCR – ABIDEC – juillet 2015 140
Figure 46 Pyramide des âges des patients dans l'annuaire ABIDEC - juillet 2015 141
Figure 47 Répartition homme/femme pour le DOCCR – ARDOC – juillet 2015 143
Figure 48 Répartition homme/femme pour l'annuaire – ARDOC – juillet 2015 144
Figure 49 Pyramide des âges des patientes dans l'annuaire ARDOC - juillet 2015 145
Figure 50 Répartition des lieux de résidence des patientes référencées dans l'annuaire ABIDEC 146
Figure 51 Pyramide des âges des patientes dans l'annuaire ABIDEC-ARDOC - juillet 2015 147
Figure 52 Comparaison de l'exécution de Jaro-Winkler sur 2 champs (nom, prénom) 148
Figure 53 Diagramme de séquence de l'importeur de patients 150
Figure 54 Représentation d'un seuil de décision plancher unique 151
Figure 55 Représentation de l'intervalle de doute légitime, pour l'identification 151
Figure 56 Répartition statistique du rapprochement des patients en fonction du score produit par l'algorithme d'identification 152
Figure 57 Graphique du temps moyen de l'import d'un patient dans la base GINSENG 156
Figure 58 Architecture de la base de connaissances distribuée 158
Figure 59 Exemple de requête SPARQL interrogeant la base sémantique de GINSENG 159
Figure 60 Page d'accueil du portail (Liferay) du projet GINSENG 160
Figure 61 Diagramme des cas d'utilisations de l'application WEB 161
Figure 62 Interface Liferay de requêtage de GINSENG 162
Figure 63 Capture de l'outil de sélection dans Zeus (v201603.1) 163
Figure 64 Capture d'une base de tests présentée dans Zeus (v201602.1) pour l'affichage des résultats issus de GINSENG avec la possibilité de récupérer les Compte-Rendus ACP 165
Figure 65 Vue Google Earth de 3 variables (simulées) attachées aux communes correspondantes 166
Figure 66 Représentation des mesures de radon moyenne en Auvergne par code postal 167

Figure 67 Représentation de la dangerosité pour l'homme du niveau de radon dans l'air 168
Figure 68 Cartographie de l'Auvergne avec représentation de l'aléa radon (BRGM) et mesure du taux moyen de radon dans l'air en Bq/m^3 170
Figure 69 Représentation du nombre de codage ADICAP poumon (en valeur absolue) depuis 1990 et cartographie de mesure du radon dans l'air en Bq/m^3, par Code Postal 170
Figure 70 Représentation du nombre de codage ADICAP poumon depuis 1990, par Code Postal 171
Figure 71 Représentation du nombre de codage ADICAP poumon et sein (normalisé), depuis 1990 et cartographie de mesure du radon dans l'air en Bq/m^3 172
Figure 72 Mise en évidente de la non corrélation entre le taux moyen de radon d'un code postal et l'incidence des codages ADICAP du poumon normalisé par le rapport nombre de codage du poumon sur nombre de codage colon 172
Figure 73 Représentation des rapports les plus élevés du nombre de codage ADICAP poumon sur nombre de codage colon depuis 1990, cartographie de la mesure moyenne du radon dans l'air en Bq/m3 en surimpression de représentation de l'aléa radon par Code Postal en Auvergne 173
Figure 74 Nuage de point représentant le rapport du nombre de codage ADICAP du poumon sur les codages colon ; en fonction de taux de radon moyen dans l'air en B/m^3 avec la boîte de Tukey de chaque série et les droites de tendance 174
Figure 75 Nuage de point représentant le nombre de codage ADICAP du poumon en fonction de la population des agglomérations pour les départements de l'Allier et du Puy de dôme 174

Table des matières

TABLE DES MATIÈRES

RÉSUMÉ II

ABSTRACT IV

REMERCIEMENTS V

INTRODUCTION GÉNÉRALE 1

CHAPITRE I

- E-SANTÉ ET ACCÈS À L'INFORMATION MÉDICALE – VERS UNE ARCHITECTURE DISTRIBUÉE ? 6

INTRODUCTION 6

1.1 **AVANT-PROPOS/CONTEXTE** 7

1.2 **L'E-SANTÉ** 8

1.2.1 L'E-SANTÉ EN FRANCE 9

1.2.2 L'E-SANTÉ EN EUROPE 27

1.2.3 L'E-SANTÉ DANS LE MONDE 30

1.3 **LES DONNÉES DE SANTÉ ET LA CONFIDENTIALITÉ EN FRANCE** 32

1.3.1 L'ASPECT LÉGISLATIF 32

1.3.2 LE SUIVI DU PATIENT, L'IDENTITOVIGILANCE 36

1.4 **LES DONNÉES DE SANTÉ ET LEUR GESTION** 37

1.4.1 DONNÉES ET MÉTADONNÉES 37

1.4.2 FLUX DE TRAVAUX DE L'UTILISATION DES DONNÉES DE SANTÉ 38

1.4.3 L'INTEROPÉRABILITÉ DES DONNÉES DE SANTÉ 39

1.5 **NORMES ET STANDARDS POUR LES DONNÉES DE SANTÉ** 41

1.5.1 HEALTH LEVEL SEVEN (HL7) 42

1.5.2 SYSTEMATIZED NOMENCLATURE OF MEDICINE (SNOMED) 43

1.5.3 INTERNATIONAL CLASSIFICATION OF DISEASES (ICD) 45

1.5.4 L'ASSOCIATION POUR LE DÉVELOPPEMENT DE L'INFORMATIQUE EN CYTOLOGIE ET EN ANATOMIE PATHOLOGIQUE (ADICAP) 46

1.5.5 LA NORME D'ÉCHANGE D'ANATOMO-CYTO-PATHOLOGIE (ACP) 47

1.5.6 AUDIPOG 48

1.5.7 INTEGRATING THE HEALTHCARE ENTERPRISE (IHE) 48

CONCLUSION 49

CHAPITRE II

– LA GESTION DES DONNÉES MÉDICALES DISTRIBUÉES 52

INTRODUCTION 52

2.1 **LE PROJET GINSENG 53**

2.1.1 GENÈSE 53

2.1.2 CONTEXTE 55

2.1.3 LES OBJECTIFS DU PROJET 61

2.1.4 LES ACTEURS DU PROJET 64

2.1.5 LA LÉGISLATION 72

2.1.6 LES DONNÉES MÉDICALES 73

2.2 **RECUEIL DES SPÉCIFICATIONS DU PROJET GINSENG 75**

2.2.1 UN RÉSEAU NON « INVASIF » 75

2.2.2 UN RÉSEAU DISTRIBUÉ ET SÉCURISÉ 75

2.2.3 STANDARDISATION DES BASES DE DONNÉES 78

2.2.4 IDENTIFICATION DES PATIENTS 79

2.2.5 AUTHENTIFICATION 79

2.2.6 INTERFACE 80

CONCLUSION 86

CHAPITRE III

– ARCHITECTURE TECHNIQUE ET COORDINATION DU PROJET GINSENG 88

INTRODUCTION 88

3.1 **UNE INFRASTRUCTURE DISTRIBUÉE ET SÉCURISÉE D'ACCÈS AUX INFORMATIONS MÉDICALES 89**

3.1.1 LES SERVEURS 89

3.1.2 RÔLE DES DIFFÉRENTES MACHINES 97

3.1.3 LES SERVICES 104

3.1.4 RÉSEAU SÉCURISÉ 106

3.2 **LA GESTION DES BASES DE DONNÉES MÉDICALES 110**

3.2.1 STRUCTURE DES BASES DE DONNÉES MÉDICALES 110

3.2.2 LES ALGORITHMES D'IDENTIFICATION DES PATIENTS 124

3.2.3 LES REQUÊTES SUR LES BASES DE DONNÉES 129

3.2.4 L'ACCÈS AUX DONNÉES MÉDICALES 131

CONCLUSION 133

CHAPITRE IV

– VALIDATION, TESTS DE PERFORMANCE ET RÉSULTATS **135**

INTRODUCTION **135**

4.1 **L'IDENTIFICATION DES PATIENTS ET IMPORTATION DES DONNÉES MÉDICALES** **136**

4.1.1 DESCRIPTIF DU CONTENU DES BASES DE DONNÉES 136

4.1.2 L'UTILISATION DE JARO-WINKLER 148

4.1.3 IMPORTATION DES DONNÉES MÉDICALES 149

4.1.4 BENCHMARK 152

4.1.5 E-SANTÉ, TRANSFERT DE FICHIERS 156

4.2 **REQUÊTES SUR LES BASES DE DONNÉES ET PRÉSENTATION DES RÉSULTATS** **157**

4.2.1 L'UTILISATION DES SPARQL ENDPOINT 157

4.2.2 L'UTILISATION DU LOGICIEL 'R' 159

4.2.3 INTERFACE WEB 159

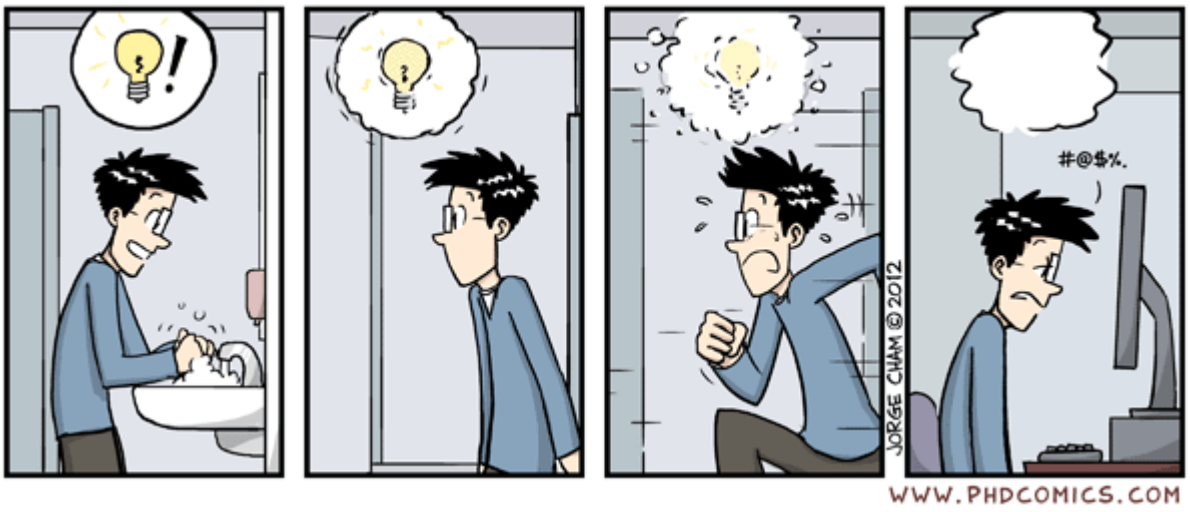
4.2.4 REPRÉSENTATIONS VISUELLES 163

4.3 **ÉTUDE ÉPIDÉMIOLOGIQUE SUR L'IMPACT DU RADON SUR LES CANCERS DU POUMON EN AUVERGNE** **168**

CONCLUSION **176**

CONCLUSION **177**

PERSPECTIVES **181**


#@$%.
JORGE CHAM © 2012
WWW.PHDCOMICS.COM

Introduction générale

> “
> Presque tous les hommes meurent de leurs remèdes, et non pas de leurs maladies.
> - Jean-Baptiste Poquelin, *le malade imaginaire*

Un système informatique d’échange de données médicales

En 2016, les systèmes informatiques se retrouvent intégrés de plus en plus dans notre système de santé. Chaque consultation de patient produit des informations médicales, personnelles et confidentielles (pathologie, identité, etc.) qui doivent être idéalement partagées avec l’accord du patient entre les différents acteurs de sa santé. Si les informations médicales des patients intéressent les médecins pour suivre l’évolution d’une maladie ou d’une pathologie, les médecins épidémiologistes qui surveillent l’état de santé des populations et l’évolution des pratiques médicales trouvent un intérêt à utiliser des dossiers médicaux électroniques si ceux-ci sont bien renseignés par les praticiens qui en ont la charge.

Actuellement, la pratique des professionnels de santé est dans une phase de transition concernant la sauvegarde des données médicales produites. L’évolution va d’une sauvegarde sur papier à une sauvegarde électronique vers des serveurs médicaux centralisés ou même encore dématérialisés sur une infrastructure de type cloud.

C’est pour répondre à des besoins d’échanges de données médicales efficaces et rapides ainsi qu’à une analyse exhaustive des données médicales des patients pour les besoins de la santé publique qu’en 2011, a été lancé le projet ANR GINSENG (Global Initiative for Sentinel E-health Network on Grid). Ce projet visait à démontrer l’efficacité d’une infrastructure complètement distribuée pour l’e-santé et l’épidémiologie en région Auvergne. Les applications directes et premières de cette infrastructure concernent le suivi des cancers du dépistage organisé (cancers du sein, du côlon et du col de l’utérus) et le suivi des actes de périnatalité dans la région. Ces deux champs applicatifs sont formalisés par deux réseaux collaboratifs : le Réseau Sentinelle Cancer Auvergne (RSCA) et le Réseau Santé Périnatalité Auvergne (RSPA). Ces réseaux collaboratifs regroupent les acteurs de santé, les structures de gestions du dépistage

organisé du cancer (SGDO) et les acteurs de santé publique de la région Auvergne. Un premier cahier des charges, synthétisant les besoins des partenaires et les premières solutions technologiques à mettre en œuvre, a été établi entre 2008 et 2011 par (DeVlieger 2011). Nous avons prolongé et approfondi cette étude au cours du projet GINSENG pour réaliser une preuve de concept fonctionnelle.

Le développement d'une politique de prévention des cancers, rend en effet nécessaire de disposer d'outils adaptés à sa mise en œuvre ainsi que son évaluation. Dans le cadre du suivi des cancers en région Auvergne, trois types de cancers (cancer du sein, du côlon et du col utérin) bénéficient de mesures de dépistage organisé dont la mise en œuvre est rendue possible par les associations de dépistage organisé ARDOC pour les départements du Puy-de-Dôme, du Cantal et de la Haute-Loire et l'ABIDEC pour le département de l'Allier.

À ce jour, les mesures de risque des cancers ou d'efficacité de leur suivi, se font à partir de recueils de données médicales spécifiques initiés par les médecins épidémiologistes. Ces recueils disposent néanmoins de certaines limites : perte d'information, biais de déclaration, absence de données pour un risque non connu, biais de mesure (par exemple pour les données de nature médico-économiques).

Depuis 2011, plusieurs partenaires publics et privés se sont associés pour créer un réseau sentinelle informatique expérimental pour l'e-santé et l'épidémiologie des cancers en Auvergne. Le projet GINSENG, financé par l'ANR et l'ARS réunit le CNRS, l'Université Blaise Pascal, l'Université d'Auvergne, le CHU Gabriel Montpied associés aux sociétés Gnúbila et Mnemotix, aux associations de dépistage des cancers ARDOC et ABIDEC et au cabinet d'anatomie et cytologie pathologiques (ACP) Sipath-Unilabs pour mettre en œuvre des solutions informatiques et d'analyses sécurisées des données médicales pour améliorer le suivi des cancers.

Le partage sécurisé de données médicales entre différentes structures médicales publiques et/ou privées est à ce jour en pleine mutation technologique. Les technologies proposées doivent rendre possible un partage électronique et sécurisé de ces données de manière à les rendre disponible à tout instant dans le cadre de l'observation sanitaire à l'évaluation de prises en charge ou de politiques de santé. Pour répondre à ces besoins, l'infrastructure GINSENG se base sur des informations produites dans le cadre des soins, sans nouvelles modalités de recueil, permettant à la fois une vitesse d'accès à l'information et une exhaustivité accrue. Ce recueil se fait par ailleurs avec de meilleures garanties d'anonymat et un chaînage de l'information médicale pour chaque patient. Grâce à une interconnexion sécurisée de bases de données

médicales existantes et distribuées géographiquement, nous pouvons alors espérer des avantages majeurs :

• un coût d'infrastructure réduit,

• les données médicales n'ont pas besoin d'être centralisées mais simplement être maintenues accessibles aux utilisateurs autorisés,

• une interopérabilité des données médicales qui reste facilitée,

• des alarmes sanitaires efficaces nécessitant moins d'étapes intermédiaires de déclaration et de traitement des données.

En plus de la mise en œuvre d'une infrastructure distribuée complète pour l'e-santé et l'épidémiologie, nous avons également travaillé à l'intégration de requêtes sur les bases de données en utilisant des ontologies, ce travail a été effectué en partenariat avec la société Mnemotix en charge de la valorisation du logiciel Corese/KGRAM[1] et du langage SPARQL. Le but était de pouvoir extraire des données statistiques depuis l'interrogation des bases de données médicales distribuées ; ces données pouvant être complétées par un croisement avec des bases de données de type DBPedia en accès libre.

Les contraintes légales en France sont importantes, particulièrement lorsque l'on traite des données relatives à la santé des personnes. De plus, la gestion de bases de données comportant des informations nominatives est extrêmement encadrée. Nous verrons l'importance des structures administratives, comme la CNIL, qui garantit le respect des informations à caractère confidentiel de chaque patient. Les établissements ont aussi des obligations de conservation sur de nombreuses années des documents qu'ils détiennent. Une autorisation de la CNIL a été octroyée à l'infrastructure informatique du projet ainsi qu'à son utilisation pour le suivi des cancers en octobre 2013.

Aujourd'hui notre prototype est en cours de validation dans des conditions de production. Nous présenterons en détails les étapes qui furent nécessaires pour mener à bien cette réalisation.

Notre objectif est de fournir de nouveaux outils, plus pratiques, plus rapides, plus sûrs, plus fiables, sans surcoût.

[1] http://wimmics.inria.fr/corese - date d'accès avril 2016

Organisation de la thèse

Après avoir dressé l'état de l'art des problématiques liées à l'e-santé que sont l'identification du patient au sein d'un système de santé hétérogène, l'interopérabilité des données, l'authentification des praticiens, pour accéder de façon sécurisée aux données des patients et l'architecture des systèmes d'informations médicaux. Nous détaillerons ensuite les différents aspects limitant de ces problématiques avant d'envisager comment débloquer ces verrous ou les contourner. Nous présenterons par la suite la mise en pratique de notre approche théorique au travers des réalisations mise en œuvre dans le réseau du projet GINSENG.

Ce manuscrit aborde dans le premier chapitre la littérature internationale relative à l'identification de patients, la sécurité des systèmes d'information et la recherche d'information utilisant des ontologies. Nous présentons la place de l'e-Santé en France, en Europe et dans le monde. Nous nous intéressons à la législation qui régit la manipulation informatique des données de santé. Avant de regarder plus en détails la façon dont ces données circulent et les différents normes et standards qui permettent de les traiter.

Nous développons le périmètre de nos actions à l'intérieur du chapitre 2. Après avoir introduit le projet GINSENG, nous dressons un recueil de spécifications qui met en exergue les points majeurs de notre recherche. Nous abordons l'authentification des usagers, l'identification des patients, la structuration des données médicales, l'analyse de ces informations avec des outils sémantiques, le stockage dans des bases de données, la distribution de l'architecture informatique pour relier des sites distants, et la sécurisation de l'ensemble du système d'information.

Le troisième chapitre propose des éléments de solution. L'architecture que nous proposons est présentée en détails, ce qui permet de comprendre comment les différents sites sont reliés. Puis nous nous intéressons au système de bases de données. Les différentes étapes du workflow sont exposées et explicitées.

Le quatrième et dernier chapitre expose les résultats de nos implémentations à travers les mesures des temps d'importation des données patients dans nos bases de données en conditions de production. Les interfaces utilisateurs qui permettent de requêter les bases et les représentations graphiques des données qui en découlent sont dévoilées à la fin de ce manuscrit. Une recherche étudiant un lien possible entre cancer du poumon et présence de radon dans l'environnement, clos ce chapitre. C'est la première étude qui croise les données produites par

GINSENG avec des informations environnementales collectées pour une recherche totalement indépendante de notre réseau.

Enfin, un bilan des résultats est présenté et des perspectives de recherche sont proposées dans la conclusion générale.

Chapitre I

- E-santé et accès à l'information médicale – vers une architecture distribuée ?

Introduction

L'e-santé vise à mettre à disposition toute information médicale relative au suivi des patients dans le respect de leurs droits au moyen des technologies de l'information. Cette mise à disposition de l'information médicale fait l'objet de contraintes légales destinées à encadrer les pratiques et protéger la vie privée du patient. Les quantités de données médicales sont, à ce jour, très conséquentes ; chaque examen médical peut être la source de dossiers virtuels allant de quelques kilo-octets pour les comptes rendus d'examens à plusieurs méga-octets lorsque des données d'imagerie sont produites. Une infrastructure médicale se doit donc d'être robuste, pour permettre de traiter des quantités importantes, mais aussi, très interactive, pour permettre à tout personnel de santé d'accéder en temps réel à l'information qui le concerne. Ces enjeux d'envergure positionnent le patient au centre du système d'information ; ses données médicales sont quant à elles, la plupart du temps, très réparties géographiquement et, à ce jour, difficilement interconnectables. Deux positions peuvent donc être adoptées concernant la gestion de cette information : décider d'héberger la totalité des données de chaque patient chez un seul et même hébergeur ou bien faire le choix de laisser les données réparties géographiquement dans les structures médicales et mettre en place un réseau fiable permettant l'accès aux données réparties dans des bases de données distribuées dédiées à ce réseau. Chacune de ces solutions à ses avantages et doit répondre à des contraintes légales retardant parfois leur mise en œuvre. Dans le contexte de la recherche publique, le choix a été fait depuis quelques années de mettre en œuvre des outils informatiques permettant l'interopérabilité de

données réparties géographiquement, les technologies de grille informatique puis de cloud, bien démocratisées maintenant, répondent aux besoins de stockage et d'interactivité des utilisateurs pour un coût de fonctionnement raisonnable. Ces technologies ont été mises en œuvre pour le projet ANR GINSENG afin de créer une infrastructure de grille pour l'e-santé et l'épidémiologie en région Auvergne dont le déploiement est opéré par l'association Réseau Sentinelle Cancer Auvergne (RSCA). Ce projet, à l'origine destiné à améliorer le suivi des cancers gérés dans le cadre du dépistage organisé (cancer du sein, du côlon et du col de l'utérus), vise également à permettre une analyse exhaustive fiable des données médicales pour les besoins en santé publique.

La première partie de ce chapitre vise à expliquer les enjeux et les contraintes liées à l'e-santé dans le panorama législatif national et européen(Clavier 2007; Dumont 2010). Dans une deuxième partie, les enjeux liés à la préservation de la confidentialité des données de santé seront abordés (Vulliet-Tavernier 2002) avant de s'intéresser à leur interopérabilité(Couvreur 2010). Une dernière partie traite des normes et standards utilisés pour la gestion des jeux de données de santé (interop'santé 2015), en particulier ceux afférents au projet GINSENG (ADICAP 2009; France-Périnat 2007).

1.1 Avant-propos/contexte

Pour bien comprendre l'apport de l'e-santé il est important de percevoir le contexte dans lequel se positionne notre recherche. En effet, le panorama médical français est à la fois hétérogène de par son degré d'équipement informatique, il est aussi très réglementé. L'effet direct est que nous évoluons dans un cadre dont le périmètre est défini avec des contraintes fortes. Avant de nous focaliser sur l'e-santé intéressons-nous brièvement à la santé en France. Nous pouvons schématiquement la cloisonner en 3 grandes entités, comme représenté par la Figure 1. Historiquement, et dans la pratique, ces 3 grandes entités sont souvent mal interconnectées. De plus, pour respecter le secret médical du patient, toutes ces informations médicales ne peuvent être communiquées entre les partenaires médicaux sans un accord explicite de celui-ci (bien qu'elles puissent être utiles à un meilleur accompagnement de la personne). Les informations possédées par les uns peuvent être importantes et utiles pour les autres. Cependant toutes les informations ne sont pas forcément utiles et une synthèse est, souvent, plus pratique à transmettre. Se pose alors le problème du contenu de la synthèse et de son mode de communication.

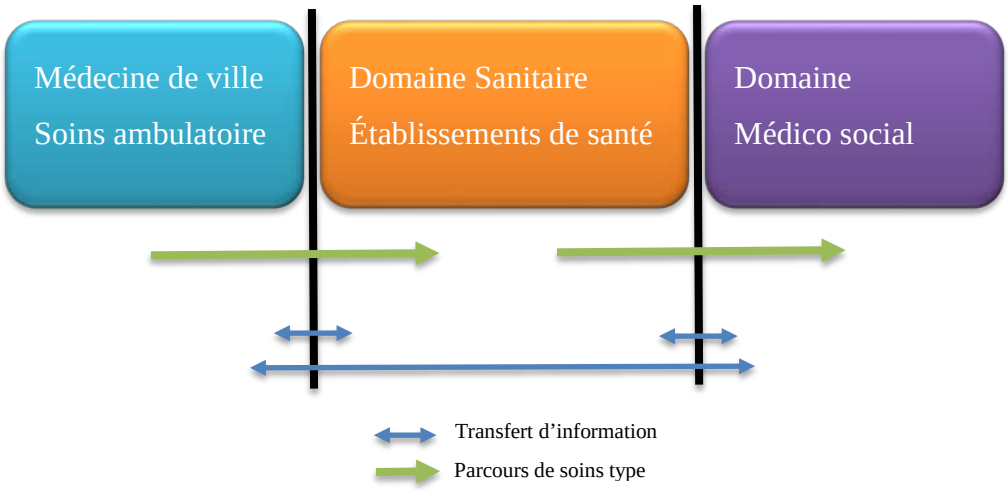


Figure 1 Les 3 grandes entités de la santé en France avec le parcours de soins et le transfert d'informations médicales

De plus, il est important de structurer les échanges et le stockage de l'information, si toutes les synthèses d'un patient se retrouvent dans le même dossier simplement ordonnées chronologiquement, il sera difficile de retrouver l'information pertinente. Nous nous proposons ici d'étudier ces problématiques qui sont rencontrées quotidiennement, au travers du regard de l'informaticien, qui les regroupera sous la terminologie « e-santé ». L'identification et l'authentification de la personne, qu'elle soit utilisateur ou patient, sont essentielles pour que le système d'information puisse savoir précisément à qui se réfère chaque donnée et à qui elle est destinée.

1.2 L'e-santé

Dans tout le manuscrit, le terme « e-santé » sera utilisé pour parler de « santé électronique », de « santé en ligne » ou « télésanté » (« *eHealth* » en anglais). Ce terme recouvre les différents instruments qui s'appuient sur les Nouvelles Technologies de l'Information et de la Communication (NTIC) pour faciliter et améliorer la prévention, le diagnostic, le traitement et le suivi médical des patients, ainsi que la gestion de la santé et du mode de vie. Dans cette partie, l'e-santé fait référence aux échanges d'informations entre les différentes structures médicales, nous ne considérons cependant pas les enjeux de la m-Santé (ou Santé Mobile) qui sera dans un futur proche, un vecteur important des données de santé.

Le législateur s'est intéressé à de nombreuses reprises à l'e-santé notamment au travers du décret n°2010-1229[2] du 19 octobre 2010 dans lequel il définit ce qu'est la télémédecine. Il

[2] http://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000022932449 - date d'accès octobre 2015

décrit la télémédecine comme un conteneur qui regroupe les actes médicaux réalisés à distance, au moyen d'un dispositif utilisant les TICs (Technologies de l'Information et de la Communications). Ainsi la télémédecine regroupe la téléconsultation, la téléexpertise, la télésurveillance, la téléassistance médicale ainsi que la coordination des services d'urgences.

1.2.1 L'e-santé en France

L'e-santé, outil au service de la santé, est sous la tutelle du Ministère des Affaires sociales et de la Santé.

Nous suivrons un parcours chronologique des avancées, des besoins et des limitations qui entourent l'e-santé dans le but de dresser une image de son état actuel, pour éclairer les réflexions sur notre démarche. Depuis 1985, date de la première utilisation d'un micro-ordinateur en médecine, la dématérialisation des données et la mise en place de systèmes d'information, n'ont cessé de croître sur le territoire français (cf. Figure 2), actuellement avec le DMP Dossier Médical Partagé ou Personnel, l'état entend généraliser l'e-santé.

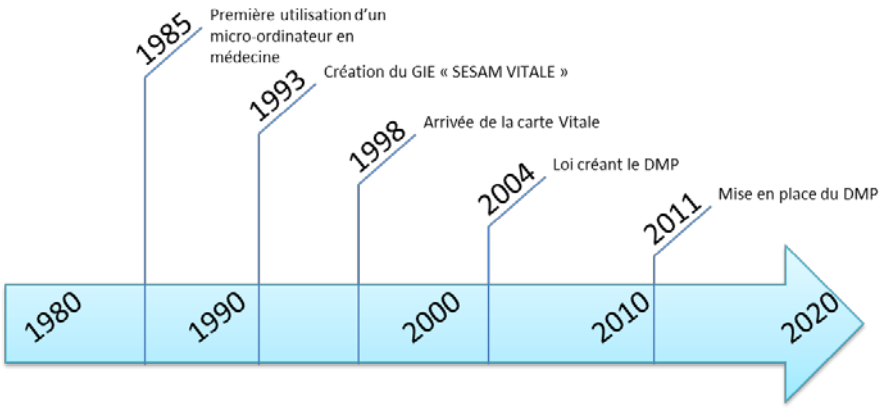


Figure 2 Frise chronologique de la mise en œuvre de l'e-santé en France

Dans un souci de simplification des démarches administratives et pour réduire les frais de traitement des dossiers, le système français a évolué dès 1990. Le Groupement d'Intérêt Économique (GIE) « SESAM Vitale » est créé en 1993, avec pour objectif la dématérialisation des feuilles de soins vers les Feuilles de Soins Électroniques (FSE).

La carte vitale

En 1998, la carte (SESAM) vitale est lancée. La carte vitale est actuellement l'un des piliers du système de santé français. Tout citoyen français de plus de 16 ans, s'il en fait la demande, dispose de sa propre carte vitale qui est une carte au format des cartes bancaires. Cette carte est mise à disposition gratuitement. Il existe différentes versions de carte vitale, en fonction de

leurs dates d'émission, actuellement les cartes de 3ème génération sont disponibles. Les cartes vitales contiennent de nombreuses informations sur l'assuré comme :

1. Des données visibles sur la carte
    1. Un numéro d'émetteur
        - Un numéro propre à la carte
    2. La date d'émission de la carte
    3. Des données d'identification du titulaire
        - Numéro d'inscription au répertoire national d'identification des personnes physiques (N° INSEE)
        - Nom de famille
        - Prénom
        - Photographie en couleur (non présente initialement)
        - Signe d'identification de la carte en relief
2. Des données inscrites dans le composant électronique de la carte :
    1. Les données visibles inscrite sur la carte (cf. 1)
    2. Les données relatives au régime de base d'assurance maladie
        - Droits aux Affection de Longue Durée ALD
        - Droits à la Couverture Maladie Universelle Complémentaire CMUC
        - Droits à l'exonération du ticket modérateur
    3. Les données relatives au rattachement des enfants
    4. Les données relatives au choix du médecin traitant
    5. Les données relatives à la protection complémentaire d'assurance maladie
    6. Les données relatives aux accidents du travail et maladies professionnelles
    7. Les données relatives à l'accès aux soins dans l'Union européenne
    8. Les coordonnées d'une personne à prévenir en cas d'urgence
    9. La mention indiquant que son titulaire a eu connaissance des dispositions de la réglementation sur le don d'organe
    10. Des données permettant :
        - Les fonctions de signature électronique
        - La protection de la carte
        - L'authentification de la carte d'assurance maladie

Les cartes vitales sont le support d'un identifiant incontournable pour le système de santé français, appelé Numéro de Sécurité Social (NSS) ou code INSEE (Institut National de la Statistique et des Études Économiques) ; son acronyme officiel est le NIR qui signifie Numéro d'Inscription au Répertoire National d'Identification des Personnes Physiques (RNIPP). Cependant, l'utilisation de ce numéro, qui semble le plus approprié pour identifier chaque patient français, est extrêmement réglementée et n'est pas accordée dans la pratique aux applications de recherche visant à améliorer le système de santé français. Deux autres solutions

sont envisagées pour fournir un Identifiant National de Santé (INS). L'un, aléatoire, l'INS-A constitué d'une suite de douze chiffres, le second, calculé, l'INS-C, selon la méthode décrite dans (ASIP-Santé 2009) qui s'appuie sur le NIR, les prénoms et la date de naissance pour le recombiner en un nouvel identifiant résultant d'un haché en SHA256. Techniquement, les cartes vitales sont des cartes à puces. Comme le précise (C Quantin et al. 2004), la carte vitale est une carte d'assuré social et non une carte individuelle ; de plus, elle ne permet pas l'identification des parents. Pour accéder aux données contenues sur la carte, un lecteur est nécessaire. Le principe est celui des cartes bancaires, mais pour plus de praticité, aucun code n'est nécessaire pour accéder aux informations contenues sur la carte. Ces lecteurs sont présents chez les professionnels de santé (médecins, pharmacien, …) ainsi que dans les caisses d'assurance maladie. Il existe différents type de lecteur avec ou sans contact, ils doivent respecter le recueil des spécifications du GIE SESAM Vitale (ASIP-Santé 2013). Certains disposent de deux fentes pour lire simultanément la carte vitale et la carte CPS (voir paragraphe suivant). En fonction de l'utilisation désirée du lecteur de carte certains composants logiciels sont nécessaires, ils sont fournis par l'ASIP-santé pour Windows, Linux et MacOS.

La Carte de Professionnel de Santé (CPS)

Comme la carte vitale, la CPS est une carte à puce. Tout professionnel de santé peut disposer d'une carte pour s'identifier sur les SI, cette carte porte tout naturellement le nom de Carte de Professionnel de Santé (CPS) (cf. Figure 3). Les cartes CPS comportent deux certificats numériques permettant pour le premier, la signature des documents émis par son possesseur et pour le second, le chiffrement des données. Ce sont deux certificats X509 (rfc2459[3] et rfc5280[4]). Les CPS permettent une authentification forte de leur porteur (identité, profession, spécialisation, lieu d'exercice) et garantissent ainsi un accès sécurisé aux services informatiques que ce soit en local ou à travers une application web (ASIP-Santé 2013).

[3] http://tools.ietf.org/html/rfc2459 - date d'accès octobre 2015
[4] http://tools.ietf.org/html/rfc5280 - date d'accès octobre 2015

Figure 3 CPS version 3 distribuée par l'ASIP santé

L'utilisation de la carte nécessite des composants matériels (un lecteur) et une partie logicielle (des pilotes et outils de management). Son utilisation implique que l'utilisateur se situe au sein d'une infrastructure à clef publique (ou PKI pour *Public Key Infrastructure*), basé sur un chiffrement par clefs asymétriques[5], dont le détail des interactions est accessible dans la thèse de (DeVlieger 2011) ainsi que dans le rapport de stage de (Passerat-Palmbach 2009). C'est l'Agence des Systèmes d'Information Partagés de santé (ASIP santé[6]) qui est en charge de la promotion des CPS[7]. Il en existe différents types en fonction de la profession et du rôle de son propriétaire, nous pouvons citer par exemple la Carte de Personnel d'Établissement (CPE) ou la Carte de Professionnel de santé en Formation (CPF). En juillet 2013, 844 594 cartes de la famille CPS (CPS, CPE, etc.) sont en circulation dont 506 450 CPS valides et 330 504 cartes CPE. Il est intéressant de noter que, 161 607 cartes sont distribuées au sein des établissements hospitaliers français dont 88 301 cartes CPS. Pour mémoire l'INSEE[8] estime à 1 100 663 le nombre de professionnels de santé en 2012 sur le territoire français, ce qui correspond à un taux d'équipement des professionnels de santé de 77%. La loi française rend obligatoire depuis la parution du "décret confidentialité" le 15 mai 2007 en application de la loi Kouchner du 4 mars 2002, l'utilisation des CPS comme moyen d'identification des professionnels de santé à l'intérieur d'un SI de santé.

[5] Les clefs asymétriques fonctionnent par paire. Une clef publique à communiquer largement, la seconde privée à conserver précieusement – Elles sont capables de se reconnaître l'une, l'autre ; et de déchiffrer les messages encodés par la clef réciproque. Les fonctions mathématiques qu'elles utilisent ne permettent cependant pas de recréer la clef privée à partir de la clef publique. Ce qui garantit authentification et confidentialité.

[6] http://esante.gouv.fr/asip-sante - date d'accès octobre 2015

[7] Mission définie par l'article 2 de sa convention constitutive approuvée par arrêté ministériel du 28 novembre 2009 modifié

[8] http://www.insee.fr/fr/themes/tableau.asp?ref_id=nattef06103 - date d'accès octobre 2015

En pratique, la majorité des échanges de documents médicaux entre professionnels de santé se réalise sans utilisation de la CPS. En effet, il n'est pas rare que deux praticiens se consultent par messagerie non sécurisée (type SMS ou MMS) au sujet d'un dossier patient en utilisant leurs téléphones personnels. Les e-mails personnels ou professionnels non cryptés sont aussi des vecteurs courants de transmission de l'information médicale en 2014. Les cartes CPS sont également souvent utilisées parfois abusivement par un groupe de personnels au nom d'un seul et même utilisateur.

Le DMP

La loi créant le Dossier Médical Partagé (DMP) (Couvreur 2010) (Bourret 2010) est votée en 2004. Le Groupement d'Intérêt Public (GIP) DMP puis l'ASIP Santé permettent en 2011 l'avènement de cet outil qui, entre temps, a été renommé en Dossier Médical Personnel. La genèse difficile[9] de ce projet est détaillée par (Manaouil 2009). Le DMP est accessible sur le site www.dmp.gouv.fr, c'est un dossier médical informatisé et centralisé, accessible via Internet. Le 12 juillet 2013 nous pouvions dénombrer exactement 348 625 dossiers DMP créés en France, 420 000 en janvier 2014 et 480 000 en septembre (alors que les objectifs fixés par l'État étaient de 5 000 000 de dossiers dès 2010). Ces chiffres sont encourageants car ils sont en constante progression. Cependant l'adhésion au projet reste faible, car fin 2014 nous ne sommes qu'à 10% des objectifs de 2010. Le coût de fonctionnement de l'hébergement du DMP pour 2014 est de 7 à 10 millions d'euros (selon les sources) et un rapport interne du Conseil national de la qualité et de la coordination des soins, chargé d'arbitrer les financements destinés à l'amélioration de la médecine de ville[10], chiffre à 500 millions d'euros le développement de cet outil depuis 2004. Le parti pris du DMP est la centralisation de l'information. C'est-à-dire que toutes les données recueillies sont regroupées et stockées chez un hébergeur de données de santé. Au niveau national le consortium associant « La Poste[11] » et « Atos Origin[12] » a été retenu par l'ASIP santé pour prendre en charge l'hébergement du DMP. Cependant, les solutions techniques retenues pour mettre en œuvre le DMP varient d'une région à l'autre. En

[9] http://philippe.ameline.free.fr/phis/ - date d'accès octobre 2015

[10] http://www.lemonde.fr/sante/article/2014/01/04/dossier-medical-partage-un-cout-excessif-pour-un-succes-mitige_4342961_1651302.htm - date d'accès octobre 2015

[11] http://legroupe.laposte.fr/date d'accès octobre 2015

[12] http://atos.net/fr-fr/accueil/nous-sommes/newsroom/communique-de-presse/2010/pr-2010_0215_04.html - date d'accès octobre 2015

Auvergne par exemple, les entreprises qui en ont la charge sont Orange[13], Covalia[14] et Almerys[15]. La coordination de la mise en place est assurée par le GCS (Groupement de Coopération Sanitaire) SIMPA.

Le DMP comporte en théorie huit rubriques :

- Les données médicales générales
- Les traitements et les soins (dont les prescriptions)
- Les comptes rendus (hospitalisation, etc.)
- L'imagerie médicale (scanners, IRM, etc.)
- Les résultats des laboratoires d'analyses et de biologie médicale
- La prévention (vaccination, rappels, etc.)
- Les certificats médicaux et déclarations d'aptitude
- L'espace personnel d'expression du patient.

Le DMP est « le » dossier stratégique de l'e-santé en France. Ses objectifs sont ambitieux et visent à unifier les transmissions de données à l'intérieur du territoire. En effet, pour être mené à bien, le DMP devra répondre de manière globale à une utilisation pour l'instant très hétérogène (en matière d'outils et de données) de l'e-santé. Le patient peut lui-même uploader des documents sur son espace DMP. Le type de document que le patient a la liberté d'ajouter à son dossier DMP est libre. Il est important de noter que le DMP n'a pour vocation que le stockage et la mise à disposition des éléments de synthèses, qui résument un événement médical du patient sans le retranscrire en détails.

L'ASIP santé met à disposition une charte graphique qui permet aux éditeurs tiers d'harmoniser leurs interfaces graphiques, avec les boutons types DMP.

Figure 4 Boutons de la charte graphique DMP, source ASIP Santé

[13] http://www.orange.fr/ - date d'accès octobre 2015
[14] http://www.covalia.com/accueil/ - date d'accès octobre 2015
[15] https://www.almerys.com/fr/ - date d'accès octobre 2015

Les Hébergeurs Agréés de Données de Santé

Le concept des Hébergeurs Agréés de Données de Santé (HADS, aussi nommé HDS), qui se sont fédérés en une Association Française des Hébergeurs Agréés de Données de Santé[16] (AFHADS), est né du constat que l'informatisation croissante du système de santé nécessite toujours plus de moyens informatiques. Les centres médicaux n'ont pas pour vocation première d'héberger les serveurs informatiques nécessaires au stockage de leurs données de santé. L'option de l'externalisation est alors un compromis acceptable, car en échange du coût financier du forfait, le prestataire se retrouve en charge de l'infrastructure informatique. Pour aider les décideurs du système de santé dans le choix crucial que représente la sélection d'un HADS, le gouvernement français a donc décidé de mettre en place une procédure d'agrément. La procédure d'accréditation devait évoluer en faveur d'une procédure de certification dans le courant de l'année 2015 (avril), mais actuellement la procédure d'agrément est toujours en vigueur.

Les HADS peuvent être des centres de ressources informatiques, qui sont des lieux sécurisés dédiés au stockage et à la gestion des données. Les méso-centres disposent de toute l'infrastructure technique nécessaire pour assurer une utilisation pérenne des ressources informatiques. Les accès sont protégés par des moyens qui permettent une authentification de la personne comme un badge d'accès. L'alimentation électrique est elle aussi protégée par des onduleurs, et des groupes électrogènes qui doivent permettre à minima un arrêt des systèmes dans des conditions normales en cas de coupure de courant. Outre le support matériel proposé par les HADS, leur offre de service peut proposer des solutions « clef en main » qui facilite les intégrations logicielles. Nous pouvons citer par exemple des services d'authentification des personnels de santé au moyen de leurs CPS. Ce type de services permet de déléguer des processus standard et commun qui évitent de recoder des solutions déjà existantes et nécessaires au bon fonctionnement des applications modernes d'e-santé. Les solutions proposées par les HADS se rapprochent donc, dans un certain sens, des services proposés par les prestataires de « cloud » (Mell and Grance 2011) qui mettent à disposition des machines (IaaS) (*Infrastructure as a Service*) (Bhardwaj, Jain, and Jain 2010) jusqu'à des solutions clefs en mains (SaaS) (*Software as a Service*) (Cusumano 2010), tout en garantissant l'accréditation de la structure comme HADS. Le cahier des charges pour le déploiement de solutions au sein d'un HADS est particulièrement rigoureux et restrictif, en raison de l'obligation légale qui leur incombe. L'une des valeurs ajoutées essentielle des HADS est leur capacité de conseils et

[16] http://www.afhads.fr/ - date d'accès octobre 2015

d'accompagnement qui permet de garantir à leurs utilisateurs une réalisation de la solution en adéquation avec les attentes législatives. Et, effectivement, même si l'acronyme HADS contient le terme hébergeur, leur activité de conseils est plus essentielle que leur capacité de stockage informatique qui pourrait, elle, être sous-traitée.

Le gouvernement accrédite des institutions publiques ou privées pour l'hébergement de données de santé. En décembre 2013, 55 décisions d'agrément[17] pour la mise en place des HADS ont été rendues, par le ministre en charge de la santé[18]. Elles sont désormais 84 en août 2015. Un livre blanc[19] (Lehalle and Sérézat 2014) a été publié par l'AFHADS disponible sur le site DSIH[20] dans lequel sont précisés les modalités et les objectifs des HADS ainsi que leurs devoirs. Le chapitre 4 de ce livre blanc permet d'obtenir des pistes pour aider au choix d'un HADS. En 2015 une réflexion est menée autour du projet de loi[21] sur la modernisation du système de santé qui vise à remplacer l'agrément ministériel des HADS par une certification.

La messagerie sécurisée de santé

Si les outils se généralisent très rapidement, les bonnes pratiques sont plus lentes à se mettre en place. Julien Dufrenne nous présente dans son manuscrit de thèse (Dufrenne 2011) un état de l'art de la Messagerie Sécurisée de Santé (MSS) utilisée par les médecins généralistes en 2011. Il constate qu'actuellement, des dossiers médicaux transitent en clair sur Internet par des messageries non sécurisées. Ainsi, dans (Dufrenne 2011), l'auteur nous expose clairement que malgré une obligation légale, de nombreuses fois répétées depuis 1996, et un accompagnement par des avantages financiers, en 2010, 15% des médecins généralistes n'utilisaient toujours pas la télétransmission des FSE. Le site officiel de la MSS[22] française, en version béta en septembre 2014, est accessible avec une carte CPS. Il est réservé aux professionnels de santé qui peuvent par ce biais activer leur compte de messagerie sécurisée. Le site web reste cependant peu convivial, et en cas de problème avec l'installation du lecteur de CPS, l'utilisateur ne dispose d'aucune information, ni aide. La messagerie n'est compatible qu'avec Thunderbird et permet d'échanger entre boîtes respectant la norme MSSanté contenue

[17] dans le cadre de la procédure d'agrément des hébergeurs de données de santé à caractère personnel précisée par le décret du 4 janvier 2006

[18] http://esante.gouv.fr/services/referentiels/securite/hebergeurs-agrees - date d'accès octobre 2015

[19] http://www.wobook.com/WBXb7Pf4BI4I - date d'accès octobre 2015

[20] http://www.dsih.fr/ - date d'accès octobre 2015

[21] http://www.legifrance.gouv.fr/affichLoiPreparation.do?idDocument=JORFDOLE000029589477&type=contenu&id=2&typeLoi=proj&legislature=14 - date d'accès octobre 2015

[22] https://cms.mssante.fr/ - date d'accès octobre 2015

dans la spécification (ASIP-Santé 2014b) . Un Webmail (une page internet permettant de consulter ses mails) est aussi disponible et les logiciels pour smartphone sont en cours de réalisation. Le but étant de proposer aux professionnels de santé un moyen simple et efficace pour partager les informations sur leurs patients de façon sécurisée. La MSSanté est régie par l'ASIP-Santé qui édite le Dossier des Spécifications Fonctionnelles et Techniques (DSFT) (ASIP-Santé 2014a). Le DSFT détaille précisément le fonctionnement de la messagerie ainsi que les interactions avec l'annuaire national MSSanté, il fournit de nombreuses recommandations mais n'incite pas à l'utilisation d'outils en particuliers. Toute structure répondant aux attentes de ce document est éligible pour devenir Opérateur de messagerie en signant un contrat avec l'ASIP Santé. En Auvergne, la mise en œuvre de la MSSanté par le GCS SIMPA s'appuie sur la solution de serveur open source Zimbra pour fournir un courriel standard auquel ont été ajoutés les mécanismes de sécurité suffisants pour protéger la confidentialité des patients, comme la mise en place d'un proxy de messagerie de MSSanté, en plus d'une politique de sécurité qui respecte les règles de bon sens. L'un des objectifs principaux de la MSSanté est de créer un espace de confiance au sein duquel les données de messagerie peuvent transiter en toute sécurité. De plus chacun pourra clairement être identifié à l'aide d'une adresse « prénom.nom@profession.mssante.fr ».

Une solution actuellement très répandue chez les professionnels de santé se nomme « Apicrypt »[23]. S'il s'agit bien d'une messagerie médicale sécurisée (propriétaire), il est important de préciser qu'elle ne répond pas au cahier des charges de la MSS française. Ainsi, Apicrypt n'intègre pas le périmètre de confiance de la MSS bien qu'il regroupait 50 000 professionnels de santé en 2014 et que (Dufrenne 2011) estimait qu'en 2008 Apicrypt rassemblait 75% des utilisateurs de messagerie sécurisée. Il n'est pas évident pour un professionnel de santé d'identifier que les solutions utilisant Apicrypt ne sont pas les solutions officielles. En effet, Apicem qui développe Apicrypt est une association de médecins fondée en 1996, et, à ce titre, est promue par les Unions Régionales des Professionnels de Santé (URPS) notamment PACA et Nord Pas de Calais comme « la » messagerie de santé à utiliser. Les projets DMP et MSSanté sont des projets récents qui n'obtiennent pas l'adhérence des professionnels de santé initialement prévue. Pour essayer d'améliorer la vitesse de leurs déploiements Madame la ministre Marisol Touraine prévoit de transférer la responsabilité de ces deux projets de l'ASIP Santé vers la CNAM à l'article 25 du projet de loi de modernisation

[23] http://www.apicrypt.org/ - date d'accès octobre 2015

de notre système de santé[24] qui a été examiné au sénat en juillet 2015. Rendant obligatoire pour les établissements de santé l'adoption d'une solution MSS compatible avant le 31 décembre 2015. L'ANSSI, a décerné à l'APICEM le Label France CYBER SECURITY pour sa solution APICRYPT 2 le 16 octobre 2015. En parallèle la demande d'agrément pour devenir HADS a été déposée la réponse devrait intervenir en novembre 2015, pour pouvoir intégrer le périmètre de confiance de la MSS. L'ASIP envisage des solutions d'interopérabilité pour faire converger les solutions vers une MSS unifiée.

Le Département d'Informatique Médical (DIM) hospitalier

À l'hôpital, pour accompagner ces changements numériques, un service a été créé, le Département d'Informatique Médical (DIM) qui organise le recueil, la circulation et le traitement des données médicales des patients hospitalisés dans l'établissement, sous l'autorité d'un médecin responsable. Son rôle, en matière de centralisation des informations médicales, est prévu par l'article L.710.5 du Code de la Santé Publique. Le DIM gère un programme national, le PMSI (Programme de Médicalisation des Systèmes d'Information), qui permet de connaître pour chaque service, l'activité réalisée et définit ainsi les budgets en fonction de cette activité dans le cadre de la Tarification À l'Activité (T2A) des établissements de santé. Le PMSI a pour but de suivre en particulier les courts séjours MCO (Médecine, Chirurgie Obstétrique), les Soins de Suite et de Réadaptation (SSR) et la psychiatrie. Le DIM n'est pas qu'un service support au PMSI. Il a un rôle prépondérant dans l'Hospitalisation À Domicile (HAD), la gestion des archives médicales, ainsi que dans la transmission des dossiers médicaux aux anciens patients ou à leur médecin traitant, le cas échéant.

De plus, l'hôpital doit déclarer son activité auprès de l'Agence Technique de l'Information sur l'Hospitalisation (ATIH), l'information saisie est extraite, rendue anonyme et envoyée sous la forme d'un fichier de RSA (Résumé Standardisé Anonymisé) sur la plateforme internet de l'ATIH grâce à divers utilitaires fournis par le Système d'information Hospitalier (SIH) et l'ATIH. Cet envoi numérique est validé par l'établissement, la région et l'ATIH. L'information ainsi déclarée et validée, a plusieurs utilités :

- Le « casemix » qui décrit l'éventail des cas traités de l'hôpital (ou de la clinique) sous forme de Groupes Homogènes de Malades (GHM), ou de Catégorie Majeures Diagnostics (CMD).

[24] http://www.legifrance.gouv.fr/affichLoiPreparation.do?idDocument=JORFDOLE000029589477&type=contenu&id=2&typeLoi=proj&legislature=14 - date d'accès octobre 2015

- La recette de l'hôpital est facilement calculable car chaque GHM ayant une valeur en euros publiée chaque année par Arrêté.
- La justification d'une activité relevant d'une autorisation particulière (ex : réanimation, soins palliatifs, néonatalogie, …)
- L'information des tutelles régionales pour permettre une vision territoriale de l'offre de soins.

Ainsi, le dossier informatisé dispose de nombreux avantages tels que: l'optimisation des soins, l'accès à une traçabilité médico-légale, l'accès à une base de description des séjours hospitaliers en code Classification Internationale des Maladies (version 10) (CIM-10 ou ICD-10) (WHO 1992) et Classification Commune des Actes Médicaux (CCAM) (Baude 2007). Le DIM est souvent personnifié, dans ce cas, on fait référence au responsable de l'information médicale et de l'évaluation. In fine, c'est lui qui décide si une transmission électronique peut ou non être réalisée. Il statue en s'appuyant sur les textes de loi et son expérience, sa décision est souveraine. Dans certaines structures le DIM peut aussi occuper le rôle de CIL. Le DIM est de facto l'une des personnes incontournables dans la réalisation des projets d'e-santé.

Les registres

De manière à réaliser des études épidémiologiques d'envergure sur des données considérées comme exhaustives, les registres ont été mis en place. La définition que fournit l'Institut de Veille Sanitaire (InVS), est la suivante :

"Un registre est défini comme un recueil continu et exhaustif de données nominatives intéressant un ou plusieurs événements de santé dans une population géographiquement définie, à des fins de recherche et de santé publique, par une équipe ayant les compétences appropriées".[25]

La France, au travers de l'arrêté du 6 novembre 1995 se dote d'un Comité National des Registres (CNR). Le CNR a pour mission d'évaluer la qualité des registres de morbidité, de donner un avis sur l'opportunité des registres existants ou en création et de proposer une politique des registres s'appuyant sur les besoins en matière de santé publique et de recherche épidémiologique. Le CNR dénombre en janvier 2013 les registres listés dans le Tableau 1.

[25] source : (arrêté du 6 novembre 1995 modifié relatif au Comité national des registres).

Tableau 1 Liste des différents registres qualifiés par le CNR en janvier 2013

Intitulé du registre	Responsable scientifique	Date de création	Lieu d'implantation	Qualification en cours
REGISTRES GENERAUX DU CANCER				
Cancers généraux - Bas-Rhin	Michel Velten	1974	Faculté de médecine – Strasbourg	2013-2016
Cancers généraux – Calvados	Anne-Valérie Guizard	1978	CLCC – Caen	2013-2016
Cancers généraux - Doubs et du Territoire de Belfort	Anne Sophie Woronoff	1977	CHU de Besançon	2010-2013
Cancers généraux – Gironde	Gaëlle Coureau	2004	ISPED- Bordeaux	2010-2013
Cancers généraux – Guadeloupe	Jacqueline Deloumeaux	2008	CHU de Pointe à Pitre	2012-2014
Cancers généraux – Guyane	Angéla Fior	2005	URML Cayenne	2011-2013
Cancers généraux - Haut-Rhin	Antoine Buemi	1988	CH de Mulhouse	2011-2014
Cancers généraux – Hérault	Jean-Pierre Daures	1983	CLCC Montpellier	2010-2013
Cancers généraux – Isère	Marc Colonna	1977	CHU de Grenoble	2010-2013
Cancers généraux - Lille et région	Karine Ligier	2005	Groupement régional de promotion de la santé	2012-2015
Cancers généraux – Région Limousin	Nathalie Léone	1998	CHU de Limoges	2013-2015
Cancers généraux - Pays de la Loire (Loire-Atlantique et Vendée)	Florence Molinié	1999	CHU de Nantes	2012-2015
Cancers généraux – Manche	Simona Bara	1994	CH de Cherbourg	2009-2012
Cancers généraux – Nouvelle Calédonie	Sylvie Laumond	1977	DASS Nouvelle Calédonie	2013-2015
Cancers généraux – Région Poitou Charente	Pierre Ingrand	2007	Faculté de médecine et de pharmacie de Poitiers	2013-2015
Cancers généraux - Somme	Olivier Ganry	1982	CHU d'Amiens	2013-2016
Cancers généraux - Tarn	Pascale Grosclaude	1982	CHS Pierre Jamet – Albi	2010-2013
REGISTRES SPECIALISES DU CANCER				
Cancers digestifs - Bourgogne	Anne-Marie Bouvier	1976	Faculté de médecine de Dijon	2013-2016
Cancers digestifs - Calvados	Guy Launoy	1978	CHU de Caen	2010-2013
Cancers digestifs - Finistère	Jean-Baptiste Nousbaum	1984	CHRU de Brest	2013-2016
Hémopathies malignes - Basse-Normandie	Xavier Troussard	2002	CHU de Caen	2010-2013
Hémopathies malignes - Côte-d'Or	Marc Maynadie	1980	Faculté de médecine de Dijon	2013-2016
Hémopathies malignes - Gironde	Alain Monnereau	2002	CLCC – Bordeaux	2013-2016
Registre national des hémopathies malignes de l'enfant	Jacqueline Clavel	1995	Hôpital Paul Brousse- Paris	2011-2014
Registre multicentrique à vocation national des Mésothéliomes pleuraux (MESONAT)	Françoise Galateau-Salle	2006	CHU de Caen	2013-2016
Registre Rhône Alpins des cancers thyroïdiens	Geneviève Sassolas	2000	Centre de recherche en cancérologie de Lyon	2013-2015
Cancers du sein et gynécologique de Côte-d'Or	Patrick Arveux	1982	CLCC Dijon	2012-2015

Tumeurs primitives du système nerveux en Gironde	Isabelle Baldi	1999	ISPED – Bordeaux	2012-2015
Tumeurs solides de l'enfant - National	Brigitte Lacour	1999	CHU de Nancy	2011-2014
REGISTRES DES PATHOLOGIES NEURO-CARDIO-VASCULAIRES				
Cardiopathies ischémiques - Bas-Rhin	Dominique Arveiler	1984	Faculté de médecine de Strasbourg	2013-2016
Cardiopathies ischémiques - Haute-Garonne	Jean Ferrieres	1984	Faculté de médecine de Toulouse	2010-2013
Cardiopathies ischémiques - Lille	Philippe Amouyel	1985	Institut Pasteur de Lille	2013-2016
Accidents vasculaires cérébraux - Dijon	Maurice Giroud	1985	CHU de Dijon	2011-2014
Accidents vasculaires cérébraux du pays de Brest	Serge Timsit	2008	CHU de Brest	2011-2013
Accidents vasculaires cérébraux - Lille	Philippe Amouyel	2010	Institut Pasteur de Lille	2011-2013
REGISTRES DES MALFORMATIONS CONGENITALES				
Centre des malformations congénitales - Auvergne	Christine Francannet	2005	CHU de Clermont Ferrand	2012-2015
Malformations congénitales d'Alsace	Bérénice Doray	2005	Hôpital de Hautepierre	2011-2014
Registre des malformations congénitales en Bretagne	Florence Rouget	2008	CHU de Rennes	2013-2015
Malformations congénitales de la Réunion	Hanitra Randrianevo	2001	CHR Félix Guyon-Saint Denis de la Réunion	2012-2015
Malformations congénitales - Paris	Babak Khoshnood	1981	Hôpital Saint Vincent de Paul	2013-2016
AUTRES REGISTRES				
Handicaps de l'enfant - Haute-Garonne	Catherine Arnaud	1999	Faculté de médecine de Toulouse	2011-2014
Handicaps de l'enfant - Isère	Christine Cans	1991	CHU de Grenoble	2010-2013
Maladies inflammatoires du tube digestif - Nord et Ouest	Corine Gower-Rousseau	1988	CHRU de Lille	2013-2016
Victimes corporelles d'accidents de la circulation routière – Rhône	Bernard Laumon	1995	IFSTTAR	2010-2013
Registre des hépatites de Côte-d'Or et du Doubs	Anne Minello	1994	Faculté de médecine de Dijon	2012-2015
Registre du Réseau épidémiologie et information en néphrologie (Rein)	Christian Jacquelinet	2002	Agence de Biomedecine	2012-2015
Registre Lorrain de la Sclérose en Plaques	Francis Guillemin	2003	CHU de Nancy	2013-2016

Comme nous le montre le Tableau 1 reprise depuis le site de l'InVS[26], la France ne dispose que de 46 registres, toutes spécialités confondues, dont les 3/5 sont dédiés aux cancers. Et aucun des registres sur le cancer n'est d'envergure nationale.

[26] http://www.invs.sante.fr/Espace-professionnels/Comite-national-des-registres - date d'accès octobre 2015

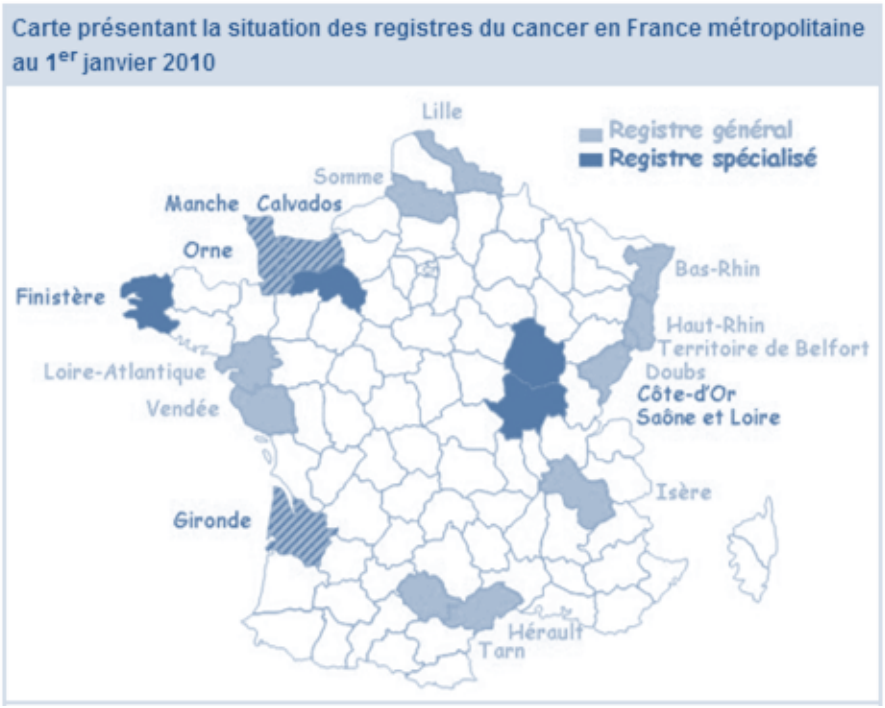


Figure 5 Distribution sur le territoire français des registres qualifiés par le CNR (source : FRANCIM)

En Auvergne, un registre des cancers a déjà été expérimenté. Cependant, le coût de fonctionnement de ce type de structures ainsi que leur mise à jour régulière par le personnel de santé, en effectuant des saisies multiples de l'information médicale, ne permettent pas de rendre pérennes ces solutions. Pourtant, l'intérêt des registres n'est plus à démontrer lorsqu'il s'agit d'obtenir des données statistiques d'intérêt pour la santé publique.

L'un des registres français le plus abouti est le registre des cancers Loire-Atlantique (ARCLA) et Vendée (AVEC) en partenariat avec l'association EPIC-PL (Épidémiologie des Cancers en Pays de la Loire). Depuis 1998, les nouveaux cas de cancers en Loire-Atlantique et Vendée sont recensés sur un bassin de population de près de 2 millions d'individus. La base de données ainsi constituée permet de générer des statistiques sur l'incidence de cancers. Il est alors possible de synthétiser ces résultats (voir Figure 6) pour une communication auprès du public.

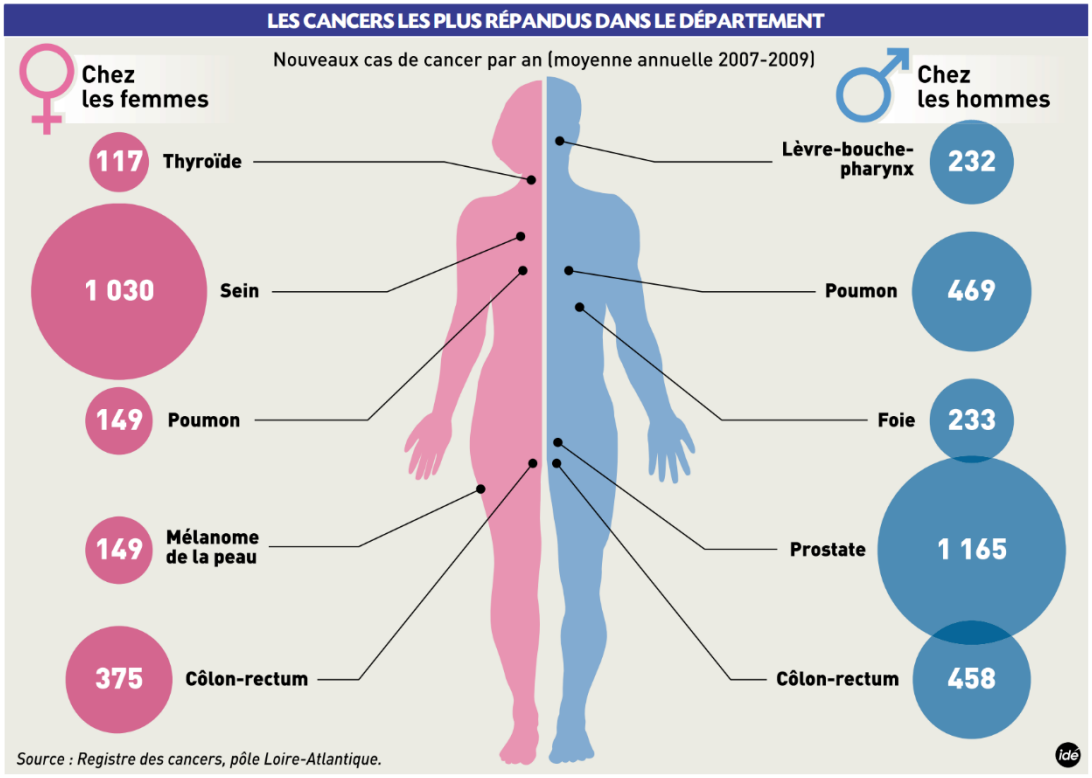


Figure 6 Nouveaux cas de cancer en Loire-Atlantique (moyenne annuelle 2007-2009)

Un registre permet la mise en lumière de différences régionales notamment lors de la comparaison avec une référence nationale. Sur l'exemple du réseau vendéen (FRÉDÉRIC BRENON 2013), le D[r] Moliné observe une augmentation de +70% et + 134% du cancer du foie respectivement chez les femmes et les hommes domiciliés en Loire-Atlantique. Résultats qu'il est ensuite possible, par exemple, de confronter aux consommations d'alcool des différentes régions françaises (Beck et al. 2005) (Lemoine 2013) pour envisager d'éventuelles corrélations. À terme, l'intérêt est de trouver la cause de telle ou telle différence régionale pour appliquer une politique de santé efficace et réduire les causes potentielles de cancer (hygiène de vie, exposition à des matières cancérigènes…). Cependant, les registres sont des structures coûteuses, c'est pour cette raison que le registre auvergnat n'a pu être pérennisé. De plus, les registres rencontrent fréquemment de nombreuses difficultés pour accéder aux données qui leurs sont nécessaires (Doussin and RAULT 2008) nécessitant des apports de données d'autres établissements, avec des délais de transmission, des coûts (financier et temporel) importants. Se pose aussi les problèmes liés à la qualité et l'exhaustivité des données manquantes ou difficiles à récupérer.

Le réseau FRANCIM est une association loi 1901 qui regroupe tous les registres qualifiés (par le CNR) des cancers en France. Ce réseau est partenaire de l'InVS et de l'Institut National

du Cancer (INCa)[27] ainsi que du département de bio statistiques des Hospices Civils de Lyon (HCL). L'objectif étant de fournir une base commune des registres de cancers en France accessible librement à partir du site de l'InVS. La base de données ainsi mise à disposition est donc de l'Open Data (Piwowar and Vision 2013; Molloy 2011), et permet la comparaison des résultats locaux avec d'autres régions et une moyenne nationale. Cependant comme le nombre de registres reste limité à quelques régions (Figure 5) ce système ne permet pas une analyse exhaustive de la population française.

Les Dossiers Communiquant en Cancérologie (DCC)

Le Dossier Communiquant en Cancérologie DCC a pour but le partage et l'échange de données médicales entre professionnels de santé, hospitaliers et libéraux, dans l'objectif d'améliorer la qualité des soins et la continuité de la prise en charge du patient sur le terrain. Le DCC est inclus dans la mesure 34 du Plan cancer 2003-2007 et a été mis en œuvre et développé par les réseaux régionaux de cancérologie (RRC). Le Plan cancer 2009-2013 prévoyait de déployer cet outil en lien avec la relance du dossier médical personnel (DMP) et d'élaborer un cahier des charges spécifique ainsi qu'un programme d'actions (mesure 18.3).

Le DCC devra permettre aux professionnels de santé :

1. D'échanger des données médicales telles que les fiches de réunions de concertation pluridisciplinaire (RCP), les comptes rendus opératoires, les comptes rendus anatomopathologiques via la télé-imagerie, les téléconférences, visioconférences…
2. De gérer les outils et les services nécessaires à l'activité de cancérologie : annuaires des RCP, gestion informatisée des RCP, élaboration du programme personnalisé de soins (PPS), accès aux recommandations de pratique clinique et aux registres des essais cliniques…

L'INCa et l'ASIP santé (Agence des Systèmes d'Information Partagés de santé) se sont associés pour faire évoluer le DCC dans le cadre du DMP. Les régions Alsace, Aquitaine, Lorraine, Pays-de-la-Loire, Picardie, Rhône-Alpes et Midi-Pyrénées ont été retenues pour la phase pilote de mise en œuvre du service DCC.

La mise en route du service DCC du DMP sur l'ensemble du territoire était prévue initialement avant la fin du Plan Cancer 2009-2013. En 2014, le DCC Auvergnat est

[27] http://www.e-cancer.fr/ - date d'accès octobre 2015

actuellement à l'étude par Oncauvergne[28] et le centre anti cancéreux clermontois, le Centre Jean-Perrin (CJP)[29]. Le rapport final du plan cancer 2009-2013 (Plan cancer 2013) mentionne que l'Auvergne avec l'aide de son ARS a pris part à l'expérimentation du dépistage organisé du cancer du col de l'utérus et qu'une Unité fonctionnelle de Gestion des Urgences en Cancérologie (UGUC) a été créée.

Exemples de projet e-santé en France

Le recueil d'informations à distance, de façon rapide, permet de générer des études statistiques difficiles voire impossibles préalablement. Le réseau de vigilance grippenet.fr[30] mis en place par des institutionnels français comme l'Institut National de la Santé Et de la Recherche Médicale (INSERM), l'Université Pierre et Marie Curie et l'InVS, permet d'obtenir la traçabilité de diverses pathologies tout au long de l'hiver, à un niveau national. Cette approche est originale car ici, les statistiques ne sont pas obtenues à partir des bases de données des hôpitaux ou des déclarations des médecins volontaires, mais directement auprès de la population. Pour se faire, les patients s'inscrivent sur le site, renseignent leur fiche signalétique, et déclarent à intervalle régulier leurs symptômes ou leur bonne santé. Cette démarche permet de remplir une base de données avec les informations déclarées par les participants. Après analyse, les données sont communiquées à la population sous forme graphique pour une meilleure lisibilité.

Un autre projet porté par les mêmes équipes françaises est le réseau « Sentinelles »[31], ce dernier permet de tracer l'activité de la varicelle, de la diarrhée aiguë et des syndromes grippaux, les actes suicidaires, les zonas, etc. Le réseau « Sentinelles » a collecté ses premières données en 1984. Ce réseau s'appuie sur 1300 médecins généralistes libéraux (MGL) (soit 2,2% de la totalité des MGL en France métropolitaine), répartis sur le territoire. Les informations sont ensuite rendues accessibles à la population sous forme de cartes (cf. Figure 7), ou des tableaux de données au format CSV (*Comma Separated Values)* qui sont facilement exploitables par les épidémiologistes au travers des logiciels de statistiques 'R[32]' ou 'SAS[33]'.

[28] http://www.oncauvergne.fr/ - date d'accès octobre 2015
[29] http://www.cjp.fr/fr/ - date d'accès octobre 2015
[30] https://www.grippenet.fr - date d'accès octobre 2015
[31] http://websenti.u707.jussieu.fr/sentiweb/ - date d'accès octobre 2015
[32] http://www.r-project.org/ - date d'accès octobre 2015
[33] http://www.sas.com/ - date d'accès octobre 2015

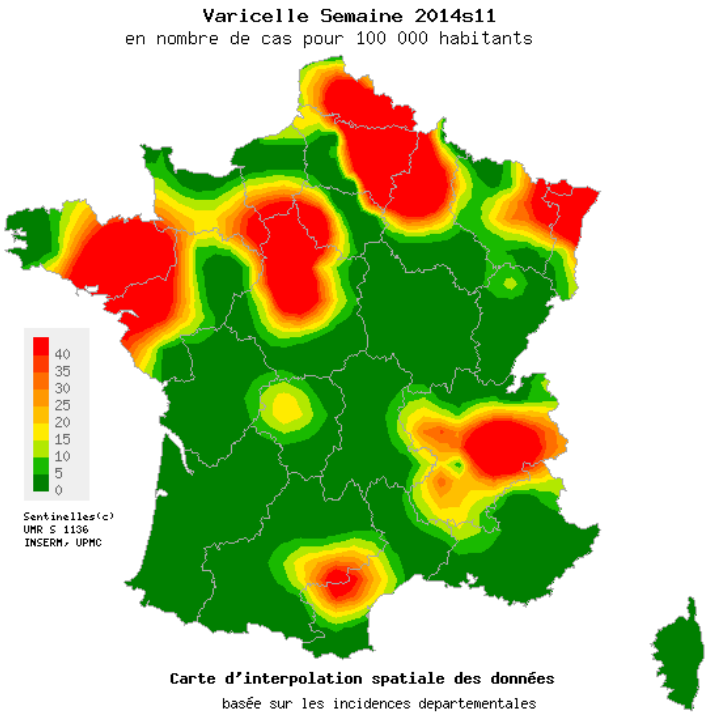


Figure 7 Carte mise à disposition par le réseau Sentinelles

Le réseau cardiauvergne[34] est un réseau basé sur la région Auvergne qui a pour but de suivre les patients atteints d'une insuffisance cardiaque en fédérant différents professionnels de santé (médecins, pharmaciens…). Ce réseau est devenu une réalité grâce notamment aux avancées techniques qui permettent de faire communiquer des capteurs simples d'utilisation comme une balance et un smartphone avec les professionnels de santé qui surveillent le parcours de santé des patients.

Le Réseau de Santé Périnatale d'Auvergne (RSPA) (RSPA 2001; RSPA 2004) est un autre réseau auvergnat qui a pour objectif d'améliorer la qualité de soin des patients avec l'aide des nouvelles technologies. RSPA regroupe 14 centres hospitaliers, cliniques et centres de périnatalité de la région Auvergne, ainsi que des professionnels de la périnatalité, capable d'accompagner la femme enceinte et son enfant tout au long de la grossesse.

Les objets connectés au service du médecin

Les médecins libéraux sont également concernés par l'informatisation de leur métier. Une étude récente réalisée par VIDAL et le Conseil National de l'Ordre des Médecins (CNOM) (Riviere 2013), met en exergue l'utilisation grandissante des smartphones et tablettes par la

[34] http://cardiauvergne.com/index.php - date d'accès octobre 2015

communauté médicale. Les chiffres présentés dans cette étude réalisée auprès de milliers de médecins généralistes et spécialistes, libéraux et/ou salariés, nous apprennent que 3138 d'entre eux possèdent un smartphone et l'utilisent couramment comme outil de travail généralement pour consulter leur agenda, ou les bases de données médicamenteuses. L'ASIP Santé prépare une application mobile pour l'accès au DMP et la messagerie sécurisée. On note également que 89% des médecins consultent notamment les bases de données médicamenteuses au travers de ces nouveaux dispositifs portatifs et connectés à Internet. De même, les premières applications pour SmartWatch sont déjà disponibles.

Nous venons de parcourir un éventail non exhaustif des différentes pistes retenues en France parmi les plus intéressantes et les plus abouties en matière d'e-santé.

Que les initiatives soient d'origine locale, régionale ou nationale, toutes vont dans le même sens, celui d'un meilleur suivi des patients. Intéressons-nous par la suite aux solutions retenues dans les différents pays de l'union européenne.

1.2.2 L'e-santé en Europe

Le Centre Européen de prévention et de contrôle des Maladies (CEPCM) (*European Centre for Disease Prevention and Control* ECDC[35]), basé à Stockholm en Suède est une agence de l'Union Européenne (UE) qui a pour mission de lutter contre les maladies infectieuses ; c'est le pendant européen du *Centers for Disease Control and Prevention* (CDC) américain.

En Europe, l'e-santé se développe à une échelle et une rapidité qui varient d'un pays à l'autre. Ces variations sont souvent liées à la taille du pays et son découpage administratif. Il est cependant important de noter que l'Europe est l'un des leaders mondiaux de l'e-santé. La France, comme nous venons de le voir, dispose de quelques projets d'envergure nationale et de nombreuses initiatives régionales. En Suisse, les solutions retenues se limitent souvent au périmètre du canton (Gnaegi, Wieser, and Dupuis 2010), (Gnaegi and Michelet 2011). Alors qu'en Belgique et au Danemark, les systèmes mis en œuvre, le sont majoritairement à l'échelle du pays (European-Commission 2010a), (Danish-eHealth-Authority 2013), (European-Commission 2010b), ce qui peut s'expliquer par la superficie des territoires.

Le Danemark est reconnu pour être l'un des pays les plus actifs dans le domaine de l'e-santé en Europe (Danish-ministry-of-Health 2012), (Nielsen et al. 2013). Un rapport anonyme[36] disponible depuis le site de l'ambassade de France au Danemark dresse un état des

[35] http://ecdc.europa.eu/en/Pages/home.aspx - date d'accès octobre 2015

[36] http://www.ambafrance-dk.org/IMG/pdf/DK-sante_au_DK.pdf - date d'accès octobre 2015

lieux du service de santé au début des années 2000. Ce rapport est intéressant car il fournit une vision globale du système de santé Danois.

Pour mieux appréhender le contexte européen, le service de statistiques de la Commission Européenne met à notre disposition différents éléments, comme le graphique de la Figure 8, qui nous éclaire sur le taux d'accès à Internet en fonction du pays d'origine. Il semble logique que l'informatisation des services de santé se développe de pair avec l'utilisation grandissante des nouvelles technologies dans les ménages.

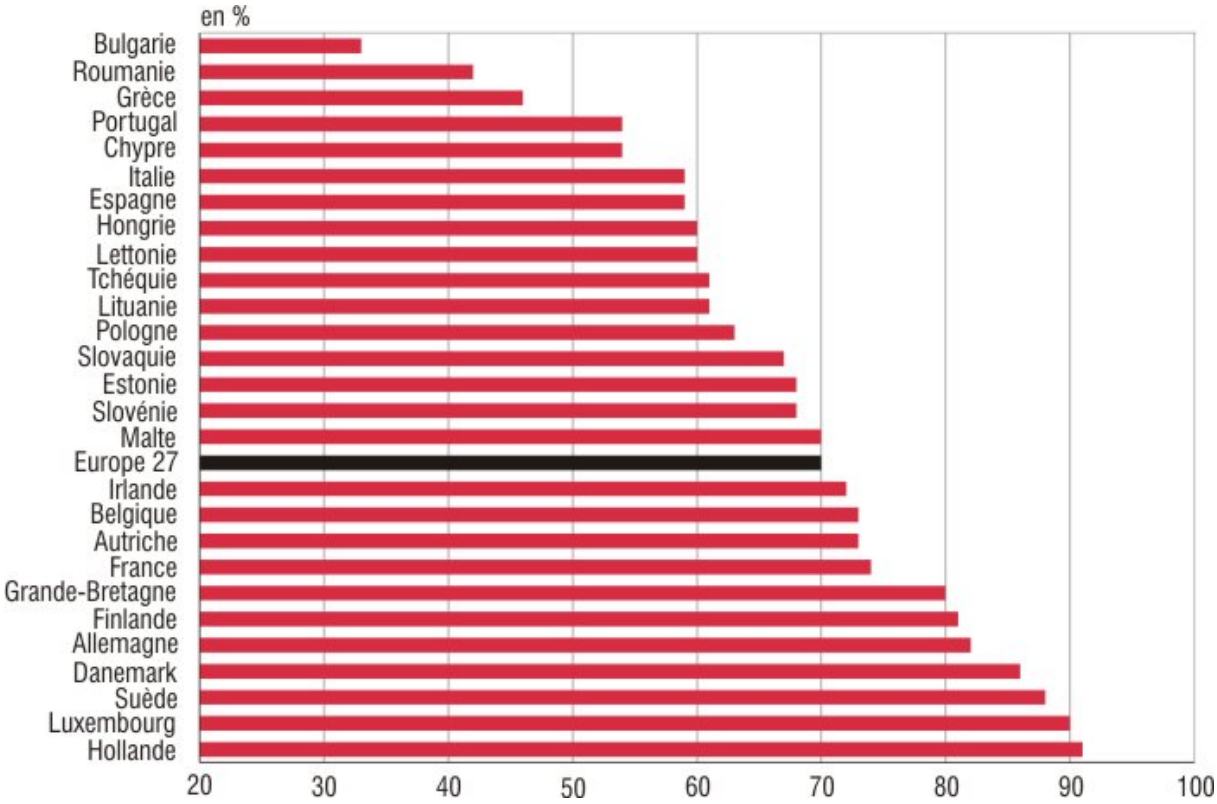


Figure 8 Proportion de ménages européens ayant accès à Internet en 2010 (source : Eurostat)

Le projet Influenzanet[37] est un projet Européen qui s'intègre dans un projet de plus grande envergure nommé Epiwork. Influenzanet fédère de nombreux projets européens dans une dizaine de pays, ayant pour objectif la surveillance de la grippe. C'est le moteur informatique sur lequel se base la version française Grippenet.fr (Marquet et al. 2006), (Debin et al. 2013). Il s'agit donc de recueil d'information par sondage auprès d'une population volontaire d'internautes. Durant l'épidémie de grippe H1N1 de 2009, ces outils ont permis d'étudier la population en temps réel. Des publications comme celles de (Tilston et al. 2010), (Paolotti, Gioannini, and Colizza 2010) concluent d'après leurs expériences nationales que de tels outils

[37] https://www.influenzanet.eu/ - date d'accès octobre 2015

ne sont pas destinés à remplacer les systèmes actuellement mis en place (retour hebdomadaire des médecins généralistes), mais sont de très bons alliés capables d'augmenter la représentativité des services de surveillance, notamment quand les services traditionnels sont submergés par l'afflux de patients lié à une pandémie. De plus, l'exemple hollandais (Marquet et al. 2006) nous démontre qu'il est possible de fédérer un grand nombre de participants derrière ce genre d'initiative et que les résultats obtenus par cet outil sont très cohérents avec les résultats obtenus par les voies plus traditionnelles et fiables comme l'étude des dossiers des patients.

Le projet Epiwork[38] financé par le FP7, 7ème Programme Cadre de la commission européenne a pour but de prévoir les futures épidémies en croisant les données existantes avec des modèles. Les données sont collectées, soit par sondage comme dans le cas français de Grippenet.fr (Influenza en général), soit par une extraction de bases de données médicales. Les sources d'informations sont ensuite associées à des modèles qui permettent, comme dans le cas de la météorologie, d'obtenir des prévisions. Le but étant d'être capable de prévenir des situations pandémiques, en prédisant l'évolution des foyers détectés, ce qui permet de mettre en œuvre les contre-mesures de confinement appropriées. Le projet EPIWORK a permis la création de plusieurs dizaines de publications depuis 2009 comme (Katriel 2010; Coughlin 2006; Tizzoni et al. 2012; Poletto et al. 2013; Parshani, Carmi, and Havlin 2010) (Cf. liste[39]) couvrant les domaines de la création de modèles de population, aux systèmes de surveillance.

D'autres projets ont essayé d'intégrer les technologies des grilles de calculs au monde médical durant les 10 dernières années (Amendolia et al. 2003; Warren et al. 2007; Breton et al. 2005; Gjermundrøda et al. 2007).

Le projet BIOPATTERN du programme FP6 (www.biopattern.org) a été créé pour proposer un réseau sécurisé d'échange de Bioprofile et fournir un moyen de les analyser pour traiter des maladies comme le cancer.

Le projet MammoGrid (Amendolia et al. 2003; Warren et al. 2007) avait pour objectif de démontrer la puissance d'une grille de calculs focalisée sur l'amélioration du diagnostic du cancer du sein. La particularité de ce projet était de transposer les technologies et outils employés dans les grilles pour comparer les mammographies des patientes avec celles stockées dans les bases de données distribuées. Les utilisateurs pouvaient ainsi accéder à de nombreuses sources, et avaient également accès à un outil d'aide à la décision capable d'identifier automatiquement certaines pathologies. Dans la même thématique, eDiaMoND (Lloyd et al.

[38] http://www.epiwork.eu/ - date d'accès octobre 2015
[39] http://www.epiwork.eu/publications/papers/ - date d'accès octobre 2015

2005) visait à construire une photothèque de mammographies à destination des cliniciens du Royaume-Uni en partenariat avec IBM.

L'initiative *Intensive Care Grid* (ICGrid) (Gjermundrøda et al. 2007) s'appuyait sur les technologies mise en œuvre dans l'infrastructure de grille européenne *Enabling Grid for E-sciencE* (EGEE) pour permettre une meilleure traçabilité des cas cliniques notables dans les unités de soins intensifs. Grâce à une interface ergonomique les médecins pouvaient accéder simplement et facilement à la grille EGEE qui gérait de façon transparente les tâches complexes de partage et de réplication de l'information.

Le projet CardioGrid (Eijo et al. 2011) se basait lui aussi sur la grille EGEE pour analyser les électrocardiogrammes. Cette solution permet de traiter sur la grille les données récupérées par des capteurs mobiles.

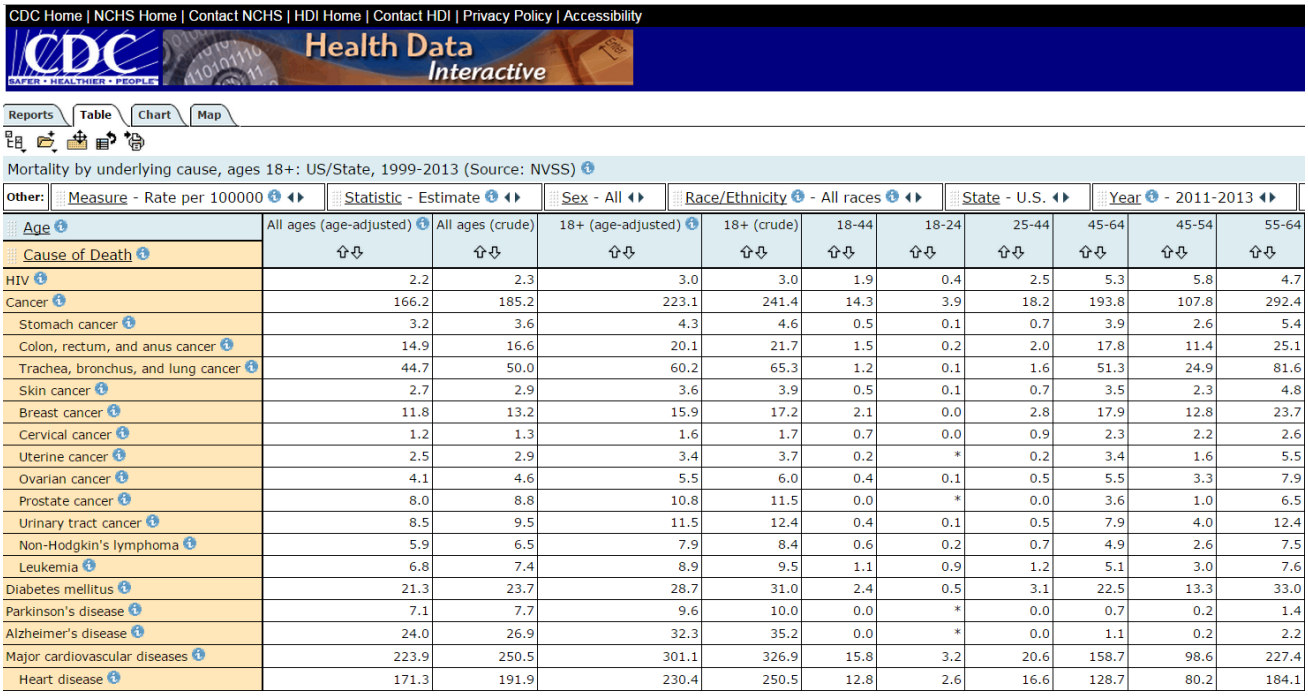
Le projet World-wide In Silico Docking On Malaria (WISDOM) a rendu possible la recherche *in Silico* de médicament grâce à la puissance de calcul répartie dans l'infrastructure de grille (Kasam et al. 2009). WISDOM a démontré l'efficacité de la méthode en accélérant et réduisant les coûts inhérents aux développements de nouvelles molécules pour lutter contre la malaria.

1.2.3 L'e-santé dans le monde

Dans (Bourquard 2007), l'auteure dresse un état de l'art des DMP à travers le monde ; les variantes que nous pouvons relever dans les différentes approches retenues sont liées à l'histoire, la culture et le niveau de maturité vis-à-vis de NTIC, de chaque pays. Des initiatives telles que *Cancer Biomedical Informatics Grid* CaBIG (Fenstermacher et al. 2005) sont à mentionner bien que son impact en dehors des États-Unis soit restreint d'après (Warden 2011). CaBIG est un projet qui a débuté en 2004, avec le soutien du National Cancer Institute[40], ainsi que du National Institues of Health[41]. CaBIG a produit un microcosme de sous logiciels qui répondent chacun à un ou plusieurs besoins particuliers ce qui rend le déploiement très modulable. Le déploiement est facilité par le fait que cette suite logicielle a été développée sous licence open source pour pouvoir plus facilement réutiliser les différents composants. Malheureusement, suite à l'arrêt du projet puis à l'arrêt de la maintenance du site web, la plupart des liens de téléchargement ne permettent désormais plus d'accéder aux fichiers[42].

[40] http://www.cancer.gov/ - date d'accès octobre 2015
[41] http://www.nih.gov/ - date d'accès octobre 2015
[42] Date de non accès octobre 2014

Le *Centers for Disease Control and Prevention* (CDC) en partenariat avec le *National Center for Health Statitics* (NCHS) ont produit le *Data Online Query System* (DOQS) qui permet de requêter les bases de données au travers d'une interface WEB[43] pour des besoins statistiques ou épidémiologiques. Un autre outil aux fonctionnalités similaires est le Health Data Interactive[44] (HDI) dont proviennent les *National Health Interview Survey* (NHIS), *National Vital Statistics System* (NVSS), *National Ambulatory Medical Care Survey* (NHAMCS), *National Ambulatory Medical Care Survey* (NHDS), *National Nursing Home Survey* (NNHS), ou du *National Health and Nutrition Examination Survey* (NHANES). Le HDI permet de consulter 24 heures sur 24, les résultats de ces études nationales qui fonctionnent pour certaines depuis 1965. Un exemple d'accès à ces données est présenté Figure 9. On notera l'une des particularités de cet outil qui est de filtrer les données par ethnie.



Age / Cause of Death	All ages (age-adjusted)	All ages (crude)	18+ (age-adjusted)	18+ (crude)	18-44	18-24	25-44	45-64	45-54	55-64
HIV	2.2	2.3	3.0	3.0	1.9	0.4	2.5	5.3	5.8	4.7
Cancer	166.2	185.2	223.1	241.4	14.3	3.9	18.2	193.8	107.8	292.4
Stomach cancer	3.2	3.6	4.3	4.6	0.5	0.1	0.7	3.9	2.6	5.4
Colon, rectum, and anus cancer	14.9	16.6	20.1	21.7	1.5	0.2	2.0	17.8	11.4	25.1
Trachea, bronchus, and lung cancer	44.7	50.0	60.2	65.3	1.2	0.1	1.6	51.3	24.9	81.6
Skin cancer	2.7	2.9	3.6	3.9	0.5	0.1	0.7	3.5	2.3	4.8
Breast cancer	11.8	13.2	15.9	17.2	2.1	0.0	2.8	17.9	12.8	23.7
Cervical cancer	1.2	1.3	1.6	1.7	0.7	0.0	0.9	2.3	2.2	2.6
Uterine cancer	2.5	2.9	3.4	3.7	0.2	*	0.2	3.4	1.6	5.5
Ovarian cancer	4.1	4.6	5.5	6.0	0.4	0.1	0.5	5.5	3.3	7.9
Prostate cancer	8.0	8.8	10.8	11.5	0.0	*	0.0	3.6	1.0	6.5
Urinary tract cancer	8.5	9.5	11.5	12.4	0.4	0.1	0.5	7.9	4.0	12.4
Non-Hodgkin's lymphoma	5.9	6.5	7.9	8.4	0.6	0.2	0.7	4.9	2.6	7.5
Leukemia	6.8	7.4	8.9	9.5	1.1	0.9	1.2	5.1	3.0	7.6
Diabetes mellitus	21.3	23.7	28.7	31.0	2.4	0.5	3.1	22.5	13.3	33.0
Parkinson's disease	7.1	7.7	9.6	10.0	0.0	*	0.0	0.7	0.2	1.4
Alzheimer's disease	24.0	26.9	32.3	35.2	0.0	*	0.0	1.1	0.2	2.2
Major cardiovascular diseases	223.9	250.5	301.1	326.9	15.8	3.2	20.6	158.7	98.6	227.4
Heart disease	171.3	191.9	230.4	250.5	12.8	2.6	16.6	128.7	80.2	184.1

Figure 9 Exemple d'accès aux données mises à disposition par le CDC, cause de mortalité aux États-Unis de 2011 à 2013 (pour 100000) en fonction de l'âge

L'Organisation Mondiale de la Santé (OMS)[45] ou *World Health Organisation* (WHO) dépend de l'Organisation des Nations Unies (ONU) ; crée en 1948, son siège se situe en Suisse. Le texte fondateur de l'OMS (ONU 1946) révisé en 2006 est très clair sur les objectifs de cette institution, qui doit faire un état de complet du bien-être physique, mental et social pour tout

[43] http://doqs.cdc.gov/NCHSDOQS/query/submit/ed/EDCntyHospED/Count.html - date d'accès octobre 2015
[44] http://205.207.175.93/HDI/ReportFolders/reportFolders.aspx - date d'accès octobre 2015
[45] http://www.who.int/fr/ - date d'accès octobre 2015

être humain. Pour atteindre cet objectif, il est important de pouvoir compter et catégoriser les individus. C'est dans ce but que l'OMS a proposé l'*International Classification of Diseases* (ICD) (cf. 1.5.3). L'OMS analyse les données de 194 états membres pour identifier les causes de maladies, d'accidents ou de décès et pouvoir cibler ses actions aux plus près des besoins. Les résultats de ces études sont publiées dans *World Health Statitics* qui agrège les résultats de chaque pays (Artmann, Giest, and Dumortier 2010; European-Commission 2010b; World Health Organization 2011). L'OMS a beaucoup œuvré contre des maladies infectieuses comme la peste, la tuberculose ou la variole, on considère cette dernière éradiquée depuis 1980 suite aux actions de l'OMS. L'OMS soutient de nombreux projets et programmes qui couvrent un éventail très large de domaines[46] comme la gouvernance et les politiques de santé ; la standardisation et l'interopérabilité ; la recherche et l'épidémiologie ; le e-learning [47]; le support de réseau sociaux et la collaboration Sud-Sud ; aussi bien que le développement et le déploiement d'applications[48].

1.3 Les données de santé et la confidentialité en France

1.3.1 L'aspect législatif

Dans l'intérêt des citoyens, de nombreuses structures telles que la Commission Nationale Informatique et Liberté (CNIL), la Haute Autorité de Santé (HAS) ou encore l'ASIP santé, parmi tant d'autres, ont été créées en France pour préserver la confidentialité des données de santé, protéger les droits fondamentaux des citoyens face au mauvais emploi des nouvelles technologies de l'information. Au plus haut niveau c'est la Ministre des Affaires sociales et de la Santé qui est la garante des politiques du système de santé de la République Française.

La loi n° 78-17 du 6 janvier 1978 – informatique et libertés

La loi n° 78-17 du 6 janvier 1978[49] modifiée, relative à l'informatique, aux fichiers et aux libertés impose notamment une information en cas de collecte et de traitement de données personnelles qui garantit au citoyen un droit de veto et de regard sur les données le concernant. En termes de confidentialité, le code de santé publique dans sa partie législative est très clair. Si les articles L1110-1, 2 et 3, s'attachent respectivement au droit fondamental, à la protection

[46] http://www.who.int/topics/fr/ - date d'accès octobre 2015
[47] http://vaccine-safety-training.org/ - date d'accès octobre 2015
[48] http://www.who.int/features/2012/malaria_drug_resistance_thailand/fr/ - date d'accès octobre 2015
[49] http://www.legifrance.gouv.fr/affichTexte.do?cidTexte=LEGITEXT000006068624&dateTexte=20080609 - date d'accès octobre 2015

de la santé, au respect de la dignité, et à la non-discrimination. C'est dès l'article L1110-4 que le législateur a tenu à préciser que « *toute personne prise en charge par un professionnel, un établissement, un réseau de santé ou tout autre organisme participant à la prévention et aux soins a droit au respect de sa vie privée et du secret des informations la concernant.* ». Les textes de lois français découlent principalement de la directive européenne 95/46/EC relative à la protection des données.

Un système informatique doit être capable d'identifier les personnes qui l'interrogent. Pour se faire, en France, c'est la CPS qui est retenue pour identifier les professionnels de santé (cf. 1.2.1). Cette CPS étant nominative elle sert de clef d'entrée unique.

La Commission Nationale Informatique et Liberté

La CNIL[50] créée par la loi n° 78-17 du 6 janvier 1978 modifiée, dite "informatique et libertés", a pour mission essentielle de protéger les données personnelles. Elle a pour slogan *« Protéger les données personnelles, accompagner l'innovation, préserver les libertés individuelles »*. La CNIL a la particularité d'être une autorité administrative totalement indépendante. Elle est dirigée par un collège de 17 membres et compte environ 160 collaborateurs. La CNIL poursuit principalement 6 missions : informer, réguler, protéger, contrôler, sanctionner, et anticiper.

Il existe plusieurs types de requêtes que nous pouvons déposer auprès des services compétents de la CNIL. Les requêtes varient en fonction de l'utilisation qui est faite des données traitées, et de l'impact vis-à-vis de la vie privée des personnes qu'elles concernent. Sur le site WEB de la CNIL[51] sont accessibles différents formulaires pour la déclaration d'utilisation de données confidentielles.

Les déclarations normales sont les procédures les plus courantes pour la plupart des traitements qui ne posent pas de problèmes vis-à-vis de la protection des libertés. Un Correspondant Informatique et Liberté (CIL) d'une institution a la délégation d'autorisation pour ce genre de demande.

Les déclarations simplifiées aussi nommées déclarations de conformité permettent la gestion des fichiers ou traitements de données personnelles les plus courants, qui ne portent pas atteinte à la vie privée ou aux libertés. Les traitements sont conformes à un modèle déjà défini par une décision CNIL.

[50] http://www.cnil.fr/ - date d'accès octobre 2015
[51] http://vosdroits.service-public.fr/professionnels-entreprises/R18458.xhtml - date d'accès octobre 2015

Pour les requêtes plus complexes la CNIL propose des démarches plus longues. Les demandes d'autorisations si les données considérées sont sensibles, comme l'utilisation de données confidentielles très « identifiantes », les demandes d'avis pour les organismes publics ou la gestion d'infractions, l'accès au RNIPP ou la biométrie.

Les deux dernières demandes sont régies par la loi informatique et liberté (chapitre IX et X) il s'agit des demandes d'autorisations de recherche médicale, ainsi que les demandes d'autorisations d'évaluation de pratiques de soins.

Certaines activités sont dispensées de déclaration pour des raisons de commodités nous pouvons ici citer les activités purement personnelles, les fichiers d'associations politiques, religieuses ou syndicales ainsi que les opérations courantes de l'entreprise comme la comptabilité.

L'Agence des Systèmes d'Information Partagés de santé (ASIP santé)

En 2009 l'Agence des Systèmes d'Information Partagés de santé (ASIP santé) est créée par les pouvoirs publics avec pour vocation d'améliorer l'accès aux soins tout en veillant au respect des droits des patients.

L'assemblée générale de l'ASIP Santé est garante de la politique générale de l'agence. Elle est composée d'un président secondé d'administrateurs représentant l'État, la Caisse nationale d'assurance maladie des travailleurs salariés (CNAMTS) et la Caisse nationale de solidarité pour l'autonomie (CNSA). L'une des principales prérogatives de l'ASIP santé est d'assurer le développement des projets de systèmes d'information de santé. Le ministère des Affaires sociales et de la Santé lui délègue donc la gestion des politiques de santé en lien avec les nouvelles technologies. L'ASIP Santé se trouve au centre d'un écosystème composé de partenaires publics et industriels, d'associations de patients, d'Ordres, de sociétés savantes, d'ARS et de GCS. La première version du DMP avait été confiée à l'ASIP Santé avant que la réalisation soit transférée à la CNAM dans l'objectif d'un hypothétique DMP 2.0. L'ASIP est aussi en charge de la distribution des cartes de la famille CPS. C'est donc un acteur incontournable de l'e-santé en France.

Les Agences Régionales de Santé (ARS)

Les Agences Régionales de Santé (ARS) ont été créées par la loi du 21 juillet 2009 dans le cadre de la réforme de l'hôpital et relative aux patients, à la santé et aux territoires (HPST), plus précisément dans l'article 118. Elles sont le pilier de la réforme du système de santé. L'ARS rassemble au niveau régional les ressources de l'État et de l'Assurance maladie, pour

renforcer l'efficacité collective et garantir l'avenir du service public de la santé. L'ARS regroupe en une seule entité, plusieurs organismes chargés des politiques de santé dans les régions et les départements : Directions Régionales et Départementales des Affaires Sanitaires et Sociales (DRASS et DDASS), Agences Régionales de l'Hospitalisation (ARH), Groupements Régionaux de Santé Publique (GRSP), Unions Régionales des Caisses d'Assurance Maladie (URCAM), Missions Régionales de Santé (MRS) et volet hospitalier de l'assurance maladie. Ce dernier est composé d'une partie du personnel des Caisses Régionales d'Assurance Maladie (CRAM), du Régime Social des Indépendants (RSI), de la Mutualité Sociale Agricole (MSA), des Directions Régionales du Service Médical (DRSM).

En unifiant des forces dispersées, les ARS permettent de mener des politiques de santé plus efficaces et de simplifier le système de santé français.

Les ARS ont pour mission d'assurer, à l'échelon régional, le pilotage d'ensemble du système de santé. Elles sont responsables de la sécurité sanitaire, des actions de prévention menées dans la région, de l'organisation de l'offre de soins en fonction des besoins de la population, y compris dans les structures d'accueil des personnes âgées ou handicapées.

Les Groupements de Coopération Sanitaire (GCS)

Chaque région française possède un Groupement de Coopération Sanitaire (GCS[52]), ils sont souvent l'un des outils opérationnels des ARS. Chaque GCS peut faciliter les rapports entre les sociétés productrices de services informatiques et les professionnels de santé, conseiller sur les aspects juridiques et techniques, coordonner les actions de différentes structures médicales pour créer plus de sens et de cohérence et éventuellement fédérer des projets similaires qui auraient pu s'ignorer sans l'intervention du GCS.

En Auvergne les différents projets d'e-santé peuvent se fédérer autour du GCS qui se nomme Système d'Information Médicale Partagée en Auvergne (SIMPA). En Auvergne, les membres du GCS sont financièrement solidaires comme le sont les membres d'un Groupement d'Intérêt Économique (GIE). Le CGS SIMPA est le relais opérationnel pour la mise en œuvre des projets d'e-santé en Auvergne. Il a vocation à fédérer les initiatives et les acteurs de télésanté et assurer la cohérence des solutions mises en œuvre. Il est la maîtrise d'ouvrage de l'Espace Numérique Régional de Santé (ENRS) [53]. L'ENRS est une interface WEB qui a pour vocation de réunir en un seul et même lieu tous les projets et outils à destination des acteurs de santé

[52] http://www.sante.gouv.fr/le-groupement-de-cooperation-sanitaire.html - date d'accès octobre 2015
[53] http://www.simpa-telesante.org/ - date d'accès octobre 2015

auvergnats. Ainsi, lorsqu'un utilisateur se connecte à l'aide de sa carte CPS, il dispose automatiquement de tous les outils auxquels il est habilité à accéder. Le GCS est porté stratégiquement par l'Agence Régionale de Santé (ARS) Auvergne, il suit et participe à la promotion des recommandations de l'ASIP santé. Il sert de support et de conseil pour les projets auvergnats d'e-santé.

1.3.2 Le suivi du patient, l'identitovigilance

La traçabilité du patient est l'un des aspects primordiaux de la gestion des données de santé. Les données assignées à un dossier doivent être relatives au patient auquel se réfèrent ces informations. Ce qui est vrai en termes de soins, l'est aussi d'un point de vue épidémiologique. Il est important de comptabiliser de façon fiable les cas pour ne pas fausser les statistiques. Pour permettre cette reconnaissance de patient au travers des différents SI dans lesquels ses informations sont collectées, il est nécessaire d'identifier les variables discriminantes d'un individu à l'autre. Comme proposé dans (C Quantin et al. 2004) qui utilise différents champs comme le nom et le prénom et où l'on apprend que l'utilisation du second prénom est à éviter. Un individu doit pouvoir être retrouvé dans les différents SI, les différents services, les différents hôpitaux, dans lesquels il a effectué une consultation, et où ses prélèvements ont été analysés.

Une fois que les champs les plus pertinents et judicieux ont été retenus il faut pouvoir les traiter en analysant la ressemblance des uns avec les autres. Ainsi, si l'on considère un champ comme le prénom, se pose la question de savoir si « Philippe » ressemble plus à « Philips » ou à « Filippe ». Il en sera de même pour les dates « 12/07/1980 » et « 1980-07-12 » ou encore « 07-12-80 ». La complexité de la problématique est évidente avec les problématiques de casse, d'accentuation, d'erreurs induites par la transposition au clavier, la phonétique ou même le type de stockage informatique, etc. Winkler a beaucoup travaillé sur cette problématique (William E Winkler 2005; Sauleau, Paumier, and Buemi 2005; W. Winkler 1990; W E Winkler 2007) et avec l'aide de Jaro (Matthew A Jaro 1989; M A Jaro 1995), a proposé l'algorithme de Jaro-Winkler (William E Winkler 1999). Les différentes méthodes d'identification des patients et plus généralement de chaînages des données sont très bien présentées dans la partie 4.4 de (DeVlieger 2011) et l'état de l'art de (Li 2015). Les différents algorithmes ont été comparés dans (Cohen, Ravikumar, and Fienberg 2003) il en ressort que c'est un algorithme hybride qui obtient les meilleurs résultats en terme de reconnaissance et de rapidité. C'est aussi la conclusion de (Li 2015) qui propose une méthode alternative pour faire coexister une très bonne reconnaissance du chaînage avec des temps de calculs courts.

Nous verrons dans un chapitre dédié comment le chaînage des différentes données médicales est possible, à partir de quelles informations, avec quels moyens et pour quel résultat.

1.4 Les données de santé et leur gestion

Dans cette section il faut interpréter le mot « distribution » en tant qu'une répartition géographique, et non en tant que propagation de l'information. Nous axons notre étude sur un système informatique distribué qui « s'oppose » philosophiquement à certaines des solutions techniques retenues pour le DMP ayant pour but de centraliser les informations dans un seul et même lieu sur des serveurs de données (centre de calculs*).*

1.4.1 Données et métadonnées

Il est intéressant de se poser la question – « Qu'est-ce qu'une donnée ? ». Cette question semble triviale, cependant prenons un instant pour la considérer. Une donnée est une information, relative à un sujet, par exemple, si nous photographions un pull bleu, l'image sera stockée dans un certain format (« .jpeg », « .png », etc.) elle constituera notre donnée principale. Les métadonnées quant à elles, viennent décrire la donnée principale. Cette dernière se trouve enrichie par l'adjonction d'autres informations « périphériques », comme le propriétaire de l'image, l'heure de création du fichier, le lieu de la prise de vue, etc.

Un autre ensemble de données qui caractérisent cette image, est celui des données sémantiques. Les mots clefs « pull », « bleu », « coton », « Sébastien », « image », « jpeg », « avril », « 2011 » ; peuvent être rajoutés aux métadonnées qui entourent cette image. Chaque mot clef, étant significatif d'un attribut relatif à cette image, et permettant une indexation ayant pour but de faciliter la recherche ultérieure de ce document. Ce qui est vrai pour une image de la vie quotidienne peu facilement se transposer aux données médicales. Ainsi, nous venons de montrer simplement que pour chaque information considérée il faudra aussi traiter les métadonnées qui l'entourent pour gagner en précision. Ces données qui apportent du sens à l'information que nous traitons sont des données sémantiques elles s'additionnent aux métadonnées, citées précédemment pour consolider notre donnée principale qui devient ainsi plus facilement exploitable.

1.4.2 Flux de travaux de l'utilisation des données de santé

Il est utile de bien comprendre le flux de travaux (*workflow*) de l'utilisation des données médicales d'un patient pour élaborer un système de partage de l'information efficace. Nous expliquons ci-dessous les étapes de ce flux de travaux (VAN DER AALST and VAN HEE 2008) (pour des raisons de praticité et de simplification les problématiques de facturation sont volontairement éludées ou simplifiées à l'extrême).

La figure 10 résume les interactions nécessaires avec le système d'information au sein d'un cabinet de médecin libéral : de la prise de rendez-vous jusqu'au règlement de la consultation.

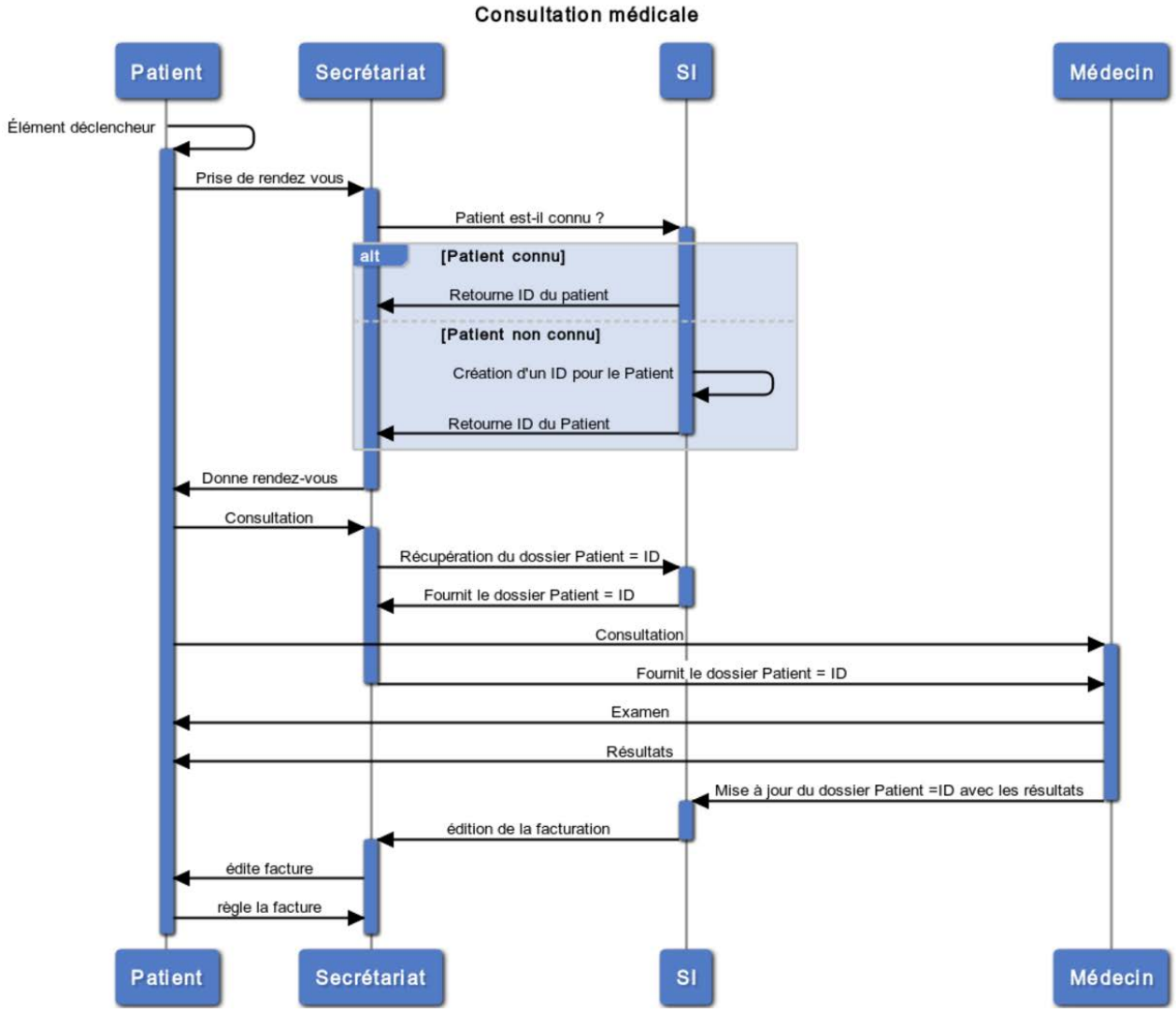


Figure 10 Diagramme de séquence UML d'une consultation médicale

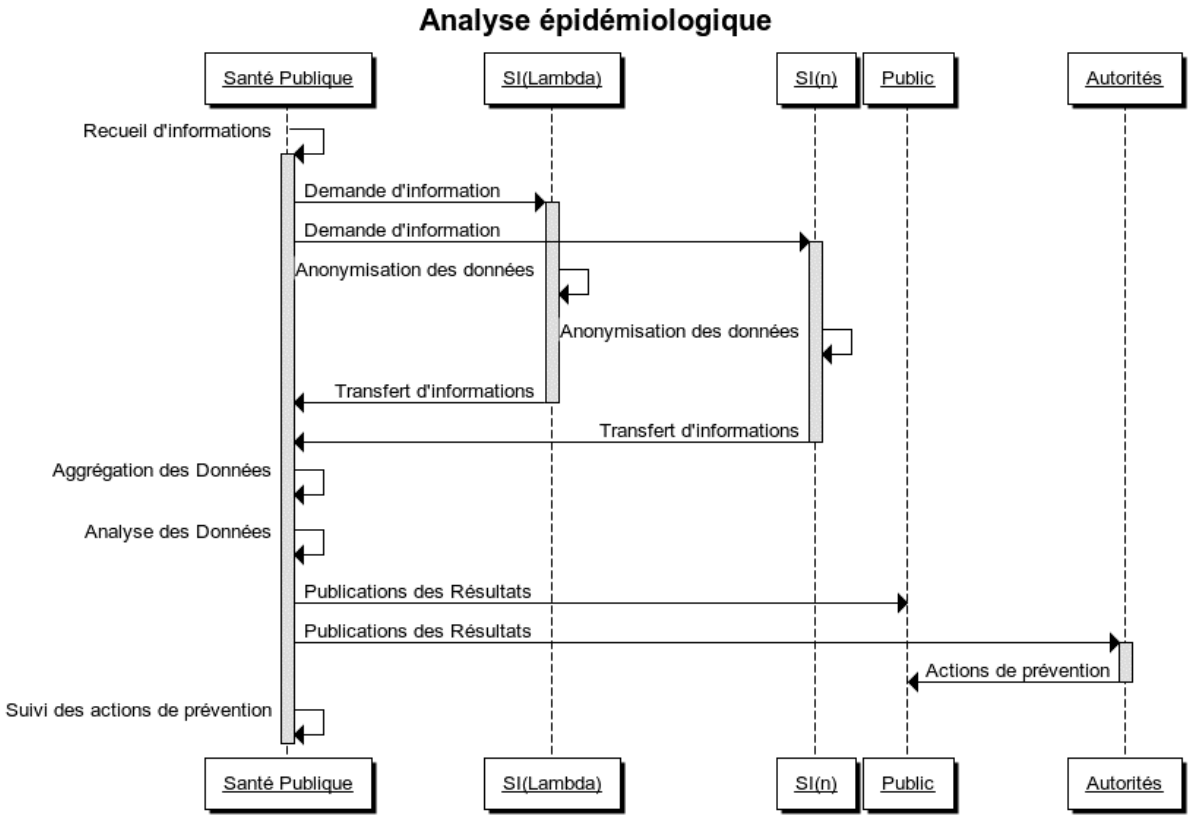


Figure 11 Diagramme de séquence UML d'une étude de santé publique

La figure 11 est relative aux différentes étapes nécessaires pour la réalisation d'une étude statistique à but épidémiologique. Depuis l'envoi d'une requête sur différents sites de la part d'un épidémiologiste, en passant par l'analyse des résultats, jusqu'aux mesures à mettre en œuvre après la communication des résultats

1.4.3 L'interopérabilité des données de santé

L'interopérabilité, est la capacité de faire travailler différents systèmes hétérogènes ensembles. Hétérogènes du point de vu physique mais surtout du point de vue logiciel et plus particulièrement au niveau des formats de stockage de l'information. À titre d'exemple, le SI d'un laboratoire d'anatomie et cytologie pathologies (ACP) doit être capable de fournir des données à une association de dépistage du cancer ou plus généralement à une Structures de Gestions des dépistages Organisés (SGDO), bien que les deux SI ne soient pas structurés de la même façon. Pour réaliser cette fonctionnalité il est nécessaire que les données soient présentées sous une forme compréhensible par chacun des sites. La solution qui semble la plus évidente pour atteindre ce but est d'avoir recours à un ou plusieurs standards, ou normes. Une fois la décision du standard effectuée il suffit d'automatiser la conformation de l'export du premier site vers le fichier normalisé. Et si besoin, lors de l'arrivée du fichier normé sur le second site

une étape supplémentaire consiste à extraire les données pour les intégrer. Nous nommons cette étape « alignement » car elle consiste à aligner les différents champs, des différentes bases de données, pour faire correspondre les informations qu'ils contiennent ; à cet effet plusieurs façons de procéder sont envisageables cf. Figure 12 et Figure 13.

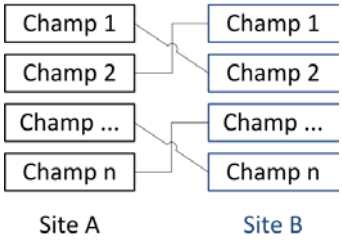


Figure 12 Alignement direct des champs de deux sites

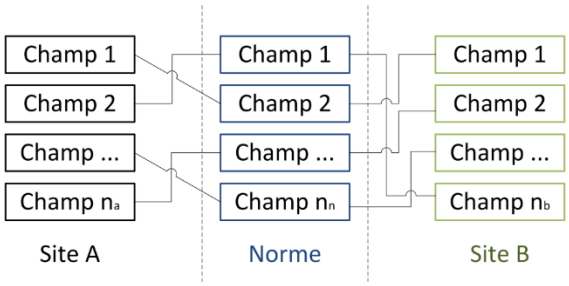


Figure 13 Alignement sur un standard ou une norme

Le principal avantage de la solution s'appuyant sur une norme est sa capacité à se transposer sur plusieurs sites à moindre coût ce qui permet de réaliser des économies d'échelle très importantes(Walker et al. 2005). En effet, si l'on ajoute une nouvelle base de données, il faut simplement aligner les champs de la base avec la norme choisie pour que cette base puisse interagir avec toutes les autres bases du système.

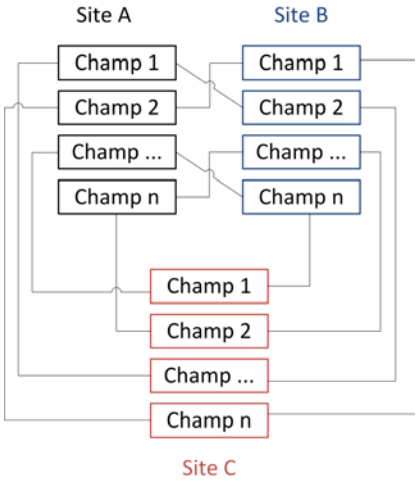


Figure 14 Alignement des champs de données de 3 sites en direct

Si nous généralisons pour ‘n’ systèmes il faudra sans le recours à une norme ou un standard, un nombre d’interfaces pour relier tous les SI entre eux, qui sera égal à la combinaison de 2 parmi n (systèmes) :

$$nombre\ d'interfaces = C_n^2 = \frac{n_{syst}!}{2(n_{syst}-2)!}$$

Alors que si nous employons une norme comme pont entre les différents SI, le nombre d’interfaces diminue à n :

$$nombre\ d'interfaces = nombre\ de\ systèmes$$

Ce qui pour un exemple avec seulement 6 systèmes d’informations à interconnecter fait passer le nombre d’interfaces de 15 à 6 ; et avec 100 systèmes de 4950 à 100. Ce nombre d’interfaces est à multiplier par le nombre de champs si l’on souhaite obtenir une idée du nombre d’alignements nécessaires.

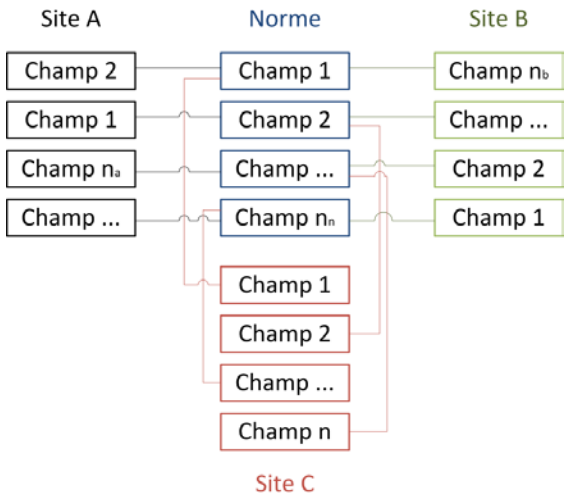


Figure 15 Alignement des données en utilisant une norme commune

Dans les paragraphes suivants nous présenterons certains des standards et normes qui peuvent servir de support pour un alignement basé sur une norme.

1.5 Normes et standards pour les données de santé

Comme nous venons de l’évoquer dans le paragraphe précédent, les standards permettent l’interopérabilité des systèmes entre eux. Mais pas seulement, ils sont aussi importants si l’on souhaite simplement comparer les données entres différents sites et même à l’intérieur d’un seul et même établissement où les différents services peuvent se comporter comme des sites distants totalement indépendants (interop’santé 2015). En effet, si les informations sont normalisées de la même façon, elles deviennent interopérables et compréhensibles par le plus

grand nombre. Nous attirons ici l'attention sur la différence qu'il existe entre norme et standard[54]. Une norme est un ensemble de règles de conformité ou de fonctionnement légiférés par un organisme de normalisation mandaté (AFNOR, ISO, etc.). Alors qu'un standard est un ensemble de recommandations ou de préférences préconisées par un groupe d'utilisateurs caractéristiques et avisés. Nombreuses sont les normes que nous utilisons en informatique comme celles éditées par le *World Wide Web Consortium* (W3C)[55] : HTML, HTTP, XML, etc. ; l'*American National Standards Institute* (ANSI)[56] (C90) ; ou l'*International Organization for Standardization* (ISO)[57] (9001). Il en est de même dans le domaine de la médecine et de la santé. Nous n'aborderons pas, dans les paragraphes suivants, les normes et standards d'imagerie comme le *Digital Imaging and COmmunications in Medicine* (DICOM) pour nous focaliser sur les différentes façons de classer et échanger les données de santé.

1.5.1 Health Level Seven (HL7)

Nommé ainsi en référence à la représentation théorique d'un réseau, (le modèle *Open Systems Interconnection* (OSI) de l'*International Organization for Standardization* (ISO)) qui comporte 7 niveaux, et notamment le 7ème qui est dédié à la couche applicative. HL7 est une structure américaine à but non lucratif, accréditée par l'*American National Standards Institute* (ANSI). Le but de HL7 est de permettre l'interopérabilité des données de santé en créant des normes et standards, comme le CDA R2 présenté dans le paragraphe suivant, et encourager leur adoption. Il existe une centaine de standards HL7 classés en 7 sous catégories, dont certains standards de plus bas niveau servent à construire les standards dédiés à l'*Electronic Health Records* (EHR) par exemple.

HL7 Clinical Document Architecture Release 2 (CDA R2)

Le *Clinical Document Architecture* est une norme soutenue par le HL7, cette norme a obtenu sa certification ISO en 2009[58] (ISO/HL7 27932:2009). Les différences entre la première version du CDA (Dolin et al. 2001) et sa seconde mouture dont il est question dans ce paragraphe sont exposées dans (Dolin et al. 2006). Cette norme est basée sur un format

[54] Question d'autant plus pertinente qu'il n'existe qu'un mot en anglais (« *standard* ») pour traduire c'est deux concepts bien différents que sont d'une part les standards et d'autres parts les normes.

[55] http://www.w3.org/ - date d'accès octobre 2015

[56] http://www.ansi.org/ - date d'accès octobre 2015

[57] http://www.iso.org/iso/home.html - date d'accès octobre 2015

[58] http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=44429 - date d'accès octobre 2015

Extensible Markup Language (XML) qui structure le document au travers de balises et d'indentation, il réutilise aussi des standards existant tels que le SNOMED CT (cf. 1.5.2) et le *Logical Observation Identifiers Names and Codes* (LOINC)[59] (très utilisé pour les résultats de biologie). Comme son nom l'indique son but est de préciser la structure et l'objet d'un document médical. Il faut ici se rapporter au paragraphe sur la donnée médicale, et comprendre que cette norme s'applique aux métadonnées, plus qu'aux données elles-mêmes. L'un des avantages du CDA R2 est la capacité d'encapsuler des données déjà existantes avec des entêtes, pour rendre les données compatibles CDA sans pour autant les modifier.

1.5.2 Systematized Nomenclature of Medicine (SNOMED)

Le terme SNOMED (Cornet and de Keizer 2008) est une extension de SNOP (*Systematized Nomenclature of Pathology*) qui date de 1965. Ce standard est maintenu avec le support du *College of American Pathologists* (CAP) et depuis 2007 du *International Health Terminology Standards Development Organisation* (IHTSDO) une organisation à but non lucratif basée au Danemark. Il existe différentes versions de SNOMED dont la version 3.5 et la version *Clinical Terms* (CT) notamment, la CT étant la version la plus récente. Les différentes versions coexistent et sont souvent référencées dans la littérature sans même préciser la version qui est considérée (Cornet and de Keizer 2008). SNOMED est une terminologie qui a pour vocation d'avoir une portée globale dans le domaine médical en couvrant un large éventail de besoins et de spécialités cliniques. SNOMED est destiné à représenter les données du patient dans un but clinique.

La version 3.5, publiée en 1998, a pour intérêt d'avoir été traduite en français, version que l'on peut donc nommer SNOMED 3.5 VF (version française). Cette traduction a rendu SNOMED 3.5 VF intéressante aux yeux des dirigeant français qui en ont fait l'acquisition en octobre 2006 par le biais du GIP DMP cette acquisition a été commentée par la cour des comptes dans son rapport annuel de 2009 (cour des comptes 2009). Cependant cet achat est controversé. C'est ce qu'évoque (cour des comptes 2009) qui remarque qu'à l'époque où la France décide d'acquérir la version 3.5, le consortium utilisant 3.5 décide-lui, de migrer vers la version SNOMED CT. Ainsi la version 3.5, dès sa livraison, pouvait déjà être considérée comme obsolète. Suite à cette démarche, l'État a mis cette version gratuitement à la disposition du public depuis 2008. Pour acquérir gratuitement la version 3.5 VF il suffit d'accepter un contrat

[59] https://loinc.org/ - date d'accès octobre 2015

de licence disponible sur le site de l'ASIP-Santé[60] avant de télécharger le fichier contenant les 11 fichiers au format Excel ('.xls').

Tableau 2 Liste des fichiers constitutifs de SNOMED 3.5 VF

Nom du fichier XLS	Taille	Commentaires
A_Agents_physiques_FR_1_1	0.18 Mo	Liste d'agents physiques
C_Chimie_Biol_Medic_FR_1_1	0.89 Mo	Liste d'éléments chimiques ou physiques
D_Diagnostics_FR_1_1	7.35 Mo	Liste de maladies et affections
F_Fonctions_FR_1_1	2.36 Mo	Liste de fonctions biologiques
G_Modificateurs_FR_1_1	0.16 Mo	Liste de concepts
J_Metiers_FR_1_1	0.40 Mo	Liste des métiers
L_Organismes_vivants_FR_1_1	2.66 Mo	Liste des organismes vivants
M_Morphologie_FR_1_1	1.07 Mo	Liste de morphologies descriptives
P_Procedures_FR_1_1	3.50 Mo	Liste de procédures
S_Social_FR_1_1	0.13 Mo	Liste des contextes sociaux
T_Topographie_FR_1_1	2.05 Mo	Liste topographique du corps humain

Chaque fichier '.xls' représente un classeur qui peut contenir plusieurs feuilles, chaque feuille représente alors un chapitre et chaque chapitre peut être subdivisé en section. Si nous considérons le fichier relatif aux diagnostics (D_Diagnostics_FR_1_1.xls), il contient 16 chapitres (cf. Tableau 3, le chapitre des « Maladies de la peau et des tissus sous-cutanés » contient lui 13 sections (cf. Tableau 4), dans lesquelles sont classés les codages (cf. Tableau 5), dans ce classeur chaque chapitre contient entre 500 et 6000 entrées.

Tableau 3 Liste des chapitres du classeur de diagnostics de la classification SNOMED 3.5 VF

Classeur	Feuille	Nom	Cellule	Valeur
D_Diagnostics_FR_1_1.xls	D0		E2	Chapitre 0 Maladies de la peau et des tissus sous-cutanés
D_Diagnostics_FR_1_1.xls	D0		E1835	Classification clinique des tumeurs (chapitre 2 de la CIM-10)
D_Diagnostics_FR_1_1.xls	D1		E2	Chapitre 1 Maladies de l'appareil locomoteur et des tissus conjonctifs
D_Diagnostics_FR_1_1.xls	D2		E2	Chapitre 2 Maladies de l'appareil respiratoire
D_Diagnostics_FR_1_1.xls	D3		E2	Chapitre 3 Maladies de l'appareil circulatoire
D_Diagnostics_FR_1_1.xls	D4		F2	Chapitre 4 Maladies congénitales
D_Diagnostics_FR_1_1.xls	D5		E2	Chapitre 5 Maladies de l'appareil digestif
D_Diagnostics_FR_1_1.xls	D6		E2	Chapitre 6 Maladies métaboliques et nutritionnelles
D_Diagnostics_FR_1_1.xls	D7		E2	Chapitre 7 Maladies de l'appareil génito-urinaire
D_Diagnostics_FR_1_1.xls	D8		E2	Chapitre 8 Diagnostics liés à la grossesse et à la période périnatale
D_Diagnostics_FR_1_1.xls	D9		E2	Chapitre 9 Troubles mentaux
D_Diagnostics_FR_1_1.xls	DA		E2	Chapitre A Maladies du système nerveux et des organes des sens
D_Diagnostics_FR_1_1.xls	DB		E2	Chapitre B Maladies du système endocrinien
D_Diagnostics_FR_1_1.xls	DC		E2	Chapitre C Maladies des systèmes hématopoïétique et immunitaire
D_Diagnostics_FR_1_1.xls	DD		E2	Chapitre D Lésions traumatiques et empoisonnements
D_Diagnostics_FR_1_1.xls	DE		E2	Chapitre E Maladies infectieuses et parasitaires
D_Diagnostics_FR_1_1.xls	DF		E2	Chapitre F Maladies de classes générales, états des victimes et décès

[60] http://esante.gouv.fr/services/referentiels/referentiels-d-interoperabilite/snomed-35vf - date d'accès octobre 2015

Tableau 4 Liste des sections du chapitre Maladies de la peau et des tissus sous-cutanés de la nomenclature SNOMED 3.5 VF

Classeur	Feuille	Nom	Cellule	Valeur
D_Diagnostics_FR_1_1.xls	D0		E3	Section 0-0 Maladies de la peau et des tissus sous-cutanés: termes généraux, types histologiques et infections
D_Diagnostics_FR_1_1.xls	D0		E408	Section 0-1 Maladies bulleuses et vésiculaires non infectieuses
D_Diagnostics_FR_1_1.xls	D0		E767	Section 0-2 Maladies érythémateuses, papuleuses et squameuses non infectieuses
D_Diagnostics_FR_1_1.xls	D0		E1001	Section 0-3 Affections vasculaires de la peau
D_Diagnostics_FR_1_1.xls	D0		E1040	Section 0-4 Affections dégénératives de la peau
D_Diagnostics_FR_1_1.xls	D0		E1123	Section 0-5 Affections des annexes de l'épiderme
D_Diagnostics_FR_1_1.xls	D0		E1456	Section 0-6 Ulcères cutanés et cicatrices
D_Diagnostics_FR_1_1.xls	D0		E1508	Section 0-7 Affections de la pigmentation, éruptions médicamenteuses et affections causées par des agents physiques
D_Diagnostics_FR_1_1.xls	D0		E1676	Section 0-8 Lésions pseudo-tumorales de la peau
D_Diagnostics_FR_1_1.xls	D0		E1711	Section 0-9 Maladies inflammatoires des tissus adipeux sous-cutanés et granulomes non infectieux
D_Diagnostics_FR_1_1.xls	D0		E1760	Section 0-A Maladies des muqueuses
D_Diagnostics_FR_1_1.xls	D0		E1788	Section 0-B Affections des points d'injection et des zones d'application
D_Diagnostics_FR_1_1.xls	D0		E1836	Section 0-F Classification clinique des tumeurs de la peau

Tableau 5 Exemples de codage issus de la nomenclature SNOMED 3.5 VF

LIGN	TERMCODE	FMOD	FCLASS	FNOMEN	REFERENCE	ICDCODE	ICD10
2251	D0-F0352		01	carcinome in situ de la peau du sein	(T-02430) (M-80102)	232.5	D04.5
2252	D0-F0352		02	carcinome intra-épithélial cutané du sein	(T-02430) (M-80102)	232.5	D04.5
2253	D0-F0353		01	tumeur maligne de la peau du sein	(T-02430) (M-80003)	173.5	C44.5
2254	D0-F0353		02	tumeur maligne cutanée du sein	(T-02430) (M-80003)	173.5	C44.5

On remarque que lorsque c'est possible les codes de SNOMED 3.5 VF ont été mis en correspondance avec les codes ICD-10 (cf. 1.5.3) ainsi que la version SNOMED *Clinical Terms* (CT) (Ihtsdo 2015) publiée pour la première fois en 2002. (Merabti et al. 2009) se sont interrogés sur les alignements qui pouvaient être faits entre SNOMED CT et SNOMED 3.5 VF ainsi qu'avec ICD-10. SNOMED CT utilise une approche sémantique des définitions, ce standard repose sur un triplet « concept-terme-composant » qui rappelle le triplet RDF « sujet-prédicat-objet ».

1.5.3 International Classification of Diseases (ICD)

International Classification of Diseases (ICD) aussi connue en France sous l'acronyme CIM pour Classification Internationale des Maladies est une initiative de l'OMS (cf. 1.2.3). Depuis sa création, plusieurs versions se sont succédées (ICD-6, ICD-9, ICD-10 et bientôt ICD-11). L'évolution du codage permet de gagner en détails, ICD-9 comportait 13 000 codes alors que ICD-10 en contient 68 000. Des ajouts important ont été effectués (AMA 2014) notamment au niveau de la latéralité, mais surtout en termes de précisions, cependant les codes ne sont pas rétro-compatibles, le passage à l'un implique l'abandon du plus ancien. Il existe aussi des variantes spécialisées comme ICD-O-3 dédiée à l'oncologie. ICD-O-3 permet de coder la topographie et l'histologie propre au compte rendu d'anatomo-cyto-pathologie (ACP) à la manière du code ADICAP que nous présenterons dans le paragraphe suivant.

1.5.4 L'Association pour le Développement de l'informatique en Cytologie et en Anatomie Pathologique (ADICAP)

Le code ADICAP porte le nom de l'Association pour le Développement de l'informatique en Cytologie et en Anatomie Pathologique[61] qui est à l'origine de cette classification. Le code de classification produit par cette association sert à qualifier d'une façon normalisée les lésions rencontrées en anatomo-cyto-pathologie.

Ce code est distribué librement par l'association, la dernière version date de novembre 2009 (ADICAP 2009). C'est un code alpha numérique de 15 caractères structuré, résumé dans la représentation de la Figure 16.

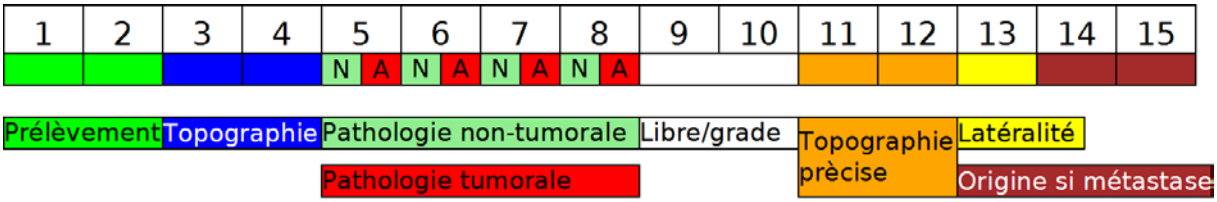


Figure 16 Représentation de la structure d'un code ADICAP

En pratique, seuls les 8 premiers caractères (obligatoires) sont utilisés pour coder les analyses. A titre d'exemple, les Frottis Cervicaux Vaginaux (FCV) sont codés de la façon suivante :

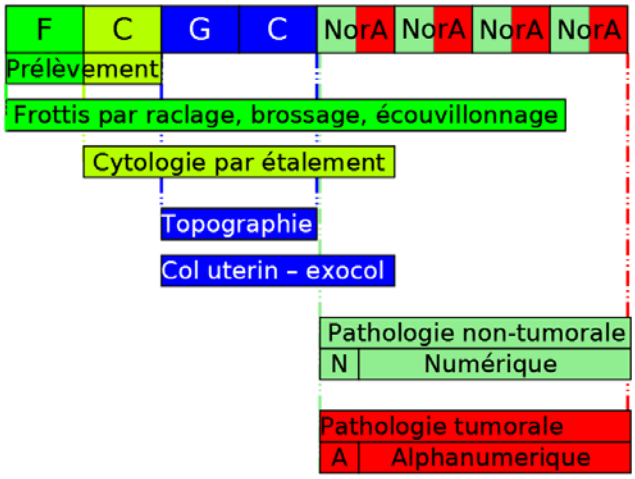


Figure 17 Représentation d'un code ADICAP codant pour un Frottis Cervical

Il existe des outils en ligne pour passer du code ADICAP en CIM-O-3 (cf. 1.5.3) comme celui proposé par la société mondeca[62].

[61] http://www.adicap.asso.fr/ inaccessible en septembre 2014

[62] http://client3.mondeca.com/node02/AdicapSimple/ - date d'accès octobre 2015

1.5.5 La norme d'échange d'Anatomo-Cyto-Pathologie (ACP)

Cette norme ACP est publiée au Journal Officiel est définie un fichier cadre pour les échanges entre les cabinets ACP et les associations de Dépistage Organisé du Cancer (DOC). Pour les besoins spécifiques de certaines structures comme l'ABIDEC (Association Bourbonnaise Interdépartementale de Dépistage des Cancers) -ARDOC, leurs partenaires (O.S.I Santé) utilise cette norme pour réceptionner les données transmises par APICRYPT depuis le cabinet Sipath-Unilabs. Cette norme codée sur 244 caractères, détaillée dans le Tableau 6, permet le transfert des données en texte brut. L'analyseur syntaxique, qui parcourt le fichier, différencie les champs en fonction de la position sur la ligne uniquement. Chaque ligne du fichier d'échange correspond à l'entrée d'une patiente, et chaque ligne est terminée par un saut de ligne (Windows).

Tableau 6 Tableau des caractéristiques des examens cytologiques de la norme d'échange ACP (ABIDEC/ARDOC- O.S.I Santé 2009)

Informations	Positions sur la ligne	Nombre de caractères	Type de données
Code Laboratoire	1-2	2	Alphabétique
N° d'enregistrement du frottis	3-12	10	Alphabétique
Nom du prescripteur	13-37	25	Alphabétique
Prénom du prescripteur	38-62	25	Alphabétique
N° prescripteur (ADELI)	63-71	9	Alphabétique
N° prescripteur (RPPS)	72-82	11	Alphabétique
Nom de la patiente	83-107	25	Alphabétique
Nom de JF de la patiente	108-132	25	Alphabétique
Prénom de la patiente	133-157	25	Alphabétique
NSS de la patiente	158-170	13	Alphabétique
Clé	171-172	2	Alphabétique
Date de naissance	173-180	8	Date
Commune Code Postal	181-185	5	Alphabétique
Commune Libellée	186-210	25	Alphabétique
Date du frottis précédent	211-218	8	Date
Date du frottis	219-226	8	Date
Type de frottis	227-227	1	Alphabétique
ADICAP	228-231	4	Alphabétique
Code EVE	232-233	2	Numérique
Mode de frottis	234-234	1	Alphabétique
Qualité du frottis	235-235	1	Alphabétique
Recherche HPV	236-236	1	O/N
Date du test HPV	237-244	8	Date

Nous pouvons dès lors constater que les standards et normes sont nombreux, et très fréquemment propriétaires. Comme nous l'évoquions dans la partie 1.4.3, le challenge est de créer des passerelles entre les différents standards et normes les plus utilisés.

1.5.6 AUDIPOG

Association des Utilisateurs de Dossiers Informatisés en Pédiatrie, Obstétrique et Gynécologie[63] (AUDIPOG) est une association de droit français loi 1901(à non but non lucratif) qui se concentre sur le développement d'un SI en périnatalité avec pour finalité la surveillance de la santé périnatale et l'évaluation des pratiques médicales. L'AUDIPOG est une association incontournable en Auvergne, et en France en termes de suivi périnatal. Ainsi l'un des leaders des logiciels métiers en périnatalité « ICOGEM » qui produit le logiciel « ICOS maternité » s'appuie sur le standard AUDIPOG pour stocker les données relatives aux patientes.

Les services que procurent AUDIPOG sont nombreux : on notera le fait que l'association a créée dès 1994 un réseau sentinelle sur un ensemble de maternité française volontaire qu'elle soit informatisée ou non. Le réseau bénéficie d'un accord CNIL depuis 1995 (avis N° 374982) qui permet d'exploiter un fichier centralisé, pour des raisons de protection vis-à-vis des services administratifs, à minima, seul les accouchements effectués en janvier sont transmis à cette base centralisée. Les dossiers de chaque accouchement sont rendus anonymes puis les 300 variables sont analysées par une étude statistique, qui permet de fournir en décembre les résultats de l'année en cours. Ces résultats peuvent ensuite être comparés aux moyennes nationales, ils permettent l'édition de rapports (France-Périnat 2007) et de graphiques. Ces données peuvent participer à l'évolution des bonnes pratiques hospitalières en permettant l'évaluation des pratiques professionnelles (EPP) décrites par (Vendittelli, Crenn-Hébert, and Tessier 2007) qui répondent aux directives de la HAS.

1.5.7 Integrating the Healthcare Enterprise (IHE)

IHE est une structure européenne qui dispose d'antennes nationales dans une dizaine de pays, dont une pour la France. Son but est de créer un réseau de professionnels et d'éditeurs de logiciels, pour améliorer la façon dont les informations sont échangées. Pour se faire des standards nommés « profils » sont définis. IHE n'est pas une autorité de certification cependant elle tient un registre des logiciels qui ont passé avec succès les différents échanges prévus par un profil précis. De grands noms des systèmes d'informations médicaux comme

[63] http://www.audipog.net/ - date d'accès octobre 2015

Philips healthcare[64], Agfa healthcare[65], General Electric healthcare[66] ou Siemens healthcare[67] sont membres de cette organisation.

L'antenne française de IHE se nomme interop'santé[68], elle a vu le jour en 2001 d'un effort conjoint entre la SFR (Société Française de Radiologie) et le GMSIH (Groupement pour la modernisation du système d'information hospitalier. Cette organisation est une association régie par la loi 1901. C'est l'antenne française qui est à l'origine des premiers Connectathons en Europe. Elle édite aussi régulièrement des guides d'interopérabilité comme (interop'santé 2015) qui permettent d'accompagner l'évolution des systèmes d'informations hospitaliers.

Un Connectathon est un séminaire regroupant de nombreux informaticiens et professionnels de santé dans le but de réaliser des tests grandeur nature pour l'interopérabilité des systèmes d'information de santé. En 2014, 500 ingénieurs ont testé en temps réel la capacité à échangé les informations entre 150 systèmes, durant 5 jours. Le but étant de valider les techniques d'échanges de nouveaux produits avec les logiciels des éditeurs.

Des formations sont aussi prodiguées à des tarifs de 360€ la journée pour les adhérents de l'association, et le double pour les non adhérents. La prochaine formation se déroulera en octobre 2015 et aura pour thème le standard HL7 version 2.5.

Conclusion

Après une présentation du large panorama de projets et des institutions impliqués en e-santé en France et dans le monde, nous avons voulu mettre en évidence dans ce chapitre, la complexité du contexte législatif français autour de la confidentialité des données médicales ainsi que les difficultés de mise en œuvre du DMP en France. La multitude de normes et de standards proposés pour la gestion des données médicales ainsi que la multitude de solutions proposées pour les rendre interopérables rendent incontournable la création d'un recueil des spécifications fonctionnelles et techniques au préalable de toute mise en œuvre d'un projet d'e-santé. Ce recueil des spécifications doit répondre aux besoins immédiats des acteurs de santé tout en respectant des solutions et des standards pérennes à moindre coût. Nous avons donc étudié la faisabilité d'une infrastructure de bases de données (BDD) médicales distribuées, sécurisées et interrogeables par des personnes disposant de différents niveaux d'accréditations depuis une interface graphique ergonomique. Le chapitre suivant se propose donc d'établir le

[64] www.medical.philips.com - date d'accès octobre 2015
[65] http://www.agfahealthcare.com/global/en/main/landing/index.jsp - date d'accès octobre 2015
[66] http://www3.gehealthcare.fr/ - date d'accès octobre 2015
[67] http://www.healthcare.siemens.com/ - date d'accès octobre 2015
[68] http://www.interopsante.org/ - date d'accès octobre 2015

recueil des spécifications technique du projet GINSENG en justifiant les choix des technologies utilisées pour la création d'une infrastructure totalement distribuée d'e-santé et d'épidémiologie pour la région Auvergne.

Chapitre II

– La gestion des données médicales distribuées

Introduction

Dans le cadre du projet ANR TECSAN GINSENG et en partenariat avec le Réseau Sentinelle Cancer Auvergne (RSCA) et le Réseau Santé Périnatalité Auvergne (RSPA), nous avons mis en place un recueil des spécifications du réseau à créer dans le projet pour répondre aux objectifs de partage de données médicales entre plusieurs partenaires médicaux ainsi que de mise à disposition de données médicales depuis des bases de données géographiquement distribuées pour la réalisation d'analyses épidémiologiques. Les partenaires impliqués dans la mise en œuvre du réseau de partage et d'analyse des données médicales pour le projet GINSENG sont l'Association Régionale des Dépistages Organisés des Cancers (ARDOC[69]), l'Association Bourbonnaise Interdépartementale de Dépistage des Cancers (ABIDEC) pour le suivi des cancers du sein et du côlon, l'association ABIDEC-ARDOC, crée en 2009, pour le suivi des cancers du col utérin ainsi que le réseau auvergnat des laboratoires d'anatomie-pathologique privés Sipath-Unilabs dont l'entité mère est localisée à Clermont-Ferrand. Un autre partenaire et bénéficiaire du projet, RSPA, regroupe pour sa part les bases de données médicales de suivi des grossesses des maternités auvergnates, elles-mêmes regroupées dans une base de données unique régionale. Dans ce chapitre, après avoir expliqué le périmètre d'action du projet, nous recensons les systèmes d'information que nous voulons interconnecter et les problématiques d'utilisation qui en découlent. Une deuxième partie est consacrée au recueil des spécifications pour le projet ; celui-ci concerne en particulier les problématiques liées au réseau, à la standardisation des bases de données médicales, à l'identification des patients,

[69] http://www.ardoc.org/ - date d'accès octobre 2015

l'authentification des utilisateurs du réseau et l'interface graphique d'accès aux fonctionnalités du réseau.

2.1 Le projet GINSENG

2.1.1 Genèse

Les technologies de grilles informatiques sont apparues au début des années 2000 pour faciliter les communications et les collaborations entre les chercheurs et pour répondre à leurs besoins spécifiques en matière de calcul et stockage. La Commission Européenne a financé les premières recherches de « grille » au travers du cinquième programme cadre de recherche (FP5), avec plus de 50 millions d'euros. À cette époque, l'objectif principal de l'*Information Society Technologies* (IST) était le projet européen Datagrid (Gagliardi et al. 2002) ayant pour but de développer un middleware capable de répondre aux défis de trois communautés scientifiques différentes : celle de la physique des hautes énergies, celle de l'observation de la Terre, et celle des applications biomédicales. En 2004, l'Union européenne a lancé plusieurs initiatives de mises en œuvre de grilles informatiques contribuant à façonner les infrastructures européennes ; le projet EGEE (*Enabling Grids for E- Science in Europe*) (Gagliardi 2005) dans le cadre du projet FP6 fut l'un des projets majeurs. Le Laboratoire de Physique Corpusculaire (LPC) a été impliqué dès 2002 dans ces projets. Il en a été de même depuis 2006, avec le lancement des initiatives de grilles informatiques régionales (AuverGrid pour l'Auvergne, l'infrastructure de calcul distribué Strasbourg Grand-Est, Grille au service de la Recherche en Ile-de-France (GRIF) ou l'interconnexion grid5000[70]) pour les besoins des calculs scientifiques des universitaires, des chercheurs ainsi que des entreprises privées. En 2008, l'initiative auvergnate LifeGrid pour l'utilisation des infrastructures réseau et des outils pour les sciences de la vie a permis de financer la première version d'un portail web dédié à l'utilisation de l'infrastructure de grille EGEE pour les besoins de planifications de traitement des cancers par simulation Monte Carlo ; il s'agissait du projet HOPE (HOspital Platform for E-health) porté par le LPC. Puis, en 2008, le LPC a commencé à investiguer un projet d'envergure pour la création d'un réseau sentinelle pour le dépistage du cancer en Auvergne. L'association RSCA (Réseau Sentinelle Cancer Auvergne) s'est constituée à la même période afin de piloter ces travaux de recherche et en confier la maîtrise d'œuvre au LPC. Cette association regroupe différents partenaires médicaux, des épidémiologistes et des

[70] https://www.grid5000.fr/ - date d'accès octobre 2015

chercheurs afin de faciliter le partage des données médicales entre les associations de dépistage du cancer, les laboratoires d'anatomo-cyto-pathologie (ACP), les centres anti cancéreux et les hôpitaux. En 2008, une thèse co-encadrée par le LPC et le département de santé publique du CHU de Clermont-Ferrand a été financée par le conseil régional d'Auvergne, afin de créer une liste d'exigences pour la création d'un réseau distribué pour le partage de données médicales (De Vlieger et al. 2009; De Vlieger et al. 2010; Catherine Quantin et al. 2009). En 2010, une proposition nommée GINSENG (*Global Initiative for Sentinel E-health Network on Grid*[71]) (Cipière, De Vlieger, et al. 2012)(Cipière, Gaspard, et al. 2012) pour créer un réseau dédié à l'e-santé et de l'épidémiologie en Auvergne a été financé par le programme TECSAN de l'Agence Nationale de la Recherche (ANR[72]). Ce projet de 3 ans avait pour objectif de montrer la valeur ajoutée d'une telle infrastructure distribuée pour le partage des données médicales et la veille sanitaire.

À l'origine de ce projet les premiers besoins pour les associations de dépistage organisé des cancers ont été manifestés par l'intermédiaire du docteur André Lautier, qui est président de l'ARDOC depuis sa création le 2 mars 2009, et qui cherchait à améliorer la façon dont étaient communiquées les informations médicales (mammographies et dossiers médicaux, notamment les compte-rendus anatomo-pathologiques) aux associations. Un des éléments moteur de la démarche des associations était de trouver un moyen peu coûteux (financièrement mais aussi en moyens humains) pour le transfert des informations médicales comparativement à un registre des cancers dont la maintenance et l'agrégation de données médicales sont plus coûteuses. De par leur capacité à générer des informations sur l'incidence des cancers en région Auvergne, les associations de dépistage organisé et les laboratoires d'anatomie-pathologique sont devenus des interlocuteurs privilégiés pour les instances régionales et nationales de veille sanitaire : l'Agence Régionale de Santé (ARS) en Auvergne et l'Institut National de Veille Sanitaire (InVS). Il s'avérait donc également nécessaire de proposer une solution informatique capable de requêter des bases de données médicales réparties afin de générer des analyses épidémiologiques d'intérêt.

Par la suite, en 2011, le projet GINSENG a été étendu au suivi des grossesses en région par l'intermédiaire d'un partenariat avec les maternités auvergnates du réseau RSPA, sous la responsabilité du professeur Françoise Vendittelli. Ce réseau regroupe au sein d'une base de

[71] https://e-ginseng.com/ - date d'accès octobre 2015

[72] http://www.agence-nationale-recherche.fr/ - date d'accès octobre 2015

données régionale, nommée ICOS, toutes les informations médicales liées aux parturientes et leurs enfants nouveau-nés.

2.1.2 Contexte

Le cancer première cause de mortalité en France

Le cancer, avec 147 500* décès par an, est en 2015 la première cause de mortalité en France, devant les maladies cardio-vasculaire (140 000* décès) et les drogues (94 000* décès). Si pour certains cancers, comme le cancer du poumons, l'intérêt d'un dépistage organisé, même chez les personnes à risque comme les fumeurs de plus de 60 ans, reste discutable, d'autres cancers, comme le cancer du sein ou le cancer colorectal, le dépistage permet de repérer les personnes présentant des anomalies évocatrices d'une lésion cancer ou d'un stade précancéreux afin de les adresser rapidement à une structure appropriée pour un diagnostic et un traitement nécessaire le cas échéant. Les programmes de dépistage sont particulièrement efficaces pour les cancers fréquents, pour lesquels on dispose d'un test économique, d'un coût abordable, acceptable et accessible pour la majorité de la population exposée. En outre, d'après l'OMS, le nombre de nouveaux cas détectés chaque année devrait augmenter de 70% dans les 20 prochaines années.

À titre de comparaison, les accidents de la circulation ont provoqué 3 388 décès en 2014[73] et la grippe[74], 18 000.

Une détection précoce des cancers permet de sauver des vies, c'est la raison de l'existence des associations de dépistage organisé du cancer comme l'ARDOC et l'ABIDEC pour l'Auvergne. Le but du projet GINSENG est de fournir des recherches et des échanges automatiques d'informations médicales pour faciliter et améliorer le suivi des patients. Pour se faire, l'outil informatique doit être capable d'identifier un patient au sein des bases de données, repérer les informations pertinentes et essentielles qui concernent ce patient, et retourner le résultat de ces recherches facilement et rapidement à l'utilisateur qui en avait fait la demande. L'ensemble de ces opérations doit s'effectuer sur un réseau sécurisé qui permet de garantir un accès réglementé aux données médicales.

* http://www.insee.fr/fr/bases-de-donnees/bsweb/serie.asp?idbank=000436394 - date d'accès octobre 2015

[73] http://www.interieur.gouv.fr/ - date d'accès octobre 2015

[74] http://www.pasteur.fr/fr/institut-pasteur/presse/fiches-info/grippe - date d'accès octobre 2015

Le cancer colorectal

Le cancer colorectal, ou cancer du côlon-rectum, touche 42 000 personnes et est à l'origine de plus de 17 500 décès chaque année en France. Les femmes comme les hommes sont concernées par le dépistage organisé. La population est classée en 3 groupes relatifs au niveau de risque de développer un cancer colorectal « moyen » (80 %), « élevé » (15 à 20 %) et « très élevé » (1 à 3 %). L'histoire personnelle ainsi que les antécédents familiaux influents beaucoup sur le positionnement de chaque individu, par le médecin généraliste, à l'intérieur d'un groupe. S'il est détecté tôt, le cancer colorectal se guérit dans 9 cas sur 10[75]. Ce qui fait donc du cancer colorectal un excellent candidat pour le dépistage organisé. Nous avons résumé les différentes étapes du dépistage organisé dans la Figure 18.

La population cible de ce dépistage est composée d'hommes et femmes âgés de 50 à 74 ans. Ils reçoivent une invitation à se rendre chez leur médecin généraliste pour retirer le kit de dépistage gratuit. Une fois le patient en possession du kit de dépistage il peut réaliser le test à son domicile. Le but étant de recueillir une infime partie des selles sur un dispositif de prélèvement. Une fois expédié par voie postale à un laboratoire, l'échantillon est analysé en vue de déceler la présence de sang dans les excréments qui est un marqueur indiquant de probables polypes qui méritent un examen plus approfondi. Ce test est analysé positif dans seulement 4% des cas. Si le test est positif, une coloscopie est programmée pour déceler et retirer les éventuels polypes avant qu'ils n'évoluent en cancer. Si les résultats sont négatifs, aucune suite n'est donnée, la démarche est renouvelée tous les 2 ans.

[75] http://www.e-cancer.fr/Comprendre-prevenir-depister/Se-faire-depister/Depistage-du-cancer-colorectal - date d'accès octobre 2015

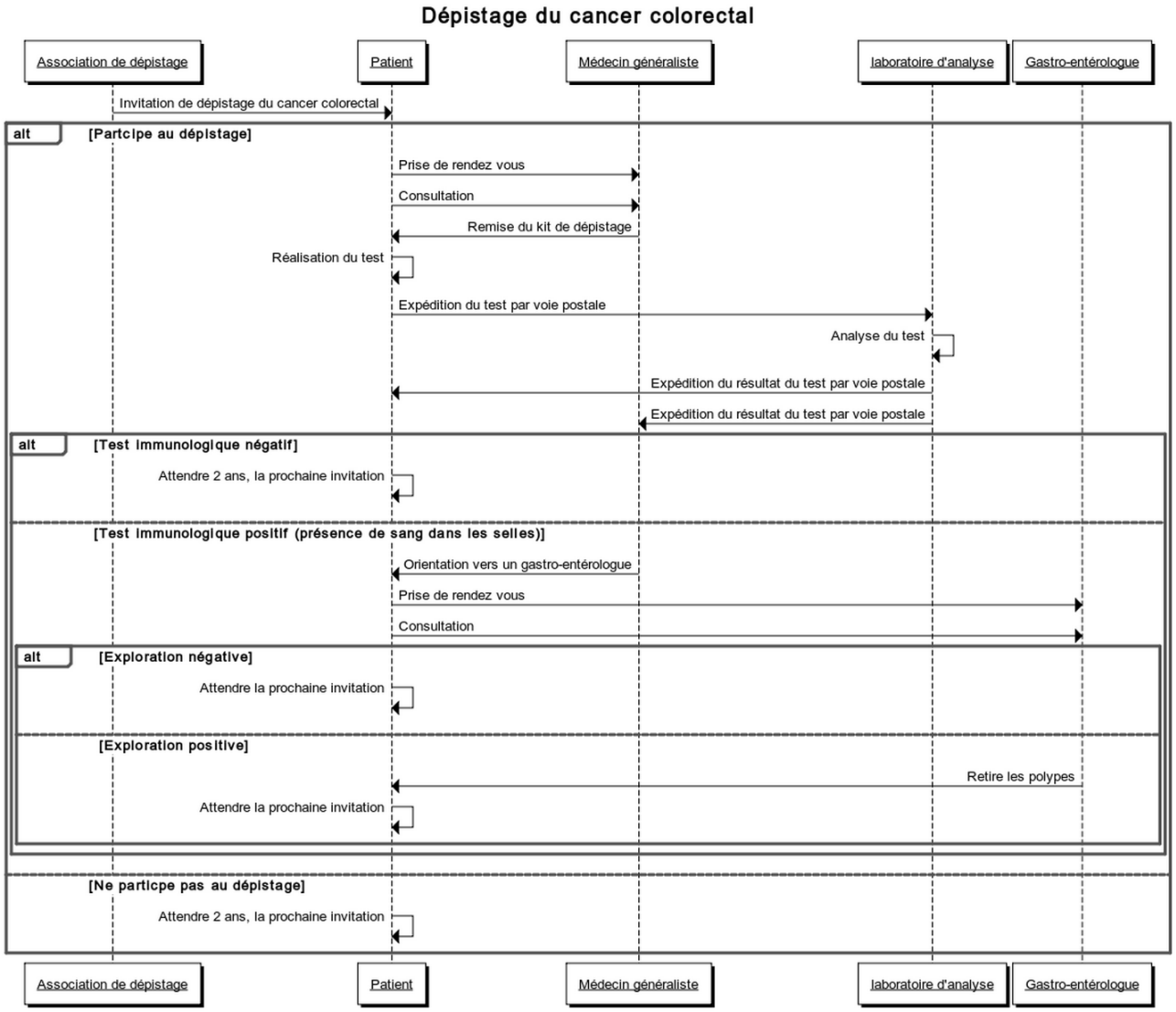


Figure 18 Déroulement d'une invitation pour le dépistage du cancer colorectal par une association de dépistage organisé du cancer

Le cancer du col de l'utérus

Le cancer du col de l'utérus touche environ 3000 personnes par an dont 1000 décèdent, ces chiffres sont stables depuis 10 ans. Le cancer du col de l'utérus est le 4ème cancer de la femme en termes de fréquence alors qu'il reste le 1er dans les pays sous-développés n'ayant pas encore mis en place de politique de dépistage organisé. Détecté tôt, ce cancer peut facilement être évité ("British Medical Journal" 1986) à hauteur d'une diminution de la fréquence de l'ordre de 93.5% en cas de dépistage annuel. Ici encore, ces résultats font du cancer du col de l'utérus un excellent candidat au dépistage organisé du cancer. Avant même les lésions cancéreuses, l'anatomopathologie peut repérer des dysplasies qui sont des lésions précancéreuses. Le prélèvement peut être réalisé par tout médecin, il consiste à recueillir des cellules chez la patiente au court d'un examen peu invasif. Ce prélèvement se nomme Frottis Cervico-Vaginal (FCV), il consiste à étaler sur une lame des cellules prélevées pour qu'elles soient analysées par un cabinet anatomo-pathologie. Le résultat de cette analyse peut être classifié à l'aide du code ADICAP (cf. 1.5.4) ou de la classification de Bethesda (cf. Tableau 7). Cette analyse permet de repérer des carcinomes in situ qui sont des lésions cancéreuses précoces qui ne touchent qu'une partie superficielle de la muqueuse vaginales ou utérine, généralement facile à circonscrire et éradiquer. Il est causé par une infection liée à un virus, de la famille des papilloma virus (HPV), et évoluant sur plusieurs années. Les femmes sont concernées dès le début de leur vie sexuelle. La fréquence des frottis est d'une fois par an les 3 premières années puis si les résultats sont normaux, les frottis sont espacés de 2 à 3 ans.

Le cancer du sein

En 2010, 52 000 nouveaux cas ont été décelés, c'est le cancer le plus fréquent chez la femme en France. En 2004, seuls 37% des 8 millions de françaises âgées de 50 à 74 ans ont participé au dépistage. On estime que 3 000 femmes pourraient ne pas décéder si le dépistage organisé mobilisait au moins 70% des femmes de 50 à 74 ans, chaque année. Ce chiffre est l'équivalent du nombre de décès par accident de la route en France chaque année. Une invitation à pratiquer une mammographie est envoyée tous les 2 ans depuis le 1er janvier 2004, date de la généralisation du dépistage organisé du cancer du sein. Une mammographie est une radiographie (rayons X) des seins (cf. Figure 19), pratiquée par un radiologue. Elle permet d'obtenir une image de l'intérieur des seins. La lecture de cette image par un spécialiste peut révéler la présence d'anomalies cancéreuses ou précancéreuses. Le résultat de cette analyse visuelle est classifié à l'aide du *Breast Imaging Reporting And Data System* (BIRADS) de

l'*American College of Radiology* (ACR[76]). Cette classification comporte 6 niveaux, de ACR 0 à ACR 5, en fonction du niveau de suspicion du cancer du lecteur qui analyse la mammographie (cf. Tableau 7). Cette classification d'une mammographie permettra d'orienter la patiente vers une biopsie si le cliché laisse supposer un risque potentiel trop important. Une biopsie consiste en en prélèvement de tissu, dans le cadre d'une biopsie mammaire c'est un radiologue qui prélève avec une aiguille une partie de la lésion que l'on souhaite catégoriser avec précision. Ce prélèvement sera envoyé dans un cabinet d'anatomie-pathologique pour examen. Un code ADICAP sera alors ajouté au dossier pour classer ce prélèvement en fonction de son niveau de gravité.

Tableau 7 Classification ACR et suites recommandées

ACR	Constatations	Recommandations
0	Exploration incomplète	
1	Mammographie normale	Ne rien faire
2	Mammographie typiquement bénigne	Ne rien faire
3	Mammographie probablement bénigne	Contrôle à 6 mois
4	Anomalie suspecte	Biopsie
5	Anomalie évocatrice de cancer	Biopsie

Actuellement, les images sont toujours transférées sous format papier pour respecter les contraintes légales, bien que l'imagerie numérique de haute définition soit disponible. Les professionnels sont actuellement en attente du législateur pour qu'il permette de faire évoluer leurs pratiques à l'aide des nouvelles technologies. En octobre 2006 la HAS[77] considère « que rien ne s'oppose à l'introduction de la mammographie numérique dans le dépistage organisé » (HAS 2006). En effet, avec l'avènement des écrans de très haute définition (du type des écrans dits 4K de 3840x2160 pixels et de 24 à 32 pouces, les mammographies numériques peuvent être étudiées avec beaucoup de précision. Les études actuelles (Cole et al. 2004; Skaane and Skjennald 2004; Pisano et al. 2005) ne montrent pas un avantage qualitatif de l'une ou l'autre des deux méthodes. L'intérêt du numérique réside surtout dans le fait de ne pas avoir d'impression à réaliser et dans ses caractéristiques de partage lié au support numérique, dont GINSENG pourrait être l'un des vecteurs.

[76] http:// www.acr.org - date d'accès octobre 2015

[77] http://www.has-sante.fr/portail/jcms/c_461657/fr/place-de-la-mammographie-numerique-dans-le-depistage-organise-du-cancer-du-sein - date d'accès octobre 2015

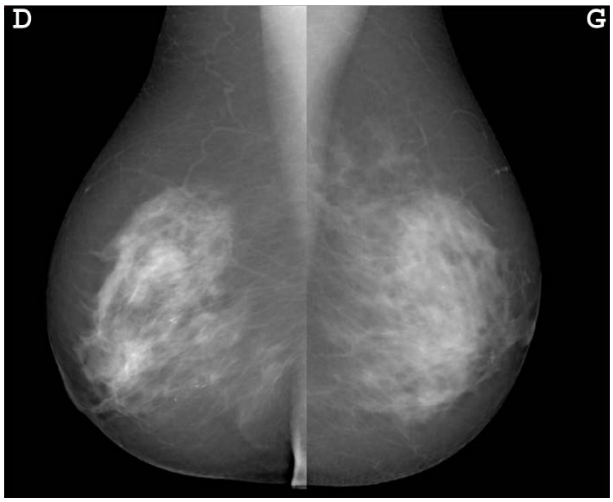


Figure 19 Mammographie seins Gauche et Droit (source : SFR)[78]

La périnatalité

Le parcours de soins d'une future mère peut l'amener à consulter dans différents services de différentes villes, et accoucher dans une maternité dont les services sont les mieux adaptés à son bien-être et à celui de son/ses enfant(s). L'Auvergne à la chance de disposer du RSPA qui est un réseau ville/hôpital, qui depuis 2008 a déployé un dossier informatisé d'obstétrique qui accompagne la patiente auprès des professionnels membres de RSPA.
Les avantages de cette solution sont multiples :

- Améliorer la qualité de la prise en charge des femmes enceintes et de leur(s) enfant(s),
- Optimiser la circulation des informations médicales dans l'intérêt des patients,
- Améliorer les pratiques professionnelles après évaluation de certains indicateurs.

GINSENG doit permettre de poursuivre cette dynamique en agrégeant de nouvelles données notamment celles de laboratoires d'analyses relatives à la détection de la trisomie 21. Là encore, on pourra noter l'encadrement légal des pratiques et son évolution permanente car l'arrêté du 23 juin 2009[79], modifié par celui du 27 mai 2013[80], fixe les règles de bonnes pratiques en matière de dépistage et de diagnostic prénatals avec l'utilisation des marqueurs sériques maternels de la trisomie 21. Les modifications apportées en 2013 visent à préciser le

[78] http://pe.sfrnet.org/Data/ModuleConsultationPoster/pdf/2004/1/329574b9-97cc-4eab-8e1c-0939a5892598.pdf - date d'accès octobre 2015
[79] http://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000020814373 - date d'accès octobre 2015
[80] http://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000027533804 - date d'accès octobre 2015

contenu et les destinataires, de la transmission des données, et attribuent à l'Agence de la biomédecine (ABM) l'évaluation du dépistage de la trisomie 21 au niveau national.

L'intégration de GINSENG au sein de l'ENRS Auvergnat pourra permettre d'offrir de nouveaux accès en consultation de la base « ICOS Régionale » hébergée par le HADS IDS. Pour le moment, cette base a été prévue pour agréger les données des différents participants au RSPA et utilisant les solutions développées par la société ICOGEM[81], mais il semble que les problématiques d'interrogation évolutive n'aient pas été considérées lors de la rédaction du cahier des charges de cette solution. En effet si le stockage des données a été considéré rien n'a été prévu pour que les données soient requêtées, par les professionnels de santé.

2.1.3 Les objectifs du projet

L'objectif principal du projet GINSENG est la mise en œuvre d'un système de médiation entre différents systèmes d'information (SI) médicale. Cet objectif peut être décomposé en de nombreuses sous-parties techniques : installation physique du matériel et des éléments réseaux, création de machines virtuelles, configuration logicielle. De la même façon nous pouvons définir des sous objectifs fonctionnels.

Objectifs fonctionnels

- Gain de temps et réduction des coûts

Ce projet a pour but d'améliorer la qualité de suivi des patients, tout en faisant gagner du temps aux différents services grâce à l'automatisation du traitement de l'information et de sa distribution. En limitant au minimum les saisies au clavier des dossiers, l'insertion d'erreur de frappe est ainsi réduite, ce qui se traduit par une augmentation de la qualité de l'information. Les données sont accessibles plus rapidement, ce qui permet un suivi des populations plus réactif, voir en temps réel. Grâce à un niveau d'exigence très élevé en matière de sécurité, les données médicales sont manipulées dans un environnement sûr.

- Amélioration de la qualité des données stockées

Une meilleure transmission de l'information des dossiers des patients entre les structures médicales permet à chaque soignant de prodiguer des soins plus appropriés, plus rapidement. Par exemple, des analyses déjà pratiquées ne sont plus prescrites une seconde fois par manque de résultats. Avec plus d'information disponible, les faisceaux de présomptions sont créés avec plus de certitudes. Les coûts financiers et humains engendrés pour une pathologie sont réduits.

[81] http://www.icogem.fr/ - date d'accès octobre 2015

L'impact économique est moindre pour le patient, l'assurance maladie et les mutuelles. Le médecin dispose de plus de temps pour se consacrer à d'autres cas.

Une transmission rapide et uniformisée de l'information permet également une meilleure comparaison des établissements de santé. Les entités de décision sont mieux informées de la situation et des pratiques. La création et le partage d'indicateur a la capacité de mettre en exergue des pratiques qui diffèrent de la norme et sont peut-être dictées par la facilité de mise en œuvre plutôt que par l'intérêt du patient.

❖ Production de nouveaux indicateurs

L'un des avantages de GINSENG par rapport aux outils déjà existants est sa capacité à croiser différentes bases de données. Ainsi, on peut facilement envisager d'intégrer des données provenant de l'Open Data ou de toute autre source de données. Dans le cadre de nos études autour du cancer nous pensons notamment intégrer à terme des cartes de l'Institut de Radioprotection de Sûreté Nucléaire (IRSN[82]) et du Bureau de Recherche Géologique Minière (BRGM[83]), notamment sur la thématique du risque d'exposition au gaz radon[84]. Ce qui nous permettrait de comparer la présence potentielle de radon avec l'incidence des cancers des poumons. Cet exemple peut être transposable avec les tracés des lignes à haute tension à proximité des lieux d'habitations ou toute autre cartographie. Dans une optique similaire qui vise à s'assurer de la sécurité des populations, la communauté d'agglomération de Clermont-Ferrand nous a contactés pour envisager l'utilisation de notre solution informatique pour suivre l'évolution des potentiels cancers suite à l'implantation d'une station de traitement des déchets. L'évolution ou la stagnation des indicateurs pourrait dans ce cas mettre en évidence l'impact ou la neutralité des émissions de cette structure sur la santé des populations.

Objectifs techniques

❖ Une solution Open Source

L'intérêt économique de s'appuyer sur des systèmes libres et gratuits permet de ne pas dépenser des centaines voire des milliers d'euros de licences pour chaque serveur équipant les partenaires du projet GINSENG comme résumé dans le Tableau 8 où sont présentées des gammes tarifaires pour un système d'exploitation, un hyperviseur et un SGBD. Les tarifs que

[82] http://www.irsn.fr/FR/Pages/Home.aspx - date d'accès octobre 2015

[83] http://www.brgm.fr/ - date d'accès octobre 2015

[84] https://www.data.gouv.fr/fr/reuses/irsn-connaitre-le-potentiel-radon-de-sa-commune/ - date d'accès octobre 2015

nous présentons dans ce tableau sont liés aux différentes options envisageables, ainsi qu'aux différents producteurs et fournisseurs de ces logiciels. A titre d'exemple, si nous souhaitions équiper chaque serveur du réseau GINSENG avec une solution de virtualisation vSphere de chez VMware à 894,50€[85] pour créer trois machines virtuelles utilisant Windows 10 pro x64 à 279€[86] x 3 = 537€ avec une installation Oracle Standard Edition One à 5 000€[87]. La somme de cette installation (hors coût du serveur) avoisine les 6 500€. Avec une solution Open Source, nous pourrons donc équiper 4 fois plus de sites pour la même somme.

Tableau 8 Économies réalisées grâce à l'utilisation de logiciels Open Source

Logiciel	Tarif en euros
Système d'exploitation	100 à 800
Hyperviseur	700 à 900
SGBD	5 000 à 18 000

De façon très similaire si l'on considère une machine virtuelle pour réaliser le rôle du routeur VPN et du firewall, une économie de 150 à 1000€ par site peut être envisagée. Il s'agira ici de remplacer un composant matériel par une solution virtualisée. Ce remplacement peut toutefois être discuté en termes de sécurité qui serait prise en charge dans son intégralité par l'hyperviseur.

Outre l'aspect financier, le choix de l'Open Source permet une meilleure distribution de notre solution. En effet, nous pouvons proposer aux professionnels de santé une solution n'obligeant aucun investissement préalable.

❖ Un réseau généralisable

Nous venons de l'évoquer la généralisation de l'infrastructure GINSENG est l'un des objectifs du projet. Réfléchi dans un premier temps pour la cancérologie et la périnatalité, le projet doit être capable d'être étendu à d'autres disciplines confrontées aux mêmes difficultés de transmission, d'échange et de partage d'information. Son périmètre d'action a aussi vocation à s'étendre à l'imagerie médicale avec les mammographies ou l'imagerie microscopique d'anatomie-pathologique.

[85] http://www.vmware.com/fr/products/vsphere/pricing.html - tarif en septembre 2015
[86] http://www.microsoftstore.com/store/msfr/fr_FR/pdp/productID.320443800 - tarif en septembre 2015
[87] http://www.oracle.com/us/corporate/pricing/technology-price-list-070617.pdf - tarif en septembre 2015

❖ Un niveau de sécurité fiable

Les aspects relatifs à la sécurité des données médicales sont essentiels pour garantir la confidentialité des échanges. Toutes les mesures nécessaires sont mises en œuvre tant d'un point de vue physique que logiciel. Des audits extérieur et indépendant seront envisagés pour valider les bonnes pratiques que nous appliquons pour répondre aux obligations légales et assurer un niveau de sécurité fiable. Contre des défaillances involontaires ou des attaques programmées nos solutions doivent permettre de maintenir les données confidentielles et inaltérables.

2.1.4 Les acteurs du projet

Pour mener à bien tous les objectifs cités précédemment de nombreux acteurs travaillent de concert à la réalisation de l'infrastructure informatique.

Les partenaires privés

Figure 20 Logos des partenaires privés du projet GINSENG

En dépit des solutions Open Source choisies pour établir l'infrastructure GINSENG, il apparaît nécessaire de commercialiser à terme la maintenance informatique de chaque site équipé. Dans le projet GINSENG, nous nous sommes appuyés sur l'expérience et l'expertise en technologie de grille de calcul de la société Gnúbila France [88] (anciennement MaatG France) impliquée notamment dans différents projets européens comme le projet NeuGRID for Users (N4U[89]) pour la mise en œuvre d'une cartographie du cerveau ou le projet MD-PAEDIGREE[90] (*Model-Driven PAediatric European DIGital REpository*) pour créer un outil de prédiction des pathologies pédiatriques.

D'autres acteurs privés ont rendu possibles la faisabilité du projet ; ceux-ci sont des prestataires de service ayant en charge la gestion logicielle et/ou réseau du parc informatique

[88] https://www.gnubila.fr/ - date d'accès octobre 2015
[89] https://neugrid4you.eu/ - date d'accès octobre 2015
[90] http://www.md-paedigree.eu/ - date d'accès octobre 2015

des partenaires du projet. Nous avons donc eu recours à ces sociétés pour la réalisation des câblages réseau, l'acheminement du courant pour l'alimentation des machines ou la réalisation de logiciels métiers.

Nous avons travaillé avec la société O.S.I Santé [91] spécialisée dans les logiciels de gestion à destination des Structures de Gestion du Dépistage Organisé (SGDO). Cette société équipe aujourd'hui 31 départements en France métropolitaine et outre-mer. L'interface logicielle développée pour les SGDO se nomme « ZEUS » ; elle est disponible sous Windows et permet la gestion administrative de la SGDO, le suivi médical des patients, l'édition de statistiques et favorise les échanges d'informations. O.S.I Santé partage avec nous son expérience en matière de gestion de bases de données médicales.

Concernant la gestion sémantique des bases de données médicales, de par l'utilisation de triplets RDF permettant de rajouter du sens aux données par l'utilisation de métadonnées ; nous avons collaboré avec la société Mnémotix[92], entreprise innovante en Ingénierie des Connaissances et du WEB Sémantique, travaillant sur la distribution des requêtes sémantiques.

Les partenaires institutionnels et associatifs

Figure 21 Logos des partenaires institutionnels et associatifs

Le projet ANR GINSENG est financé pour partie par l'ANR dans le cadre du programme Technologies pour la santé et l'autonomie (TecSan) 2010, sur une durée de 44 mois, d'avril 2011 à décembre 2014. Avec un budget global d'environ 800 000 euros.

[91] http://www.osi-sante.fr/ - date d'accès octobre 2015
[92] http://mnemotix.com/blog/ - date d'accès octobre 2015

Le Centre National de la Recherche Scientifique (CNRS[93]) et les deux universités clermontoises, l'Université Blaise Pascal (UBP[94]) et l'Université D'Auvergne (UDA[95]) qui hébergeant les laboratoires et équipes auvergnats parties représentent une part importante des personnels impliqués dans le projet avec :

Le laboratoire de Physique Corpusculaire (LPC[96]) UMR 6533-UBP, a coordonné le projet et mis à disposition les ressources techniques et humaines nécessaires à la bonne gestion de ce projet.

Le Laboratoire d'Informatique, de Modélisation et d'Optimisation des Systèmes (LIMOS[97]) Unité Mixte de Recherche UMR 6158-UBP qui a fourni son support en matière d'informatique et de sciences de l'information. L'apport du LIMOS a été essentiel durant la phase de prototypage grâce à la mise à disposition de matériels permettant de simuler les différents sites à interconnecter.

Le Centre Régional de Ressources Informatiques CRRI, qui a mis à notre disposition des solutions réseaux entre les différents laboratoires du campus des Cézeaux pour simuler le prototype réseau du projet GINSENG.

Le service de Santé Publique du CHU de Clermont-Ferrand – UDA en charge d'une partie des études épidémiologiques du projet.

Les équipes de l'unité recherche Périnatalité, grossesse, Environnement, PRAtiques médicales et DÉveloppements (PEPRADE[98]) Établissement d'Accueil EA 4681 - UDA-CNRS, en charge des études épidémiologiques sur le suivi des femmes enceintes.

L'Institut des Sciences de l'Image pour les Techniques interventionnelles ISIT[99] UMR 6284-UDA-CNRS, qui s'est intéressé aux algorithmes d'identification des patients à l'intérieur du SI, en considérant notamment leur vitesse et leur précision sur les traits d'identification retenus.

Le Centre de Recherche en Acquisition et Traitement de l'Image pour la Santé (Creatis[100]) UMR 5220 est l'un de nos partenaires depuis mai 2013. Les collaborateurs issus de ce laboratoire se sont principalement focalisés sur les requêtes de type SPARQL EndPoint au

[93] http://www.cnrs.fr/ - date d'accès octobre 2015
[94] http://www.univ-bpclermont.fr/ - date d'accès octobre 2015
[95] http://www.u-clermont1.fr/ - date d'accès octobre 2015
[96] http://clrwww.in2p3.fr/ - date d'accès octobre 2015
[97] http://limos.isima.fr/ - date d'accès octobre 2015
[98] http://www.u-clermont1.fr/peprade.html - date d'accès octobre 2015
[99] http://isit.u-clermont1.fr/fr - date d'accès octobre 2015
[100] http://www.creatis.insa-lyon.fr/site/ - date d'accès octobre 2015

travers du portail VIP (www.creatis.insa-lyon.fr/vip/) et intégré au sein du site Internet du projet www.e-ginseng.com en Liferay.

Le laboratoire d'Informatique, Signaux et Systèmes de Sophia Antipolis (I3S[101]) partage avec nous son expertise en matière de traitements sémantiques des bases de données notamment au travers des développements menés en coopération avec la société Mnémotix qui développe des entrepôts de données pouvant être requêtés en langage SPARQL.

L'Agence Régional de Santé (ARS) Auvergne[102] qui gère la politique de santé en région, nous assure de son soutien stratégique et opérationnel et depuis 2015 prend en charge une partie des frais de personnel, pour poursuivre la mise en œuvre du projet. C'est notamment grâce à l'ARS que nous avons pu nous rapprocher du GCS SIMPA pour pérenniser notre solution. Le GCS a partagé son expérience et ses contacts avec nous, notamment lors du choix du HADS.

L'ARDOC, l'ABIDEC, l'ABIDEC-ARDOC, les 3 associations auvergnates en charge du dépistage organisé du cancer qui sont à l'origine du projet. Ces organisations se focalisent sur les cancers de sein, le cancer colo rectal, ainsi que le cancer du col de l'utérus. Elles couvrent la totalité de la population auvergnate.

Les patients

Les populations cibles du projet sont d'une part les femmes enceintes et d'autre part les personnes ciblées par la prévention du cancer. À terme ces populations pourront être élargies et étendues à d'autres tranches d'âges ou d'autres pathologies, jusqu'à la totalité de la population d'un territoire.

Le RSPA se concentre sur le suivi des femmes enceintes en région Auvergne. Le but étant de pouvoir faire suivre les dossiers entre les différents lieux de consultations et la maternité où la patiente accouche. Des études sociodémographiques telles que l'incidence de la précarité en zone rurale sur le suivi d'une grossesse Prugnancy (Precariousness in RUral areas during preGNANCY) pourront être menées dans la suite de ce volet dédié à la périnatalité. Le RSCA s'intéresse quant à lui au dépistage organisé du cancer du col de l'utérus, du cancer du sein ainsi que du cancer du côlon.

Les outils informatiques que nous développons ont pour but d'améliorer les processus de transmission des dossiers et la connaissance de répartition géographique de certaines pathologies. Une meilleure connaissance des bassins de population et des risques potentiels (on

[101] http://www.i3s.unice.fr/I3S/index.php - date d'accès octobre 2015
[102] http://www.ars.auvergne.sante.fr - date d'accès octobre 2015

peut imaginer la mise en évidence d'un agent cancérigène) permettrait de mieux protéger la population. En Auvergne on peut envisager la proximité d'une industrie toxique, ou une surexposition à des contaminants naturels comme le gaz radon qui pourrait augmenter la fréquence de cancers des poumons en cas d'une exposition prolongée à une concentration importante.

Tout est mis en œuvre pour protéger les droits et l'anonymat du patient selon les recommandations de la CNIL. Le patient peut à tout moment demander à ce que les informations relatives à ces dossiers médicaux ne soient pas traitées par GINSENG. Pour se faire nous mettons à sa disposition une notice d'information leur permettant d'exprimer leur droit de retrait du projet la Figure 22 présente le recto de cette plaquette, alors que les Figure 23 et Figure 24 représentent les versos dédiés respectivement aux applications cancer et périnatalité.

SURVEILLANCE DU CANCER

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE

ANR

Tout savoir sur GINSENG :

Site internet :

e-ginseng.org

Renseignements :

Lydia Maigne

☎ : +33 4 73 40 51 23

✉ : Laboratoire de Physique Corpusculaire
24, Avenue des Landais
BP 80026
63171 Aubière Cedex

@ : Lydia.Maigne@clermont.in2p3.fr

GINSENG
SENTINELLE
GRID NETWORK

Projet GINSENG
Informations Patients

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE

ANR

Figure 22 Visuel de l'accord patient recto RSCA

Le projet GINSENG

Le projet GINSENG vise à mettre en place une plateforme informatique permettant l'interrogation et le croisement de bases de données mises en œuvre dans le cadre du suivi médical des patients à des fins d'évaluation des pratiques de soins et de recherche épidémiologique.

Le suivi du cancer du sein en région Auvergne est l'une des applications pilotes du projet.

A ce titre, les informations recueillies lors de votre prise en charge, feront l'objet, sauf opposition de votre part, d'un enregistrement informatique réservé à l'usage des professionnels de santé adhérents au projet. Ces professionnels de santé, par l'intermédiaire du Réseau Sentinelle Cancer Auvergne (RSCA), se tiennent à votre disposition pour vous communiquer ces renseignements ainsi que toutes informations nécessaires sur votre état de santé*. GINSENG garantit que toutes les mesures sont prises pour garantir la confidentialité de vos données de santé.

Toutes les informations relatives à ce projet sont à votre disposition sur :
Site internet : e-ginseng.org
Renseignements téléphoniques : 04.73.43.06.61

*conformément à l'article 32 de la loi du 6 janvier 1978 modifiée

Vous avez le choix !

Si toutefois, vous ne souhaitez pas que vos données de santé soient utilisées dans ce projet, vous avez la possiblilité à tout moment de refuser la transmission de vos données sans conséquence pour votre prise en charge médicale. Merci de renvoyer par courrier papier ou email à l'association RSCA le coupon réponse ci-dessous.

Je refuse que mes données de santé soient utilisées dans le cadre du projet GINSENG pour le suivi du cancer du sein.

NOM :.....

Prénom : .....

Date de naissance :.....

Signature :

Coupon réponse à renvoyer à :

✉ : Association RSCA
7 rue Edith Piaf
63100 Clermont-Ferrand

@ : chantal.mestre@ardoc.org

Figure 23 Accord patient RSCA

Le projet GINSENG

Le projet GINSENG vise à mettre en place une plateforme informatique permettant l'interrogation et le croisement de bases de données mises en œuvre dans le cadre du suivi médical des patients à des fins d'évaluation des pratiques en santé et de recherche épidémiologique.

La pertinence des césariennes en région Auvergne est l'une des applications pilotes du projet.

A ce titre, les informations recueillies lors de votre prise en charge, feront l'objet, sauf opposition de votre part, d'un enregistrement informatique réservé à l'usage des professionnels de santé adhérents au projet. Ces professionnels de santé, par l'intermédiaire du professeur F. Vendittelli, se tiennent à votre disposition pour vous communiquer ces renseignements ainsi que toute information nécessaire concernant votre état de santé*. GINSENG garantit que toutes les mesures sont prises pour garantir la confidentialité de vos données de santé.

Toutes les informations relatives à ce projet sont à votre disposition sur :
Site internet : e-ginseng.org
Renseignements téléphoniques : 04 73 75 06 04

*conformément à l'article 32 de la loi du 6 janvier 1978 modifiée

Vous avez le choix !

Si toutefois, vous ne souhaitez pas que vos données de santé soient utilisées dans ce projet, vous avez la possiblilité à tout moment de refuser la transmission de vos données sans conséquence pour votre prise en charge médicale. Merci de renvoyer, par courrier papier ou email, à Audrey Lelong le coupon réponse ci-dessous.

Je refuse que mes données de santé soient utilisées dans le cadre du projet GINSENG pour l'évaluation de la pertinence des césariennes.

NOM :.....

Prénom : .....

Date de naissance :.....

Signature :

Coupon réponse à renvoyer à :

: Projet Ginseng
Service de Santé publique
7 place Henry Dunant
63000 Clermont-Ferrand

@ : admin.rspa@chu-clermontferrand.fr

Figure 24 Accord patient RSPA

Les professionnels de santé

Chaque professionnel de santé (médecin, biologiste, épidémiologiste…) doit disposer d'un profil de droits qui lui est propre : un accès en lecture ou écriture, sur la totalité d'une base ou seulement une partie d'autres filtres doivent être envisagés en fonction du niveau d'accréditation de chaque professionnel. Les utilisateurs à l'origine de la création d'une donnée médicale doivent posséder les droits en écriture et en lecture. Les personnels consultant les informations médicales doivent quant à eux disposer des droits en lecture seule. Chaque utilisateur peut être associé à un ou plusieurs groupes. Comme pour un système d'exploitation, le groupe permet de gérer les droits des utilisateurs composant le groupe. Les actions appliquées aux groupes seront répercutées sur les profils des utilisateurs.

Le comité de pilotage

Le comité de pilotage permet de garantir l'éthique autour de l'utilisation des données médicales ainsi que les nouveaux cas d'utilisation de l'infrastructure. Ce comité sera donc en charge de décider si un utilisateur a légitimement le droit d'accéder aux informations d'un patient. Le comité pourrait être constitué de médecins, de représentants des patients, et d'au moins un informaticien en charge de répercuter les décisions sur les droits d'accès.

2.1.5 La législation

Chaque partenaire médical déjà détenteur d'une base de données doit faire une déclaration normale auprès de la CNIL pour expliquer les raisons de la détention des données et leurs utilisations. Dans le cadre du projet GINSENG, les associations de dépistage organisé des cancers, les cabinets d'anatomie-pathologique ou encore les maternités disposent tous de déclarations normales auprès de la CNIL. Dans le cadre d'échanges de données médicales à caractère nominatif et du fait de la construction d'une infrastructure informatique destinée au transit et au stockage de ces informations, il a été nécessaire d'établir des demandes d'autorisations auprès de la CNIL pour faire la preuve d'une infrastructure sûre et garantir que les données confidentielles des patients ne seront pas utilisées à mauvais escient. Cf. Annexe autorisations (p. 193) CNIL. Les autorisations CNIL[103] 1515344 & 1519026 permettent à travers l'outil GINSENG ; l'identification du patient (nom, prénom, date de naissance, sexe, code postal de résidence) à des fins de validation des identités et de chaînage des données. Dans

[103] http://www.legifrance.gouv.fr/affichCnil.do?id=CNILTEXT000028268603 - date d'accès octobre 2015

le but de mettre en œuvre un traitement automatisé de données à caractère personnel ayant pour finalité la mise en place, à titre expérimental, d'une plateforme informatique permettant l'interrogation et le croisement de bases de données mises en œuvre dans le cadre du suivi médical des patients à des fins d'évaluation des pratiques de soins et de recherche épidémiologique.

2.1.6 Les données médicales

Les données médicales sont très nombreuses et très variées, elles sont cependant classables en deux grand types ; les variables et les fichiers. Les fichiers pouvant être des compte rendu '.doc', '.pdf' ; des enregistrements audio ou des images Digital Imaging and COmmunications in Medicine (DICOM) '.dcm' ou '.jpg', etc. Les images peuvent être fixes ou animées, une vidéo est une succession images fixes. Les images peuvent ainsi regrouper les résultats d'échographie, IRM, etc. Les variables peuvent contenir toutes les informations comme les valeurs de résultats de type prélèvement (biopsie, analyse, …), de mesures (températures, pression, etc.), mais aussi commentaires, remarques et analyses relatives aux diagnostiques. Nous pouvons généraliser en considérant d'une part les variables susceptibles d'être stockées dans une base de données et les fichiers qui nécessitent un Gestion Électronique des Documents GED comme le Picture Archiving and Communication System PACS. Actuellement nous ne travaillons qu'avec des variables textuelles destinées à être stockées dans un SGBD, et des compte-rendus au format '.doc' (Microsoft Word 1997-2003) qui sont le résultat d'un export depuis les bases métier de nos partenaires. L'imagerie est prévue pour une phase ultérieure du projet.

Trois cas distincts d'utilisation ont été abordés :

1. Un premier cas est dévolu aux études épidémiologiques. Il s'agit alors de fournir un accès à des données médicales non nominatives à partir desquelles pourront être basées leurs études. Il doit alors être possible de proposer une identité numérique unique pour chaque patient permettant de ne pas divulguer l'identité réelle des patients. L'infrastructure devra également permettre une analyse des bases de données médicales non nominatives en temps réel.
2. Le second cas concerne le transfert de données médicales entre sites distants. Les partenaires médicaux réalisent actuellement ce transfert de fichiers médicaux par Messagerie Sécurisé de Santé (MSS), par clef USB ou même par transfert de fichiers papiers. La solution informatisée et automatisée possède de nombreux avantages. L'absence de ressaisie des dossiers évite l'introduction d'erreurs, les

données sont donc réputées plus fiables. La mise à disposition de façon automatique rend les tâches moins chronophages pour les personnels de sites médicaux, et réduit les temps de mise à disposition des dossiers patients. Les différents protocoles que nous détaillerons dans la partie 0 de ce manuscrit permettent une transmission numérique extrêmement sécurisée.

3. Le troisième cas doit permettre à un praticien de consulter des dossiers distants stockés sur un autre site sans avoir à les recopier physiquement.

Pour répondre à ces besoins les informations seront poussées depuis les bases de données métier vers les serveurs du réseau GINSENG. Après traitement, les informations qu'elles contiennent seront consultées et/ou transmises aux utilisateurs habilités.

Base de données médicales Sipath-Unilabs

Le cabinet d'anatomie et cytologie pathologiques (ACP) Sipath-Unilabs regroupe la majeure partie des analyses histologiques et cytologiques en Auvergne avec de 3 000 à 5 000 analyses effectuées par semaine et stockées dans leur système d'information. Le SGBD est fourni et géré par la filiale santé de la société Infologic[104], le logiciel « métier » utilisé par le laboratoire se nomme DIAMIC ; il fonctionne sous Windows et s'appuie sur une base Oracle. Ce logiciel permet de stocker les compte-rendus d'analyse en format '.doc' avec les informations confidentielles de chaque patient. Des exports des données médicales, sous la forme de fichiers '.XML', quotidiens et personnalisés, peuvent être réalisés par la société Infologic vers le serveur du réseau GINSENG. Le contenu des fichiers '.XML' sera détaillé dans la partie 3.2.

Base de données médicales ARDOC, ABIDEC et ABIDEC-ARDOC

Les associations de dépistage organisé du cancer ARDOC et ABIDEC poursuivent les mêmes objectifs. La différence majeure entre ces deux structures est leur périmètre d'actions. L'ABIDEC couvre l'Allier et depuis peu une partie de la Nièvre, alors que l'ARDOC gère le reste de l'Auvergne. Leurs systèmes d'informations sont très similaires. La société O.S.I Santé leur fournit la solution logicielle « Zeus » pour la gestion de leur base de données. Zeus s'appuie sur un serveur SQL Server et fonctionne donc sous une version de Windows Server.

La société O.S.I Santé est en charge de fournir les exports de variables médicales d'intérêt des bases de données des associations pour la réalisation d'analyses épidémiologiques ; la

[104] http://www.infologic-sante.com/ - date d'accès octobre 2015

description détaillée de cette base et les exports pris en charge pour les analyses sera faite dans la partie 3.2.1

2.2 Recueil des spécifications du projet GINSENG

2.2.1 Un réseau non « invasif »

Les données de santé sont des informations précieuses. Le parti pris de l'état français avec le projet DMP, est de regrouper toutes les informations relatives à un patient en un seul et même endroit extrêmement sécurisé. L'approche de GINSENG quant à elle consiste à laisser les données là où elles sont produites. Les partenaires du projet sont garants de la validité des données médicales qu'ils produisent ainsi que de leur caractère confidentiel. Il est essentiel pour chacun des acteurs de pouvoir garantir aux patients que les données qui lui sont relatives sont gérées dans le respect de ses droits et en parfait accord avec la législation française. Dans le but de minimiser les risques de perte d'information il est convenu de ne pas interférer avec la base métier du site médical dans lequel les services GINSENG sont déployés. Ainsi en aucun cas les infrastructures GINSENG ne pourront modifier, altérer, perturber ou effacer les bases « métier » de nos partenaires.

Après avoir présenté le projet GINSENG ainsi que ces particularités, intéressons-nous au recueil des spécifications qui permettra de répondre aux attentes des différentes parties prenantes.

2.2.2 Un réseau distribué et sécurisé

Dans un premier temps, doivent être raccordées au réseau les associations de dépistage du cancer ARDOC, ABIDEC et ABIDEC/ARDOC, qui disposent chacune d'un réseau (*Symmetric Digital Subscriber Line*) SDSL de 4 Mb., la version professionnelle de l'ADSL qui permet un débit symétrique en émission et en réception. Le laboratoire d'ACP Sipath-Unilabs qui est connecté par une liaison optique gérée par le service informatique de la société et le HADS IDS hébergeant un annuaire central dans lequel les informations confidentielles liées au patient sont liées à un identifiant unique GINSENG. Ces cinq entités sont les premières à être connectées au réseau GINSENG pour valider la preuve de concept et la faisabilité de la solution. Par la suite, des laboratoires de biologie médicale, ainsi que des hôpitaux CHU ou CH, des centres anti cancéreux comme le CJP ; ou bien des médecins spécialistes ou généralistes pourront être connectés à l'infrastructure.

Les étapes suivantes devront être effectuées pour chaque nouvelle structure à rattacher au réseau GINSENG. Les plans d'adressages des réseaux internes seront communiqués par les prestataires de services qui maintiennent les solutions informatiques des structures. La solution sera capable de s'adapter au cas où une adresse IP publique dédiée ne serait pas disponible pour supporter les échanges prévus par GINSENG. Il faudra obtenir les autorisations nécessaires pour demander l'ouverture des ports auprès des administrateurs réseaux de chaque site. La même démarche permettra d'obtenir l'adresse IP interne nécessaire pour que le serveur d'identification puisse obtenir l'export depuis le serveur métier de la structure. La solution sera prévue pour cohabiter avec les différents pare-flammes (*firewall*) et routeurs ainsi que toutes mesures de sécurité déjà présente sur le nouveau site à raccorder à GINSENG. Si possible toutes les actions doivent se dérouler de façon transparente pour les utilisateurs actuels et ne pas impacter défavorablement les solutions métiers. Pour une meilleure modularité et évolutivité le plan d'adressage interne de GINSENG permettra de gérer chaque site indépendamment des autres en conservant une cohérence et une facilité de lecture et de déplacement au sein de la solution.

La sécurité et la confidentialité des données sont essentielles à l'intérieur des systèmes d'informations médicales. Il faut que les données soient et reste accessible, mais seulement par les personnes qui disposent de la légitimité pour consulter ces informations. L'information doit rester « disponible et confidentielle ». Nous trouvons ci-après une explication pour chacun des termes et concepts nécessaires dans notre contexte.

- L'intégrité et la validité de l'information : l'information que l'on consulte est effectivement celle que l'on souhaite consulter et n'a pas été modifiée ou altérée volontairement ou involontairement.
- La confidentialité : l'information est accessible uniquement et exclusivement aux personnes auxquelles elle est destinée. Comme nous l'exposions dans le paragraphe Les professionnels de santé de 2.1.4 relative aux droits d'accès et à l'usage des groupes d'utilisateurs ; il est primordial que les données que nous mettons à disposition soit seulement celles qui ont été prévues par les étapes préalables à la consultation des dossiers. Ce qui ne peut pas être rendu possible sans une authentification forte des individus.
- L'authentification : elle va de pair avec la confidentialité elle permet de s'assurer que l'identité électronique d'un utilisateur est liée à une personne physique unique.

- La disponibilité : les utilisateurs peuvent accéder aux informations qu'ils souhaitent consulter depuis les lieux initialement prévus et durant les plages horaires permettant ces interrogations.
- La non répudiation : c'est la garantie de la traçabilité, infalsifiable des actions. Ainsi le système crée des journaux qui permettent de tracer les utilisateurs et leurs actions, en son sein.

Ainsi la sécurité est ici entendue au sens large du terme, et il faut bien comprendre « les sécurités », terme que nous pouvons étudier sous différents aspects :

1. La sécurité physique

Les actions à mettre en œuvre pour garantir la sécurité physique des installations de serveurs GINSENG sont nécessaires. Il s'agit d'empêcher ou de freiner au maximum les actions pouvant physiquement être nuisibles aux matériels. Nous pouvons envisager : le vol, l'effraction, les incendies, les inondations, les radiations, que les origines soient volontaires ou involontaires. Les serveurs doivent donc être situés dans des salles dédiées, dans lesquelles la température est régulée. Ces salles machines doivent être à accès limité et les serveurs doivent disposer d'antivols arrimés à un point fixe de la salle. De plus, l'accès aux composants de la machine doit être protégé par un verrou de sécurité.

2. Une sécurité matérielle (hardware)

Pour se prémunir d'une défaillance électrique, les serveurs doivent être équipés d'au moins deux blocs d'alimentation redondants. Chaque alimentation étant reliée à une phase différente. Des onduleurs doivent, en outre, être systématiquement déployés en même temps que les serveurs.

Concernant les communications réseau, chaque serveur doit être équipé de deux cartes réseau physiquement distinctes configurées en mode répartition de charge. Chaque câble réseau se connectant à un équipement différent qui le relie à un réseau distinct jusqu'à la jonction avec le réseau externe.

Concernant le stockage des données médicales, il faudra utiliser des composants redondants. La technologie RAID (*Redundant Array of Independent Disks*) concernant l'agencement redondant de disques indépendants. Il existe différents types de configuration RAID, les plus intéressants pour garantir la pérennité des données sont les RAID1 ou RAID5.

La technologie RAID permet d'utiliser plusieurs disques simultanément en les groupant, pour palier à la défaillance de l'un d'eux (cas qui nous intéresse particulièrement) ou augmenter les débits des lectures, écritures sur les disques. Le RAID 1 utilise 2 disques vers lesquels les

écritures seront effectuées de façon similaire, les deux disques sont donc une image conforme l'un de l'autre. Dans ce cas si l'un des disques devient illisible ; il suffit de le remplacer, le disque sain sera recopié sur le nouveau disque. Et le système pourra continuer de fonctionner. L'inconvénient est que seulement 50% du volume total des disques est utilisable à cause de la réplication. Le RAID 5 mobilise au minimum 3 disques mais le volume utilisable est celui de la somme des disques moins 1, donc 67% pour 3 disques et 80% pour 5 disques. Ici ce n'est pas une réplication qui est effectuée mais une somme de contrôle. Pour un groupe de 'n' disques, les informations sont écrites en parallèle sur n-1 disques, le $n^{ième}$ contient une valeur de parité qui permet de reconstruire 1 des disques du groupe en cas de défaillance. Si deux disques tombent en panne simultanément la totalité du groupe devient irrévocablement inexploitable.

On peut en plus des principes évoqués dans le paragraphe électricité adjoindre à la carte contrôleur du RAID une batterie qui permet d'assurer l'intégrité des données durant de courte coupure électrique. De plus, il est conseillé d'utiliser des disques durs de provenance hétérogène pour éviter un souci qui pourrait toucher une série de disque du même fournisseur ou fabricant. Les disques doivent être homogènes en termes de volumétrie, sinon, chaque disque sera considéré comme possédant un volume égal au disque ayant la plus petite capacité.

3. Une sécurité logicielle
    a. Les disques durs des machines doivent être chiffrés
    b. Les réseaux sur lesquels transite l'information sont des VPN
    c. Lorsque l'information transite dans les VPN elle doit être chiffrée

2.2.3 Standardisation des bases de données

Dans l'absolu si votre médecin souhaite accéder à vos antécédents, ce qu'il souhaite c'est entrer votre identifiant dans la machine, pour que le système lui révèle la totalité exhaustive de vos différents compte rendu retranscrivant vos consultations médicales, sans filtre géographique, ni professionnel. C'est la vision idéalisée de l'objectif, l'utopie. Dans la réalité, les solutions mises en œuvre comme le projet DMP vise à mettre à disposition des synthèses des consultations que le patient a réalisées. Une synthèse est par définition un résumé, elle ne contient pas l'ensemble exhaustif des données. Pour réaliser des comparaisons ou des analyses il faut que les données soient comparables. Ce qui implique d'utiliser un vocable commun. C'est ici l'un des grands problèmes du monde médical. Cette standardisation du vocabulaire commun n'est pas encore avérée. Ainsi tous les cabinets d'ACP de France n'utilisent pas le codage ADICAP par exemple certains utilisent la variante CIM-O-3 qui est une évolution de la

CIM-O-2 de CIM-10. De plus la classification CIM-O-2 est devenue totalement obsolète, alors qu'il existe des correspondances CIM-O-3 / ADICAP. Il faut donc considérer que les codages actuels ne sont pas immuables et envisager leurs remplacements futurs. C'est à cause de ces nombreux standards concurrents que l'interopérabilité dans le milieu de la santé est aussi difficile. Notre solution optera donc pour la compatibilité multi standard. Un effort supplémentaire sera nécessaire au moment de la rédaction du composant logiciel en charge de l'importation des données dans notre système. Il faudra considérer des tables d'alignement entre les différents standards que nous pourrons rencontrer.

2.2.4 Identification des patients

L'identification des patients est l'un des points primordiaux de notre solution. Si un patient consulte dans différents centres médicaux, il doit pouvoir être identifié par le système comme une seule et même personne. Les accords CNIL obtenus par le projet GINSENG nous permettent d'effectuer des recherches nominatives aux seins des différentes bases du projet. Dès qu'un nouveau patient est identifié il est inséré dans un annuaire centralisé hébergé chez un HADS ainsi que dans l'annuaire local de la structure dans laquelle le patient vient de consulter. Lors de son inscription dans l'annuaire le patient reçoit un numéro unique qui sera désormais son identifiant à l'intérieur du réseau GINSENG. L'identification nécessite de comparer chaque nouveau patient avec les patients connus et stockés dans un annuaire. Il est préférable d'effectuer cette opération la nuit sur un créneau dédié à la structure de cette façon nous évitons au maximum les collisions de requêtes et ne consommons pas la bande passante nécessaires aux besoins de la structure. Le chaînage correct des informations médicales est d'autant plus primordial pour la réalisation d'analyses épidémiologiques significatives.

2.2.5 Authentification

L'authentification vise à s'assurer de l'identité des personnes utilisant le système pour accéder au contenu des bases de données. Les données traitées ont un caractère confidentiel et sont donc très sensibles. Elles doivent donc pouvoir être partagées à l'intérieur de groupes de travail accrédités et préalablement définis. Il est donc primordial de s'assurer de l'identité des utilisateurs du système aux moyens de solutions robustes et fiables. Les mécanismes d'authentification doivent permettre de repérer et de bloquer les tentatives de connexions frauduleuses. L'utilisation de listes blanche et noire d'adresses IP associées à des adresses MAC peut être envisagées.

Les utilisateurs devront s'identifier et se connecter au réseau par l'intermédiaire d'une interface WEB en utilisant leur carte CPS comme le recommande l'ASIP Santé. Il est préférable que l'interface web du réseau GINSENG soit accessible depuis l'ENRS auvergnat SIMPA de manière à rendre son utilisation la plus aisée et la plus pérenne possible.

Il est au nécessaire de conserver toutes les traces des connexions et des actions effectuées par chaque utilisateur. Ces traces seront archivées dans des journaux avec des horodatages précis. La consultation de ces journaux pourra permettre à postériori de retracer le fils des événements d'un ou plusieurs accès au système.

L'utilisateur est très précisément identifié, et dispose de droits très précis, qui délimitent son périmètre d'action. Périmètre qui ne peut être étendu que, et uniquement, par une requête formulée et motivée, à destination du comité de pilotage du projet, que ce dernier devra au préalable valider. En fonction des droits dont dispose un groupe ou un utilisateur, il sera plus ou moins autorisé à requêter les différentes bases de données agrégées qui constituent le système. Les garants de la confidentialité des données du patient sont les membres du comité de pilotage qui sont les seuls habilités à rendre accessibles les informations médicales.

2.2.6 Interface

Pour que la solution trouve sa place auprès des utilisateurs finaux il ne faut pas seulement qu'elle soit rapide et efficace, il faut aussi qu'elle soit ergonomique et facile d'accès. Deux interfaces principales sont souhaitées. La première, s'appuyant sur les solutions Web et accessible à travers un navigateur Web. La seconde, intégrée directement dans l'interface métier des associations de dépistage organisé du cancer.

Interface Web

L'interface Web aura pour objectif d'être accessible au travers des navigateurs Web les plus courant (Internet Explorer (IE), Firefox, Safari, Chrome, Opera et désormais Edge (le remplaçant d'IE dans Windows 10)). Une attention particulière sera apportée sur la compatibilité avec les versions précédentes de ces logiciels de navigation, sans se limiter à leurs dernières versions. En effet, certaines structures sont encore sous une version 8 d'IE. Le but étant de rendre l'interface accessible au plus grand nombre. Si possible, ce service de présentation des informations sera hébergé (ou à minima simplement présenté) par le portail de l'ENRS auvergnat SIMPA. Cette interface Web sera régie par les droits d'accès présentés plus tôt. Trois grands volets seront disponibles en fonction des autorisations accordées à l'utilisateur.

1. Informations généralistes

Les premières pages hébergées par l'ENRS accessibles par un onglet dédié GINSENG du domaine cancer pour RSCA, et natalité pour les informations RSPA, seront consacrées à une information généraliste sur GINSENG telle que nous l'avons présenté au début de ce chapitre. Accessible sans authentification préalable elles auront vocation à informer le grand public de la démarche mise en œuvre. Il sera aussi possible de télécharger les notices de non-participation à l'étude GINSENG (Figure 23 et Figure 24) qui sont dédiées respectivement aux applications relatives au cancer et à la périnatalité ; par la suite, un formulaire en ligne permettra d'effectuer ces démarches directement à travers cette interface WEB. En plus de la présentation du projet, il sera également possible de rendre accessibles des cartes d'informations comme le fait le réseau sentinelles (Figure 7) adapté pour notre région Auvergne. On peut envisager la carte de l'incidence des cancers des poumons en surimpression d'une cartographie de la présence de radon. Il n'est pas envisagé que le patient puisse consulter ses dossiers, ce n'est pas un objectif à moyen terme d'autant que nous ne disposons pas de l'habilitation pour lire les cartes vitales, qui pourraient servir de moyen d'authentification des patients. Par respect des contraintes légales le patient pourra consulter ses informations en se déplaçant sur rendez-vous dans une des structures prévues à cet effet.

2. Création des requêtes

Après s'être authentifié à l'aide d'une CPS ou carte apparentée, les utilisateurs qui disposeront de droits suffisants (cf. 2.2.5) pourrons générer et tester des requêtes statistiques à but épidémiologique en s'appuyant sur les données auxquelles ils ont accès en lecture. Une interface d'aide à l'écriture de requête, dont la maquette est présentée Figure 25, a été entièrement réalisée. Cet outil d'aide à la création de requêtes permettra de choisir les bases que l'utilisateur souhaite interroger en précisant la (ou les) structure(s) ainsi que le service ou la base de données ciblés. Il sera possible de préciser la période temporelle de l'étude, le sexe des patients concernés, ainsi que leur tranche d'âges. De plus, si cette étude a pour but de suivre l'évolution d'une cohorte déjà constituée, il sera possible de préciser à l'outil logiciel les identifiants des patients la constituant. La requête sera identifiée par son nom qui permettra de la rechercher ultérieurement pour modification ou simplement pour une mise à jour des résultats. Un champ descriptif est mis à disposition du créateur de la requête pour qu'il puisse laisser ses commentaires à l'intérieur du système et décrire les moyens mis en œuvre ainsi que les résultats souhaités. Les commentaires permettent une meilleure maintenabilité à plus long terme de la requête (notamment par des personnes tierces). Il sera en effet possible de permettre une réutilisabilité des requêtes que l'on aura créées, en cochant la case prévue à cet effet. En

fonction des bases sélectionnées les champs auxquels l'utilisateur a accès seront affichés, il pourra sélectionner les champs pertinents pour son étude ou sa requête. Il aura l'opportunité d'associer chaque champ à une étiquette qui lui permettra une navigation plus rapide lors de ses prochaines études ; il pourra, s'il le désire, naviguer en direction des étiquettes qu'il aura préalablement générées au travers de cet outil. Pour les champs de type date, un calendrier permettra d'affiner les requêtes grâce à l'interface graphique. Les requêtes créées seront écrites ou traduites en SQL ou en SPARQL pour être appliquées à des bases MySQL fédérées ou un SPARQL EndPoint sur un serveur hébergé chez IDS, en fonction du langage retenu. L'une des requêtes les plus simples pourrait être « nombre de cancers par code postal » et on peut envisager toute sorte de requêtes plus complexes tel que « quel est le code ADICAP le plus négatif chez les participants au dépistage organisé âgé de 50 à 75 ans et n'ayant jamais eu d'antécédent ». Les analyses des épidémiologistes pourront s'exprimer dans le cadre du contrat qui les lie aux propriétaires des données. Les résultats seront mis à disposition de l'utilisateur ou du groupe désiré sous forme de fichiers '.CSV' téléchargeable depuis le site. Une fois ces requêtes validées, elles pourront à terme être automatisées et leurs résultats serviront pour la création de cartes rendues publiques.

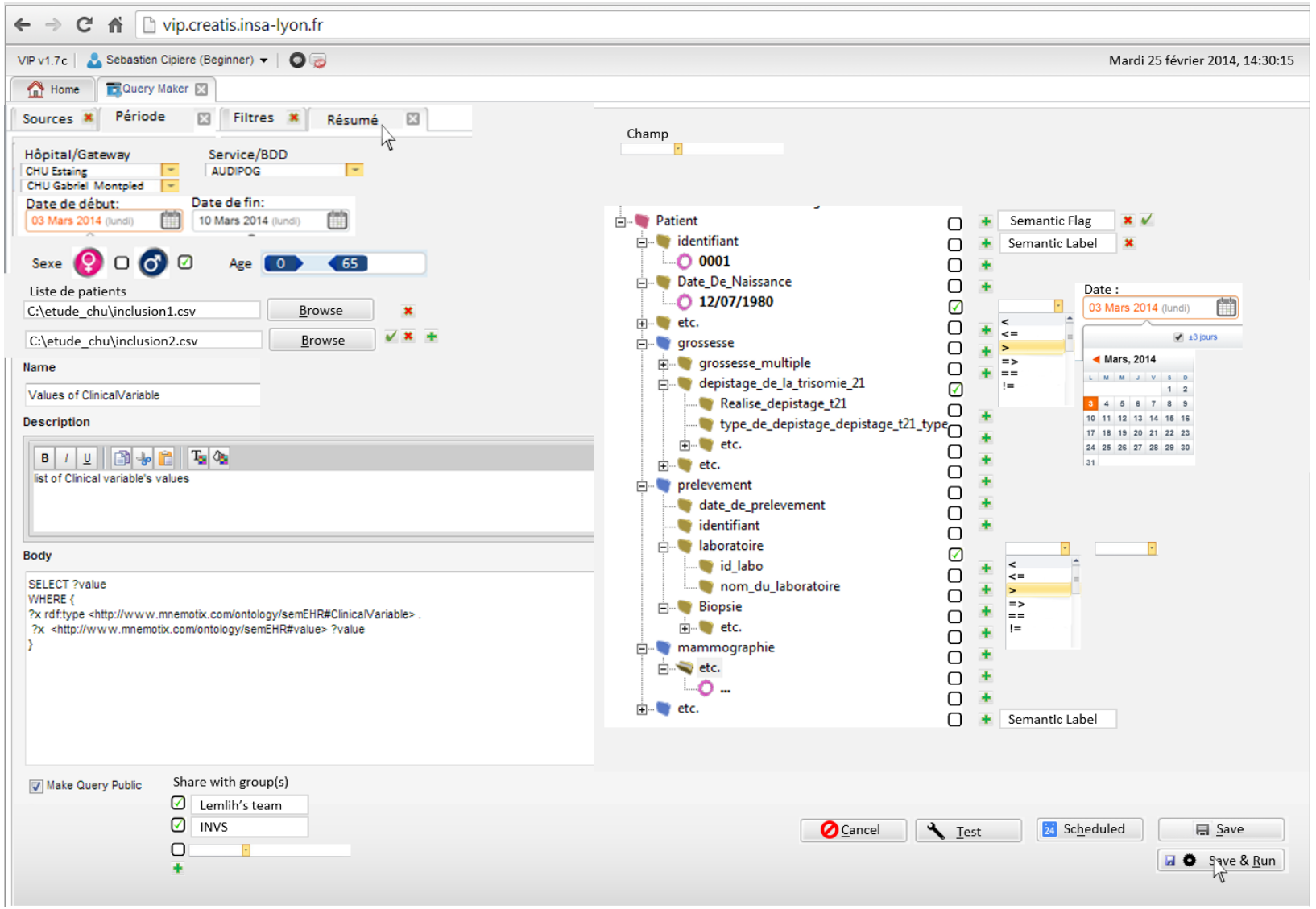


Figure 25 Maquette de l'interface d'aide à la création de requête pour GINSENG à destination des épidémiologistes

3. Accessibilité des résultats de requêtes

L'accès aux résultats (à but confidentiel) sera possible pour les utilisateurs disposant de carte de la famille CPS. Ces résultats n'ont pas, dans un premier temps, de vocation à devenir publiques. Plusieurs cas sont à envisager comme une remonté d'information à l'InVS si tous les acteurs ont validé la démarche. Un épidémiologiste habilité à créer des requêtes, aura la possibilité de les partager avec ses collaborateurs habilités. Il est envisagé une automatisation des requêtes lorsque celles-ci auront été testées et validées. Ainsi chaque semaine, chaque mois, ou chaque année un nouveau fichier de résultats sera mis à la disposition des utilisateurs. Les résultats présentés dans cette interface sont totalement dépendants des requêtes créées dans la partie 2. C'est à l'auteur des requêtes de définir à qui seront accessibles les résultats. Dans une certaine mesure, il sera également intéressant de partager la requête pour une meilleure interprétation des résultats.

Interface intégrée aux outils métiers

Une autre approche consiste à présenter les résultats directement à l'intérieur du logiciel métier des associations de dépistage du cancer. Cette interface, nommée « Zeus », gérée par la société O.S.I Santé, est écrite en Access pour fournir une interface graphique à un MySQL server (Windows server 2008 ou 2012 en fonction de la structure). Actuellement, après l'envoi d'une invitation, si l'association reçoit un résultat positif lors du dépistage. Elle se retrouve obligée, de mobiliser du personnel pour entreprendre la recherche des suites données, auprès des différentes structures susceptibles d'avoir traitées le patient. L'objectif est de pouvoir trouver et rapatrier ces informations afin qu'elles soient stockées et analysées. Pour mieux comprendre le devenir des patients détectés grâces aux SGDO.

Conjointement avec la société O.S.I Santé, le LPC a travaillé à l'importation des compte-rendus d'anatomie pathologique (au format '.doc' ou '.pdf') dans l'interface « Zeus » dès que la recherche des informations du patient s'est avérée concluante depuis l'interface WEB GINSENG. La complétion du code ADICAP, directement dans la base de données des associations, sera également mise en œuvre.

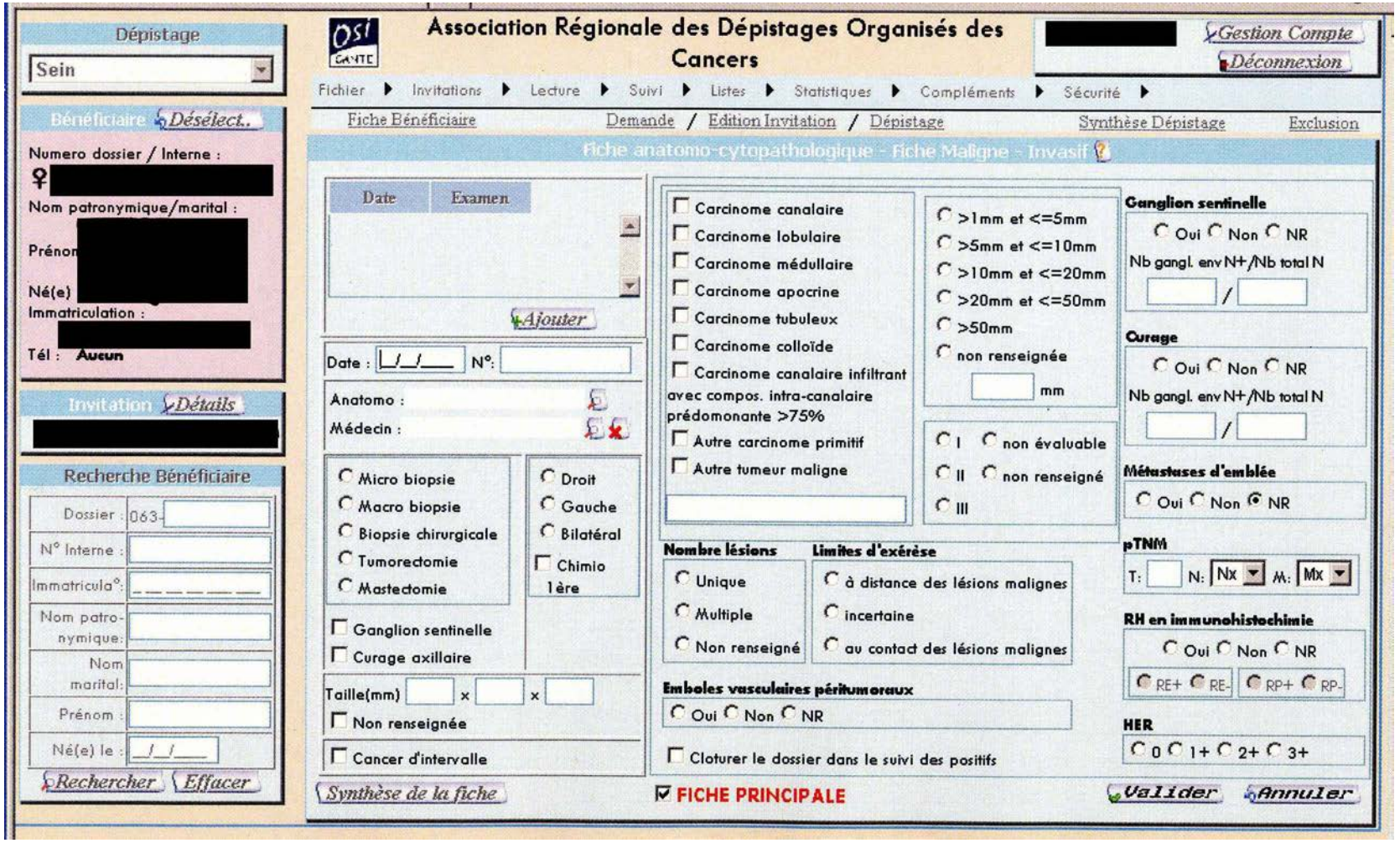


Figure 26 Exemple d'une vue de l'interface Zeus pour l'ARDOC

Conclusion

Dans ce chapitre, nous avons présenté en détails le projet GINSENG ainsi que le recueil des spécifications qui a été établi pour sa mise en œuvre. Après avoir présenté les différents acteurs du réseau GINSENG, nous avons exposé leurs besoins en termes de base de données et d'échanges d'informations. Pour faciliter l'adoption de notre système par nos partenaires nous avons cherché à minimiser à la fois le coût financier des installations des sites médicaux mais également l'implication du personnel médical et technique de chaque partenaire. Les conclusions de notre recueil des spécifications font état de la mise en place d'un réseau extrêmement sécurisé qui doit relier les SGDO, les cabinets ACP, les hôpitaux et le HADS. L'architecture du système doit être distribuée. Cette approche permet une meilleure tolérance aux fautes et fournit des propriétés intéressantes comme la non centralisation des données en un seul et même lieu. Ce réseau devra pouvoir être exploité depuis des connexions généralistes de type connexions ADSL. L'authentification des utilisateurs s'effectuera au moyen de cartes de type CPS, pour respecter les contraintes légales et les recommandations de l'ASIP santé. L'identification du même patient dans des bases différentes est l'un des points importants de notre étude. En partenariat avec l'équipe de recherche de l'ISIT dans ce domaine, nous choisirons l'algorithme et l'implémentation la plus adaptée à nos besoins, après avoir soigneusement comparé les différentes solutions à notre disposition. L'interopérabilité des systèmes d'information que nous avons à interfacer a été étudiée. Nous avons recherché quelle pourrait être l'ontologie capable de couvrir la totalité du périmètre de notre étude. Nous avons exploré des pistes dans lesquelles nous proposions notre propre structure hiérarchique de données, pour garantir une généralisation optimale de notre infrastructure. Concernant l'interface utilisateur, deux cas d'utilisation doivent être pris en compte : le premier est l'utilisation d'un portail accessible grâce à un navigateur WEB à travers internet ; le second est l'implémentation de fonctionnalités de partage de documents à l'intérieur des logiciels métiers utilisés par les professionnels de santé. Dans le prochain chapitre, nous présentons les choix retenus pour la mise en œuvre concrète du réseau de partage d'informations. Puis nous nous intéressons aux résultats de la mise à l'épreuve du système en conditions réelles.

Chapitre III

– Architecture technique et coordination du projet GINSENG

Introduction

La mise en œuvre de l'infrastructure du projet à partir du recueil des spécifications établi dans le chapitre précédent s'est déroulée en deux grandes étapes. Durant la première partie du projet, correspondant à une période de 2 ans, les choix technologiques ainsi que leurs implémentations ont été majoritairement imposés par le partenaire industriel du projet, la société Gnùbila, dans le but de permettre une valorisation directe des développements informatiques vers d'éventuels clients de l'infrastructure déployée. Au cours de cette période, le travail de thèse s'est focalisé sur la gestion des données médicales distribuées, en particulier leur standardisation, ainsi que sur l'implémentation des algorithmes d'identification des patients. Suite au départ du projet de la société partenaire, une complète refonte de l'infrastructure a été réalisée pour proposer des solutions techniques open source et plus facilement maintenables. La nouvelle infrastructure a été déployée au cours de la quatrième et dernière année du projet ANR. Nous avons donc souhaité retranscrire dans ce chapitre les implémentations réalisées pour la mise en place de l'infrastructure distribuée au cours des deux dernières années du projet.

Dans ce chapitre, nous commencerons donc par présenter l'infrastructure du système distribué mise en place et permettant des échanges et des analyses de données médicales sécurisés entre chaque site. Nous détaillerons par la suite les bases de données et leurs structures, avant de nous intéresser aux mécanismes d'identification des patients et de chaînage de l'information. Pour finir, nous expliquerons l'authentification des utilisateurs qui sera mise en place sur une interface web ergonomique ainsi que les mécanismes de requêtes sur les bases de données médicales distribuées.

3.1 Une infrastructure distribuée et sécurisée d'accès aux informations médicales

Nous avons retenu une architecture distribuée pour réaliser le projet GINSENG, et ce pour de nombreuses raisons. L'un des intérêts de s'appuyer sur une technologie de type « grille » est qu'il n'est pas nécessaire de centraliser sur un seul site d'importants moyens informatiques. Ainsi, le seul point central de notre solution est le site du HADS IDS de par l'obligation légale exigeant des garanties de sécurité et de traitement de la donnée médicale nominative. Ce dernier HADS fournit le support pour le serveur Web hébergeant le site internet qui permet de visualiser les résultats des requêtes épidémiologiques. Un autre serveur, hébergeant l'annuaire de correspondance des données patients confidentielles avec leur identifiant GINSENG unique, est aussi localisé dans les locaux du HADS. Tous les serveurs hébergeant les données médicales des patients, que nous nommerons « nœuds », restent situés sur chaque site médical produisant ces données. Ce système permet au producteur de données médicales de garder le contrôle sur ses données afin de les partager ou non. La tolérance aux fautes profite aussi de cette solution car si un site est coupé du réseau momentanément alors les autres sites ne sont pas impactés et les autres données médicales restent toujours accessibles.

3.1.1 Les serveurs

Les serveurs utilisés dans le cadre du projet GINSENG déployés chez nos partenaires sont « virtualisés » permettant ainsi une économie du nombre de machines avec un unique serveur physique secondé par un onduleur sur chaque site. Les avantages des machines virtuelles sont nombreux : le gain de place, les facilités d'administration des différentes machines virtuelles, et les économies d'énergie, tant en consommation qu'en refroidissement.

Architecture de la grille GINSENG

Les objectifs de GINSENG, contrairement à un projet type DMP qui vise à regrouper toute l'information en un seul et même lieu, sont de laisser les données là où elles sont produites. Chaque site partenaire dispose d'un serveur dédié au projet GINSENG. Généralement il s'agit d'un serveur rackable mais l'on peut envisager un serveur tour et même plus généralement il suffit de déployer plusieurs machines virtuelles si l'infrastructure en place le permet. Nous pouvons considérer la Figure 27 comme une vue d'ensemble de l'infrastructure.

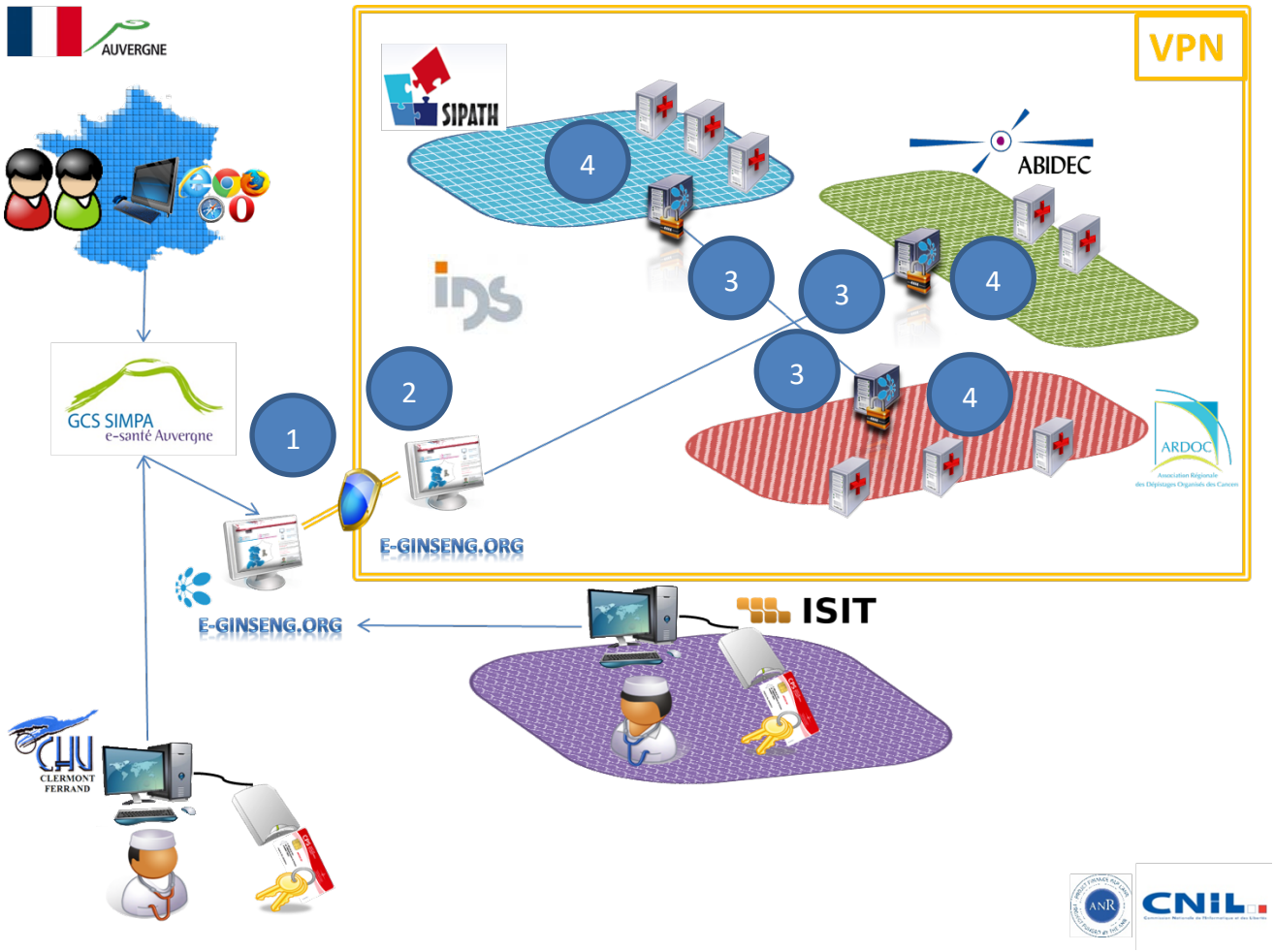


Figure 27 Architecture informatique de GINSENG – vue globale simplifiée RSCA

Le réseau est composé de différents types de serveurs

1. Le serveur Web qui présente l'information.
2. Le serveur d'authentification qui permet l'accès aux personnes autorisées.
3. Les Nœuds qui communiquent entres eux et répondent au serveur Web.
4. Les serveurs d'identification et d'anonymisation.

Comme nous l'avons défini dans le paragraphe sur la virtualisation, chaque site ne dispose que d'un seul serveur de virtualisation qui contient plusieurs machines virtuelles. Ces machines virtuelles peuvent être des serveurs. Ainsi, le serveur Web et le serveur d'authentification peuvent se trouver sur la même machine.

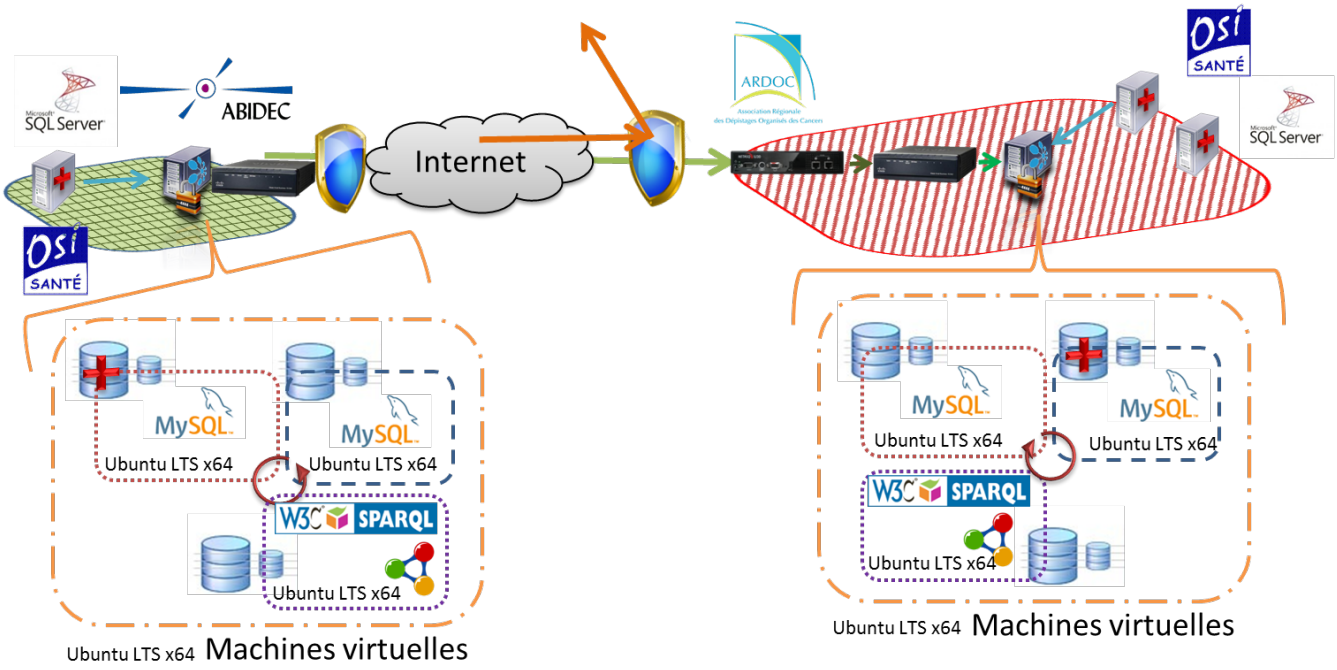


Figure 28 Architecture informatique GINSENG – ABIDEC/ARDOC-
vue composants réseaux et machines virtuelles

Il y a autant de nœuds dans GINSENG que de producteurs de données, il est même possible de trouver plusieurs nœuds dans un même hôpital si les différents services utilisent des SI très différents. C'est une solution très utile car elle permet l'interopérabilité des données entre les différents services. Les machines achetées avec le marché CNRS sont des Dell R710 pour les salles serveurs et T610 pour les laboratoires ne disposant pas de salle dédiée à l'informatique. La solution matérielle retenue possède 4 cœurs à plus de 2 GHz, 16 à 32 Go de RAM pour permettre la virtualisation des différentes machines et au moins 1 To de disque en RAID 1 pour plus de sécurité. Les machines disposent de carte iDrac Enterprise pour l'administration à distance des serveurs physiques. De plus la version Entreprise de l'iDrac Dell permet d'augmenter la sécurité en déportant le management d'iDrac sur un connecteur réseau RJ45 dédié. Les versions standards de iDrac propagent (automatiquement) l'interface de management du serveur sur le premier connecteur RJ45 du serveur à l'adresse 192.168.0.120 en parallèle du trafic dédié à l'OS ou l'hyperviseur (avec un couple login/password standard (root/calvin)).

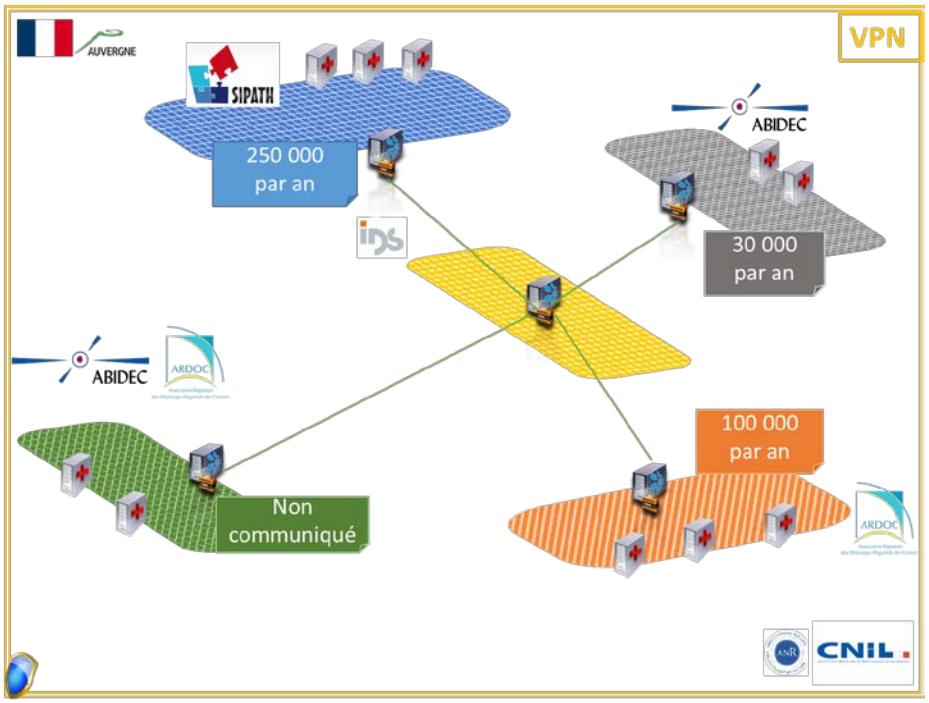


Figure 29 Volume de fichiers patients (en attente/traité/produit) par site intégré dans GINSENG

Installation physique

Différentes machines ont été utilisées au cours du projet. Les machines les plus anciennes datent de l'étude de faisabilité de GINSENG en 2010 alors que les machines les plus récentes ont été achetées à la fin 2014 pour équiper les nouveaux participants au projet en 2015. Pour les associations de dépistage du cancer ainsi que pour le cabinet Sipath-Unilabs, les serveurs retenus sont majoritairement des DELL (marché CNRS) dont les fiches techniques sont présentées ci-après.

Tableau 9 Fiche technique des machines testées durant la phase de prototypage version 2012

Description
CPU : Intel® Core(TM) i7-2600K CPU @ 3.40GHz (4 cœurs HT x2)
GPU : NVIDIA GeForce GTX 590
RAM Corsair 16Go DDR3 1866MHz 9-10-9-27
Carte mère ASUS P8P67
Disque Western Digital 1To WD black SATA 64Mo cache 7200 RPM

Tableau 10 Fiche technique des machines équipant les sites Sipath-Unilabs/ARDOC/ABIDEC version 2010

Description
DELL PowerEdge T610
CPU : Intel® Xeon® Processor E5507 (4M Cache, 2.26 GHz, 4.80 GT/s Intel® QPI) (4 cœurs non HT)
RAM 16Go
Disque RAID 1 300Go SAS 15000 RPM

Tableau 11 Fiche technique d'un serveur rackable GINSENG version 2015

Description
Dell PowerEdge R520
Alimentation redondante 495W (1+1)
48Go en 6 x 8Go 1600 MHz RDIMM voltage standard
Processeurs Intel Xeon E5-2407v2 à 2.20 GHz à 4 cœurs non HT, 10 Mo cache, 6.4 GT/s, 80W – x2
3.5" Chassis with up to 8 Hard Drives and Hardware RAID
Carte RAID H710 avec 512 MB nvram
Disques dur 300 Go 15000 tr/min SAS - x2
Disques dur 2 To SATA 7200 Tpm format 3,5" - hotplug - x5
On-Board Broadcom 5720 à 2 ports 1Gb
Carte réseau Broadcom 5719 quatre ports gigabit
iDrac 7 Enterprise
Lecteur graveur interne SATA 8X DVD +/-RW
Pas de système d'exploitation
3 ans de service Pro Support et d'intervention sur site le jour ouvrable suivant

Les machines disposent d'alimentations redondantes pour limiter l'impact d'une défaillance sur l'une d'elles. Le volume de RAM de 48 Go (version 2015) permet de créer quatre machines virtuelles disposant de 12 Go de RAM chacune. Les deux processeurs fournissent au serveur huit cœurs physiques, ce qui permet d'allouer deux cœurs « réels » à quatre machines. Chaque machine peut disposer d'une interface réseau dédiée et de 2 To d'espace disque pour le stockage ainsi que de 75 Go pour le système.

Les bonnes pratiques d'installation de machine sont respectées, comme le verrouillage du BIOS par mot de passe, l'attribution d'une adresse IP dédiée à l'iDrac ainsi que le changement du login et du mot de passe par défaut. Les partitions RAID 1 pour le système sur les deux disques de 300 Go SAS, et RAID 5 pour les cinq disques de 2 To sont créées cela permet de disposer d'un volume de stockage, de 300 Go pour le système et 8 To effectifs pour les données.

Les configurations RAID assurent une redondance des données permettant de pallier la défaillance d'un disque de façon temporaire, sans interrompre l'activité du serveur.

Le système d'exploitation retenu est un Ubuntu serveur x64 LTS pour minimiser les mises à jour. Après avoir testé les solutions de virtualisation de Citrix et VMware c'est l'outil de virtualisation KVM qui a été retenu pour des raisons économiques. KVM a en effet l'avantage d'être libre et gratuit.

Une version de Ubuntu x64 LTS est donc installée sur le serveur vierge, les disques sont encryptés lors de cette installation et protégés par un mot de passe. Nous utilisons le FDE (*Full Disk Encryption*) fournit par Canonical[105] avec les configurations par défaut. Les paquets installés sont listés dans le Tableau 12.

Tableau 12 Liste des paquets installés sur le serveur de virtualisation

Paquets installés sur la machine serveur de virtualisation
iptables
vim
ssh
qemu-kvm
libvirt-bin
ubuntu-vm-builder

Après l'installation du serveur de virtualisation, nous créons des machines virtuelles sur cette machine hôte. Ces machines ont l'avantage d'être cloisonnées à l'aide des règles d'iptable. Pour garantir un parc de machines homogène le système d'exploitation des machines virtuelles est, lui aussi, un Ubuntu LTS x64. Pour assurer des mises à jour avec des paquets fiables, nous maintenons un dépôt des paquets testés et validés au préalable chez notre HADS. Les disques virtuels sont encryptés une seconde fois lors de l'installation d'Ubuntu LTS dans la machine virtuelle.

Tableau 13 Liste des paquets installés sur la machine « serveur de données » et la machine « importeur »

Paquets installés sur la machine serveur de données
iptables
ssh
mysql-server-5.6
perl

[105] http://www.canonical.com/ - Date d'accès mai 2016

La machine virtuelle dont le rôle est « importeur » dispose d'un serveur vsftpd (Very Secure File Transfer Protocol Daemon) qui lui permet de recevoir les exports de données médicales provenant de nos partenaires. Cette installation est complétée par la mise en place des éléments réseau et d'un onduleur si nécessaire.

D'autres machines se situent chez le HADS, ces machines sont provisionnées comme sur un cloud du type Amazon EC2[106] qui nous permettent de disposer de machines virtuelles sur mesure en termes de puissance de calcul, de volumes de stockage et de RAM. Les serveurs sont disponibles sur un mode PaaS (*Platform as a Service*) : il suffit de définir la version du système d'exploitation (ou *OS- Operating System* que nous retiendrons par la suite) désirée et une installation permettant la connexion en SSH. Le dimensionnement de ces serveurs peut varier en fonction des besoins et de l'utilisation, sans nécessiter de réinstallation.

La virtualisation

Nous avons testé les principales solutions de virtualisation disponibles sur le marché, pour au final n'en retenir qu'une. Nous avons eu l'occasion d'administrer nos serveurs avec trois solutions distinctes : Citrix[107], VMware[108] et Qemu-KVM *(Kernel Virtual Machine)*[109].

La virtualisation consiste à partager une machine physique, entre plusieurs systèmes. La machine physique peut être un ordinateur portable, un pc familial ou un serveur, les solutions restent identiques. Les solutions de virtualisation peuvent être appelées hyperviseurs. Il existe deux grandes familles d'hyperviseurs : la première se substitue complétement à l'OS traditionnel et s'installe sur une machine vierge (Citrix XenServer et VMware Vsphere), la seconde s'installe à l'intérieur de l'OS (Oracle Virtualbox[110], VMware Player, KVM). Il est ainsi possible de créer des machines à l'intérieur du système d'exploitation (Windows, Mac OS ou Linux), à l'aide d'un VMware Player ou d'Oracle VirtualBox, par exemple. Il existe d'autres solutions sur le marché, nous avons retenu les leaders du marché et les logiciels qui nous semblaient les plus pertinents pour répondre à nos attentes. Nous pouvons aussi citer Docker[111] qui est une solution souvent retenue depuis 2013 mais à notre avis pas encore suffisamment mature pour répondre à nos attentes. Nous considérons ici la virtualisation d'un serveur. Les interfaces des différentes solutions de virtualisation sont plus ou moins conviviales et couvrent

[106] https://aws.amazon.com/fr/ec2/ - date d'accès octobre 2015
[107] https://www.citrix.fr/date d'accès octobre 2015
[108] https://www.vmware.com/fr - date d'accès octobre 2015
[109] http://www.linux-kvm.org/page/Main_Page - date d'accès octobre 2015
[110] https://www.virtualbox.org/ - date d'accès octobre 2015
[111] https://www.docker.com/ - date d'accès octobre 2015

un large spectre depuis l'interface graphique (utilisable à la souris), jusqu'à la ligne de commande. Les trois solutions que nous avons retenues pour nos expérimentations sont équivalentes en termes de performances pour nos besoins. En effet, nous n'avons pas noté de surconsommation de la RAM ou des CPUs par l'un des hyperviseurs, ce qui laisse nos ressources disponibles pour nos machines métiers « virtualisées ». Les solutions propriétaires de Citrix et VMware disposent d'interfaces de gestion que l'on peut installer sur une machine tierce équipée d'un OS Windows. Une fois installé, le gestionnaire graphique se connecte au travers du réseau à la machine à administrer. Les avantages sont alors une meilleure représentation visuelle du serveur administré avec une arborescence dans laquelle le serveur se subdivise en Machines Virtuelles (VM *(Virtual Machine)*). La solution de Citrix permet dès la version d'évaluation d'administrer plusieurs serveurs à travers une seule et unique fenêtre ; alors que chez VMware vSphere il faudra ouvrir une fenêtre par serveur ce qui est moins pratique à l'usage. Il existe un centre de pilotage centralisé de serveur de virtualisation disponible dans le catalogue VMware mais il nécessite de s'acquitter de la licence d'utilisation en plus de la licence serveur pour l'utiliser. La solution Citrix propose le centre d'administration librement accessible et ne facture que les licences serveurs. Le système de licence est basé sur le nombre de CPUs, il est donc plus avantageux pour des CPUs disposant d'un nombre très important de cœurs. Il faut de plus considérer le coût de support et maintenance à renouveler chaque année.

Bien que les solutions propriétaires soient très fiables et disposent d'une interface complète, leur coût élevé ne trouve pas de justification suffisante comparé aux solutions OpenSource. En effet, la troisième solution que nous avons explorée est celle de Qemu-KVM dans une version LTS de Ubuntu serveur 64 bits. L'avantage principal de cette solution est sa gratuité complète. En partant d'une machine vierge nous installons dans un premier temps une version minimaliste de Ubuntu[112] LTS 64 bits. Nous avons retenu la version LTS *(Long Term Support)* car elle permet de n'effectuer que des mises à jour légères sur une durée entendue, avant une mise à jour plus lourde qui fera passer le système à la future version LTS. C'est pour cette pérennité et légèreté de maintenance que nous avons retenu Ubuntu face à Debian[113] ou CentOS[114]. Dans un second temps à l'intérieur de l'OS il suffit d'installer les paquets qemu-kvm et libvirt-bin pour installer l'hyperviseur. Si une interface graphique est souhaitée il faudra rajouter des paquets pour la partie graphique d'Ubuntu non présent dans une

[112] http://www.ubuntu.com/ - date d'accès octobre 2015
[113] https://www.debian.org/index.fr.html - date d'accès octobre 2015
[114] https://www.centos.org/ - date d'accès octobre 2015

installation minimaliste, qui sont des dépendances qui seront téléchargées automatiquement lors de l'installation de virt-manager. Virt-manager permet de bénéficier d'une couche d'abstraction qui rend l'administration des machines virtuelles plus conviviales et rapproche l'expérience utilisateur au niveau d'un Oracle VirtualBox en conservant le contrôle complet permis par l'utilisation d'un Linux. Un paquet qui peut aussi être considéré lors de l'adoption de KVM associé à Ubuntu est ubuntu-vm-builder qui permet la création de machine virtuelle en mode 100% automatisé en se référant à un fichier de paramètres qui contient les informations de configuration nécessaires à la création et personnalisation de la machine à créer. Ubuntu-vm-builder se charge de la création des utilisateurs et de la configuration réseau ainsi que du téléchargement et de l'installation des paquets nécessaires aux personnalisations demandées. Pour résumer ce paragraphe sur la virtualisation, chaque machine installée chez l'un de nos partenaires peut être représentée par la Figure 30 le nombre de machines virtuelles variant en fonction des besoins spécifiques du site.

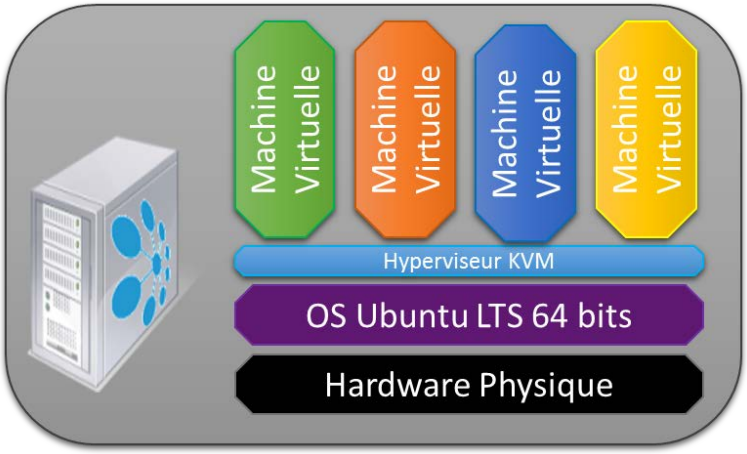


Figure 30 Représentation d'un serveur de virtualisation GINSENG

3.1.2 Rôle des différentes machines

L'infrastructure GINSENG utilise différents serveurs pour mener à bien les différentes tâches nécessaires aux partages des informations. Que ces machines soient virtuelles ou non elles se comportent vis-à-vis du système comme des machines totalement indépendantes. À l'exception du routeur (pfsense), les systèmes d'exploitation des machines suivantes sont des Ubuntu x64 LTS.

L'importeur

Chacun des sites fournissant de l'information dispose d'un serveur « importeur ». L'importeur doit récupérer les informations provenant de la base métier et intégrer ces informations à l'intérieur des bases GINSENG. Les informations originales sont récupérées selon un format pré défini au cas par cas avec la structure qui les fournit. Les données sont poussées sur le serveur FTP sécurisé (vsftpd) de l'importeur. Si nécessaire, l'archive transmise est décompressée. Puis, les informations sont parcourues une à une par un algorithme d'authentification qui recherche si le patient auquel se réfèrent les données est déjà connu, d'abord localement, puis sur les sites distants. À l'issue de cette recherche, l'algorithme retourne un identifiant déjà assigné au patient au préalable, ou nouvellement créé dans le cas contraire. En se référant à cet identifiant, les informations sont stockées dans le nœud GINSENG local. Lorsque toutes les informations ont été importées dans le nœud cela conclut la phase d'importation. L'importeur effectuera les mêmes étapes lorsque le prochain fichier sera poussé sur son serveur vsftpd.

Le nœud GINSENG

Le nœud GINSENG est la machine qui héberge la base de données distribuée. Chacun des sites fournissant de l'information dispose d'un nœud GINSENG. Le nœud GINSENG fonctionne de pair avec le serveur « importeur ». Cette machine se compose de tables MySQL *federated* qui permettent de répondre à des requêtes distantes en agrégeant les résultats de chacun des nœuds. C'est ce mécanisme qui permet actuellement le fonctionnement de façon distribuée. Nous avons auparavant expérimenté une autre approche qui se basait sur AMGA, au travers des solutions proposées par la société Gnúbila. L'algorithme d'importation du serveur « importeur » est exécuté lors de l'upload d'une base de données sur Le nœud GINSENG ; qui est ensuite interrogée par un nœud central situé chez le HADS qui distribue les requêtes aux tables *federated.*

Le routeur VPN

Le routeur est une machine optionnelle qui a pour vocation de remplacer un boitier VPN Cisco RV042 V3, si l'infrastructure ne permet pas le déploiement de ce boitier. Dans cette optique nous pouvons envisager de « virtualiser » ce composant en s'appuyant sur la solution pfsense[115] qui, dans sa version 2.2.4, permet de s'interfacer facilement avec les boitiers Cisco RV042. C'est cette solution qui a été retenue par IDS pour se connecter au sein du réseau VPN de GINSENG. Il suffit de configurer les informations nécessaires aux différentes authentifications requises lors de la négociation de la création du VPN et cette machine virtuelle se comporte de la même façon que l'un des boitiers Cisco.

Le serveur d'authentification

Le serveur d'authentification accorde l'accès à l'infrastructure GINSENG pour les utilisateurs disposant d'une carte de la famille des CPS et disposant des autorisations permettant de consulter les bases de données. Nous avons expérimenté des solutions se basant sur CAS et schibboleth en suivant les recommandations de l'ASIP Santé. Suite au rapprochement avec le GCS SIMPA et le HADS, les serveurs d'authentification, préalablement mis en œuvre, ne sont plus nécessaires. En effet, le service d'authentification des CPS fait partie des solutions que procure IDS pour l'ENRS du GCS. Ce service est de plus couplé à un outil de gestion des utilisateurs et des groupes qui permet d'assurer précisément la validité de l'authentification par la carte CPS, ainsi que l'identité de la personne et les groupes auxquels il appartient. Cette solution permet de déléguer au HADS IDS toutes les problématiques de gestion des certificats, et des révocations.

Le serveur WEB

Le serveur WEB héberge les services nécessaires à la présentation des résultats d'analyse des bases de données à travers une interface accessible à l'aide d'un navigateur WEB. Nous avons expérimenté des solutions basées sur Liferay[116]et Drupal[117] pour supporter la présentation de nos résultats. L'objectif étant de fournir un environnement graphique convivial aux utilisateurs. Pour une meilleure visibilité au niveau régional et faciliter la façon dont les usagers accèdent aux différents services numériques, nous essayons de faire partie intégrante

[115] https://www.pfsense.org - date d'accès octobre 2015
[116] http://www.liferay.com/fr - date d'accès octobre 2015
[117] https://www.drupal.org/ - date d'accès octobre 2015

de l'ENRS Auvergnat[118] dans les rubriques « Cancer » et « Grossesse ». En plus de l'homogénéité de cette solution cela permettrait de profiter pleinement des solutions proposées par IDS en matière de sécurité et d'authentification. Les alternatives permettant de profiter des solutions IDS sont accompagnées de garanties de sécurités importantes et sont agréées par l'état Français ce qui est un gage de qualité. Dès que ces solutions seront en place nous demanderons l'obtention du HONcode[119] afin de faire valider notre démarche par une organisation indépendante.

L'annuaire central

L'annuaire central se situe chez le HADS IDS. Son rôle est de maintenir une liste de correspondance qui lie les patients avec un identifiant unique pour masquer l'identité des patients. Cet annuaire est interrogé lors de l'importation des données sur « l'importeur ». Si le patient est présent dans l'une des bases de données, son identifiant est retourné, celui-ci permet de créer une identité unique non nominative (sans l'annuaire) à laquelle rattacher toutes les informations se rapportant au même patient. Si le patient n'est pas préalablement présent dans cette liste une nouvelle entrée est créée, après avoir généré un nouvel identifiant. L'annuaire permet d'assurer une bijection entre les identités numériques non nominatives et l'identité réelle des patients, dans le but d'éviter les doublons ou les duplications de pathologies liées à une seule personne physique.

Le nœud GINSENG central

La principale fonction du nœud GINSENG central est d'agréger les tables *federated* pour fournir le résultat des requêtes générées par l'interface WEB. Ce serveur se situe chez le HADS ; il accepte les requêtes créées par l'outil de création de requêtes et les exécute au nom de l'utilisateur authentifié par sa CPS. La requête SQL est distribuée par MySQL sur les différents nœuds qui sont concernées par les tables visées par la requête.

Les machines de support, dont le but est de faciliter l'administration du système, tout en augmentant le niveau de sécurité, n'ont pas un rôle direct dans le système, mais peuvent être considérées comme indispensables. Elles permettent une meilleure traçabilité, des analyses régulières et des alarmes en cas de défaillance ou de comportement anormal des installations.

[118] https://www.esante-auvergne.fr/ - date d'accès octobre 2015

[119] http://www.hon.ch/HONcode/Patients/Visitor/visitor_f.html - date d'accès octobre 2015

La gestion des anomalies est plus rapide en conservant la documentation sur le système au même endroit que les rapports sur la gestion des erreurs précédentes.

Le serveur de log et monitoring

Le but de cette machine est d'agréger les *logs* des autres serveurs. Les *logs* sont les journaux, au sens journal de bord, qui sont générés automatiquement par certains services et applications. Notre objectif est de les centraliser sur un seul serveur pour pouvoir les consulter plus aisément. Les informations inscrites dans ces journaux peuvent signaler une défaillance ou mettre en évidence les attaques subies par la machine qui les produit. Le fait de les stocker de façon distante leurs assure une pérennité de stockage même en cas d'intrusion. L'outil de capture des journaux que nous utilisons est RSYSLOG[120] il est secondé dans sa tâche par logrotate qui nous permet d'archiver au fur et à mesure les fichiers journaux, tant sur le serveur de surveillance, que sur chaque machine surveillée.

De plus, à l'aide d'outils spécialisés, comme Nagios puis Icinga 2, nous disposons de tableaux de bords permettant de superviser le comportement des différents services et composants, dès qu'une sonde est disponible, le composant est susceptible d'être surveillé. Ce serveur de surveillance est donc un élément important de la politique de sécurité du système, assimilé à une vigie qui s'assurerait du bon fonctionnement de tous les rouages. L'interface de surveillance doit être affichée en permanence sur l'écran d'un administrateur qui est en charge de maintenir le système opérationnel avec un haut niveau de disponibilité. Rappelons à toutes fins utiles qu'une garantie de 99% de disponibilité signifie qu'à l'échelle d'une année le service n'est potentiellement pas fonctionnel durant plus de 3 jours.

Figure 31 Capture d'écran d'un tableau de bord Icinga2

[120] http://www.rsyslog.com/ - date d'accès octobre 2015

Le serveur de mises à jour

Un serveur optionnel, mais vivement conseillé, est le serveur de mise à jour. En modifiant les fichiers de configurations de nos serveurs nous redirigeons leurs requêtes de mise à jour vers un serveur dont nous avons le contrôle total. Cela nous permet de valider toutes les mises à jour qui sont appliquées, ce qui limite le risque de mises à jour qui détériorent la qualité de service. Les mises à jour sont téléchargées sur les dépôts officiels puis testées en dehors des machines de production. Si le test est concluant, l'archive de la mise à jour validée est téléchargée sur notre serveur de mises à jour. Ensuite, les machines sont mises à jour à partir de notre dépôt.

Le serveur de documentation et de gestion des tickets

Pour une gestion organisée des différentes actions d'installations, de mises à jour et de réponses aux incidents, un service de gestion de tickets est mis en place. La solution Flyspray[121] disponible sous licence GPL 2.1 est utilisée sur un serveur hébergé chez notre HADS. Les avantages du système de tickets sont nombreux. Nous pouvons garder une trace de toutes les demandes effectuées qui comportent la date, l'objet, le type, la gravité de la demande, ainsi que l'identifiant de son créateur. Par la suite le gestionnaire se charge de la répartition des tâches aux différents personnels et équipes. Les personnes à qui la tâche a été dévolue peuvent reporter l'avancement de leur démarche dans cet outil en rajoutant des commentaires et des liens vers les documents de références, captures d'écrans et autres fichiers utiles à la compréhension et reproductibilité de leurs actions. Jusqu'à la clôture de la tâche le ticket reste actif, puis il sera archivé. Le fait de conserver un historique permet en cas de défaillance identique ou similaire de pouvoir se reporter aux actions déjà effectuées pour reproduire dans les meilleurs délais les solutions les plus pertinentes.

[121] http://www.flyspray.org/ - date d'accès octobre 2015

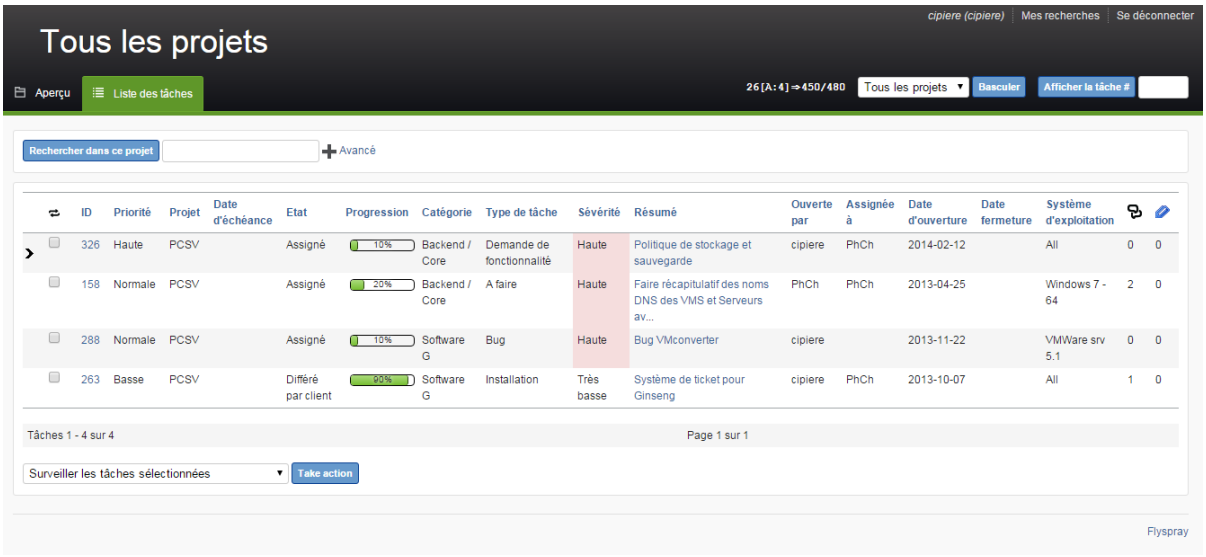


Figure 32 Capture d'écran d'une interface de gestion de ticket Flyspray

Nous n'avons pas déployé de solution de GED particulière, pour stocker la documentation relative aux différents serveurs, nous nous appuyons simplement sur une arborescence de dossier du type :

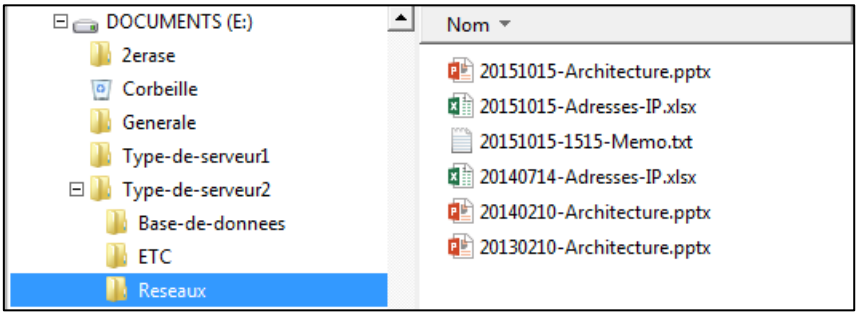


Figure 33 Illustration de la structure du stockage de la documentation

Les documents stockés sont nommés en fonction de leurs dates de création ou de modifications, ce qui permet d'avoir rapidement accès à la dernière version et de gérer facilement l'archivage.

3.1.3 Les services

Les services diffèrent forcément entre les 3 types de serveurs.

L'authentification est assurée par un Schibboleth qui permet de transférer un ticket SAML aux sites partenaires lorsque l'authentification s'est bien déroulée.

Le site Web est un site en Liferay et intègre une iFrame pour servir l'interface VIP qui est utilisée pour requêter les données, notamment au moyen d'un SPARQL EndPoint, mais il

est aussi possible d'interroger les bases grâce à du SQL² qui interroge les bases stockées avec FedEHR.

Les nœuds peuvent être abordés comme étant une somme de machines. D'une part, avec un système en charge d'importer la donnée et de la stocker sous un format FedEHR. Ce système sera aussi en charge de l'anonymisation des données pour l'épidémiologie ; c'est aussi lui qui transférera les dossiers entre les sites (si cette solution est une attente de nos partenaires). D'autre part un SPARQL EndPoint est en charge de répondre aux questions sémantiques des autres SPARQL EndPoint et de Coreze/DQP.

Le traitement des fichiers sources est effectué à l'aide de script Perl qui nécessitent la présence de certains composants supplémentaires listés dans le Tableau 14

Tableau 14 Liste des composants Perl nécessaires aux scripts GINSENG

Commandes administrateur d'installation de Perl pour GINSENG
aptitude install perl gcc make expat libexpat1-dev cpanminus
cpanm XML::Parser
cpanm XML::Twig
cpanm Text::Unaccent::PurePerl
cpanm Text::Unidecode
cpanm Text::JaroWinkler

L'importeur sémantique « Crawler FedEHR »

Afin de collecter les données disponibles dans FedEHR, nous avons construit un crawler de l'API d'accès aux données médicales, afin de les représenter au format de description de ressources des technologies du Web Sémantique : RDF. Le crawler, via un fichier de configuration, permet d'extraire indépendamment les données de chaque hôpital afin de respecter l'autonomie de leur base de connaissances respective. Pour tenir compte du grand volume de données à acquérir, le crawler propose un traitement parallèle paramétrable afin d'accélérer les traitements, tout en conservant un contrôle sur la charge des serveurs FedEHR. Une API REST a été réalisée afin de piloter l'import via une requête HTTP. L'API REST est implémentée sous la forme d'une application « Play » qui ne nécessite aucune installation de serveur et permet donc au crawler de fonctionner de façon autonome.

3.1.4 Réseau sécurisé

Pour connecter les sites médicaux distants, un réseau est nécessaire. Nous avons envisagé un médium dédié en passant par des prestataires qui pouvaient nous garantir ce service, cependant le coût trop élevé de cette solution nous a contraints à l'abandonner rapidement. De plus la volonté de permettre une accessibilité au plus grand nombre de professionnels de santé n'était pas en accord avec la solution de la connexion dédiée et spécialisée. Les sites sont actuellement équipés avec des solutions professionnelles voir même grand public d'ADSL *(Asymmetric Digital Subscriber Line)* ou SDSL *(Symmetric Digital Subscriber Line)* (4 Mbps) souscrit auprès de FAI (Fournisseur d'Accès Internet) français. Les données médicales sont des informations très sensibles. Un niveau de sécurité très élevé doit donc être appliqué à tous les stades de notre solution. Pour que chaque serveur GINSENG puisse fonctionner, il lui faut une connexion au réseau privé du serveur de données pour accéder aux informations de la base de données médicale. Sur ce canal, les données transitent unilatéralement du serveur métier vers le serveur GINSENG, sur un unique port dédié. C'est le propriétaire des données qui décide quand il souhaite partager ses informations, en les « poussant » vers le serveur d'identification et d'anonymisation. Une seconde connexion nécessaire est celle permettant de communiquer avec l'extérieur. Plusieurs configurations sont alors possibles. Lorsque que cela est possible, un boitier Cisco RV042 V3 est déployé avec chaque serveur GINSENG ; c'est lui qui sera en charge d'établir une connexion VPN IPSEC entre les différents sites comme représenté sur la Figure 34. S'il n'est pas possible de déployer ce boitier, nous pouvons le remplacer par un routeur virtuel pfsense[122], c'est la solution qui a été retenue chez notre HADS IDS. Le lien entre un CISCO RV042 et un pfsense disposant de bons paramètres de configuration, fonctionne très bien. À terme, nous pourrons peut-être considérer de n'utiliser que des routeurs virtuels pfsense pour réduire encore les coûts de déploiement. On peut ainsi envisager une solution GINSENG pour la médecine de ville qui consisterait à déployer uniquement des machines virtuelles (routeur, nœud, serveur d'identification) en parallèle de la solution métier du médecin. La troisième connexion est optionnelle, il s'agit du management de l'iDrac pour les machines DELL qui s'avère très pratiques en cas de nouvelles installations ou si l'OS n'est pas encore démarré. Dans le cas des disques encryptés, c'est la seule solution pour renseigner le mot de passe de décryptage des disques du système avant que celui-ci ne devienne opérationnel. Lorsque le site dispose de suffisamment d'adresses IP publiques ou qu'un modem spécifique a été dédié par la structure

[122] https://www.pfsense.org/ - date d'accès octobre 2015

à notre solution, la connexion vers l'extérieur dispose de sa propre IP. Si ce n'est pas possible notre solution cohabite sur l'IP utilisée par la structure en utilisant une redirection de ports spécifique. Que l'IP soit propre ou partagée la redirection des ports occupe une partie importante de la configuration, que la motivation soit purement d'ordre sécuritaire comme pour le port 22 (SSH) ou nécessaire pour cause de multi utilisation pour le port 443 (HTTPS). Nous continuons de nous appuyer sur IPV4 *(Internet Protocol Version 4)* car les réseaux avec lesquels nous nous interfaçons sont toujours en production avec ce standard et rarement en double pile IPV6.

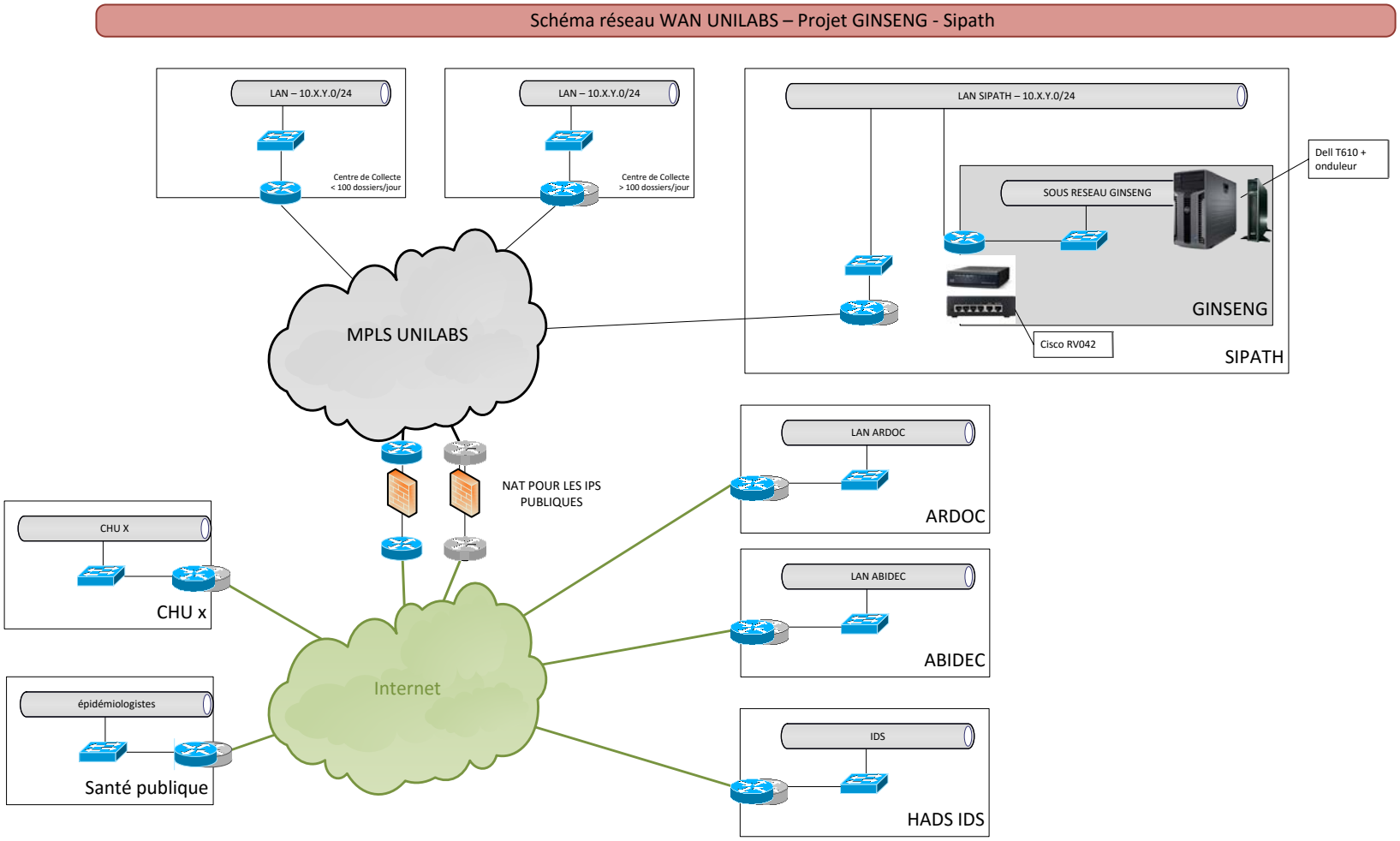


Figure 34 Vue du réseau « interne » GINSENG, exemple donné pour le site Sipath-Unilabs

En effet certains ports comme celui attribué au SSH sont parmi les premiers ports à être visités par les personnes mal intentionnées. C'est pourquoi plusieurs précautions sont nécessaires, comme interdire le SSH pour l'utilisateur root. Cette parade oblige un attaquant à trouver le nom d'utilisateur en plus du mot de passe, ce qui complexifie sa tâche. Nous pouvons aussi laisser une machine *« honey pot »* (pot de miel) écouter SSH sur le port 22 pour servir de leurre. Grâce à cette machine, il sera plus facile de capturer et étudier les tentatives de connexions frauduleuses. Pour notre trafic légitime un autre port tel que 11022 est affecté au SSH et sera redirigé vers le serveur SSH à l'intérieur de notre réseau privé. Ces solutions étant réservées aux administrateurs systèmes en dehors du VPN pour des cas très rares. Dès que cela est rendu possible, les connexions s'effectuent à l'intérieur du tunnel VPN. Pour les ports usités par différents services comme le HTTPS qui permet d'accéder à l'interface de l'iDrac, l'administration du routeur et éventuellement l'administration du vSphere ou du firewall, il est nécessaire d'utiliser une table de redirection des ports comme dans le Tableau 15.

Tableau 15 Table de redirection des ports d'un site GINSENG

Port d'entrée	*IP de sortie*	*Port de sortie/Protocole*	*Commentaires*
10443	192.168.240.99	443 / HTTPS	Management iDrac
9443	192.168.240.98	443 / HTTPS	Management Routeur
10643	192.168.240.97	443 / HTTPS	Management hyperviseur
10622	192.168.240.97	22 / SSH	SSH hyperviseur
11022	192.168.240.101	22 / SSH	SSH VM1
12022	192.168.240.102	22 / SSH	SSH VM2
13022	192.168.240.103	22 / SSH	SSH VM3

Le réseau GINSENG n'a pas pour vocation d'être totalement imperméable. Certains utilisateurs ont un besoin légitime d'accéder à certaines informations. Pour se faire, ils doivent s'authentifier à l'aide de leur carte CPS. La lecture des cartes CPS est une fonctionnalité déjà présente sur le site ENRS du GCS Simpa hébergé par IDS. C'est l'une des raisons qui nous a amené à choisir IDS comme HADS. Ainsi, nous n'avons pas eu à réécrire un composant d'authentification des CPS comme prévu dans l'une des versions antérieures du projet. Nous pouvons nous appuyer sur des solutions éprouvées mise à disposition par un professionnel agréé par l'état français. De plus, le GCS dispose d'une solution logicielle couplée au service d'authentification des CPS qui permet une gestion fine de l'appartenance du professionnel à

des groupes d'utilisateurs. Ce partenariat avec le GCS SIMPA qui est une structure qui a pour rôle l'accompagnement des projets d'e-santé en région, nous permet de déléguer l'authentification par CPS à l'ENRS et de facto à IDS.

Nous avons vu comment sont installées les machines ainsi que la façon dont le réseau est architecturé. Intéressons-nous désormais à l'architecture des données qui transitent sur ce réseau et à la façon dont elles sont traitées et stockées.

Notre solution s'appuie sur des bases de données MySQL *Community Edition – Server 5.6.26* sous licence GNU GPL[123] *(General Public Licence)*, lors de certains traitements nous avons eu recours à des bases Oracle (10 ou 11), mais le coût élevé des solutions Oracle nous a fait préférer les solutions gratuites à notre disposition. Nous utilisons InnoDB comme moteur de stockage pour MySQL notamment pour sa gestion des clefs étrangères.

3.2 La gestion des bases de données médicales

3.2.1 Structure des bases de données médicales

Chaque structure dispose de sa propre base métier, dont le schéma de la base de données lui est propre. La philosophie de notre approche est de nous adapter à l'existant pour que l'effort demandé au professionnel pour utiliser notre solution soit aussi minime que possible voir nul.

Sipath-Unilabs

Nous avons la chance de travailler avec Sipath-Unilabs, l'un des plus grands cabinets ACP de la région Auvergne. Ce qui nous permet au travers d'un seul interlocuteur d'obtenir la grande majorité des dossiers d'intérêt pour nos études.

Avec l'accord de la direction du cabinet Sipath-Unilabs nous avons pu, en collaboration avec le prestataire informatique INFOLOGIC, obtenir un fichier d'export hebdomadaire qui est poussé par sFTP chaque dimanche sur notre serveur d'identification. Ce fichier d'export contient tous les nouveaux dossiers traités par le cabinet durant la semaine (~3 000 patients/semaine). Cet export est constitué des comptes rendus textuels qui pourront être partagés avec les associations de dépistage du cancer, et un fichier XML du type présenté Figure 35 et Figure 36. La Figure 35 représente le contenu réel du fichier XML tel qu'il est agencé. La Figure 36 est une vue conceptuelle du contenu du même fichier, c'est une traduction arborescente de la Figure 35.

[123] http://www.gnu.org/licenses/gpl.html - date d'accès octobre 2015

```
<?xml version="1.0" encoding="iso-8859-1"?>
<EXPORT>1092015072605:00:24
 <NURES ID="31514">
 <LIGNE>
 <NUDDEEXT>15N404</NUDDEEXT>
 <DATPREL>2015/07/06</DATPREL>
 <DATENREG>2015/07/08</DATENREG>
 <NUPAT>-4242</NUPAT>
 <NOMPAT>CIPIERE</NOMPAT>
 <PRENOM>ELEANOR</PRENOM>
 <ADRESSE1>1 RUE DES PEUPLIERS</ADRESSE1>
 <ADRESSE2></ADRESSE2>
 <ADRESSE3></ADRESSE3>
 <CODPOSTAL>63800</CODPOSTAL>
 <VILLE>COURNON</VILLE>
 <CODPAYS></CODPAYS>
 <NOMFILLE></NOMFILLE>
 <SEXE>F</SEXE>
 <DATNAISSANCE>1980/07/12</DATNAISSANCE>
 <NOMLEC1>Florence MAURY</NOMLEC1>
 <INSEELEC1>631702255</INSEELEC1>
 <NOMLEC2>Martine GAZHALI</NOMLEC2>
 <INSEELEC2>Inconnu</INSEELEC2>
 <NOMMED>ZLOWODZKI</NOMMED>
 <INSEEMED></INSEEMED>
 <NURES>31514</NURES>
 <RESULTATCR>C:\dossier\CRdudossier.doc</RESULTATCR>
 <DATVALIDATION>2015/07/20</DATVALIDATION>
 <ADICAPS>
 <ADICAP>FCGX0000</ADICAP>
 <ADICAP>FCGX0005</ADICAP>
 </ADICAPS>
 <PTNMs></PTNMs>
 <NOMORIG>PATIENT</NOMORIG>
 </LIGNE>
 </NURES>
</EXPORT>
```

Figure 35 Exemple réduit à un patient fictif d'un export hebdomadaire Sipath-Unilabs

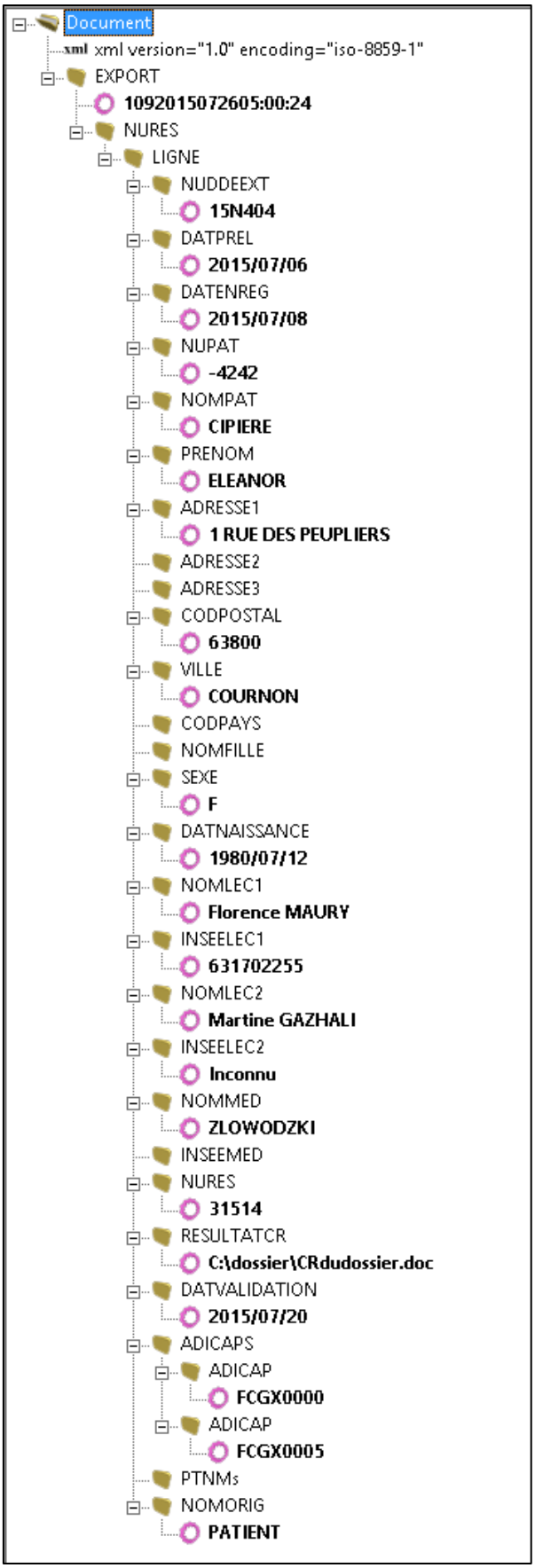


Figure 36 Représentation graphique de l'exemple de la Figure 35 (XML viewer)

En fonction de l'export que nous considérons la structure du fichier d'échange peut différer. Nous venons de présenter un fichier '.XML' issu d'un script dédié à sa création. Dans une optique de pérennisation du système d'information de Sipath-Unilabs nous avons eu accès aux cartouches magnétiques de sauvegarde. Les cartouches qui sont encore utilisées aujourd'hui sont des *Linear Tape-Open* version 2 (LTO-2) qui permettent de sauvegarder la totalité du serveur Windows chaque nuit. Après nous être procuré le lecteur adéquat et après avoir trouvé un gestionnaire de sauvegarde capable de lire la cartouche, nous avons pu « virtualiser » le serveur qui héberge la base métier. Avec l'accord de la direction Sipath-Unilabs nous avons extrait du serveur Oracle un fichier '.CSV' contenant une ligne par code ADICAP dont la liste des variables est présentée. Cette solution nous a permis d'obtenir la totalité de la base historique qu'il n'était pas possible de récupérer en utilisant le script d'export hebdomadaire.

Nous disposons donc d'un export hebdomadaire réalisé chaque dimanche depuis juillet 2014 au format '.XML'. Nous avons aussi réalisé un export historique en mai 2015 qui sera détaillé dans la partie 4 de ce manuscrit.

Après la phase d'identification, les données sont insérées dans les bases de données GINSENG. Pour structurer l'information que nous stockons nous avons utilisé des annuaires pour les patients, les variables et les « événements médicaux ». Chacune des tables stockant les annuaires portent le suffixe (_dir) pour *directory*.

Nous disposons des annuaires suivants :

-event_dir l'annuaire des « événements médicaux » (consultation, radiographie, etc.)

-var_dir l'annuaire des variables (température, poids, etc.)

-patient_dir l'annuaire des patients

Les annuaires du nœud central font référence aux tables event_dir et var_dir de chaque nœud, ce sont des copies effectuées quotidiennement.

L'annuaire patient_dir est propre à chaque site et la table du nœud central est la résultante de la somme des nœuds de chaque site.

Les données contenues dans les tables annuaires serviront de clefs étrangères pour les autres tables.

Tableau 16 Liste et signification des variables de la base métier Sipath-Unilabs

var1	NUDDE	ID ANALYSE INTERNE
var2	IDANALYSEEXT	ID ANALYSE EXTERNE
var3	IDSEJOUR	ID SEJOUR
var4	DATERECEPTION	DATE RECEPTION
var5	DATEREPONSE	DATE REPONSE
var6	IDPATIENT	ID PATIENT INTERNE
var7	NOM	NOM
var8	PRENOM	PRENOM
var9	SEXE	SEXE
var10	IDPATIENTEXT	ID PATIENT EXTERNE
var11	ADRESSE1	ADRESSE1
var12	ADRESSE2	ADRESSE2
var13	ADRESSE3	ADRESSE3
var14	CODEPOSTAL	CP
var15	VILLE	VILLE
var16	DATENAISSANCE	DATE DE NAISSANCE
var17	IDEXAMEN	ANALYSE
var18	ABRREVIATIONEXAM	ABREVIATION ANALYSE
var19	IDORIGINE	ORIGINE
var20	ABREVIATIONORIGINE	ABREVIATION ORIGINE
var21	IDMEDECIN	MEDECIN
var22	ABREVIATIONMEDECIN	ABREVIATION MEDECIN
var23	COMMENTAIRE	COMMENTAIRE
var24	ADICAP	ADICAP
var25	IDVALIDEUR	ID VALIDEUR
var26	ABREVIATIONVALIDEUR	ABREVIATION VALIDEUR

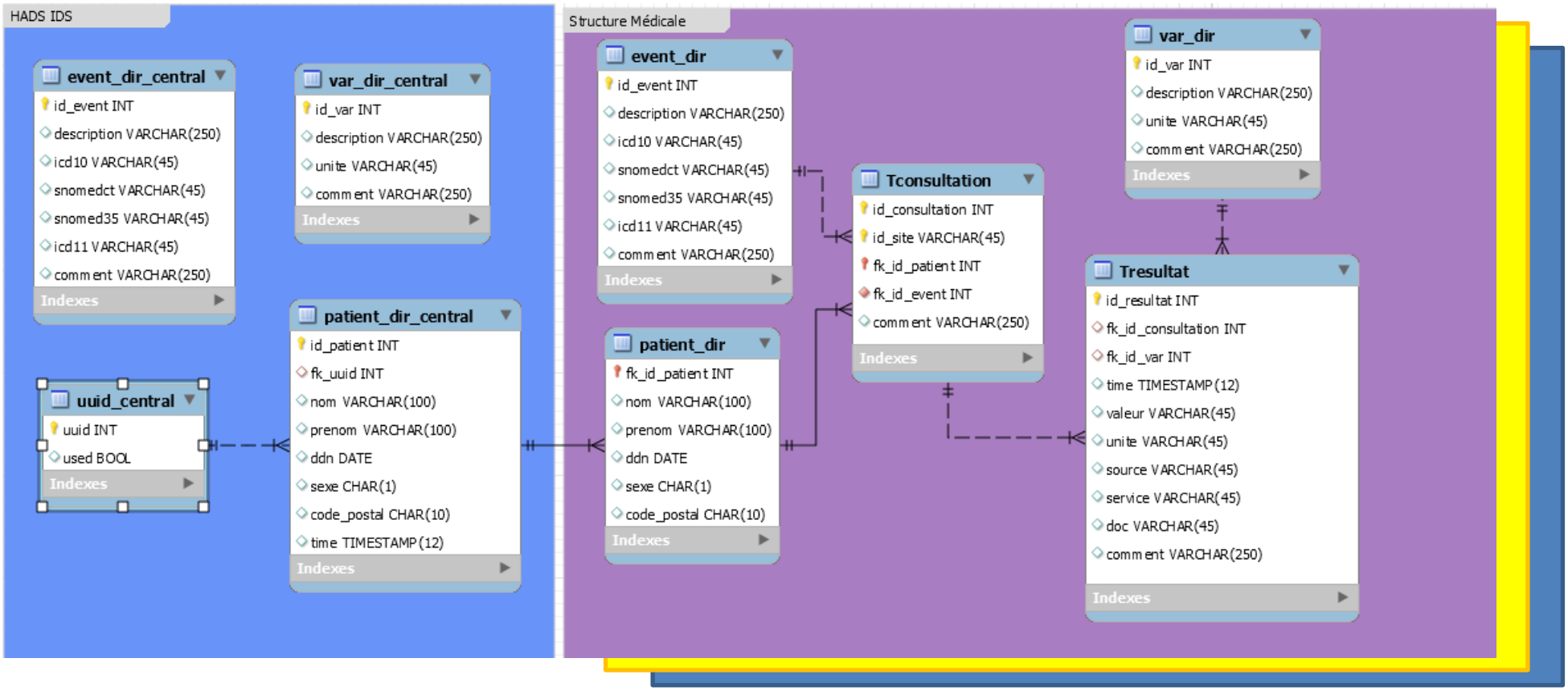


Figure 37 Représentation d'une partie des bases de données MySQL de l'infrastructure GINSENG

Notre objectif est de rendre les informations contenues dans les bases de données accessibles aux autres membres du réseau partenaire de ce cabinet. Pour se faire, le fichier va être parcouru par une routine en Perl. Cette routine parcourt le fichier patient par patient grâce aux balises 'IDANALYSE' qui délimitent les informations relatives à un nouveau patient. Nous utilisons ensuite les informations des balises NOM, PRENOM, CODEPOSTAL, SEXE ainsi que MOIS_ANNEE_NAISSANCE (mois et année), traitement rendu possible grâce à nos accords CNIL (cf. 2.1.5). Ces champs sont ceux que nous avons retenus pour identifier un patient. Ils sont alors comparés avec les différentes bases auxquelles le système a accès à l'aide de différents algorithmes. Si après avoir parcouru toutes les bases « annuaires » aucune correspondance n'est trouvée, alors une nouvelle entrée pour ce patient est créée dans l'annuaire central. Cette entrée sera associée à un numéro unique qui permet une traçabilité non nominative à l'intérieur des résultats à vocation épidémiologique. En effet, si lors de la recherche dans les annuaires, une correspondance est trouvée alors c'est cet identifiant qui serait retourné.

Les structures de gestion du dépistage organisé des cancers

La philosophie de notre approche consiste à réutiliser lorsque possible les outils déjà existant pour ne pas générer de surcoût inutile. Les SGDO dispose d'un exporteur déjà fonctionnel réalisé pour l'InVS, après avoir identifié cette solution, nous avons convenu avec OSI-santé de nous appuyer sur ces fichiers pour construire notre importeur. Ainsi notre littérature de référence est produite par l'InVS qui définit ses standards pour le dépistage du cancer colorectal dans (InVS 2013), ainsi qu'une partie fournie directement par OSI-santé.

D'après l'étude des standards nous avons pu récupérer un export au format *.CSV pour chacun des 3 types de dépistages DOCS, DOCCR et DOCU comportant les champs que nous présentons dans les paragraphes suivants (cf. Tableau 17, Tableau 18 et Figure 38).

❖ Dépistage organisé du cancer du sein

Tableau 17 Liste des variables (avec leur description) InVS contenues dans l'export du DOCS – ABIDEC 2015

dep	Numéro de département de la structure de gestion
INSEE	code INSEE de la commune de résidence
Numéro	Numéro d'enregistrement
date_nai	Date de naissance (Mois et année de naissance)
ss	Régime de sécurité sociale
atcd	Antécédents déclarés mammographie antérieure
date_atcd	Antécédents déclarés date de la mammographie antérieure
date_DO_ant	Date de la mammographie précédente dans le dépistage organisé
date_mammo	Date de la mammographie de dépistage organisé
lieu_DO	Lieu de réalisation de la mammographie de dépistage organisé
type_mammo	Type de Mammographie
vague	Rang de la mammographie de dépistage organisé dans le département (rang = vague)
THS	Traitement hormonal substitutif
clinique	Examen clinique des seins
lesion1	Tuméfaction palpable
lesion2	Lésion eczématiforme du mamelon
lesion3	Rétraction
lesion4	Inflammation
lesion5	Écoulement du mamelon
lesion6	Adénopathie
lesion7	Autre type de lésion
L1	Résultat de la première lecture avant bilan
densite	Densité mammaire
echo_hp	Échographie sur mammographie acr1 ou acr2
BC_L1	Bilan diagnostic effectue par le radiologue en l 1 = bilan immédiat 5bdi°
motif_BC	Motif du bilan
agrandiL1	Agrandissement effectue par le radiologue en l 1
echoL1	Échographie effectuée par le radiologue en l 1
cytoL1	Cytoponction effectuée par le radiologue en l 1
result_BC1	Résultat final de la première lecture après bilan l1
CAT_L1	Conduite à tenir proposée par le radiologue en l 1
cti	Cliché techniquement insuffisant (CTI)
L2	Résultat de la deuxième lecture avant bilan
bc_L2	Bilan diagnostic demandé par le deuxième lecteur
result_BC2	Résultat de la deuxième lecture après bilan différé
CAT_L2	Conduite à tenir finale proposée suite au bilan diffère
delai	Délai entre le dépistage et le résultat envoyé a la femme
surv	Mise sous surveillance
date_surv	Date de l'examen de contrôle effectue dans le cadre de la mise sous surveillance
cyto	Cytoponction

micro	Microbiopsie
macro	Macrobiopsie (ou mammotome)
biopsie	Biopsie chirurgicale ou exérèse de la tumeur
avis	Avis spécialisé ou autres examens réalises (ex IRM)
diag	Situation finale de la procédure de dépistage diagnostic
date_biop	Date de résultat de la biopsie chirurgicale (ou date de réalisation)
date_macro	Date de résultat de la macrobiopsie (ou date de réalisation)
date_micro	Date de résultat de la microbiopsie (ou date de réalisation)
date_cyto	Date de résultat de la cytoponction (ou date de réalisation)
tumeur	Tumeur primitive
taille	Taille de la tumeur primitive
type	Type de la tumeur primitive
ganglion	Classification de l'envahissement ganglionnaire
gg_senti	Ganglion sentinelle
metastase	Métastases à distance
clas_grade	Classification du grade histopronotique utilisée
grade	Grade
date_chir	Date du traitement chirurgical
date_rx	Date du traitement par radiothérapie
date_chimio	Date du traitement par chimiothérapie
date_horm	Date du traitement par hormonothérapie
date_immuno	Date du traitement par immunothérapie
k_inter	Cancer d'intervalle
date_k_inter	Date de diagnostic du cancer de l'intervalle
cdc	Cahier des charges
CAD	Utilisation du CAD pour la première lecture
type_cli	Type de clichés de deuxième lecture

Notre approche qui consiste à créer une table quasi identique aux données qui sont exportées nous permet lors de l'import de ne perdre aucune des informations communiquées. Par la suite nous pouvons les stocker sous d'autres formes en s'alignant avec le ou les standards retenus. De plus les informations d'intérêt pour l'InVS sont forcément à priori intéressantes pour les épidémiologistes qui traitent les données statistiques au sein du réseau GINSENG.

❖ Dépistage organisé du cancer du col de l'utérus

Pour le dépistage organisé du col de l'utérus (DOCU), les données sont moins nombreuses. Il s'agit essentiellement des dates importantes de la campagne de dépistages et du frottis ainsi que les résultats sous forme de code ADICAP et Bethesda.

Tableau 18 Liste des variables (avec leur description) InVS contenues dans l'export du DOCU – ABIDEC/ARDOC 2015

PER_INTID :	Numéro de Dossier Zeus
INV_DTMEDITION :	Date d'édition de l'invitation
INV_DTMRELANCE :	Date de Relance
INV_BITFICHIERCOL :	Faux = Invitation envoyé par la SG
BITPREMIERFROTTIS :	Vrai = 1er Frottis de la personne.
DTMREMBFROTTIS :	Date de remboursement du frottis
DTMFROTTIS :	Date du Frottis
TINEXAMENGYNECO :	Type d'examen gynécologique
TINMODALITEFROTTIS :	Modalité du frottis
REA_DTMDATE :	Date d'interprétation du Frottis
TINQUALITEPRELEV :	Qualité du prélèvement
VCHCODEADICAPBENIN :	Code ADICAP Bénin
VCHCODEADICAPGLANDULAIRE :	Code ADICAP Glandulaire
VCHCODEADICAPMALPIGHIEN :	Code ADICAP Malpighien
VCHCODEADICAPAUTRELESION :	Code ADICAP autre lésion cancéreuse
INTIDBETHESDABENIN :	Identifiant bénin du Bethesda
INTIDBETHESDAGLANDULAIRE :	Identifiant Glandulaire du Bethesda
INTIDBETHESDAMALPIGHIEN :	Identifiant malpighien du Bethesda
CodeSpecialiteMedPrescipteur :	Code spécialité du médecin prescripteur
DATEFROTTISCONTROLE :	Date de Frottis de contrôle
ConclusionFrottisCtrl :	Conclusion du frottis de contrôle
INTIDBETHESDABENINFrottisCtrl :	Identifiant bénin du Bethesda
INTIDBETHESDAMALPIGHIENFrottisCtrl :	Identifiant malpighien du Bethesda
INTIDBETHESDAGLANDULAIREFrottisCtrl :	Identifiant glandulaire du Bethesda

❖ Dépistage organisé du cancer colorectal

La Figure 38 présente la requête de création de la table `` `Torigin_doccr_abidec` `` de la base de données situées à l'ABIDEC. Cette table a pour vocation de récupérer sans perte l'export fournit par le prestataire de service informatique de l'ABIDEC. Cet export est à l'origine prévu pour l'InVS. Il est adapté à nos besoins pour nous permettre de charger les bases du réseau GINSENG.

```
create table `Torigin_doccr_abidec` (
  `PER_INTID` bigint(20) NOT NULL,
  `INV_INTID`  varchar(11) DEFAULT '99999999999',
  `Ivs_dtmNA` date DEFAULT '1900-01-01',
  `Ivs_dtmTest` date DEFAULT '1900-01-01',
  `RES_INTID`  varchar(1) DEFAULT '9',
  `intAnnee`  varchar(4) DEFAULT '1900',
  `intCampagne`  varchar(1) DEFAULT '9',
  `PER_DTMDECES` date DEFAULT '1900-01-01',
  `dep`  varchar(3) DEFAULT '999',
  `insee`  varchar(5) DEFAULT '99999',
  `numero`  varchar(11) DEFAULT '99999999999',
  `date_nai` date DEFAULT '1900-01-01',
  `sexe`  varchar(1) DEFAULT '9',
  `ss`  varchar(2) DEFAULT '00',
  `date_1ere_invit` date DEFAULT '1900-01-01',
  `date_relance_sans_test` date DEFAULT '1900-01-01',
  `date_relance_avec_test` date DEFAULT '1900-01-01',
  `Date_1er_test` date DEFAULT '1900-01-01',
  `date_DO_ant` date DEFAULT '1900-01-01',
  `rang`  varchar(2) DEFAULT '00',
  `type_test`  varchar(1) DEFAULT '9',
  `nb_NA`  varchar(2) DEFAULT '99',
  `cause_na`  varchar(1) DEFAULT '0',
  `result_test`  varchar(1) DEFAULT '9',
  `date_result_test` date DEFAULT '1900-01-01',
  `date_envoi_depistee` date DEFAULT '1900-01-01',
  `date_envoi_med` date DEFAULT '1900-01-01',
  `coloscopie`  varchar(1) DEFAULT '9',
  `date_colo` date DEFAULT '1900-01-01',
  `prepa_colo`  varchar(1) DEFAULT '9',
  `quali_colo`  varchar(1) DEFAULT '9',
  `motif_quali`  varchar(1) DEFAULT '9',
  `result_colo`  varchar(1) DEFAULT '9',
  `topographie`  varchar(2) DEFAULT '99',
  `asp_macro`  varchar(1) DEFAULT '9',
  `lavb`  varchar(1) DEFAULT '9',
  `result_lavb`  varchar(1) DEFAULT '9',
  `imagerie`  varchar(1) DEFAULT '9',
  `result_imag`  varchar(1) DEFAULT '9',
  `autre_exam`  varchar(1) DEFAULT '9',
  `result_aut`  varchar(1) DEFAULT '9',
  `acc_colo`  varchar(1) DEFAULT '9',
  `type_acc`  varchar(1) DEFAULT '9',
  `duree_hosp`  varchar(1) DEFAULT '9',
  `deces`  varchar(1) DEFAULT '9',
  `taille_macro`  varchar(3) DEFAULT '999',
  `prelevement`  varchar(1) DEFAULT '9',
  `nb_polype`  varchar(2) DEFAULT '99',
  `nb_polype10`  varchar(2) DEFAULT '99',
  `histo`  varchar(2) DEFAULT '99',
  `dysplasie`  varchar(1) DEFAULT '9',
  `diag`  varchar(1) DEFAULT '9',
  `pec`  varchar(1) DEFAULT '9',
  `date_pec` date DEFAULT '1900-01-01',
  `RadioT_preop`  varchar(1) DEFAULT '9',
  `date_radiot_preop` date DEFAULT '1900-01-01',
  `ttt_postop`  varchar(1) DEFAULT '9',
  `T`  varchar(1) DEFAULT '9',
  `N`  varchar(1) DEFAULT '9',
  `date_prlvt` date DEFAULT '1900-01-01',
  `type_prelev`  varchar(1) DEFAULT '9',
  `tumeur`  varchar(1) DEFAULT '9',
  `ganglion`  varchar(1) DEFAULT '9',
  `gg_preleve`  varchar(2) DEFAULT '99',
  `metastase`  varchar(1) DEFAULT '9',
  `cim10`  varchar(4) DEFAULT '9999'
) ENGINE=InnoDB DEFAULT CHARSET=latin1;
```

Figure 38 Requête SQL de la création de la table stockant les données originales du DOCCR de l'ABIDEC – version 2015

Le Réseau de Santé Périnatale Auvergne

Le RSPA est l'un des partenaires historiques du projet GINSENG, mais son intégration dans les CH et CHU rend l'accès aux bases de données particulièrement difficile. Il existe cependant une base régionale « centralisée » du RSPA, hébergée chez un HADS et soutenue par le GCS Simpa. Cette base est en cours de migration vers le même HADS que celui que nous exploitons actuellement IDS. Cela pourrait simplifier nos démarches, car pour accéder à la totalité de la base une seule autorisation serait alors nécessaire. La base de RSPA est gérée par la société ICOGEM qui s'interface avec le logiciel métier ICOS maternité. Les développements de ICOS respectent les standards édités par l'AUDIPOG (cf. 1.5.6). Le standard AUDIPOG permet de structurer des milliers de variables d'une façon hiérarchique autour de 20 points majeurs :

1. Identification
2. Renseignements généraux
3. Antécédents médicaux
4. Antécédents obstétricaux
5. Début de grossesse
6. Clôture du dossier (sans accouchement à la maternité)
7. Examen complémentaire de la grossesse
8. 1er contact à la maternité
9. Transfert maternel vers un autre établissement
10. Synthèse de la grossesse
11. Conduite à tenir pour l'accouchement
12. Admission
13. Accouchement (niveau mère)
14. Accouchement (niveau enfant)
15. Nouveau né en salle de naissance
16. Résumé du séjour de l'enfant à la maternité
17. Décès (synthèse)
18. Résumé du séjour de la mère
19. Hospitalisation postnatale de la mère
20. Transfert néonatal (ou mutation)

Le traitement des autres types de fichiers provenant de structures de santé différentes est identique. Dans un premier temps, une récupération des données nécessaire à l'identification, suivie d'une phase d'identification qui peut se conclure par la création d'une nouvelle fiche annuaire avec un identifiant unique pour le patient (sauf cas particulier que nous évoquerons dans la partie 3.2.2 dédiée aux algorithmes d'identifications). Puis insertion des données dans la base de stockage des résultats. Et ainsi de suite jusqu'à ce que tous les patients du fichier à importer soient traités.

Nous venons de présenter d'une façon générique l'identification des patients, intéressons-nous plus précisément aux différentes étapes de ce processus avant de nous consacrer aux détails des algorithmes nécessaires à l'identification du patient.

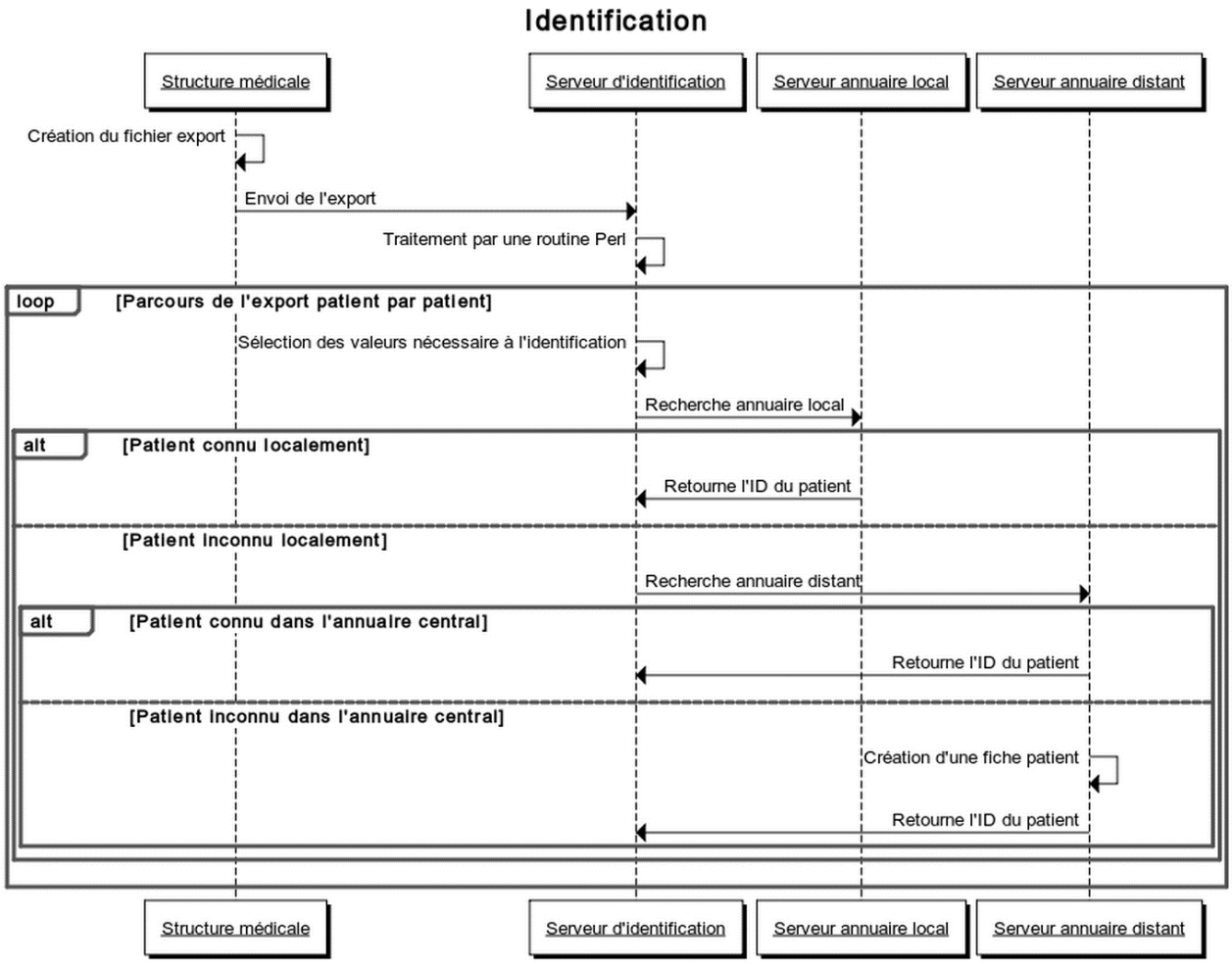


Figure 39 Processus d'identification d'un patient lors de l'importation des données

Une fois l'ID (l'IDentifiant) du patient récupéré les données sont importées dans la base de données résultats qui se situe sur une autre machine virtuelle. C'est une routine Perl qui

effectue cette opération grâce notamment au composant DBI[124] – (*Database Independent interface for Perl*) qui permet d'accéder aux fonctionnalités de la base de données, dans le cas des bases MySQL (cf. 3.2.3).

3.2.2 Les algorithmes d'identification des patients

Les données confidentielles d'identification simulées

Nous avons travaillé entre 2012 et 2015 sur des données simulées en attendant les accords CNIL, nous permettant d'utiliser les données réelles des patients. Ce travail a permis de valider certaines de nos hypothèses résumées dans la thèse (Li 2015). Ces données simulées sont primordiales pour la validation des algorithmes d'identification ; en effet comme nous contrôlons la génération des couples d'individu que nous cherchons à appareiller en ajustant le niveau de bruit sur les informations permettant l'identification, nous pouvons ajouter un champ contenant un identifiant unique pré calculé pour chaque individu. Cet identifiant unique permet à postériori de vérifier le bon déroulement des différentes étapes du chaînage. Pour simuler les données des patients nous nous sommes procuré les listes INSEE des prénoms et noms les plus répandus en France. Le but étant que les données simulées soit le plus vraisemblable possible. Nous avons généré les prénoms pour leur associer un sexe adéquat. Les dates de naissances sont générées aléatoirement pour obtenir une population dont l'âge minimal était 15 ans, jusqu'à 120 ans. Les codes postaux proviennent des bases INSEE. La base de référence contient donc les champs :

Nom ; prénom ; date de naissance ; sexe ; code postal de résistance

Ces données correspondent aux données demandées et par la suite autorisées par nos accords CNIL.

À chaque entrée de la base un identifiant unique nommé Control Code (CT) permet d'identifier avec certitude le patient. La base de référence a été bruitée de nombreuses manières, permutation, ajout et effacement de caractères (sur l'ensemble des champs et pour chacun des champs); pour créer des bases de tests ; tout en conservant l'identifiant de contrôle. L'identifiant de référence permet de juger avec certitude si le rapprochement a été ou non effectué correctement. La Figure 40 nous présente le résultat d'une itération d'un de nos algorithmes d'identification. Nous recherchons les individus de notre base de référence (listés dans la colonne de gauche par leur CT ID), dans une base bruitée, les individus susceptibles d'être la

[124] http://dbi.perl.org/ - date d'accès octobre 2015

même personne que celle recherchée sont listés dans la colonne centrale. Les scores de rapprochements apparaissent à droite.

Sur la première ligne nous recherchions CT6480 et nous avons trouvé CT6480 avec une certitude de 100% (la donnée n'avait pas été bruitée). Sur la troisième ligne nous recherchons CT10903 et nous n'avons qu'un seul candidat qui est bien CT10903 mais cette fois avec une certitude plus faible de 0.978979. La recherche se complique lors de la recherche de CT9204 qui lui a été particulièrement impacté par le bruitage de cette série. En effet la bonne personne à rapprocher dans cette base ne se trouve qu'en 11ème position avec un score de 0.787145. L'algorithme pense ici que 10 patients différents sont plus susceptibles d'être rapprochés avec CT9204 que lui-même.

Cette étape a été importante pour définir les seuils à partir duquel nous considérons qu'une donnée même légèrement bruitée peut être traitée comme identique.

Par la suite nous avons eu accès aux données provenant de nos partenaires, ce qui nous a donné l'opportunité de confronter nos solutions au monde réel.

Patient Recherché	Patient(s) Rapproché	Score de rapprochement
CT6480	CT6480	1
CT1734	CT1734	1
CT10903	CT10903	0.978979
CT8196	CT8196	0.990048
CT11757	CT11757	1
CT1069	CT1069	0.977199
CT9204	CT4236	0.853297
	CT328	0.819835
	CT7742	0.818288
	CT8227	0.814525
	CT8799	0.808626
	CT6757	0.804295
	CT10380	0.803964
	CT11884	0.793799
	CT7362	0.791016
	CT4294	0.789049
	CT9204	0.787145
CT2792	CT2792	1

Figure 40 Exemple de résultats pour la validation des algorithmes d'identification sur données simulées

La pondération des données d'identification

Pour comparer si deux données correspondent bien à un même patient nous décomposons les champs qui le caractérisent, puis non appliquons la comparaison à l'aide de l'algorithme d'identification, avant d'appliquer une pondération à chaque partie précédemment décomposée. Cette suite d'opération nous permet au final d'obtenir une distance totale normalisée (cf. Figure 41), Cette valeur représente le degré de similarité des deux jeux de données comparés. Plus cette distance est proche de 1, plus grande sont les chances que les informations soient relatives à une seule et même personne comme nous le montre la Figure 40.

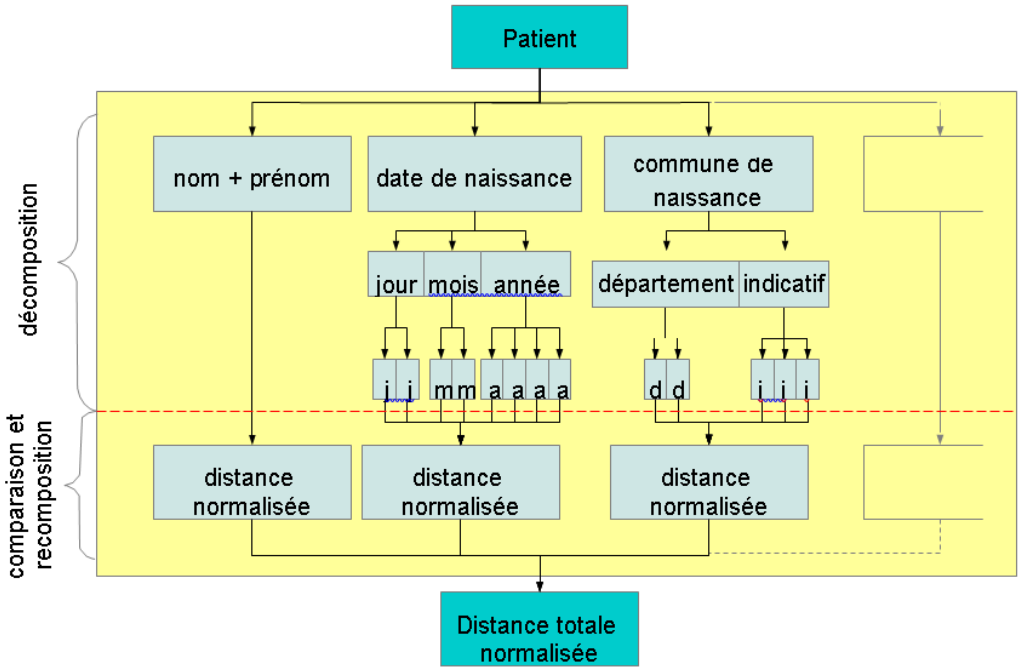


Figure 41 Représentation graphique la mesure de distance entre les patients

De plus, chaque champ n'a pas la même valeur discriminante. Nous cherchons à comparer nom de famille, prénom, code postal, sexe ainsi que le mois et année de naissance des patients. Le nom d'une femme peut évoluer durant sa vie notamment en cas de mariage. Le sexe de l'individu codé sur une seule lettre 'F' pour les femmes 'H' (ou 'M') pour les hommes. Ces deux lettres sont séparées par une seule et unique touche sur un clavier français standard (la touche 'G'). Nous pouvons donc considérer qu'une erreur sur le sexe est bien plus vite arrivée que sur l'un des autres champs lorsque la saisie n'est pas limitée à deux champs. Ainsi lors de la comparaison dans l'annuaire chaque champ se verra attribué un poids relatif à sa valeur discriminante. Ces poids sont le résultat d'heuristiques qui ont permis de différencier de façon assez catégorique les patients de nos échantillons de tests. Cette pondération est la

résultante de tests réalisés sur des bases de données simulées et bruitées (entre 1 000 et 500 000 patients).

Il est possible que ces poids nécessitent un étalonnage sur des bases dont le contenu ne soit pas similaire aux bases que nous considérons, nous pouvons penser à des bases russes ou hispaniques dont les caractères diffèrent des caractères français. De la même façon nous avons porté une attention toute particulière aux accents qui peuvent facilement changer. Ainsi, notre solution supporte la lecture des accents mais pour une meilleure compatibilité, ceux-ci ne sont pas considérés. Les accents sont remplacés par la lettre non accentuée qui les compose, comme présenté dans le Tableau 19, que le caractère soit en majuscule ou non le mécanisme reste le même. Il en est de même pour la cédille ou l'e dans l'o.

Tableau 19 Tableau de correspondance entre les caractères particuliers et leur caractère de substitution

Lettres accentuées ou avec cédille	Lettre support
à	a
é, è, ë, ê	e
i, ï, î	i
ù, û	u
ô	o
ç	c
œ	Oe

Nous avons vu que notre système repose sur des serveurs MySQL pour conserver les données des patients. Lors de la phase d'identification des patients nous allons donc comparer les informations d'un fichier source qui peut être un '.CSV' ou '.XML' qui sera parcouru à l'aide d'un script Perl et comparé au contenu d'un annuaire qui est une table MySQL qui peut être hébergée sur un serveur local ou distant chez notre HADS. Pour l'identification nous avons principalement étudié deux requêtes, l'une dite « simple », la seconde basée sur l'algorithme Jaro-Winkler.

La première requête dite « exacte » consiste à comparer caractère par caractère les différents champs que nous avons retenus pour l'identification. L'écriture de la requête en SQL consiste à vérifier que tous les champs sont strictement identiques les uns avec les autres. Aucune erreur n'est permise avec cette solution, le résultat est donc binaire. Cependant, le moindre caractère qui diffère comme une apostrophe inversée « ‘ » ou même un espace « » sur un nom composé, ou n'importe quel autre champ considéré, entraine une mauvaise identification. Cette solution simple a l'avantage d'un temps d'exécution réduit, mais est évidemment trop simpliste pour être utilisée seule. Nous avons donc besoin d'un mécanisme

capable d'identifier certaines typographies et si possible de le corriger ou à minima de les signaler à un opérateur qui sera plus à même d'évaluer la situation, et pourra trancher.

La seconde méthode que nous considérons, est basée sur les travaux de Jaro et Winkler que nous avons présenté dans le chapitre 1. L'algorithme créé par ces deux chercheurs porte naturellement le nom de Jaro-Winkler ; il est souvent considéré comme donnant de très bons résultats, mais reste peu utilisé car couteux en temps de calcul. Nos recherches nous ont permis de trouver une version de cet algorithme portée en MySQL. Nous n'avons pas réussi à trouver l'auteur de ce travail. Ayant plusieurs champs à comparer, nous aurons donc besoin de faire appel plusieurs fois à cette fonction à l'intérieur de notre routine d'identification.

Les approches hybrides et alternatives

Nous avons étudié d'autres algorithmes basés sur la sonorité des mots pour essayer de détecter des nom ou prénom qui seraient mal orthographiés suite à un quiproquo lié à une incompréhension phonétique. Pour se faire nous sommes partis de l'algorithme « Phonex » qui peut se décomposer en 2 parties : le codage phonétique et la comparaison de ces codes. Le codage comme décris sur l'algorithme original nous est apparu inutilement lourd. Il a donc été réorganisé pour effectuer le moins de passage possible sur la chaîne à identifier.
Ceci fait, le nombre de passages est réduit de 16 à 3, décris par :

1. remplacer les lettres 'y' par des 'i'
2. effectuer les transformations sur tous les sons sauf les 's'
3. remplacer les sons 's' et supprimer les lettres doubles

Dans l'algorithme originel le premier passage traitait le caractère 'y' utilisé pour coder un son, que nous avons rapproché du caractère 'i'. Dans notre approche le second passage demandait une analyse plus pointue du caractère courant et de ses voisins, ce passage synthétise quatorze passages de la version originelle. Enfin le dernier passage devait être séparé car il utilisait le codage de l'étape précédente. Pour la comparaison des codes, l'algorithme de Jaro-Winkler est utilisé comme discuté précédemment. Nous avons ici allégé un code existant et remplacer sa partie comparaison par l'algorithme de référence Jaro-Winkler.

3.2.3 Les requêtes sur les bases de données

Les requêtes peuvent être liées à l'identification et à l'importation du patient comme nous l'avons décrit dans le paragraphe précédent. Elles permettent de créer les bases de données GINSENG. Une fois ces bases générées les requêtes qui donnent le plus de valeur ajoutée à l'information, sont les requêtes exploratoires des épidémiologistes. Nous détaillerons ces requêtes dans le chapitre 4 qui nous permettra de juger de leurs temps d'exécution. Ce sont des requêtes qui permettent d'identifier des bassins de population à risque, que l'on souhaite étudier sur un laps de temps particulier. On peut imaginer la recherche d'incidence ou non, sur les populations, de l'implantation d'un incinérateur à proximité de bassins urbains. Un autre avantage de GINSENG est sa capacité à être multi sources, ainsi, on peut envisager d'intégrer des données issues de l'Open Data ou d'autres sources d'information, comme la cartographie radon de la région Auvergne. On peut ainsi comparer les cas de cancers par code postal en surimpression de la cartographie radon à l'échelle d'un département.

Le dernier grand type de requêtes qui permet de faire gagner du temps aux différents secrétariats consiste à rechercher un patient dans les différentes structures de santé. Par exemple dans le cas où une personne a été invitée pour un dépistage du cancer, mais dont nous ne disposons pas des suites données à cette invitation. Désormais, il suffit de demander au système si cette personne dispose de résultats relatifs au dépistage du cancer dans une de structure partenaire, à une date postérieure à l'invitation. Le système se charge alors de distribuer la demande à toutes les structures partenaires et répond en fonction de la présence ou non de résultats pour ce patient.

Techniquement les requêtes reposent soit sur du SQL qui interroge des bases de données MySQL (federated) soit du SPARQL basé sur les travaux de la société Mnemotix.

La société Mnemotix a développé un entrepôt de données sémantiques distribuées et autonomes basées sur le socle technologique suivant :

- langages et protocoles du Web Sémantique[125] standardisés par le W3C[126]
- le moteur sémantique Corese/KGRAM développé par l'équipe Wimmics de l'INRIA[127]
- le framework d'applications Web "Play"[128]

[125] http://www.w3.org/2001/sw/ - date d'accès avril 2016
[126] http://www.w3.org - date d'accès avril 2016
[127] http://wimmics.inria.fr/corese - date d'accès avril 2016
[128] http://www.playframework.com/ - date d'accès avril 2016

Ontologie "Semantic Electonic Health Record" (SemEHR)

Sur la base du modèle de données FedEHR, Mnemotix a créé une ontologie RDFS[129] modulaire et extensible permettant de représenter et fédérer des données médicales. Cette ontologie se veut suffisamment générique pour modéliser des données « patient » de divers domaines médicaux. Par exemple, cette ontologie permet de modéliser des données médicales traitant aussi bien des cancers du sein et de la pratique des césariennes. La Figure 42 représente les principales classes et propriétés de cette ontologie.

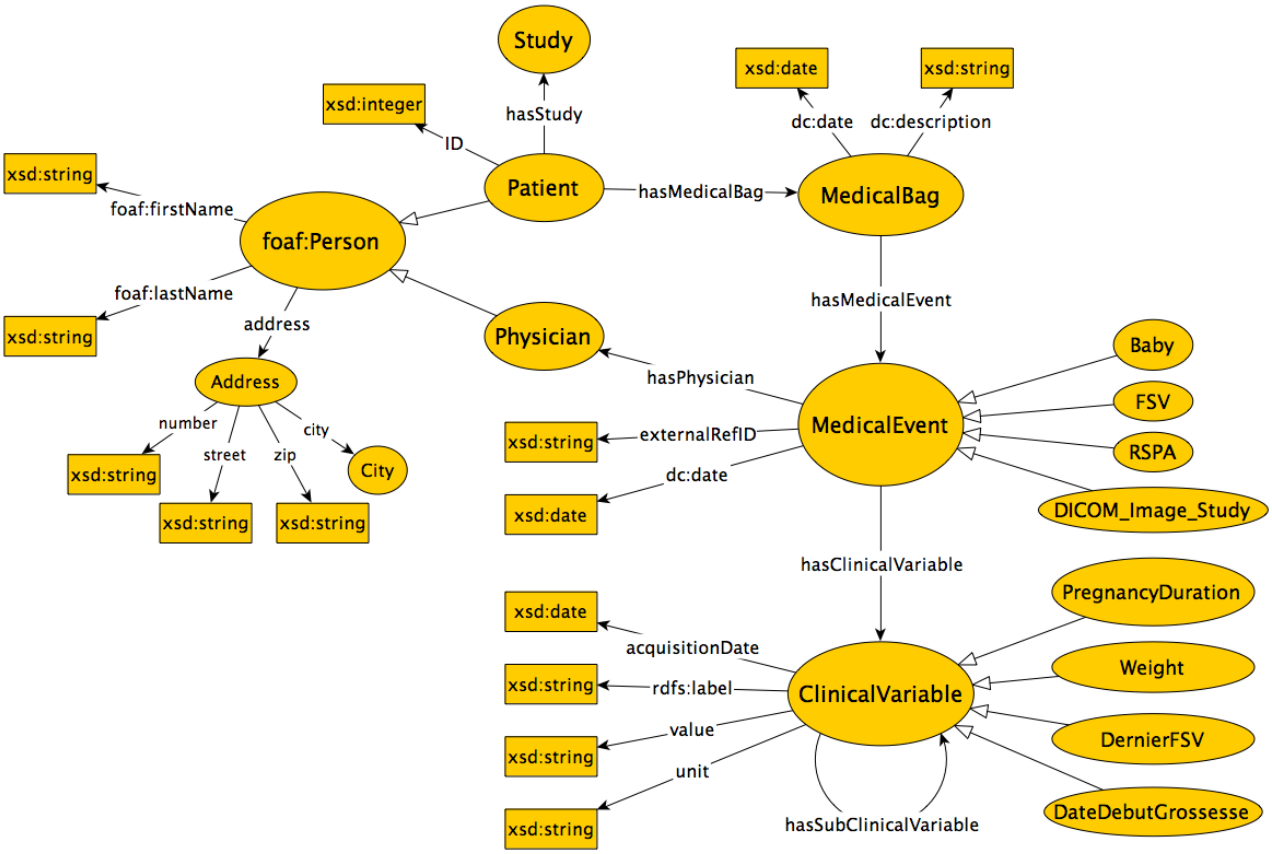


Figure 42 Noyau de l'ontologie SemEHR

Les classes `Patient` et `Physician` étendent la classe `Person` de l'ontologie `FOAF`[130] pour représenter les patients et les praticiens. Les patients sont associés à des mallettes médicales (`MedicalBag`) permettant de leur associer des ensembles d'évènements médicaux (`MedicalEvent`). La classe `ClinicalVariable` permet de représenter les variables cliniques associée aux évènements médicaux. Diverses propriétés, ainsi que leurs contraintes d'applications, permettent de décrire les propriétés des instances de ces classes et les relier.

[129] http://www.w3.org/TR/rdf-schema/ - date d'accès avril 2015

[130] Friend Of A Friend ontology http://xmlns.com/foaf/spec/ - date d'accès avril 2015

Cette base est extensible pour représenter différents domaines médicaux et raisonner sur les données correspondantes, en définissant des sous classes de chacun de ces concepts. En particulier, les données extraites à partir de la base FedEHR nous permettent de définir une taxonomie d'évènements médicaux, et de variables cliniques. Par exemple, la classe `MedicalEvent` peut être étendue pour créer une taxonomie des différents types de prélèvements (ex : `Prelevement`, `Prelevement-cervico-vaginal`, etc.). Similairement la classe `ClinicalVariable` est étendue pour créer une taxonomie de variables cliniques liées à une grossesse (ex : `Poids-du-bebe`, `type-accouchement`, etc.).

3.2.4 L'accès aux données médicales

Nous venons de voir, dans les paragraphes précédents, l'importance des requêtes SQL dans notre solution. Pour nous assurer de la portée de chacune de nos requêtes, des utilisateurs spécifiques sont créés en fonction du besoin, avec un périmètre d'action se limitant exclusivement à ses besoins. Par exemple le script Perl effectuant une requête du type « connaissez-vous ce patient ? » n'a besoin que des droits en lecture sur l'annuaire. Ce script pourra donc interroger l'annuaire au nom d'un utilisateur « script_perl_identification », ce dernier ne disposant que des droits de lecture sur l'annuaire visé par la requête.

À l'intérieur du projet nous avons identifié deux groupes d'utilisateurs, le premier souhaite partager ses données avec d'autres structures médicales qu'il a préalablement définies, le second, est orienté sur les analyses statistiques. Ces deux groupes partagent l'accès aux données en deux unités fonctionnelles : dans la première, les données sont nominatives et les dossiers peuvent être transférés d'un site à l'autre ; alors que dans la seconde les patients sont des numéros et les différentes bases ne sont interrogées que lors d'une requête épidémiologique. Dans notre approche, les producteurs des données restent toujours « maîtres » des données qu'ils produisent et uniquement si le patient ne s'est pas exprimé contre leur utilisation (cf. Figure 22, comme nous l'autorisent nos accords CNIL (cf. 2.1.5).

Si deux ou plusieurs sites souhaitent partager leurs données, ils définissent clairement leurs besoins, attentes et limitations qui sont consignés dans un cahier des charges, qui est annexé à un accord signé entre les différentes parties. Puis la solution technique qui répond à ce cahier des charges est mise en œuvre. Pour les enquêtes statistiques, la démarche est quasi identique. Le but et les moyens de l'étude sont présentés aux différents intervenants qui se prononcent sur leur envie ou non de prendre part à l'étude en mettant à disposition des chercheurs une partie de leurs bases de données non nominatives. Les tenants et aboutissants sont clairement détaillés dans un accord qui une fois signé lie les chercheurs aux structures

médicales. Ces accords peuvent contenir des clauses de confidentialité, de non publication sans accord préalable, ou au contraire des obligations de rendre publiques les résultats pour mieux informer les populations.

Cette partie d'authentification par CPS n'est pas encore concrétisée. Le fait que le porteur du projet GINSENG ne soit pas une structure de santé a ralenti notre intégration dans l'ENRS géré par le GCS SIMPA, car nous sommes soumis à un accord préalable de l'ARS et ne pouvons dès lors intégrer le GCS que lors d'une AG (Assemblée Générale) qui ne se produit qu'une à deux fois par an. Une fois les accords signés, des comptes utilisateurs sont créés, les utilisateurs sont forcément détenteurs d'une carte de la famille CPS. Notre partenariat avec le GCS SIMPA et le HADS IDS nous permet de profiter de leur système de gestion des CPS qui comporte, en plus du module d'authentification des CPS, un module de gestion des utilisateurs permettant une gestion fine des droits d'accès de chaque utilisateur.

Les résultats sont présentés sous une forme accessible par un navigateur web. Là encore nous aimerions pleinement faire partie du portail ENRS pour une meilleure lisibilité. Jusqu'à présent nous avons expérimenté une solution basée sur Liferay qui était proposée par nos partenaires de Gnùbila France, mais à la fin du partenariat prévu par le projet ANR les coûts de maintenance de leur solution se sont avérés trop importants pour être supportés par nos partenaires médicaux.

Conclusion

Nous avons présenté la mise en œuvre des solutions OpenSource pour la création d'une infrastructure distribuée hautement sécurisées pour le partage de données médicales. Depuis la configuration des nœuds serveurs, les services qu'ils hébergent, leur mise en réseau jusqu'à la structuration des bases de données médicales utilisées. L'infrastructure est entièrement opérationnelle entre le cabinet ACP Sipath-Unilabs, les SGDO et le HADS. Le réseau est sécurisé par l'utilisation d'un tunnel VPN formé par des boitiers physiques et des routeurs virtuels, auxquels s'ajoutent des pare-feu matériels et logiciels. L'interface WEB n'a pas encore été développée dans sa version finale, mais elle a été conceptualisée, et des versions de tests ont été expérimentées. La connexion par CPS a été testée et validée depuis une interface WEB (Liferay) de tests ; et sera mise en œuvre dans un futur proche avec le support de notre HADS. Il a été validé que l'infrastructure pourrait être étendue à d'autres partenaires (laboratoires de biologie, centres hospitaliers et/ou centres de lutte contre le Cancer, etc.) ainsi qu'à une utilisation plus avancée des requêtes sémantiques de manière à enrichir les données médicales avec des données en accès libre. Le tableau 18 récapitule tous les développements réalisés pour le projet, leur état ainsi qu'un descriptif. Dans le chapitre 4 est présenté une validation des choix techniques présentés dans ce chapitre ainsi que différents tests de performance.

Tableau 20 État d'avancement des différents sous projets constituants le réseau GINSENG

Sous projet	État	Commentaire
Accords CNIL	Obtenus	Renouvelable chaque année Cf. Annexe
Réseau sécurisé	Fonctionnel	VPN Cisco et Pfsense
Connexion avec le HADS	Fonctionnelle	IDS
Import/Export des données	Fonctionnel	Sipath-Unilabs ARDOC ABIDEC ABIDEC/ARDOC
Import/Export des données	En attente	Laboratoire de biologie Base ICOS régionale CHU CJP
Interface graphique	Fonctionnel	OSI santé, Zeus
Interface graphique Web	En attente	GCS Simpa
Interface graphique Web	En cours	IDS
Automatisation totale	En cours	Transfert Décompression Import
Monitoring et reporting complet	En cours	Icinga hyperviseurs Icinga VMs Analyse journaux routeurs
Identification des patients	En cours de validation	
Authentification par CPS	En attente	GCS Simpa
Authentification par CPS	En cours	IDS
Virtualisation des serveurs	Fonctionnelle	
Virtualisation des composants réseaux	Fonctionnelle	Si nécessaire (exemple HADS)
Base de données distribuées	En cours de validation	MySQL fédéré

Chapitre IV

– Validation, tests de performance et résultats

Introduction

Les validations et tests de performance réalisés au cours du travail de thèse ont d'abord concerné l'identification des patients présents dans les différentes bases de données médicales. Il est en effet primordial de pouvoir garantir une exhaustivité maximale des données concernant chaque patient par un chaînage bien établi entre les dossiers présents dans les différentes bases de données. La validation des algorithmes d'identification a dans un premier temps été réalisée sur des bases de données simulées avant de les mettre à l'épreuve des bases de données réelles du projet. Ensuite, l'importation des données médicales depuis chaque base de données métier vers chaque serveur de l'infrastructure a été testée pour garantir des temps de réponse raisonnable pour une utilisation en production. Nous présentons ensuite l'état d'avancement concernant les requêtes sémantiques qui ont pu être développées pour le projet. Finalement, nous présentons les résultats d'une étude épidémiologique réalisée grâce à l'infrastructure déployée, concernant l'incidence du radon sur la santé de la population auvergnate par l'étude de cas de cancers du poumon.

4.1 L'identification des patients et importation des données médicales

4.1.1 Descriptif du contenu des bases de données

La base de données ACP

Le cabinet d'ACP Sipath-Unilabs, partenaire du projet GINSENG, réalise une grande partie des analyses dont les résultats sont nécessaires aux SGDO pour suivre les patients invités.

Un export synthétisant les analyses « de la semaine » a été mis en place depuis juillet 2014 ; il est effectué chaque dimanche, puis exporté sur notre serveur pour traitement. Nous disposons également d'un export « historique » regroupant la totalité des archives et les dossiers courants depuis la création de la structure (1982, cf. 0) jusqu'à avril 2015. Ce dernier export complet est intéressant car il permet de réaliser des analyses sur une longue période. Pour mieux comprendre le contenu de la base de données (BDD) nous avons réalisé une analyse descriptive.

Cette base de données que nous étudions compte **1 768 033 patients** et comptabilise **9 647 383 résultats**. Les résultats sont des codes ADICAP. Une analyse peut produire plusieurs codes ADICAP stockés de façon indépendante. La répartition du genre des patients est résumée dans la Figure 43 nous avons identifié :

72.41% de femmes (1 280 313),

24.48% d'hommes (432 879), et donc

3.10% d'indéterminés (54 840).

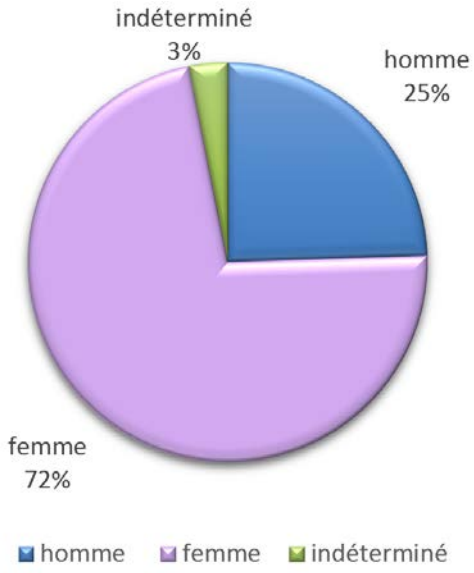


Figure 43 Répartition du genre des patients dans la base Sipath-Unilabs avril 2015

Le genre du patient est codé sur un caractère ‘F’ pour les femmes et ‘M’ pour les hommes. Si l’entrée ne correspondait pas à ce critère le genre du patient a été considéré comme indéterminé. Durant l’étude de l’âge des populations concernées, nous avons trouvé une moyenne de 58,39 ans.

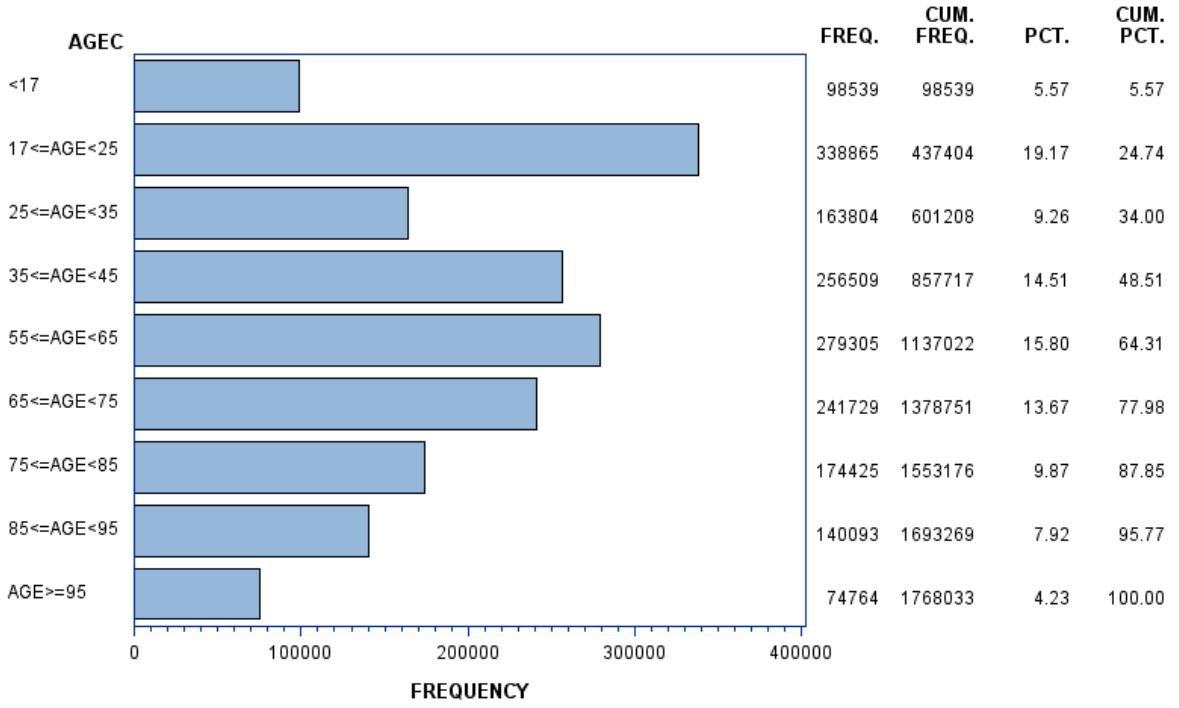


Figure 44 Pyramide des âges des patients dans la base Sipath-Unilabs avril 2015

Nous avons trouvé comme date du premier enregistrement, le 01/01/47, information qui semble erronée, il semble plus crédible que les premières expérimentations datent de 1982 et que le système d’enregistrement électronique ait réellement commencé en 1988 pour passer en routine. L’enregistrement le plus récent que nous considérerons pour cette étude date du 31/03/2015. Les imports hebdomadaires ont été traités en plus de l’import de la base « historique » ici analysée. Actuellement, nous avons noté que trois exports par an ne sont pas réalisés correctement. Les exports sont identifiés par l’incrémentation d’un entier qui est bien incrémenté chaque semaine. Cependant certains fichiers ‘n’ n’ont pas été poussé sur notre serveur de fichiers alors que nous disposons bien des fichiers ‘n-1’ et ‘n+1’. Notre hypothèse actuelle est que la tâche est tuée avant d’avoir eu le temps de se terminer, ce qui peut se produire pour des semaines de fortes activités. Nous n’avons pas noté de récurrence ou de schéma particulier reliant les semaines manquantes.

Tableau 21 Nombre d'enregistrements par an dans la base Sipath-Unilabs - avril 2015

Année	Fréquence	Pourcentage	Fréquence cumulée	Pourcentage. cumulé
1947	1	0.00	1	0.00
1982	336	0.00	337	0.00
1983	20	0.00	357	0.00
1984	36	0.00	393	0.00
1985	64	0.00	457	0.00
1986	125	0.00	582	0.01
1987	481	0.00	1 063	0.01
1988	44 484	0.46	45 547	0.47
1989	86 752	0.90	132 299	1.37
1990	176 133	1.83	308 432	3.20
1991	175 464	1.82	483 896	5.02
1992	224 103	2.32	707 999	7.34
1993	249 799	2.59	957 798	9.93
1994	273 344	2.83	1 231 142	12.76
1995	267 034	2.77	1 498 176	15.53
1996	269 376	2.79	1 767 552	18.32
1997	283 249	2.94	2 050 801	21.26
1998	284 641	2.95	2 335 442	24.21
1999	269 726	2.80	2 605 168	27.00
2000	307 691	3.19	2 912 859	30.19
2001	**719 754**	**7.46**	3 632 613	37.65
2002	385 592	4.00	4 018 205	41.65
2003	405 255	4.20	4 423 460	45.85
2004	374 163	3.88	4 797 623	49.73
2005	341 688	3.54	5 139 311	53.27
2006	365 368	3.79	5 504 679	57.06
2007	465 858	4.83	5 970 537	61.89
2008	497 808	5.16	6 468 345	67.05
2009	495 309	5.13	6 963 654	72.18
2010	502 481	5.21	7 466 135	77.39
2011	497 044	5.15	7 963 179	82.54
2012	491 135	5.09	8 454 314	87.63
2013	517 507	5.36	8 971 821	93.00
2014	542 554	5.62	9 514 375	98.62
2015	133 008	1.38	**9 647 383**	100.00

Si nous considérons l'origine géographique des patients 99.83% sont français (1 765 006) et 0.17% (3 027) proviennent d'un autre pays. À un niveau de granularité plus fin l'Auvergne agrège 57.71% (1 020 291) des résultats dont 36.80% (650 603) pour le Puy-de-Dôme.

Concernant les codes ADICAP (cf. Figure 16 & Figure 17 - p. 46), à l'aide des caractères numérotés 3 et 4 du code ADICAP nous pouvons focaliser notre étude sur des parties précises de l'anatomie. Par exemple la base contient :

173 809 résultats codant pour le sein (code « GS ») : 1.80%,

2 636 280 en relation avec le col utérin (code GC, GE) : 27.33% et

569 295 se référant au colon (code DC) : 5.90%.

Ces 3 régions du corps humains sont mises en exergue car ce sont celles visées par le dépistage organisé du cancer en région Auvergne par les associations ARDOC, ABIDEC et ABIDEC/ARDOC.

Les proportions importantes de femmes par rapport aux hommes, ainsi que de code ADICAP codant pour le col de l'utérus s'explique par le fait que les frottis cervicaux vaginaux chez les femmes sont des prélèvements très fréquents.

La base ABIDEC

Pour faciliter les démarches avec nos partenaires de l'ABIDEC nous utilisons des exports initialement destiné à l'InVS (Jezewski Serra and Salines 2013) (cf. 3.2.1). Cette approche est vraie pour les trois types de dépistages que les structures de gestion du dépistage organisé (SGDO) auvergnates gèrent actuellement. Cette démarche évite de créer un nouvel exporteur pour GINSENG. C'est l'importeur de GINSENG qui s'adapte aux choix des producteurs de données. Cette approche permet de ne pas générer de résistance à l'adoption de GINSENG liée à un surcoût en termes de préparation de la donnée lors de la mise à disposition des informations.

- Dépistage Organisé du Cancer Colo Rectal (DOCCR)

Pour le DOCCR le fichier d'export comporte 66 variables (InVS 2013). Ces variables ont des visées statistiques et couvrent neuf domaines :

1. Les caractéristiques sociodémographiques du patient.
2. Les résultats et les dates relatives au test de dépistage.
3. Les différentes informations de consultation spécialisée suite à un test positif.
4. Les caractéristiques de la lésion la plus péjorative.
5. Les informations sur les examens complémentaires.
6. Si nécessaire des données relatives aux incidents lors de la coloscopie.
7. Les résultats anatomo-cytopathologiques.
8. Les dates et types de prise en charge.
9. Les descriptions et classifications des structures cancéreuses.

Suite à l'intégration dans notre solution nous avons identifié 150 127 enregistrements liés à des patientes, 119 537 liés à des hommes, pour un total de 269 664 enregistrements, qui correspondent à 116 036 patients uniques. Soit une moyenne de 2.32 enregistrements par patient.

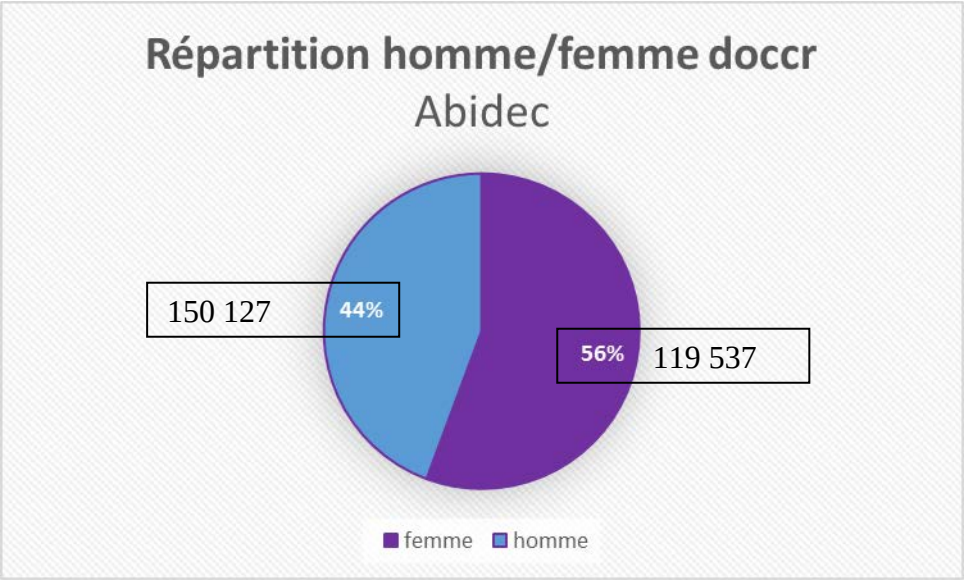


Figure 45 Répartition homme/femme pour le DOCCR – ABIDEC – juillet 2015

Tableau 22 Nombre d'enregistrements par an dans la base DOCCR ABIDEC - juillet 2015

Année	Fréquence	Fréquence cumulée	Pourcentage	Pourcentage cumulé
2004	13 719	13 719	5.09	5.09
2005	25 571	39 290	9.48	14.57
2006	16 583	55 873	6.15	20.72
2007	22 495	78 368	8.34	29.06
2008	20 168	98 536	7.48	36.54
2009	24 144	122 680	8.95	45.49
2010	**36 728**	**159 408**	**13.62**	59.11
2011	26 216	185 624	9.72	68.84
2012	28 950	214 574	10.74	79.57
2013	26 210	240 784	9.72	89.29
2014	26 332	267 116	9.76	99.06
2015	2 548	**269 664**	0.94	100.00

❖ Dépistage Organisé du Cancer du Sein (DOCS)

Pour le DOCS nous disposons de 73 variables lors de l'export organisé come suit :

1. Les caractéristiques sociodémographiques de la patiente.
2. Les résultats de l'examen clinique des seins, et de la première lecture.
3. Les résultats de la deuxième lecture.
4. La situation finale.
5. Les dates clefs, des étapes du dépistage, et du traitement.

La base de patients du DOCS est constituée à 100% de femme. Nous disposons de 301 049 entrées, les plus anciennes datent de 2003 comme le montre le Tableau 23.

Tableau 23 Nombre d'enregistrements par an dans la base DOCS ABIDEC - juillet 2015

Année	Fréquence	Fréquence cumulée	Pourcentage	Pourcentage cumulé
2003	17 341	17 341	5.76	5.76
2004	19 950	37 291	6.63	12.39
2005	21 341	58 632	7.09	19.48
2006	22 859	81 491	7.59	27.07
2007	23 976	105 467	7.96	35.03
2008	25 425	130 892	8.45	43.48
2009	24 989	155 881	8.30	51.78
2010	26 791	182 672	8.90	60.68
2011	26 104	208 776	8.67	69.35
2012	**27 343**	**236 119**	**9.08**	78.43
2013	26 186	262 305	8.70	87.13
2014	26 791	289 096	8.90	96.03
2015	11 951	**301 047**	3.97	100.00

❖ Annuaire

Après avoir effectué l'import des bases ABIDEC notre annuaire contient 319 848 patients uniques.

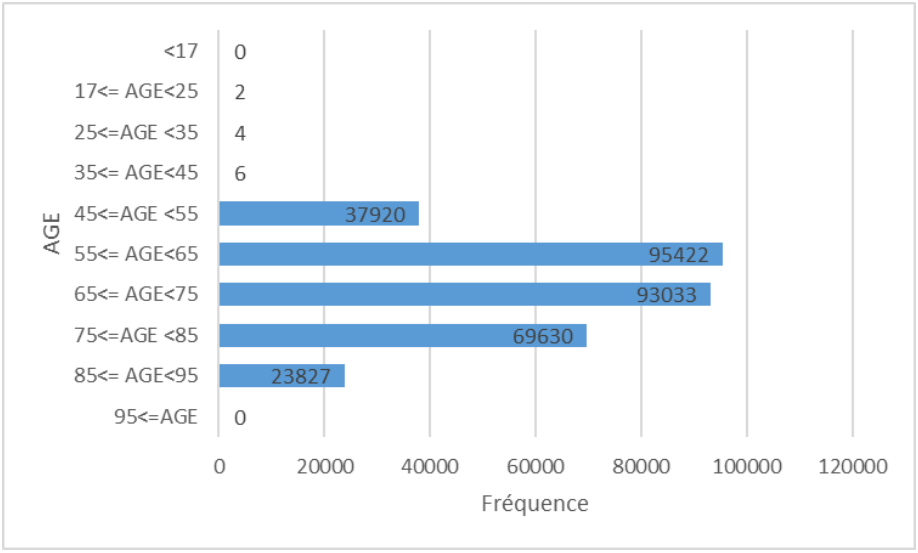


Figure 46 Pyramide des âges des patients dans l'annuaire ABIDEC - juillet 2015

Tableau 24 Nombre de patients par tranches d'âge dans l'annuaire ABIDEC-juillet 2015

Age	Fréquence	Fréquence cumulée	Pourcentage	Pourcentage cumulé
<17	0	0	0	0
17<= AGE<25	2	2	0.00	0.00
25<=AGE <35	4	6	0.00	0.00
35<= AGE<45	6	12	0.00	0.00
45<=AGE <55	37 920	37 932	11.86	11.86
55<= AGE<65	**95 422**	**133 354**	**29.83**	41.69
65<= AGE<75	93 033	226 387	29.09	70.78
75<=AGE <85	69 630	296 017	21.77	92.55
85<= AGE<95	23 827	319 844	7.45	100.00
95<=AGE	0	**319 844**	0.00	100.00

La base ARDOC

La base ARDOC est structurée de la même façon que le base ABIDEC, en effet ces deux associations poursuivent les mêmes objectifs (DOCS & DOCCR), sur deux territoires différents et avec des moyens similaires (O.S.I Santé), ce qui permet de mutualiser les développements.

- Dépistage Organisé du Cancer Colo Rectal (DOCCR)

L'analyse des données DOCCR de l'ARDOC nous indique que nous disposons d'un total de 452 133 affectées à 251 325 femmes et 200 808 hommes cf. la Figure 47. Une recherche de doublon nous a permis d'estimer le nombre de patient « unique » à 198 391. Ce qui signifie que cette base contient en moyenne 2.28 enregistrements par personne. Le plus ancien enregistrement est daté de 2004, nous avons représenté le nombre d'enregistrement annuel dans le Tableau 25.

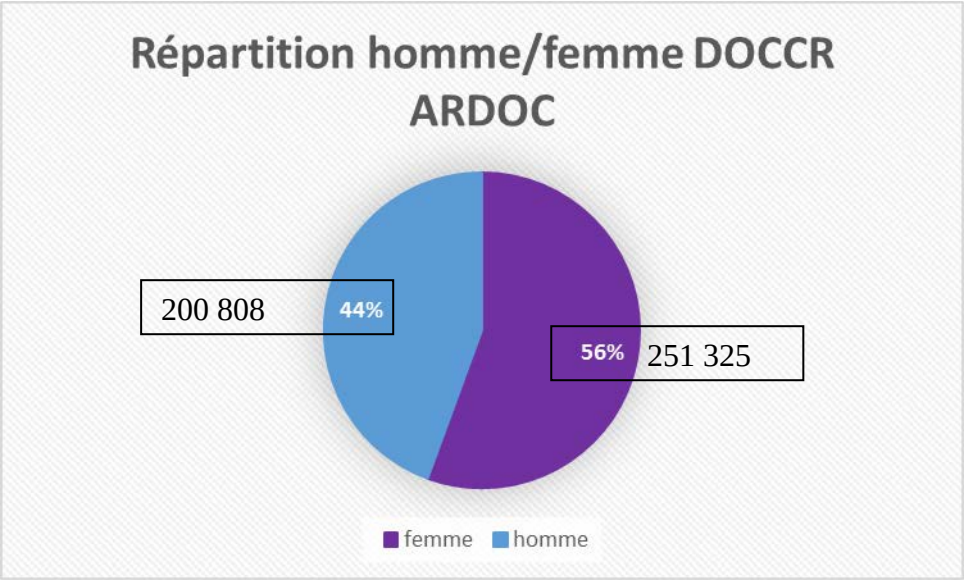


Figure 47 Répartition homme/femme pour le DOCCR – ARDOC – juillet 2015

Si nous comparons les données DOCCR de l'ABIDEC et de l'ARDOC la répartition homme, femme est similaire même si le volume total varie quasiment du simple au double.

Tableau 25 Nombre d'enregistrements par an dans la base DOCCR ARDOC - juillet 2015

Année	Fréquence	Fréquence cumulée	Pourcentage	Pourcentage cumulé
2004	7 640	7 640	1.69	1.69
2005	36 714	44 354	8.12	9.81
2006	27 848	72 202	6.16	15.97
2007	37 504	109 706	8.29	24.26
2008	51 435	161 141	11.38	35.64
2009	**54 656**	**215 797**	**12.09**	47.73
2010	51 395	267 192	11.37	59.10
2011	46 508	313 700	10.29	69.38
2012	49 519	363 219	10.95	80.33
2013	43 074	406 293	9.53	89.86
2014	43 920	450 213	9.71	99.58
2015	1 920	**452 133**	0.42	100.00

❖ Dépistage Organisé du Cancer du Sein (DOCS)

Nous disposons de 517 267 entrées dont 166 093 patientes distinctes, soit une moyenne de 3,11 résultats par patiente. Le premier enregistrement date de 2003, les apports annuels sont listés dans le Tableau 26.

Tableau 26 Nombre d'enregistrements par an dans la base DOCS ARDOC - juillet 2015

Année	Fréquence	Fréquence cumulée	Pourcentage	Pourcentage cumulé
2003	26 953	26 953	5.21	5.21
2004	37 735	64 688	7.30	12.51
2005	34 668	99 356	6.70	19.21
2006	41 240	140 596	7.97	27.18
2007	40 521	181 117	7.83	35.01
2008	40 797	221 914	7.89	42.90
2009	**49 033**	**270 947**	**9.48**	52.38
2010	41 583	312 530	8.04	60.42
2011	48 838	361 368	9.44	69.86
2012	42 686	404 054	8.25	78.11
2013	46 229	450 283	8.94	87.05
2014	44 855	495 138	8.67	95.72
2015	22 129	**517 267**	4.28	100.00

❖ Annuaire

L'annuaire ARDOC regroupe les patients susceptibles d'être invités pour un dépistage DOCS ou DOCCR actuellement dans la tranche d'âge ciblée par le dépistage (50 à 74 ans). Les personnes plus âgées que la fenêtre cible, mais ayant déjà été invitées au moins une fois, continuent de peupler cet annuaire. C'est pourquoi il est normal que les moins de 50 soient quasi inexistants dans cet annuaire, et les tranches de plus de 74 ans se peuplent par glissement du public visé, mais sont conservées à titre d'archives (cf. Figure 49 et Tableau 27). Nous retrouvons un ratio homme/femme de 45/55 (cf. Figure 48) très similaires à ceux observés dans le cadre du DOCCR (cf. Figure 45 et Figure 47).

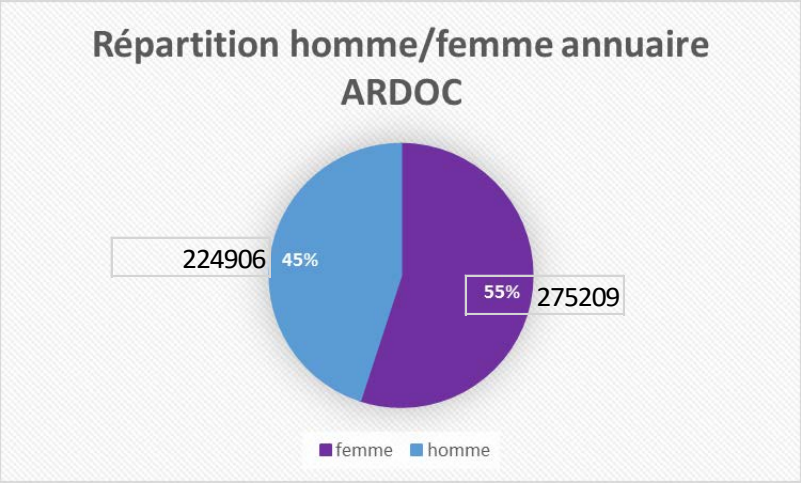


Figure 48 Répartition homme/femme pour l'annuaire – ARDOC – juillet 2015

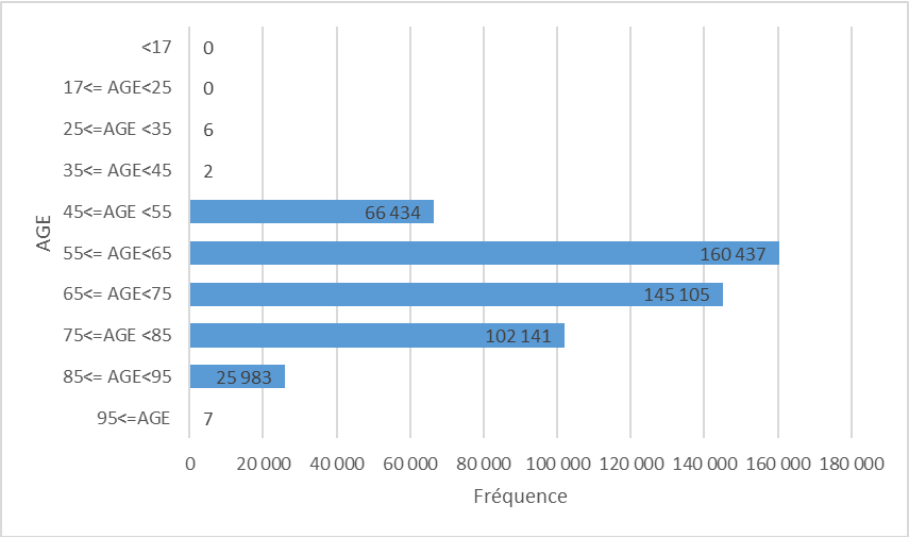


Figure 49 Pyramide des âges des patientes dans l'annuaire ARDOC - juillet 2015

Tableau 27 Nombre de patients par tranches d'âge dans l'annuaire ARDOC - juillet 2015

Age	Fréquence	Fréquence cumulée	Pourcentage	Pourcentage cumulé
<17	0	0	0.00	0.00
17<= AGE<25	0	0	0.00	0.00
25<=AGE <35	6	6	0.00	0.00
35<= AGE<45	2	8	0.00	0.00
45<=AGE <55	66 434	66 442	13.28	13.29
55<= AGE<65	**160 437**	**226 879**	**32.08**	45.37
65<= AGE<75	145 105	371 984	29.01	74.38
75<=AGE <85	102 141	474 125	20.42	94.80
85<= AGE<95	25 983	500 108	5.20	100.00
95<=AGE	7	**500 115**	0.00	100.00

La base ABIDEC-ARDOC

L'ABIDEC-ARDOC se consacre au Dépistage Organisé du Cancer Utérin, les bases ne contiennent pas les mêmes informations que pour l'ARDOC et l'ABIDEC.

- Dépistage Organisé du Cancer Utérin (DOCU)

La population est composée à 100% de femmes, comme pour le dépistage organisé du cancer du sein. Le public visé est plus jeune en effet, la tranche d'âge du DOCU est 25 à 65 ans. La pyramide des âges Figure 51 est clairement décalée vers les tranches d'âges plus jeunes que celles de l'ARDOC ou de l'ABIDEC Figure 46 et Figure 49.

Nous disposons de 608 431 entrées pour 281 352 patientes, ce qui correspond à une moyenne de 2.16 résultats par patiente. Chaque entrée dispose de 24 variables qui permettent la réalisation d'études statistiques. Parmi ces valeurs se trouve notamment les dates clefs du dépistage, les codes ADICAP du frottis et les conclusions qui en découlent.

Tableau 28 Nombre d'enregistrements par an dans la base DOCU ABIDEC-ARDOC - juillet 2015

Année	Fréquence	Fréquence cumulée	Pourcentage	Pourcentage cumulé
2005	210	210	0.03	0.03
2006	171	381	0.03	0.06
2007	5 231	5 612	0.86	0.92
2008	52 863	58 475	8.69	9.61
2009	82 590	141 065	13.57	23.19
2010	93 544	234 609	15.37	38.56
2011	**95 849**	**330 458**	**15.75**	54.31
2012	95 440	425 898	15.69	70.00
2013	90 852	516 750	14.93	84.93
2014	69 336	586 086	11.40	96.33
2015	22 336	**608 422**	3.67	100.00

❖ Annuaire

L'annuaire que nous avons constitué comporte 430 833 patientes qui sont toutes des femmes. Nous nous sommes intéressés à leur lieu de résidence nous avons trouvé une répartition géographique de 109 104 pour l'Allier (03), 43 659 pour le Cantal (15), 65 793 pour la Haute-Loire (43) et 202 976 pour le Puy-de-Dôme (63) cf. Figure 50.

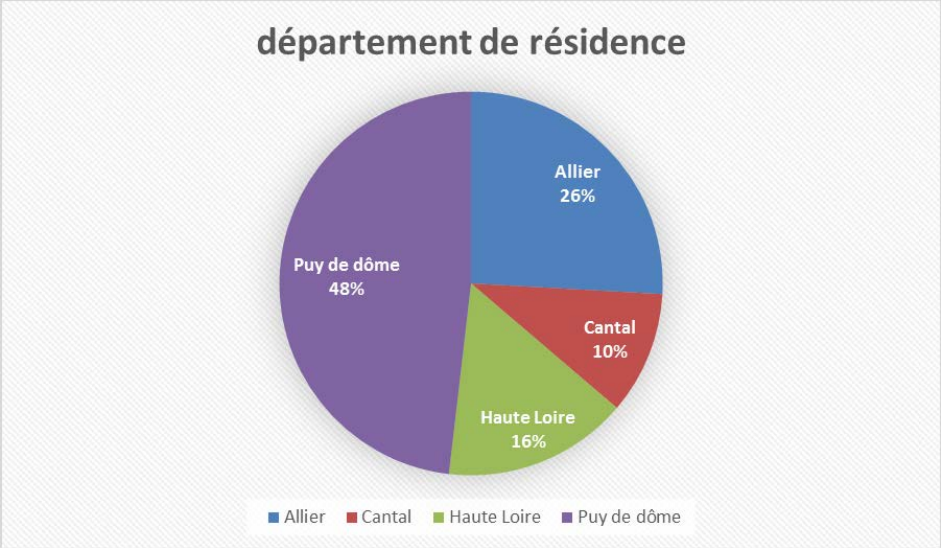


Figure 50 Répartition des lieux de résidence des patientes référencées dans l'annuaire ABIDEC-ARDOC – juillet 2015

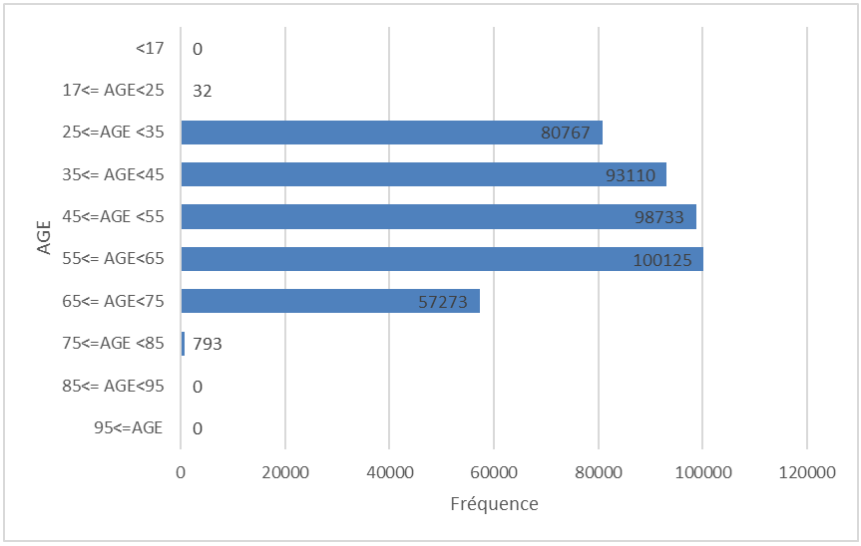


Figure 51 Pyramide des âges des patientes dans l'annuaire ABIDEC-ARDOC - juillet 2015

Tableau 29 Nombre de patients par tranches d'âge dans l'annuaire ABIDEC-ARDOC - juillet 2015

Age	Fréquence	Fréquence cumulée	Pourcentage	Pourcentage cumulé
<17	0	0	0	0
17<= AGE<25	32	32	0.01	0.01
25<=AGE <35	80 767	80 799	18.75	18.75
35<= AGE<45	93 110	173 909	21.61	40.37
45<=AGE <55	98 733	272 642	22.92	63.28
55<= AGE<65	**100 125**	**372 767**	**23.24**	86.52
65<= AGE<75	57 273	430 040	13.29	99.82
75<=AGE <85	793	430 833	0.18	100.00
85<= AGE<95	0	430 833	0.00	100.00
95<=AGE	0	**430 833**	0.00	100.00

Pour mémoire la population de l'Auvergne au dernier recensement INSEE de 2014 était de 1 359 000 personnes, nous venons de présenter les bases des 4 sites dans lesquels notre solution est implantée. La plus grosse des bases comporte 1 768 033 patients uniques. Ainsi lors de la phase d'import de nouvelles données, le système doit déterminer si les données qu'il va intégrer sont relatives ou non à un patient qu'il connait déjà. Si oui les données seront référencées avec l'identifiant unique de ce patient. Sinon un nouvel identifiant sera créé et ajouté à l'annuaire centralisé hébergé chez le HADS, avant que les données puissent être stockées. Cette phase d'identification nécessite l'utilisation d'algorithmes capables de comparer des chaînes de caractères.

4.1.2 L’utilisation de Jaro-Winkler

Nous avons vu dans 3.2.2, que l’algorithme de Jaro-Winkler permet de mesurer une distance. La durée de la comparaison est importante, plus la distance est calculée rapidement plus le système sera convivial à utiliser. Les temps de calcul varient beaucoup en fonction de l’implémentation et de processeurs utilisés. Nous avons expérimenté de façon non concluante des versions implémentées par nos soins en Cuda, pour Nvidia Tesla M2050, et une version pour Intel Xeon-Phi 5110p.

En effet le surcoût de l’investissement matériel et humain que nécessite l’approche par GP-GPU est important. L’achat d’une carte NVIDIA capable de traiter le CUDA, nécessite un budget. De plus l’architecture CUDA nécessite l’emploi de syntaxe et structures spécifiques au matériel qui oblige à considérer les développements avec un paradigme particulier, qui permet de tirer parti du matériel NVIDIA.

Pour l’approche Xeon Phi, le coût d’entrée dans la technologie existe aussi, mais le poids du portage est moins important car les cœurs sont compatibles x86. Une recompilation, suffit à porter de façon simple du code C++ existant.

Vu les coûts nécessaires en temps et en homme, nécessaires pour réaliser nos tests qui pour certains apportaient des accélérations intéressantes (x17) (Bisson 2012).

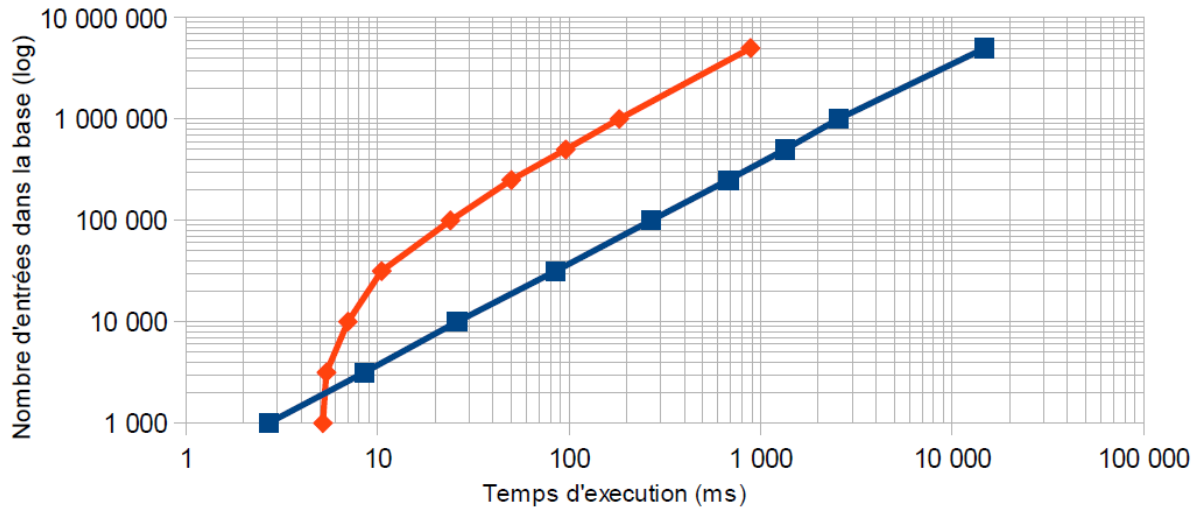


Figure 52 Comparaison de l’exécution de Jaro-Winkler sur 2 champs (nom, prénom), sur architecture CPU(en bleu) et GP-GPU(M2050) (en rouge).

Nous avons cependant préféré nous orienter vers des CPU traditionnels pour poursuivre nos expérimentations. Plusieurs approches ont été comparées en 'C', en 'R', en 'Perl' et en 'SQL'[131],. En effet le budget pour équiper tous les sites de processeurs particuliers aurait été trop élevés et à l'encontre de la philosophie du cahier des spécifications qui demande une solution la plus libre et la moins onéreuse possible.

Nous avons été supportés pour la partie identification par l'équipe de l'ISIT et nous avons particulièrement collaboré aux travaux de thèse de Xinran Li. Ces travaux nous ont permis de comparer différents processus d'identification des patients. En nous intéressant au *record linkage*, le chaînage des enregistrements.

4.1.3 Importation des données médicales

L'importation est l'étape ou l'export fourni par le partenaire médical est intégré dans notre système. Pour permettre un maximum de souplesse le format de l'export est discuté au cas par cas avec notre partenaire, c'est la solution la plus légère pour le partenaire qui est retenue. Une fois cet export en notre possession, nous appliquons une routine sur ce fichier pour identifier les patients auxquels ces données se rapportent, le but étant d'éviter les doublons ou de lier ensemble des données d'homonymes. Nous avons présenté précédemment les différentes étapes de cette importation (cf. Figure 39) nous pouvons ici détailler encore plus cette procédure (Figure 53) en précisant les algorithmes et les codes utilisés. Notre script en charge de la phase d'import est écrit en langage perl. Nous présentons ici l'import dans le SI GINSENG d'un fichier *.csv. Après une phase d'initialisation du code, le script accède au fichier contenant les champs relatifs aux patients chargé en RAM. Les bases de données locales et distantes sont monopolisées le temps de l'import pour garantir l'unicité des transactions. Les lignes chargées en RAM sont lues les unes après les autres pour être identifiées. Dans un premier temps une recherche SQL est effectuée pour trouver une correspondance avec une entrée de la base « annuaire ».

```
SELECT * FROM Tannuaire WHERE nom= « nom_recherché » AND prenom= « prénom_recherché » …
```

Cette requête est d'abord effectuée localement ce qui évite de mobiliser le réseau si le patient est déjà connu sur le site auparavant. Si ce n'est pas le cas, la requête SQL est réalisée sur le serveur annuaire distant. Si un identifiant n'est pas envoyé en retour, la requête basée sur la pondération des champs et utilisant Jaro-Winkler est exécutée sur le serveur distant.

[131] https://androidaddicted.wordpress.com/2010/06/01/jaro-winkler-sql-code/ - date d'accès octobre 2015

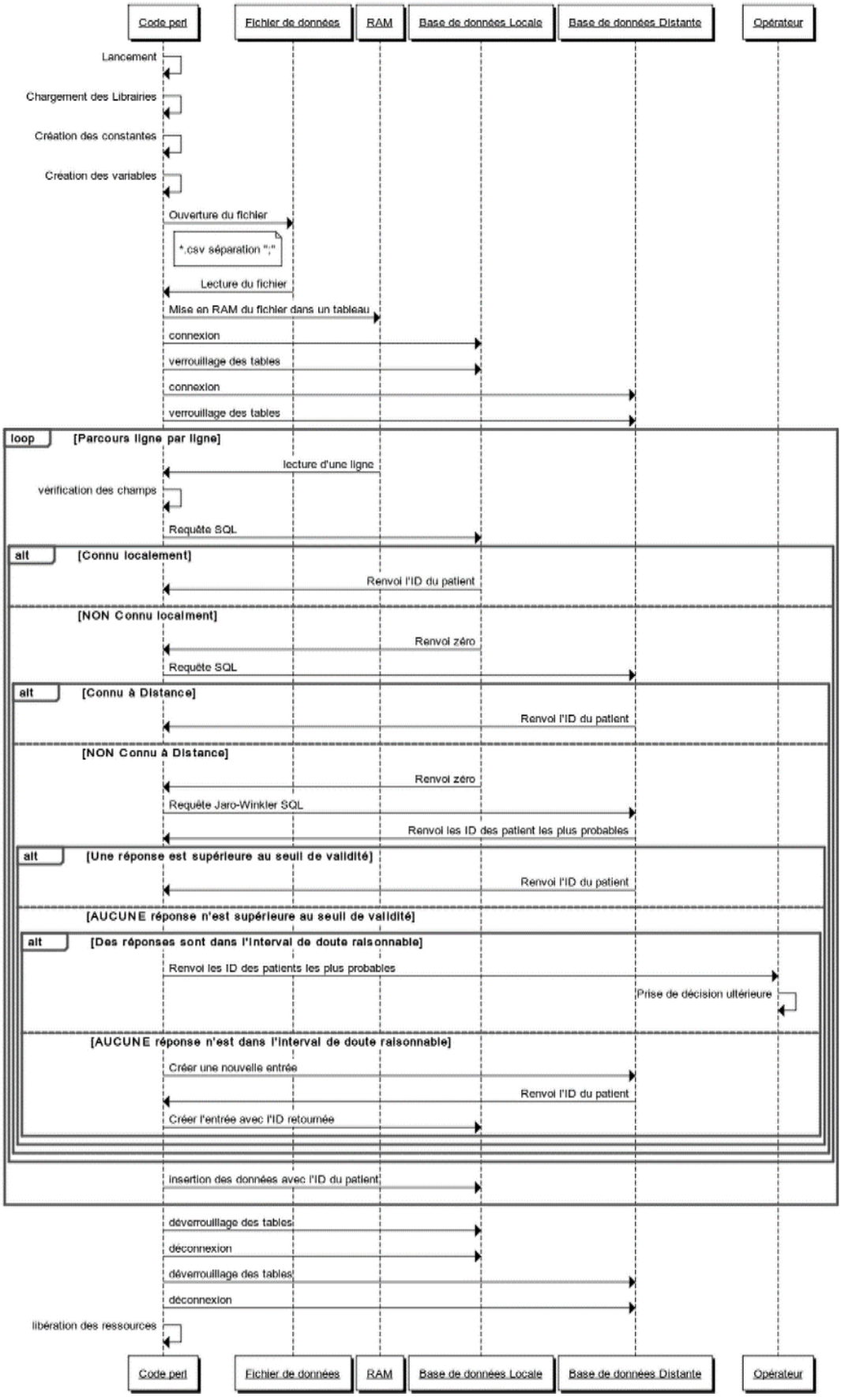


Figure 53 Diagramme de séquence de l'importeur de patients

Trois situations sont alors envisageables. Les deux plus simple à traiter sont quand l'algorithme est certain que le patient est connu, ou alors que le patient est inconnu du système. Ainsi, dans ces deux premiers cas l'annuaire distant retourne un identifiant qui caractérise le patient, si besoin après avoir instancié cet identifiant. Le troisième cas intervient lorsqu'il existe un doute raisonnable que le patient soit connu. En effet, notre méthode d'identification retourne une valeur comprise entre 0 et 1. La valeur 1 signifiant la correspondance parfaite. Nous pourrions définir une seule valeur plancher au dessus de laquelle la correspondance entre les champs soit considérée comme validée (Figure 54).

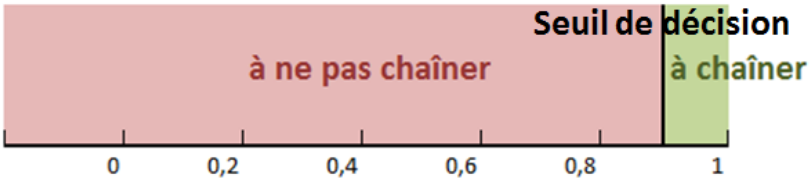


Figure 54 Représentation d'un seuil de décision plancher unique

Dans un souci d'apporter un maximum de valeur ajoutée aux utilisateurs nous avons considéré deux seuils de décision qui divise l'intervalle de zéro à un en trois (Figure 55).

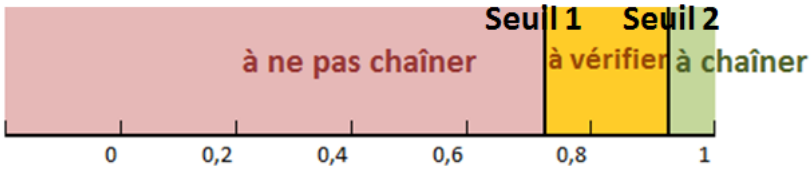


Figure 55 Représentation de l'intervalle de doute légitime, pour l'identification

Les valeurs en dessous du seuil 1 sont considérées comme ne permettant pas l'identification du patient. Les valeurs supérieures au seuil 2 permettent de récupérer l'identifiant du patient. Les valeurs comprises en le seuil 1 et le seuil 2 nécessitent un traitement ultérieur qui sera soumis à l'appréciation d'un opérateur humain. Notre but étant de couvrir notre raisonnement théorique et la présence de faux négatifs produits par l'algorithme d'identification (Figure 56a). Après avoir réalisé des tests sur données simulées bruitées nous avons défini les deux seuils permettant cette approche (Figure 56b). Et si nous ne considérons que le seuil supérieur à 0.9111, nous pouvons nous en servir comme seuil plancher avec les pondérations actuelles.

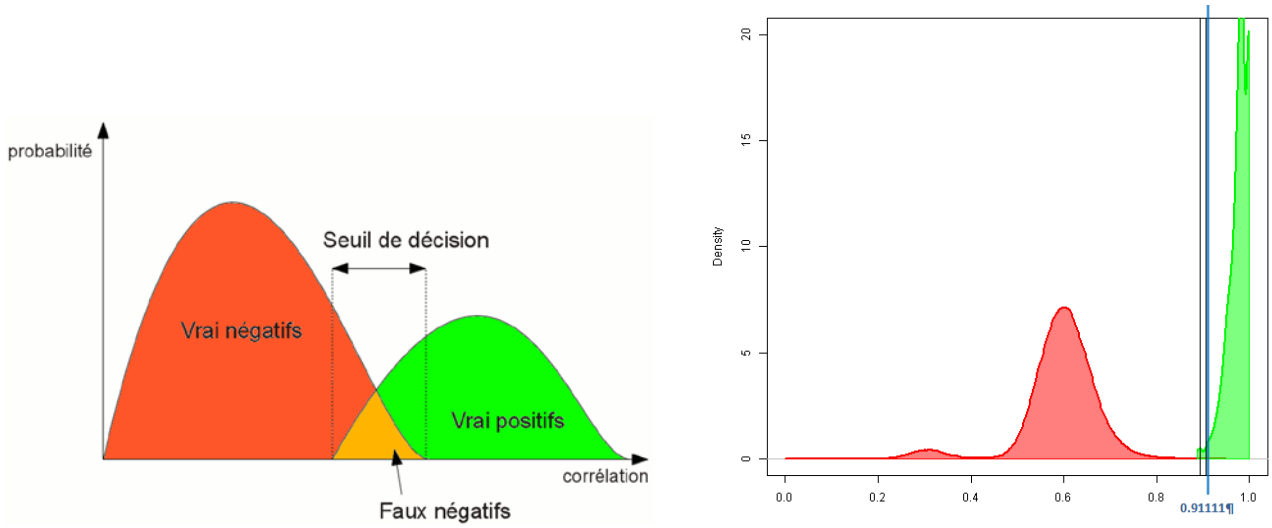


Figure 56 Répartition statistique du rapprochement des patients en fonction du score produit par l'algorithme d'identification
(a) approche théorique – (b) résultat sur données simulées bruitées

4.1.4 Benchmark

Nous présentons dans cette partie les temps nécessaires au chargement de notre système à partir des fichiers exportés par nos partenaires.

Dans la pratique le réseau inter sites sur lequel nous travaillons nous permet des débits effectifs de l'ordre de 400 Ko/s ce qui induit par exemple un temps de transfert de 21:34 minutes en utilisant la commande 'scp' pour un fichier de 464 MB à 367.4 KB/s.

À l'initialisation du système les bases de données sont vierges : elles ne contiennent que la structure des tables. Chaque semaine un partenaire comme le cabinet Sipath-Unilabs fournit un fichier d'export contenant entre 3 000 et 5 000 analyses.

Si une seconde est nécessaire pour importer une entrée, pour 10 000 000 d'entrées, il faudrait environ 116 jours. Ce chiffre est important mais nous devons cependant relativiser avec le fait que la base que nous considérons est celle de notre plus important partenaire (Sipath-Unilabs). La même opération effectuée sur une base de seulement 300 000 entrées sera réalisée en moins de 3 jours ½ sur une base vierge. Il faut aussi considérer que les choix effectués privilégient les faibles coûts, ainsi le MySQL nous limite à l'utilisation d'un seul cœur. De plus les calculs sont effectués sur les premières machines installées dans le cadre du projet avec des processeurs Intel E5504 à 2 GHz.

Sur une base vierge ou faiblement peuplée

Lors de l'initialisation de GINSENG toutes les bases sont par définitions vierges et vides de toute données. Les premières exécutions des importeurs sont plus rapides car lors de l'étape de vérification, qui vise à savoir si le patient est déjà connu, l'annuaire est vide ou faiblement peuplé. Nous avons réitéré ces premiers imports plusieurs fois avec un effacement complet des tables entre chaque exécution (DROP TABLE + CREATE TABLE). En fonction du site, des machines et des fichiers à importer nous avons constaté une vitesse d'import de l'ordre de 10 à 15 dossiers par seconde.

Pour les imports hebdomadaires qui proviennent de notre partenaire Sipath-Unilabs pour 4060 patients dans la première semaine d'octobre 2015 l'import a nécessité 4 minutes 37 (1 patient toute les 0.068 seconde). Si nous considérons un transfert inter-site de cette archive avec notre réseau dont le débit tend vers 400 Ko/s, 20 minutes sont alors nécessaires, à titre d'exemple.

D'autre part nous avons commencé l'import de la base historique de Sipath-Unilabs en octobre 2015 nous approchons de 6 000 000 d'entrées traitées. Nous estimons que l'import complet prendra plus de 4 mois pour traiter la totalité des 10 000 000 millions de ligne qui stockent autant de code ADICAP. Nous présenterons par la suite une étude préliminaire descriptive de cette source, opérée directement sur le fichier '.csv' original de 1,7 Go.

Après 5 000 000 d'entrées

Après 5 millions d'entrées, la table « patient » contient 2,5 millions de patients « uniques » les temps de traitements sont stables. Il faut compter 1.2 sec pour importer un patient supplémentaire, la solution est en cours de déploiement nous n'avons pas encore les valeurs pour une base plus peuplée. Ce temps est une moyenne et prend en compte le parcours de la base annuaire locale pour trouver le patient, dans le cas contraire, la recherche dans la base annuaire distante est aussi comptabilisée, et si le patient n'est pas identifié une nouvelle entrée est créée (cf. Figure 53 & 4.1.3). La reproductibilité (temps identiques) autours des 5 millions d'entrées est très bonne. Si l'on considère des lots de patients de taille similaire. En effet nous avons planifiés différents imports pour comparer leurs temps d'exécution. Nous pouvons remarquer que si les lots sont de taille trop différente des variations importantes sont relevées. Ce phénomène met en exergue un temps incompressible d'initialisation qui se dilue très bien dans des imports sur 500 000 personnes, mais ralentit d'environs 5% les opérations sur petit lot rapportées au temps par individu. Pour un import de 4 000 fiches ACP, 120 Mo de RAM sont utilisés par rapport à une consommation au repos variant entre 750 et 800 Mo. Ces chiffres sont mesurés sur une configuration du type des premiers déploiement en 2010 (cf. Tableau 10) en Ubuntu LTS 14.04, dont la mise à jour vers LTS 16.04 interviendra dans un futur proche.

Tableau 30 Temps d'exécution du script d'import en fonction du remplissage préalable de la base

Nb de codages dans la base	Nb de codages à importer	Date de début	Date de fin	Temps écoulé	Temps par patient en s	RAM nécessaire en Mo
150 000	4 552	11 :14,12.361787837	11 :27,17.112181678	0 :13,05	0.1724	922 (uptime 163j)
155 000	3 135	16 :40,59.540882801	16 :50,58.266631629	0 :09,59	0.1910	
160 000	4 292	15 :44,33.763096339	15 :58,41.304885378	0 :14,08	0.1975	809 (uptime 177j)
165 000	4 420	17 :13,22.866851378	17 :28,07.711759665	0 :14,45	0.2002	995 (uptime 184j)
170 000	3 139	13 :03,39.370278468	13 :15,03.971627526	0 :11,24	0.2179	868 (uptime 191j)
175 000	4 281	10 :56,52.898425165	11 :12,22.544503603	0 :15,30	0.2172	
180 000	3 099	15 :14,58.019357037	15 :26,37.533302346	0 :11,39	0.2255	
185 000	4 119	14:54,26.956966277	15:10,05.973790613	0 :15,39	0.2279	
5 000 000	1 000	10 :17,58.602743403	10 :37,40.222986644	0 :19,42	1.1826	634
5 002 000	10 000	10 :47,11.518138961	14 :03,15.967847461	3 :16,04	1.1764	666
5 012 000	6 000	14 :19,36.360913227	16 :18,51.778353725	1 :59,15	1.1925	
5 018 000	51 000	17 :03,49.182811908	J+1 09 :41,36.904764809	16 :37,47	1.1738	
5 500 000	500 000	18:27,46.800349685	J+8 04 :00,44.360195078	177 :32,57	1.2783	
6 000 000	1 000	16:27,38.841877011	16 :49,59.376401299	0 :22,21	1.3405	
6 200 000	400 000	09 :58,09.535179169	J+6 08 :15,58.551011299	142 :17,49	1.2806	
6 600 000	400 000	16 :01,03.244418231	J+6 18 :57,30.148283498	146 :56,54	1.3225	
7 000 000	500 000	15 :30,32.933817287	J+8 12 :52,34.359093685	189 :22,02	1.3634	
7 500 000	500 000	17 :33,20.153236689	J+8 23 :56,25.889226967	198 :23,36	1.4284	
8 000 000	500 000	09 :42,28.162443370	J+8 23 :00,58.758469478	205 :18,30	1.4782	
8 500 000	500 000	16 :25,41.114489792	J+9 12 :11,59.763712854	211 :46,17	1.524	
9 000 000	500 000	10 :07,58.802978519	J+9 13 :15,18.285570253	219 :7,19	1.5776	1769 (uptime 90j)

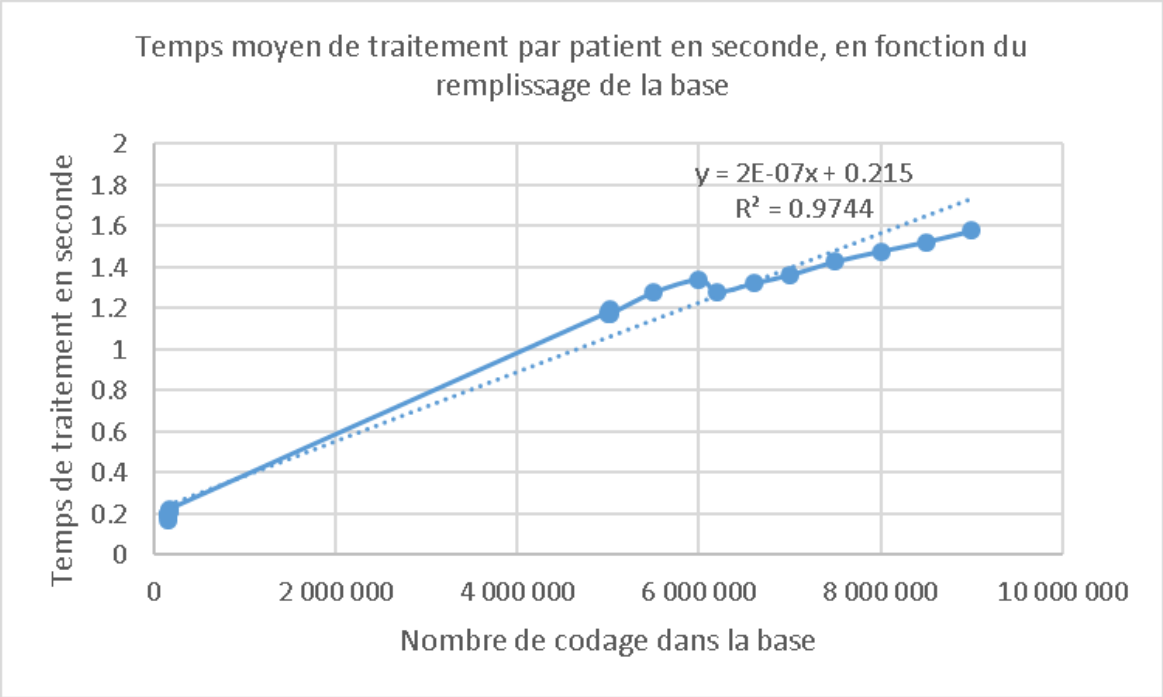


Figure 57 Graphique du temps moyen de l'import d'un patient dans la base GINSENG en fonction du nombre de code ADICAP dans la base

Nous avons donc constaté que plus la base est remplie plus les temps nécessaires à l'identification et donc à l'importation sont élevés. Ceci est tout à fait normal car à chaque nouveau patient l'annuaire à parcourir augmente de volume.

Nous pouvons remarquer que pour les imports de petites séries (1000 codages) les temps par patients sont beaucoup plus élevés à taux de remplissage de la base similaire. Ce qui met en lumière un temps incompressible lié au mécanisme d'import et qui est dilué dans les plus gros volumes (500 000 codages)

Il faut de plus considérer les temps d'accès au réseau qui peuvent fluctuer en fonction de la charge. Pour réaliser ces tests nous nous sommes assuré que la base distante de l'annuaire centralisé serait de facto disponible. En production lorsque plusieurs sites souhaitent accéder simultanément à cet annuaire il se verrouille le temps de traiter la première requête puis gère la seconde sur un principe de sémaphore à 1 ressource.

4.1.5 E-Santé, transfert de fichiers

Pour les associations de dépistages du cancer, leur objectif fonctionnel premier est la récupération du compte-rendu ACP. Notre système leur économise la démarche de recherche auprès des différentes sources susceptibles de détenir l'information. Dès que de nouvelles informations intéressantes sont disponibles, elles sont automatiquement présentées aux

opérateurs qui vont pouvoir mettre le dossier du patient à jour, suite à la lecture du compte-rendu. Pour récupérer le compte-rendu, il suffit à l'utilisateur de cliquer sur un bouton, ce qui va télécharger le compte rendu et le rendre accessible à la lecture. Il pourra par la suite être stocké dans la gestion électronique de documents (GED) de ZEUS, pour une consultation ultérieure (si les accords entre les différentes entités l'autorisent). Les différents sites communiquent entre eux avec un débit de 300 à 400 Ko/s à travers un réseau privé virtuel (tunnel VPN) sur des connexions ADSL ou SDSL ce qui est convenable pour nos besoins actuels. Le temps de réponse à une requête ping est de 60 à 90 ms.

4.2 Requêtes sur les bases de données et présentation des résultats

4.2.1 L'utilisation des SPARQL EndPoint

La publication et l'accès aux données RDF respecte les standards du Web Sémantique avec la mise en place d'un serveur SPARQL construit au-dessus du moteur Corese/KGRAM pouvant fonctionner à la fois dans un mode centralisé et dans un mode fédéré. Ce serveur est entièrement accessible et paramétrable via une API REST sous la forme d'une application « Play ». Trois ensembles de fonctions sont disponibles pour

(1) charger des données RDF,

(2) envoyer des requêtes,

(3) configurer le mode de fonctionnement du serveur.

La particularité de ce serveur SPARQL par rapport à un serveur SPARQL classique, est qu'il dispose d'un fédérateur de bases de connaissances autonomes, permettant d'interroger et de fédérer les données autonomes de chaque hôpital tout en bénéficiant de l'expressivité et de la puissance de raisonnement des moteurs sémantiques. Ainsi ce même serveur permet à chaque hôpital de publier sa propre base de connaissances autonome en mode centralisé, et de publier un web service permettant de fédérer ces bases de connaissances tout en respectant leur autonomie. De plus, la fédération de données RDF permet également de croiser les données des partenaires avec des sources externes respectant les standards du Web Sémantique telles que DBPedia[132] et d'ainsi augmenter les données disponibles. La Figure 58 montre l'architecture d'une telle base de connaissances distribuée avec différents partenaires médicaux disposant chacun de leur propre base de connaissances autonome.

[132] DBPedia Sparql Endpoint http://fr.dbpedia.org/sparql -

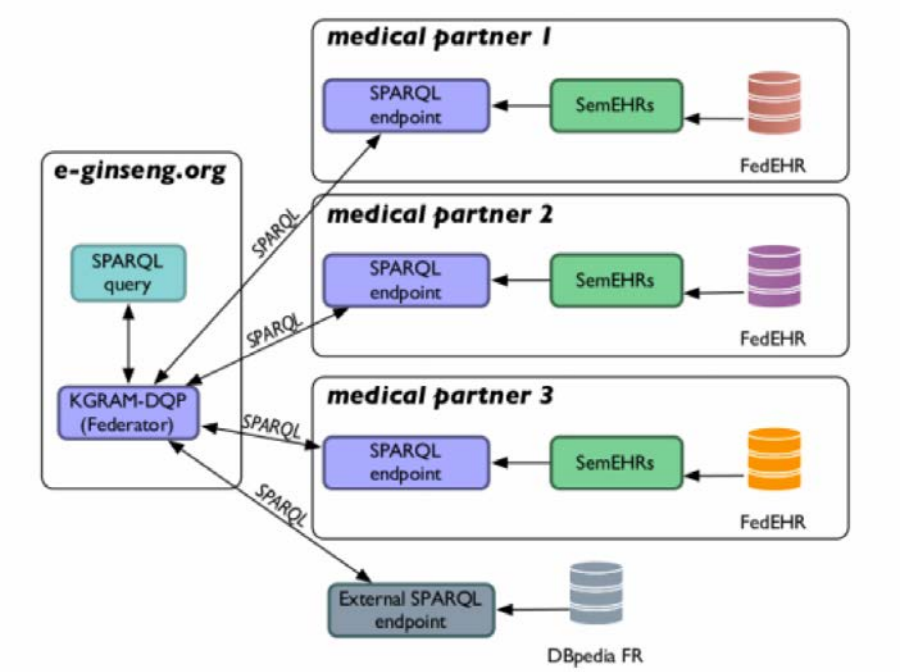


Figure 58 Architecture de la base de connaissances distribuée

Cette fédération de données a été testée sur différents scénario permettant d'extraire des données statistiques sur l'ensemble des partenaires médicaux. La requête suivante (cf. Figure 59) montre une fédération des données GINSENG avec des données démographique d'une source externe, DBPedia, afin de compléter le résultat de la requête avec les populations des villes des patients.

```
PREFIX semehr: <http://www.mnemotix.com/ontology/semEHR#>
PREFIX dbpediaowl:
<http://dbpedia.org/ontology/>
SELECT (count(distinct ?patient) as ?nbPatients) (sum(?pop) as ?totalPop) ?postalCode
WHERE {
 ?cv semehr:value "BHGSA3F0"^^xsd:string .
 ?patient semehr:hasMedicalBag/semehr:hasMedicalEvent/semehr:hasClinicalVariable ?cv .
 ?patient semehr:address/semehr:postalCode ?postalCode .
 SERVICE <http://fr.dbpedia.org/sparql> {
  SELECT DISTINCT (str(?cp) as ?postalCode) ?pop WHERE {
   ?s dbpediaowl:region <http://fr.dbpedia.org/resource/Auvergne>
   ?s dbpediaowl:postalCode ?cp .
   ?s dbpediaowl:populationTotal ?pop
  }
 }
}
GROUP BY ?postalCode
```

Figure 59 Exemple de requête SPARQL interrogeant la base sémantique

4.2.2 L'utilisation du logiciel 'R'

Nous avons expérimenté de nombreuses solutions pour analyser les données à des fins épidémiologiques. Certaines étaient propriétaires comme 'SAS', et d'autres disponibles de façon libre comme 'R'. Dans le cadre des analyses statistiques du projet, nous avons effectué un export au format '.csv', qui a été passé en paramètre à 'R', puis nous avons sauvegardé cette base de travail en binaire, facilement interrogeable par le logiciel. Cette solution nous a permis de charger plus rapidement les données à analyser, elles sont stockées sous une forme plus compacte et de la même façon moins de mémoire est nécessaire pour traiter la même quantité de données. Cette méthode permet des analyses statistiques qui n'impactent aucunement la base de données de production, car le traitement peut être déporté sur une machine située en dehors de l'hébergeur agrée de données de santé.

4.2.3 Interface WEB

L'un des objectifs initiaux de GINSENG est de mettre en ligne, accessible par internet, des informations. Pour répondre à cette problématique nous avons exploré différentes pistes. Avec l'expertise de nos partenaires nous avons envisagés d'utiliser une instance Liferay [133]couplée au framework Vaadin[134]. Le style graphique du site est présenté dans la Figure 60.

[133] https://www.liferay.com/fr/ -
[134] https://vaadin.com/home -

Figure 60 Page d'accueil du portail (Liferay) du projet GINSENG

Ce portail WEB doit remplir plusieurs fonctions :

1. Publier des informations généralistes aux visiteurs lambda
2. Gérer les utilisateurs qui souhaitent se connecter avec leur CPS
3. Présenter des services tiers comme ceux hébergés par VIP
4. Offrir une interface ergonomique aux utilisateurs de GINSENG
5. Présenter les résultats sous une forme pertinente aux utilisateurs

Nous pouvons synthétiser les cas d'utilisations de l'application de requêtage dans le diagramme de la Figure 61.

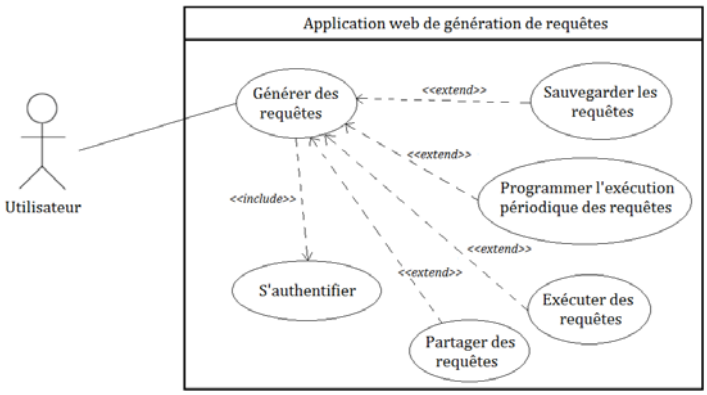


Figure 61 Diagramme des cas d'utilisations de l'application WEB

L'interface de requêtage intégrée dans le portail Liferay (Figure 62) est constituée de 3 portlets (Filtres généraux ; gestion des conditions fines ; affichage de la requête en langage de programmation). Cette interface est destinée aux épidémiologistes pour qu'ils puissent de façon simple requêter les données stockées dans GINSENG sans maitriser la structure de toutes les bases. De plus cette interface abstrait le code SQL qui sera in fine consultable et modifiable à travers le 3ème portlet pour vérification et/ou affinage de leurs requêtes dans la limite de leurs droits de consultation.

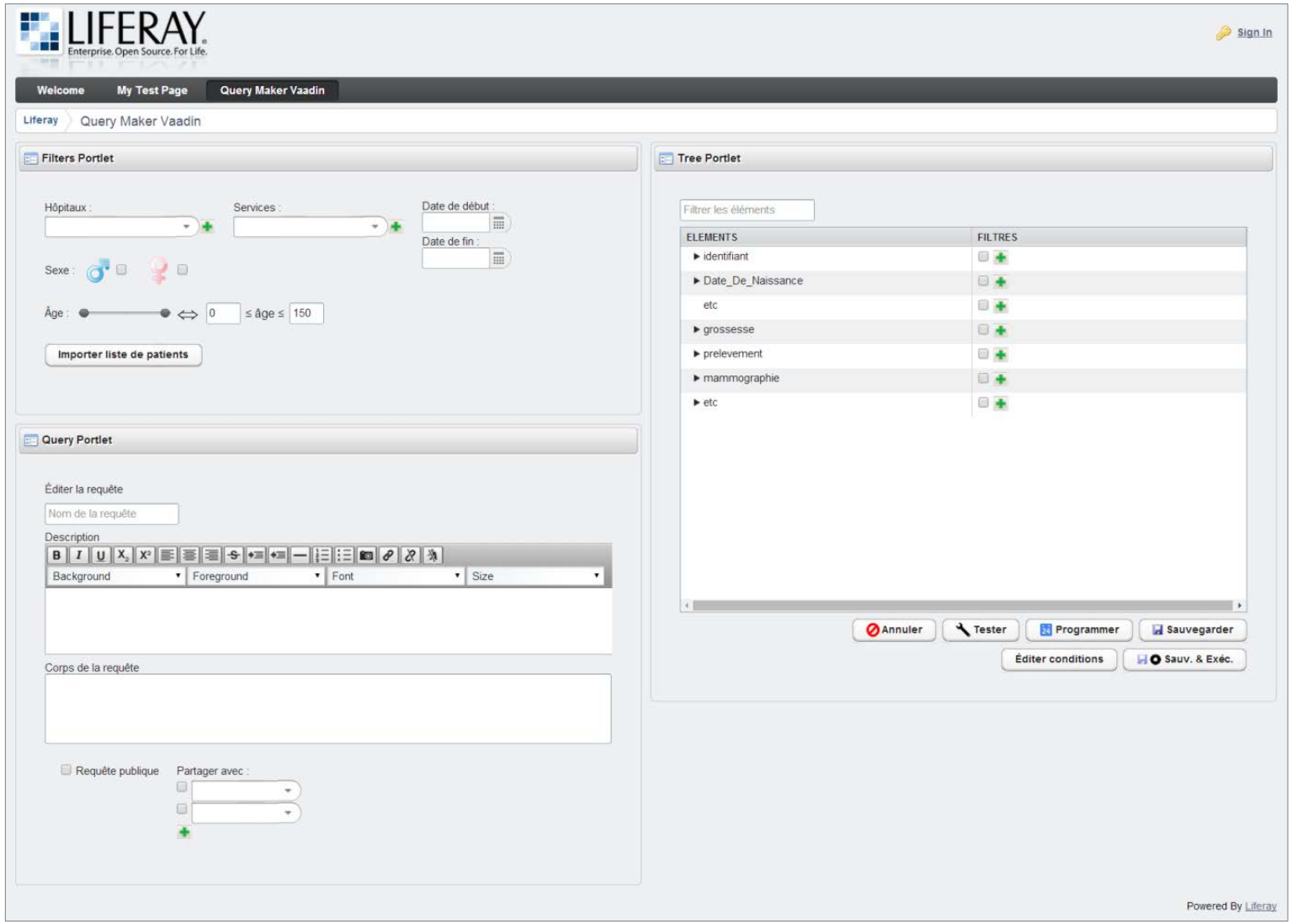


Figure 62 Interface Liferay de requêtage de GINSENG

4.2.4 Représentations visuelles

Intégration dans Zeus de O.S.I. Santé

Pour répondre aux attentes de nos partenaires, dans des délais cohérent avec le déroulement de la thèse, tout en respectant les contraintes légales, et au vu des temps de gestions des différentes administrations dont l'accord est nécessaire nos ambitions en termes d'interface ont dû être revues à la baisse. Initialement envisagée en termes de service WEB (Liferay), comme présenté durant les démonstrations avec des données synthétiques, puis profitant des solutions offertes par le GCS Simpa, la solution la plus pérenne à mettre en œuvre depuis octobre 2015 est l'intégration de l'interface dans la solution Zeus D'O.S.I. Santé.

Actuellement cette interface est fonctionnelle. Le secrétariat de l'ARDOC peut requêter la base GINSENG à travers son logiciel métier.

Différentes approches ont été implémentées en fonction du type de recherche. Si on recherche une personne précise dont le statut est « en attente de suivi » l'opérateur peut effectuer une recherche par numéro de dossier. Zeus lie automatiquement le numéro de dossier du SGDO avec les informations personnelles du patient. Puis Zeus interroge la base GINSENG en fonction des informations sélectionnée dans l'interface. La Figure 63 est un zoom de la Figure 64 dans laquelle les résultats sont présentés.

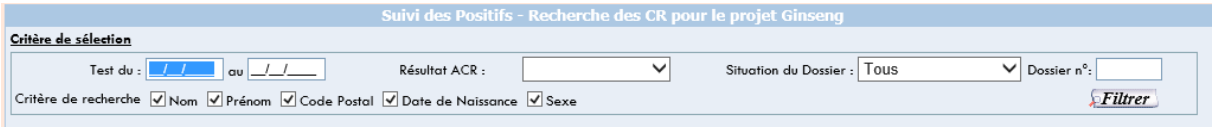


Figure 63 Capture de l'outil de sélection dans Zeus (v201603.1) permettant de choisir les champs du patient que l'on souhaite requêter dans GINSENG

Cette solution permet une recherche très rapide des comptes-rendus (CR) ACP qui sont disponibles chez Sipath-Unilabs et en attente de traitements chez les SGDO. Il faut au système une dizaine de seconde pour traiter les informations nécessaires et fournir une liste des CR pertinents pour l'intégration au sein de la SGDO. Auparavant cette recherche mobilisait une secrétaire qui se focalisait sur un dossier et effectuait des démarches (téléphoniques ou courriers) pour identifier si oui ou non la patiente avait effectuée des examens pouvant produire un compte rendu ACP et si oui dans quel cabinet ACP. Puis il était encore nécessaire de récupérer le CR par courrier, mail ou fax. Ces actions sont maintenant réalisées quasiment automatiquement, mais c'est bien l'opérateur du logiciel métier qui in fine choisit de lier un CR

à un dossier et traduit les informations du CR pour Zeus. C'était l'une des demandes fortes du cahier des charges que l'étape d'intégration reste contrôlée humainement.

L'interface GINSENG-Zeus est actuellement en béta test, il est difficile de quantifier précisément le gain engendré. Mais à chaque test de CR qui étaient inconnus du SGDO sont trouvés et intégrés dans Zeus. Pour le moment nous n'avons lié qu'une base consolidée chaque semaine, depuis juillet 2015. Mais une base similaire et remontant jusqu'en 1982 est prête à être exploitée de la même façon.

De plus, toutes les fonctionnalités de GINSENG ne sont pas exploitées car elles sont actuellement bridées par cette interface. Comme nous le verrons par la suite les représentations graphiques intéressantes pour le public ou les responsables de santé publique, n'intéressent pas les SGDO. Elles trouveront plus naturellement leurs places dans l'interface WEB. Dans la phase béta actuelle nous fournissons les résultats avec une sécurité maximale. Nous venons de le dire ce niveau de résultat est pertinent car lors de chaque test de nombreux CR ACP sont récupérés. Il convient de garder à l'esprit que l'outil permet de proposer des résultats considérés comme étant pertinents même s'ils ne correspondent pas à 100% à la requête demandée. C'est ce que permet d'une certaine façon le sélecteur de la Figure 63, mais une autre représentation est prévue à terme, qui permettra une meilleure analyse de l'identification des patients. Le but final étant de récupérer des CR de patients ayant déménagé ou pour lesquels certains champs n'ont pas été renseignés de la même façon dans les différents SI. On peut ici aisément inclure les noms/prénoms, composés ou exotique, ou simplement les typos lors de l'enregistrement du dossier.

Nous continuons les développements pour proposer une interface WEB hébergée chez un HADS. Actuellement nous considérons une solution basée sur Liferay après avoir envisagé Drupal, un *content management system* CMS OpenSource, toujours dans l'optique de fournir une solution la moins onéreuse possible. Mais Drupal ne nous permet pas de réaliser un portail au sens où nous l'entendons.

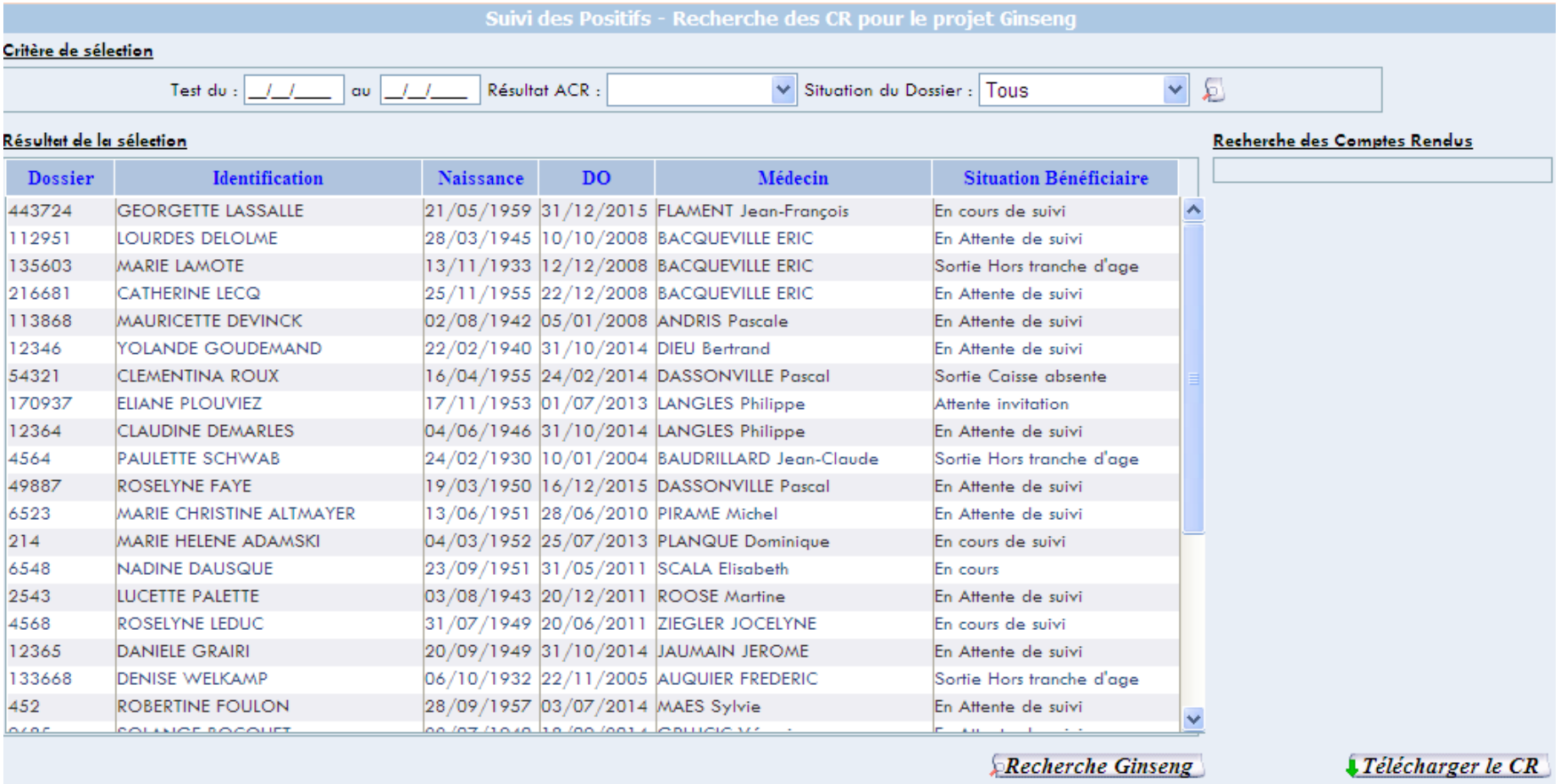


Figure 64 Capture d'une base de tests présentée dans Zeus (v201602.1) pour l'affichage des résultats issus de GINSENG avec la possibilité de récupérer les Compte-Rendus ACP

Représentation Google Earth

Pour informer les populations les représentations visuelles sont très pratiques car elles permettent de véhiculer un flux important d'informations très rapidement, ainsi il est plus agréable de regarder la Figure 65 que le tableau de valeurs qui a permis de la construire. Google utilise le format de fichier *Keyhole Markup Language* (KML) qui comme son nom le laisse deviner est basé sur le formalisme XML. KML fait partie des Systèmes d'Information Géographique (SIG) dont certaines bases sont gratuites[135] et proviennent de sources fiables comme l'INSEE ou l'IGN. L'un des objectifs de KML est de pouvoir représenter des objets géométriques sur des coordonnées de géolocalisation en surimpression d'une représentation cartographique, qu'elle soit 2D ou 3D. Pour créer une forme il suffit de renseigner les coordonnées (longitude et latitude) de chacun des points qui forment la base de l'objet puis d'assigner une hauteur à chacun de ces sommets. Pour les parallélogrammes rectangles de la Figure 65, 4 coordonnées avec une hauteur identique sont donc nécessaire. Pour traduire nos données vers ce type de représentation, nous avons écrit un script Perl qui prend en entrée une coordonnée (celle d'une ville) et une hauteur, et crée le parallélogramme correspondant en retournant le fichier KML correspondant. Une fois cette technique maîtrisée nous pouvons adapter le niveau de détails de la représentation à nos données, en transposant à l'échelle d'un pays ou d'un quartier si besoin (n'importe où sur le globe).

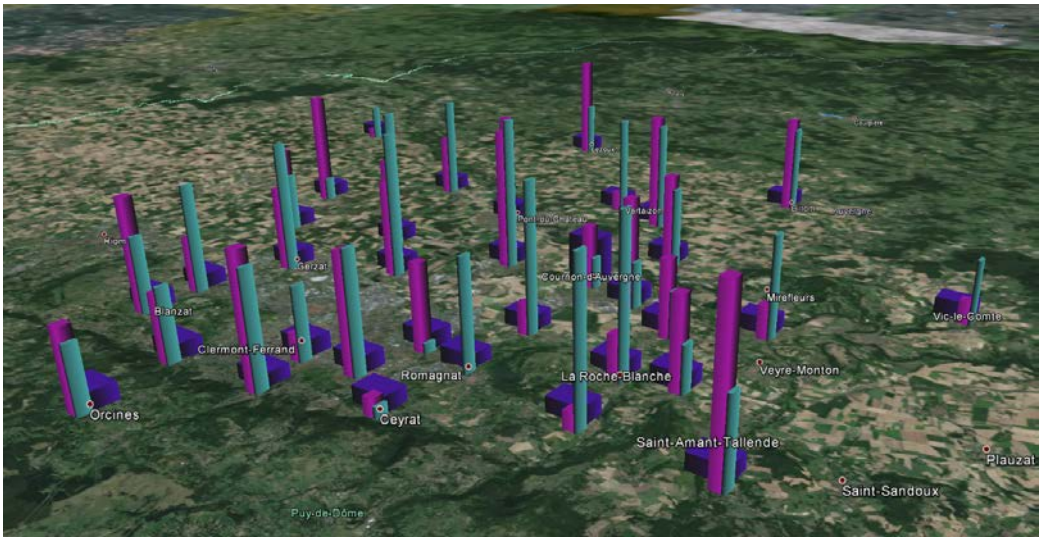


Figure 65 Vue Google Earth de 3 variables (simulées) attachées aux communes auxquelles elles correspondent

[135] http://www.infosig.net/les-donnees-sig/donnees-sig-gratuites - date d'accès octobre 2015

En effet Google fournit son API Google maps[136] librement, ce qui permet d'embarquer ce système de cartographie simplement dans notre propre site WEB (dans la limite de 1 000 requêtes par jour). Par la suite des forfaits permettent d'améliorer l'expérience utilisateurs et le nombre de requêtes quotidiennes acceptées par les serveurs Google.

Représentation géographique Excel 2016

Le tableur Excel 2016 de Microsoft intègre depuis son passage à la version 2016 une fonctionnalité de représentation cartographique. Cette solution permet de présenter rapidement des valeurs en fonction d'un code postal français. La Figure 66 ainsi que les cartographies de (Cipière 2016) ont étés réalisées avec cet outil. Pour un tableau de deux colonnes la première nommée « Code Postal » la seconde « valeur » ; il suffit de sélectionner le contenu du tableau que l'on souhaite représenter. Si les codes postaux sont au format « texte » ils seront automatiquement identifié et géo-localisé, il est ensuite possible de leur associer les valeurs associées à la même ligne dans le tableau. L'outil se situe dans « insertion » → « Carte 3D » de MS Excel 2016.

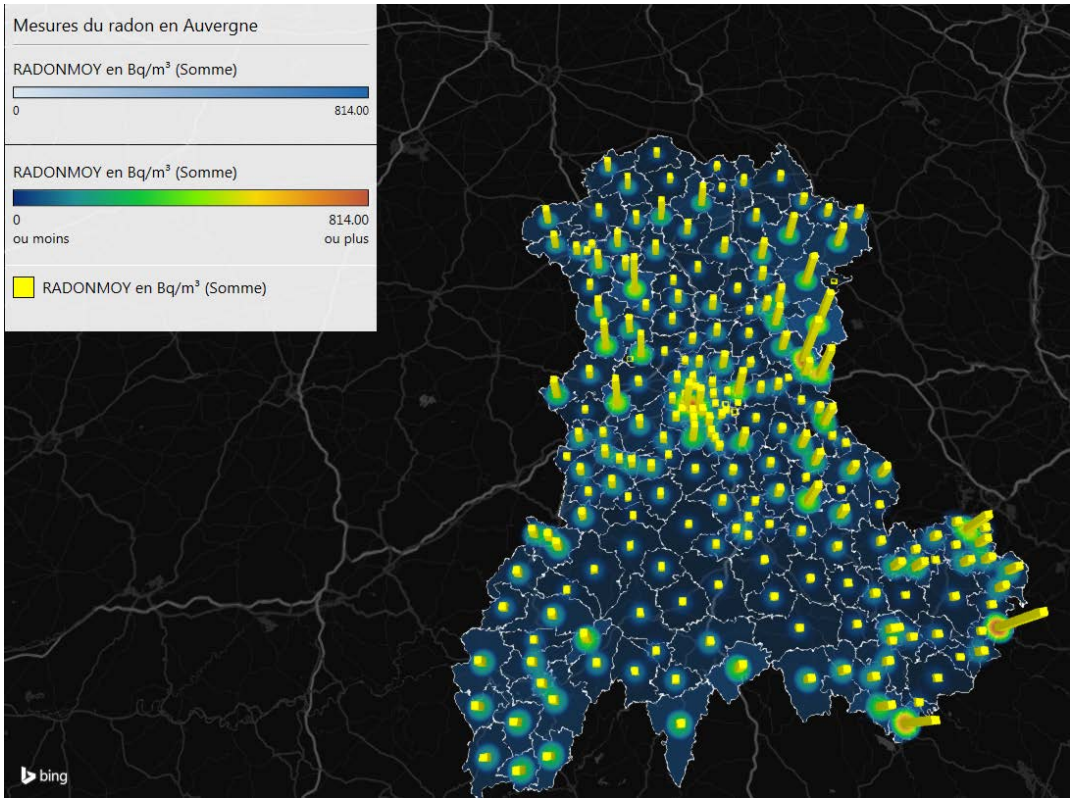


Figure 66 Représentation des mesures de radon moyenne en Auvergne par code postal (Source : des données BRGM 2007)

[136] https://developers.google.com/maps/ - date d'accès mai 2015

L'avantage de cette approche est sa facilité d'utilisation par rapport à l'approche « Google » qui nécessite de récupérer les coordonnées souhaitées pour chaque histogramme puis appliquer un script qui calculera la surface et la hauteur. De plus l'outil Excel propose de réaliser facilement des vidéos de la cartographie ainsi réalisées. Cependant la contrepartie de cette facilité est pour l'instant un manque de personnalisation des représentations.

4.3 Étude épidémiologique sur l'impact du radon sur les cancers du poumon en Auvergne

Dans cette partie nous présentons une étude réalisée à l'aide des données collectée par notre outil. L'impact du radon sur la santé a déjà été mis en évidence par l'Institut National de Veille Sanitaire (INvS) (Paillard et al. 2014; Franke and Pirard 2006; Darby et al. 2005) notamment sur l'incidence des cancers du poumon. Nous nous sommes intéressés à cet impact sur la population de la région Auvergne.

Nous avons cherché à relier les mesures relatives à la présence de radon dans l'air et les nombres de cas de cancer du poumon en Auvergne. Nous avons effectué une étude descriptive de toutes nos séries de données pour en exclure les valeurs atypiques, comme les absences de mesure renseignées à zéro, etc. Les données concernant la concentration de radon nous ont été communiqué par le Bureau de Recherches Géologiques et Minières (BRGM[137]) (Tourlière, Rouzaire, and Bertin 2007), les mesures datent de 2007. Nous avons considéré des mesures de radon dans l'air en Bq/m^3. Nous avons calculé une moyenne pondérée, par le nombre de mesures effectuées dans les communes, pour chaque code postal. En 2009 l'OMS a réduit d'un facteur 10 sa recommandation sur le seuil de radon dans l'air des locaux fermés en abaissant de 1000 Bq/m3 à 100 Bq/m3.

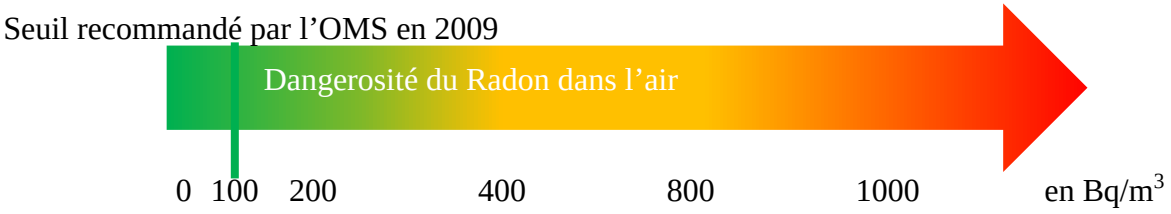


Figure 67 Représentation de la dangerosité pour l'homme du niveau de radon dans l'air d'un local fermé, avec seuil de recommandation de l'OMS 2009

[137] http://www.brgm.fr/ - date d'accès décembre 2015

Pour identifier les cancers du poumon nous avons utilisé le codage ADICAP[138] (cf. Figure 16), nous avons retenu les codages « RP » en position 3 et 4. En nous appuyant sur la base de données Sipath-Unilabs contenant le plus d'archives, non avons identifié 1407 cas de cancers du poumon sur un total de 9,5 millions d'entrées sur la période 1990-2015 pour les 4 départements auvergnats.

Les accords CNIL qui régissent le projet GINSENG nous limitent à une granularité géographique de l'ordre du code postal. Nous avons donc regroupé les mesures radon et les comptages des cancers au niveau du code postal en nous appuyant sur les données INSEE[139].

Les traitements que nous avons effectués permettent de répondre aux questions de la répartition femme/homme, par code postal, par département, par type de cancers et même plus spécifiquement pour un code ADICAP donné ainsi qu'à chacune de ces questions ; qui peuvent être composées entre elles ; par année. Pour une analyse plus rapide des données nous avons effectué un export de la base GINSENG non nominatif vers un fichier « .CSV » (9,5 millions de lignes) (1,5 Go), que nous avons importé dans le logiciel 'R'. Toutes les analyses statistiques ont été réalisées avec 'R', la représentation sous forme de cartographies 3D a ensuite été réalisées sur Excel 2016 (pc), qui intègre désormais la représentation géographique par code postaux.

Dans un premier temps en nous appuyant sur les données BRGM nous avons édité la carte d'Auvergne de l'aléa radon et des mesures moyennes et maximales du radon par codes postaux Figure 68 (les zones jaunes sont les plus exposées aux risques liés au radon).

[138] http://medphar.univ-poitiers.fr/registre-cancers-poitou-charentes/documents_registre/adicap_version5_4_1_2009.pdf - date d'accès décembre 2015

[139] https://www.data.gouv.fr/fr/datasets/correspondance-entre-les-codes-postaux-et-codes-insee-des-communes-francaises/ - date d'accès décembre 2015

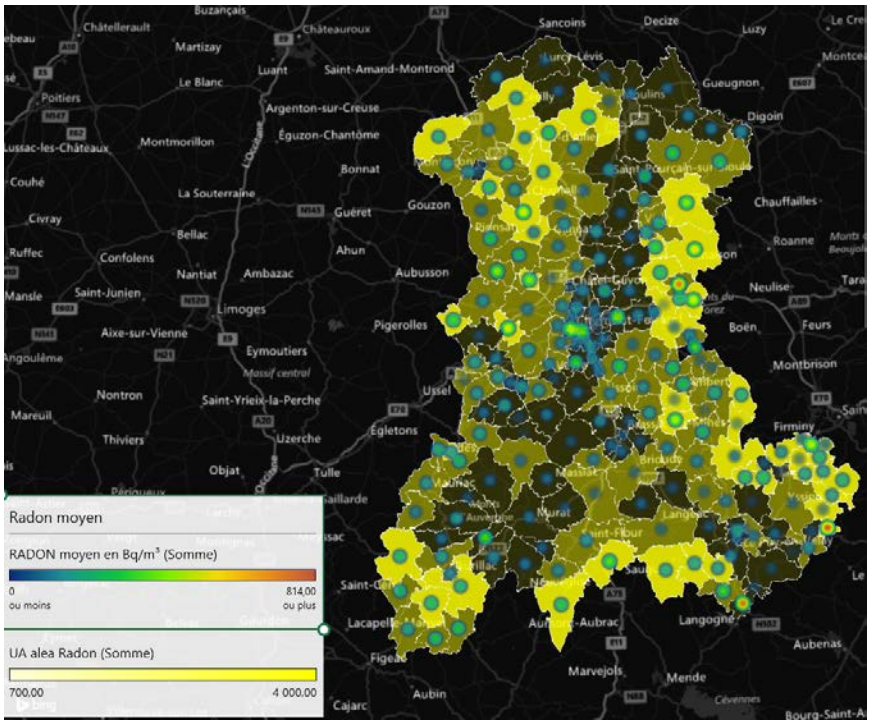


Figure 68 Cartographie de l'Auvergne avec représentation de l'aléa radon (BRGM) et mesure du taux moyen de radon dans l'air en Bq/m^3

La répartition de la population en Auvergne est particulièrement hétérogène, avec une très forte concentration de population à Clermont-Ferrand et son agglomération ; de ce fait, les représentations des données en valeurs absolues montrent une incidence prédominante dans cette zone géographique cf. Figure 69.

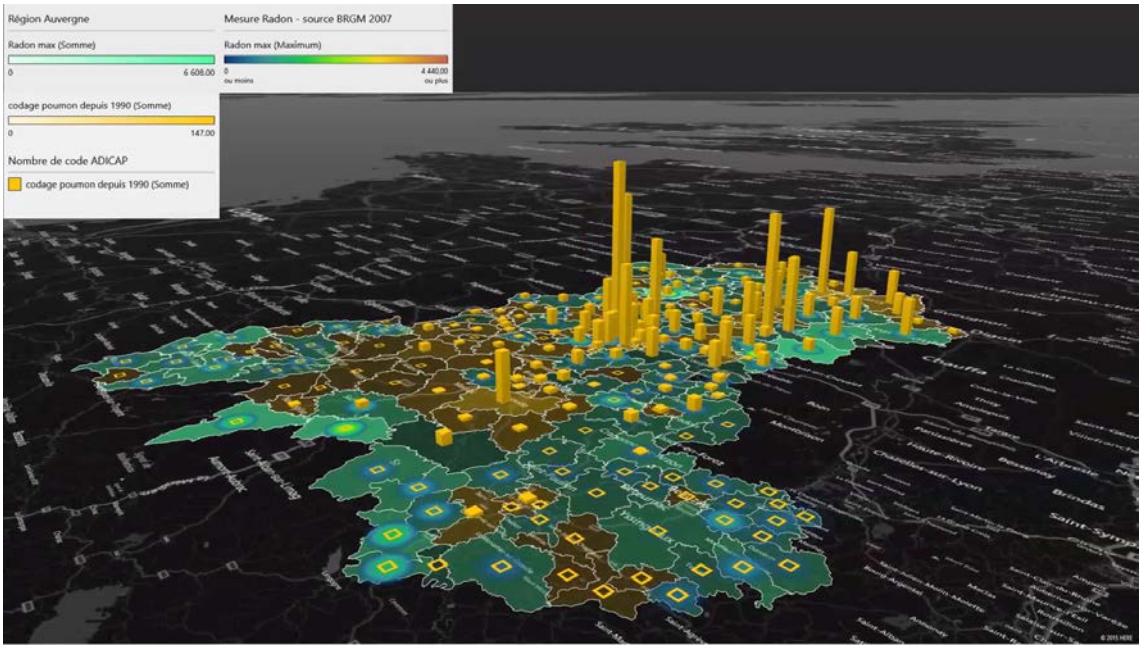


Figure 69 Représentation du nombre de codage ADICAP poumon (en valeur absolue) depuis 1990 et cartographie de mesure du radon dans l'air en Bq/m^3, par Code Postal

Nous avons donc normalisé nos données en fonction de la population grâce aux données INSEE qui fournissent un recensement mis à jour en septembre 2014 (Figure 70).

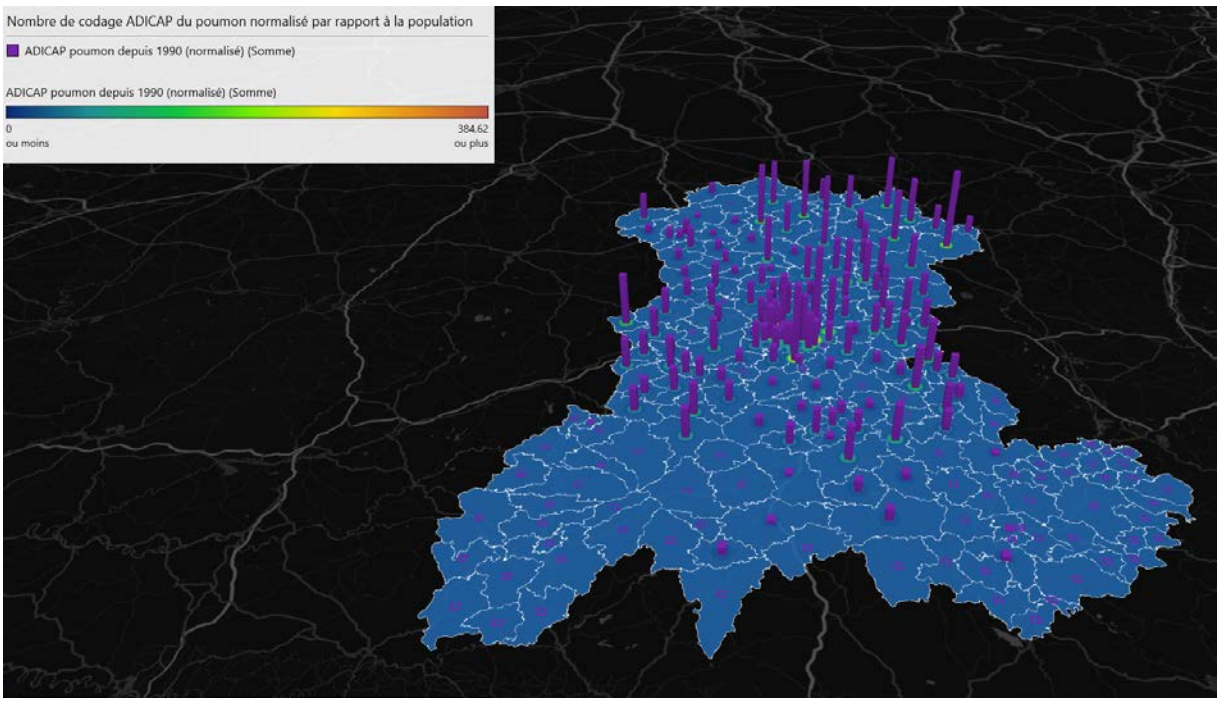


Figure 70 Représentation du nombre de codage ADICAP poumon (normalisé par rapport à la population) depuis 1990, par Code Postal

De plus, pour essayer de mettre en évidence une incidence significative de cancers du poumon dans une zone géographique, nous avons considéré le rapport du nombre de cas de cancers du poumon sur le nombre de cas de cancers du sein pour les femmes uniquement (Figure 71), ainsi que le rapport du nombre de cas de cancers du poumon sur le nombre de cas de cancers du côlon pour toute la population.

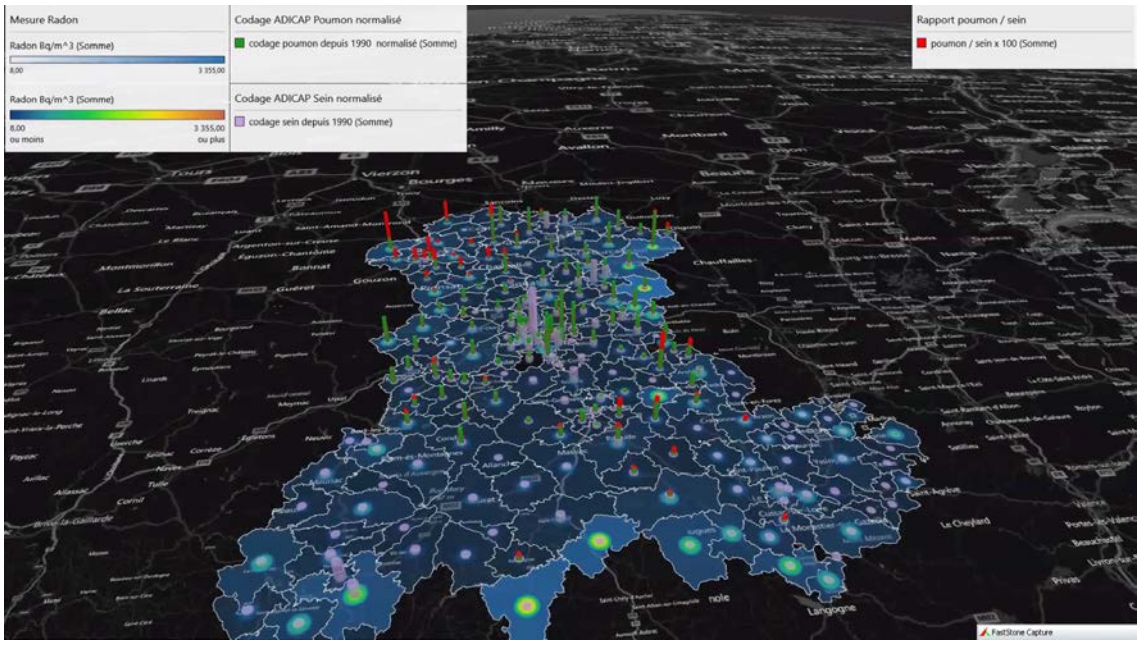


Figure 71 Représentation du nombre de codage ADICAP poumon et sein (normalisé) ainsi que le rapport nombre de cas « poumon » sur nombre de cas « sein » depuis 1990 et cartographie de mesure du radon dans l'air en Bq/m^3

Nous pourrions nous attendre à ce que ce rapport soit constant et homogène sur l'ensemble de la région. Cependant sur une dizaine de codes postaux majoritairement dans l'Allier (6 sur 9) ce rapport s'est avéré plus élevé que dans le reste de l'Auvergne (Figure 72).

	Radon	Poumon x 10 000 / colon
Code Postal	RADON moyen en Bq/m³	Rapport poumon/colon
03130	271.25	511.5511551
63111	0	446.4285714
03330	76.77777778	412.371134
03300	218.7209302	388.9886296
03440	274.3636364	364.3724696
63290	261.6428571	341.0641201
03120	409.125	334.7280335
03110	78.58333333	310.4786546
63890	274	309.2783505
63550	716.3333333	193.236715
43400	814	0
43420	716.9411765	0

Figure 72 Mise en évidente de la non corrélation entre le taux moyen de radon d'un code postal et l'incidence des codages ADICAP du poumon normalisé par le rapport nombre de codage du poumon sur nombre de codage colon

Nous avons étendu notre recherche aux 28 rapports « poumon/colon » les plus élevés que nous avons superposé dans la Figure 73 avec les mesures radon, et l'aléa radon. Néanmoins, cette prédominance de cas n'a pu être mise en coïncidence avec un taux de radon supérieur à 400 Bq/m^3.

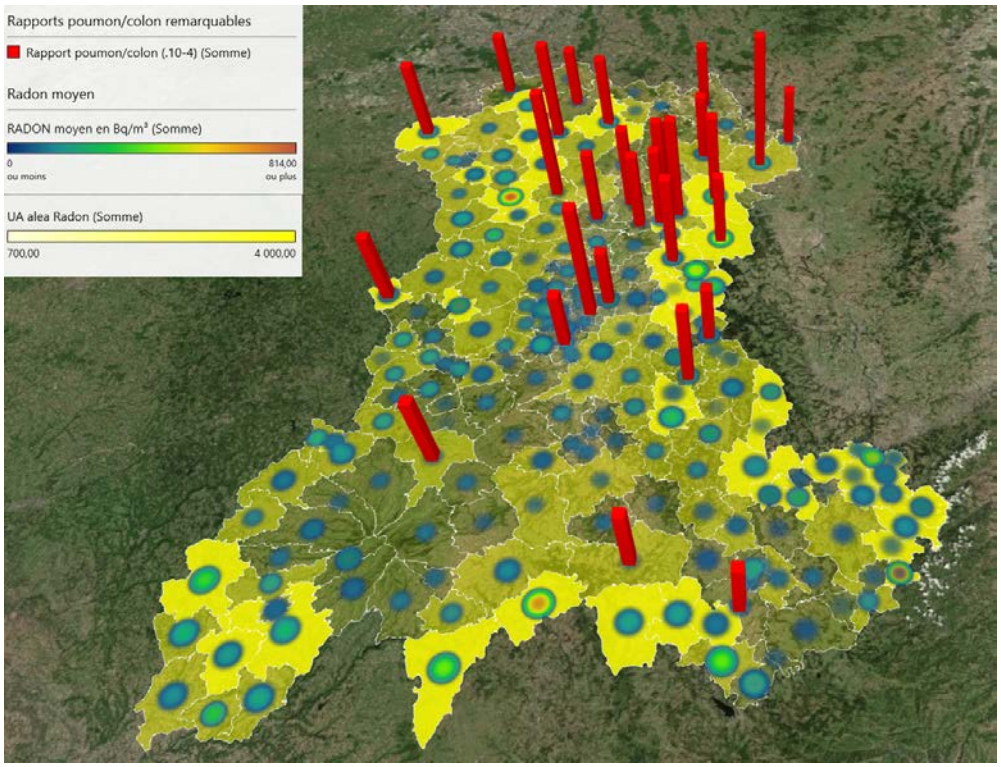


Figure 73 Représentation des rapports les plus élevés du nombre de codage ADICAP poumon sur nombre de codage colon depuis 1990, cartographie de la mesure moyenne du radon dans l'air en Bq/m3
en surimpression de représentation de l'aléa radon par Code Postal en Auvergne

Nous avons effectué une étude descriptive de nos données pour en exclure les valeurs atypiques notamment liées à l'absence de mesure, nous avons représenté nos données dans la Figure 74, les boîtes de Tukey nous renseignent sur la distribution de nos données et nous avons représenté deux droites de tendance sur la Figure 75 pour montrer l'évolution du nombre de codage ADICAP en fonction de la population des agglomérations. Le coefficient de corrélation entre les deux séries (rapport cancer poumon/colon et mesure du radon) est de 0.26, ce qui traduit un impact du taux de radon dans l'air sur l'incidence des cancers du poumon, même si la relation est faible.

Cependant il est important de remarquer qu'actuellement les données ACP en notre possession sont peu représentées pour les départements du Cantal (15) et de la Haute-Loire (43). (Figure 69 et Figure 70).

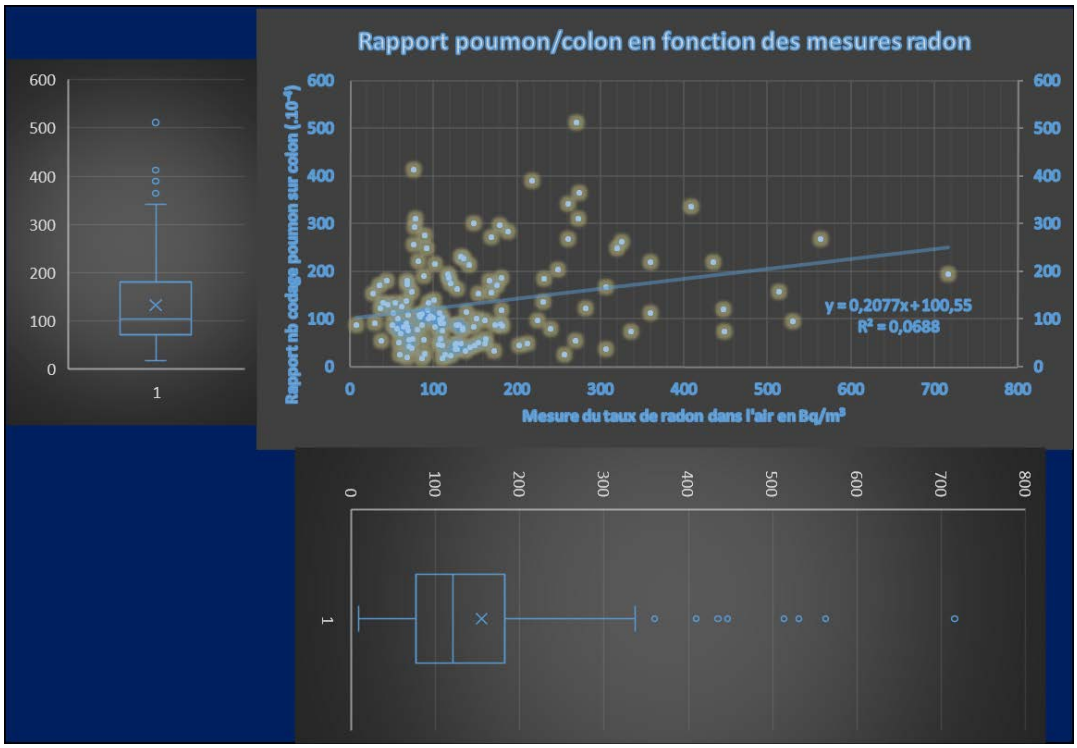


Figure 74 Nuage de point représentant le rapport du nombre de codage ADICAP du poumon sur les codages colon ; en fonction de taux de radon moyen dans l'air en B/m^3 avec la boîte de Tukey de chaque série et les droites de tendance

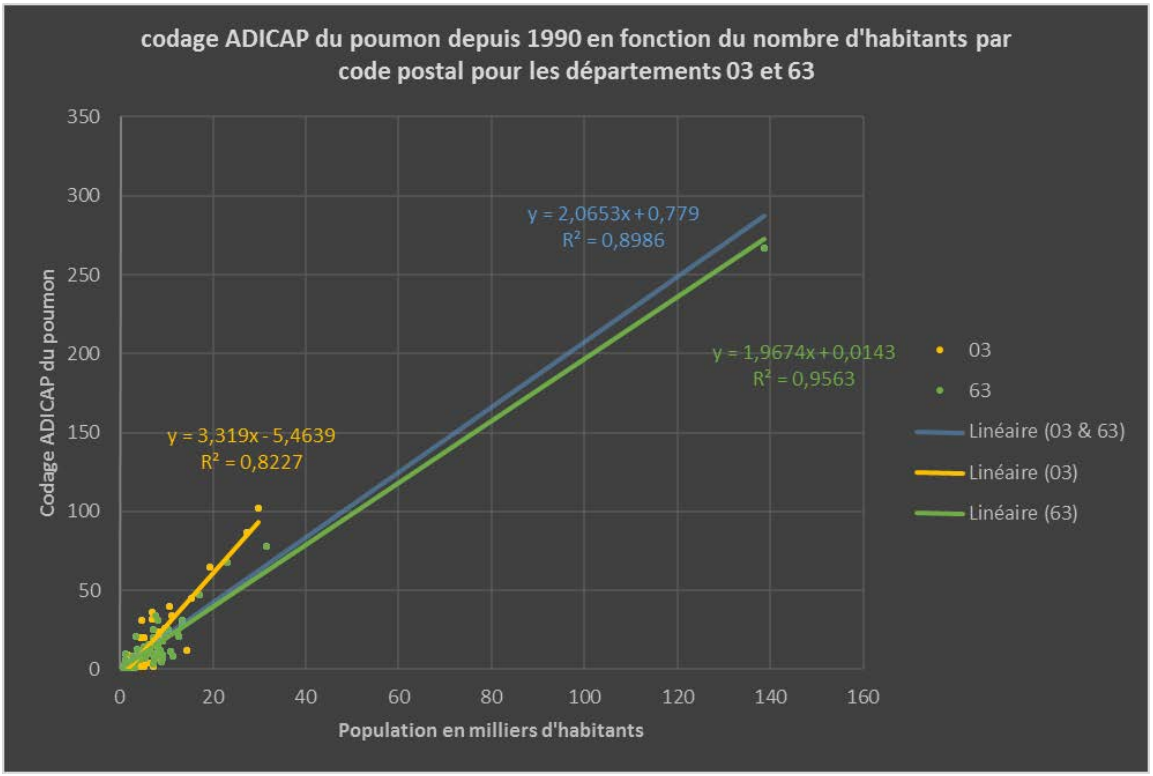


Figure 75 Nuage de point représentant le nombre de codage ADICAP du poumon en fonction de la population des agglomérations pour les départements de l'Allier et du Puy de dôme

En utilisant les données ACP du cancer du poumon de Sipath-Unilabs et les données issues des bases ARDOC et ABIDEC pour les cancers du sein et du côlon ainsi que les données INSEE relatives à la population en Auvergne en 2014, il ne nous a pas été possible de mettre en évidence une prévalence des cancers du poumon en fonction du taux de radon dans l'air. Une plus grande précision géographique avec l'utilisation des codes INSEE au lieu des codes postaux ainsi que l'utilisation de données ACP significatives pour les départements 15 et 43 où la présence de radon est importante avec des moyennes supérieures à 700 Bq/m^3 et des mesures maximales au-dessus de 4000 Bq/m^3, nous permettrait des analyses plus pertinentes.

Pour consolider nos données cancers, il serait intéressant de consulter d'autres sources qui ne font pas encore partie du réseau GINSENG comme les données ACP du Centre Jean Perrin ou du CHU de Clermont-Ferrand. Un projet INCa vient d'être proposé dans ce sens. Nous pouvons envisager d'automatiser cette étude annuellement. Il semble nécessaire de demander à la CNIL l'autorisation d'exploiter les codes INSEE en plus des codes postaux si nous voulons par la suite essayer d'étudier l'impact d'un équipement industriel sur la santé des populations.

Conclusion

Nous avons présenté dans ce chapitre les tests de performance et les validations de l'infrastructure distribuée. Celle-ci s'avère opérationnelle entre le laboratoire Sipath-Unilabs et les associations de dépistage des cancers pour le transfert de comptes-rendus médicaux d'anatomie-pathologique. Les requêtes réalisées sur les bases de données médicales peuvent être établies sans utiliser d'interface graphique. La validation du réseau dans le cadre d'une étude de cas en épidémiologie concernant l'impact du radon sur l'état de santé de la population en Auvergne a permis de mettre en exergue graphiquement et en temps réel une potentielle corrélation entre des niveaux de radon élevés et une incidence du cancer du poumon. Une action spécifique concernant la mise en œuvre d'une interface web pour l'épidémiologie sera menée à la suite de cette thèse ; il sera intégré à cette interface toutes les fonctionnalités permettant des requêtes sémantiques. Des démarches ont été entreprises pour élargir les applications médicales du réseau GINSENG à des études en périnatalité ainsi qu'en pharmacovigilance.

Conclusion

Le travail de thèse transcrit dans ce manuscrit avait pour but de faire la preuve de concept d'un prototype d'infrastructure distribuée et sécurisée de bases de données médicales pour l'e-santé et l'épidémiologie. En effet, la plupart des réflexions menées à ce jour sur la mise à disposition des documents électroniques de santé pour le suivi des patients, s'orientent vers une centralisation des données médicales. Nous avons donc cherché à proposer des solutions techniques OpenSource pour mettre à disposition des données de santé, dans le respect des droits des patients, auprès de toute structure administrative et/ou médicale partenaire afin d'améliorer leur suivi.

Après avoir discuté dans le premier chapitre sur l'état de l'art de l'e-santé dans le contexte du transfert d'information, nous avons présenté le projet ANR GINSENG dans lequel ce travail de thèse a pu être effectué. Nous avons ensuite dressé le recueil des spécifications de ce réseau sécurisé de partage d'informations médicales. Dans le troisième chapitre nous avons présenté les solutions mises en œuvre pour la construction matérielle et logicielle de toute l'infrastructure ainsi que l'environnement réseau adapté et sécurisé. Nous avons ensuite plus particulièrement travaillé sur les actions à mener pour une gestion distribuée des bases de données médicales en insistant sur leur structuration standardisée, les algorithmes d'identification des patients pour permettre un bon chaînage de l'information médicale et enfin sur les requêtes distribuées permettant l'interrogation en temps réel des bases de données intégrées au réseau.

Bien sûr, les champs applicatifs dans le domaine de l'e-santé étant vastes, nous avons orienté notre travail sur une mise en œuvre des développements auprès de partenaires privilégiés en région Auvergne, regroupés en réseaux collaboratifs structurés, le Réseau Sentinelle Cancer Auvergne (RSCA) et le Réseau de Santé Périnatale Auvergne (RSPA). Concernant le suivi des cancers du dépistage organisé, il était primordial de faciliter les transferts d'informations médicales vers les SGDO pour permettre un meilleur suivi des cancers du sein, du côlon et du col en région et améliorer le travail quotidien des personnels de ces structures. Il était aussi nécessaire de proposer un outil permettant une mise à disposition sécurisée des données médicales pour des analyses épidémiologiques en temps réel, exhaustives et à grande échelle.

Tous les développements pourront être étendus et validés à d'autres disciplines médicales et/ou médico-sociales.

Nous avons établi la liste des spécifications nécessaires à la mise en œuvre d'une infrastructure capable d'interconnecter les SGDO, les cabinets ACP et les hôpitaux. La

particularité de notre réseau par rapport aux réflexions actuelles en matière d'e-Santé est la non centralisation de l'information. En effet, il nous est apparu essentiel que les producteurs de données de santé restent maîtres de l'information qu'ils produisent. Nous avons donc considéré un réseau de bases de données distribuées. Ce réseau a été configuré de manière à pouvoir être exploité avec des connexions standards de type ADSL. Les utilisateurs peuvent être identifiés par leur carte de professionnels de santé. Les données confidentielles sont quant à elles hébergées chez un professionnel du traitement de l'information médicale un HADS, comme nos partenaires travaillent en région Auvergne nous avons choisi IDS, le même prestataire que pour le GCS référant dans notre région. Le réseau que nous créons gère des données de santé, il doit être extrêmement sécurisé c'est pourquoi il est contraint par un VPN. Les informations doivent être consultables depuis le logiciel métier des SGDO nommé « Zeus ». Cette démarche a l'avantage de ne pas perturber le mode de fonctionnement des personnels effectuant le suivi des patients dans les SGDO. Un second mode de consultation des données collectées par le réseau est également possible depuis un portail WEB hébergé chez le HADS. En fonction des autorisations accordées à chaque utilisateur, le portail WEB est en mesure de donner accès aux informations et aux services correspondants. Les requêtes sur les données médicales peuvent être lancées automatiquement ou manuellement depuis le portail. De plus, les résultats peuvent être présentés sous une forme graphique ou sous forme de fichiers texte qui peuvent être exportés pour un traitement ultérieur. Le cœur de ce système repose sur une identification correcte des patients à travers les différentes bases de données de l'infrastructure. L'un des avantages concurrentiels de notre étude est la capacité à interconnecter des bases de données très différentes. Cette étape est rendue possible grâce à l'utilisation d'algorithmes de recherche de chaînes de caractères performants.

Le choix de l'algorithme d'identification des patients, sa configuration et son exploitation ont été des éléments déterminants de notre solution. C'est uniquement à la suite de nombreuses campagnes de tests que nous avons pu décider lequel des algorithmes de comparaison de chaînes de caractères était le plus approprié à nos besoins. En s'appuyant sur nos recherches communes avec (Li 2015) nous utilisons donc l'algorithme issu des travaux de Jaro et Winkler. La méthodologie mise en œuvre pour intégrer les données confidentielles et médicales à l'infrastructure est la suivante : une fois les données de nos partenaires exportées sur chaque « nœud » (un nœud par site partenaire) sans prérequis spécifique sur le format des données, celles-ci sont triées de manière à identifier chaque patient. Pour tout nouveau patient sur chaque site, un identifiant unique est attribué, cet identifiant et sa correspondance avec les données confidentielles sont copiées sur une base de données « annuaire » sécurisée hébergée chez un

HADS. Les données de chaque patient sont ainsi chaînées entre les différents sites partenaires. L'authentification et les niveaux d'accréditation des utilisateurs du réseau est importante et doit contraindre l'accès aux données. C'est le HADS qui fournit les composants logiciels capables d'authentifier les certificats X509 contenus sur les cartes CPS. La gestion des groupes d'utilisateurs est également réalisée par le HADS.

Les tests de temps nécessaires à l'importation des données avec les solutions que nous avons retenues concernent le temps d'identification d'un patient dont le dossier médical est déjà connu dans l'un des sites participants au projet, ainsi que le temps d'intégration des nouvelles données médicales dans notre SI. Différents paramètres comme la volumétrie des données, les taux de transfert de données, ainsi que les temps nécessaires à l'analyse des données ont été testées. Certaines alternatives n'ont pas encore été confrontées aux données réelles, comme la gestion des données au moyen de la sémantique stockée dans des triplets RDF et interrogés en SPARQL. La sémantique permet d'ajouter du sens à la donnée et peut faciliter l'interconnexion des données de santé avec l'Open Data de manière à consolider les informations médicales. D'une façon similaire nous exposons les traitements que nous avons réalisés avec le logiciel d'analyses statistiques 'R', nous proposons aussi une représentation graphique des données statistiques sur des cartes géographiques.

L'infrastructure déployée a été testée pour une étude épidémiologique concernant l'impact du radon dans l'air et la survenue des cas de cancers du poumon. Pour se faire, nous avons croisé les données provenant d'une étude du BRGM caractérisant l'aléa radon en région Auvergne, avec les données ACP recueillies par le réseau pour les pathologies du poumon notamment sur la période 1990 - 2015. Le niveau de gravité et le type de pathologie n'a pas été pris en compte dans cette étude. Ces cas ont été normalisés par rapport au nombre de résidents par code postal. Ceci nous permet d'obtenir un nombre de pathologies du poumon pour mille habitants mis en relation avec la moyenne de radon en Bq/m^3 pour chaque code postal en Auvergne. Une deuxième normalisation des données a été réalisée en comparant les pathologies du poumon à celles du côlon sur la même période. Cette étude statistique préliminaire n'a pas révélé de corrélation entre le taux moyen de radon dans l'air et les pathologies liées au poumon. Cette étude nécessite d'être complétée grâce à une intégration future des bases de données ACP du centre de lutte contre le cancer Jean Perrin et du CHU de Clermont-Ferrand, ceci nous permettrait de pouvoir traiter les informations de 11500 patients atteints d'un cancer du poumon depuis les années 1990.

En proposant une solution peu onéreuse, facile à maintenir et à transposer à d'autres domaines de santé, modulable, et dont la mise en œuvre nécessite peu de développements de la

part des fournisseurs de logiciels métier ; nous maximisons les chances d'adoption de notre outil. Les autorisations de la CNIL, reconductibles chaque année, garantissent la bonne marche de notre infrastructure. Les associations de dépistage du cancer en Auvergne sont interconnectées avec le plus grand cabinet ACP de la région ainsi qu'avec un hébergeur agréé de santé. Les requêtes des données ACP par les SGDO en Auvergne fonctionnent et permettent de faciliter l'intégration et la mise à jour des données médicales dans le cadre du dépistage organisé des cancers. L'accès à ces bases interconnectées permet des études épidémiologiques ciblant le cancer colorectal, multi sources ; les premières publications exploitant ces résultats sont en cours de rédaction.

Des développements ultérieurs seront à prévoir pour affiner et automatiser totalement les processus en utilisation de routine l'ergonomie liée à l'utilisation des interfaces utilisateurs devra être également améliorée en fonction de l'utilisation grandissante de l'outil.

Notre étude s'est focalisée sur les échanges dans les domaines de la périnatalité et du dépistage organisé du cancer en région Auvergne. Nous avons gardé à l'esprit lors de nos réflexions la possibilité d'étendre nos solutions à d'autres régions, ainsi qu'à d'autres domaines de la santé. Nous ne traitons actuellement que des documents textuels mais la solution peut fonctionner de la même manière en intégrant des données d'imagerie médicale.

Perspectives

La pérennisation de notre solution et sa consolidation seront les deux objectifs majeurs à court terme. Outre la recherche de financement nécessaire à la maintenance des serveurs des développements sont encore nécessaires notamment au niveau de l'interface utilisateurs pour faire évoluer l'infrastructure de la preuve de concept vers un outil capable de passer en production. Certains processus sont à affiner autour de l'identification des patients lorsque le logiciel retourne un score qui permet un doute raisonnable et non pas une réponse franche.

Concernant le déploiement et l'usage de notre infrastructure dans la nouvelle région Auvergne Rhône Alpes nous sommes actuellement en discussion d'une part avec l'ARS et le GCS, et d'autre part nous envisageons d'étendre notre solution chez les SGDO utilisant les produits de O.S.I. Santé.

La recherche d'un standard capable de coder la totalité des informations médicale est certainement une utopie, il est toutefois intéressant d'œuvrer en faveur des standards même s'ils sont plusieurs et de considérer leur interopérabilité.

Actuellement nous partageons des documents textuels au format « .docx » ainsi que des informations sous forme de caractères. Les domaines que nous considérons et la santé en général s'appuie sur d'autres sources ; imageries, sons ou même vidéos. Il serait intéressant de considérer ces médias dans les prochaines évolutions de l'infrastructure, avec les besoins qui en découleront en termes de stockage et de débit réseau.

Nous continuons d'améliorer notre solution. Nous pouvons envisager de concevoir une version basée sur des serveurs totalement virtualisés capable de fonctionner sur une machine courante utilisée par un médecin généraliste par exemple. Nous poursuivons aussi nos démarches pour intégrer l'ENRS du GCS Simpa qui est dans notre vision la place naturelle de cet outil.

L'informatique est une science jeune, si la loi de Moore continue d'être respectée, la technologie va continuer de révolutionner nos façons d'être et de penser pendant les prochaines décennies. Les prémices de la réalité virtuelle, de l'IoT, et de la robotique humanoïde sont très prometteuses. Plus près de nos préoccupations le principe de container introduit par Docker n'est pas encore mature, notamment à cause des problèmes de sécurités liés à l'étanchéité des containers entre eux. La version de Ubuntu LTS 16.04 au travers de LXD intègre un support de containers compatible avec Docker ce qui traduit clairement l'engouement des administrateurs système en direction de cette technologie qui si elle ne remplace pas la virtualisation comme nous la connaissons pourra permettre une utilisation plus dense des ressources informatiques.

De plus LXD peut être combiné avec Open Stack et permet de profiter de toute la modularité de cette solution qui a vocation à remplacer l'approche KVM que nous utilisons actuellement dans GINSENG, si toutes les conditions de sécurités sont réunies. D'une façon similaire un système de fichier de type CephFS pourra être considéré dans l'optique d'améliorer les échanges de fichiers à l'intérieur du cloud privé, au sein de notre VPN.

La version actuelle du réseau GINSENG répond à nos attentes comme preuve de concept et elle mérite d'être améliorée et consolidée. Elle nous permet de démontrer que des solutions sont possibles et qu'elles permettent des économies de temps et d'argent pour le suivi des invitations du dépistage organisé en Auvergne. Le prochain challenge est la pérennisation de cette approche pour lui permettre d'arriver à maturité.

Déclaration d'absence de conflit d'intérêts :

Sébastien Cipière, déclare n'avoir aucun lien ou aucune affiliation, qu'elle soit de nature personnelle ou professionnelle, qui pourrait avoir une influence réelle, potentielle ou apparente sur son jugement ou ses actions.

Lexique

Lexique

A

ACR (classification) :

American College of Radiology, classification en mammographie, adaptée par l'HAS, est une classification internationale en fonction du degré de suspicion, sur une échelle de 0 à 6.

ADICAP : l'Association pour le Développement de l'Informatique en Cytologie et Anatomie Pathologique.

ACP Anatomie et cytologie pathologiques :

Souvent abrégé à l'oral par « anapath » est en fait un néologisme qui nous permet de condenser d'une part l'anatomo-pathologie étudie les lésions macroscopiques et microscopiques de tissus prélevés sur des êtres vivants malades par biopsie, frottis ou biopsie extemporanée et d'autre part la cytopathologie qui s'intéresse à un étalement et non plus à une coupe, qui permet de mieux apprécier la morphologie cellulaire plutôt que les caractéristiques tissulaires (domaine de l'histologie).

ARS : Agence Régionale de Santé.

B

BDD : Base De Données.

C

CAP : le Collège des Pathologistes Américains (College of American Pathologists)

CAS : Central Authentification Service c'est une implémentation d'un SSO souvent couplé avec Schibboleth.

CCTIRS : **C**omité **C**onsultatif sur le **T**raitement de l'**I**nformation en matière de **R**echerche dans le domaine de la **S**anté, rend notamment des avis sur la pertinence des données nominatives à caractère personnel par rapport à l'objectif de la recherche.

CERTA : **C**entre d'**E**xpertise gouvernemental de la **R**éponse et de **T**raitement des **A**ttaques informatiques http://www.certa.ssi.gouv.fr/

Châssis : Armoire informatique servant à accueillir les ressources informatiques dans les salles machines.

Chiffrement : Application d'une fonction à un message clair, qui devient alors illisible sans la clef qui permettra de le déchiffrer et de lui redonner son sens initial.

Clef : Il existe différent type de clef symétrique ou non. Elles servent à ouvrir et fermer des verrous informatiques qui peuvent dans l'exemple de la cryptographie rendre un message lisible ou non.

CNIL : **C**ommission **N**ationale de l'**I**nformatique et des **L**ibertés - http://www.cnil.fr/

CNOM : **C**onseil **N**ational de l'Ordre des Médecins, organisme de droit privé chargé d'une mission de service public, au service des médecins, dans l'intérêt des patients.

CR : **C**ompte-rendu.

Cryptage (!) : ce pourrait être l'action de crypter. Cependant ce mot n'existe pas dans la langue Française. Cf. chiffrement.

Crypter (!) : ce mot n'existe pas dans la langue Française, il est employé en lieu et place de chiffrement

Cryptographie : c'est l'une des disciplines de la cryptologie, elle a pour vocation de brouiller le message aux yeux des destinataires non légitimes (qui ne possèdent pas la clef de déchiffrement)

Cryptologie : la science du secret, elle englobe la cryptographie et la cryptanalyse

D

Déchiffrement : action qui permet en appliquant une clef appropriée à un message chiffré de leur rendre à nouveau compréhensible.

Décryptage : action qui si elle s'avère fructueuse aboutit au même résultat que le déchiffrement, mais sans posséder la clef nécessaire. C'est une attaque qui vise à casser la sécurité qui entoure le message.

DGS : Direction générale de la santé.

DMP : Dossier Médical Partagé, initiative de 2004 concrétisée en 2011 au travers du portail dmp.gouv.fr son but est de centraliser les dossiers médicaux des Français.

E

Épistémologie : Théorie de la connaissance en général.

Extemporané (examen) :

Le médecin pathologiste apporte, par ses techniques d'examen rapide, une aide importante au chirurgien pendant l'intervention chirurgicale, par exemple en lui confirmant ou non le caractère cancéreux d'une tumeur, en lui garantissant l'ablation totale de la lésion ou en l'aidant à préciser le stade d'extension du cancer.

Événement indésirable grave

F

FOAF : Friend Of A Friend, traduit par l'ami d'un ami, est un vocabulaire RDF permettant de décrire des personnes, et les relations qu'elles entretiennent entre elles.

FCV : Frottis cervico-vaginal est un examen banal, effectué dans le cadre du dépistage du cancer du col de l'utérus. Cet examen consiste à prélever des cellules superficielles au niveau du col de l'utérus. Il est pratiqué au cours d'un examen gynécologique. Le prélèvement est ensuite examiné au microscope dans un laboratoire de cytologie. Selon l'aspect des cellules, on peut supposer que le col utérin est normal, ou bien qu'il présente une infection, une lésion précancéreuse ou un cancer. La plupart des lésions décelées par le frottis sont liées à une infection par le virus HPV. Cinq à six millions de frottis sont effectués chaque année en France.

FSE : Feuille de Soins Électronique, a pour but de remplacer la version papier de la feuille de soin, et ainsi faciliter des formalités administratives.

G

GIE : Groupement d'Intérêt Économique.

GCS : Groupement de Coopération Sanitaire.

H

Hachage (fonction de) :

Son but est d'être une fonction injective et non réversible. Ainsi chaque information en entrée produira un haché que l'on peut considérer comme une empreinte qui doit lui être propre, et cette empreinte ne doit pas pour autant permettre de remonter à l'information originelle. Si on constate deux hachés identiques cela signifie que les informations à l'origine

de ces hachés sont rigoureusement identiques.

Haché : résultats d'une fonction de hachage.

HADS : Hébergeur Agrée de Données de Santé à ne pas confondre avec *Hospital Anxiety and Depression Scale* dont l'acronyme est aussi HADS

HAS : La Haute Autorité de Santé, autorité publique indépendante à caractère scientifique dotée de la personnalité morale (remplace l'ANAES depuis 2005).

Hash : haché en français, cf. fonction de hachage.

HDS : Hébergeur de Données de Santé.

I

IaaS : *Infrastructure as a Service*, Infrastructure en tant que service, concept proposé par les hébergeurs de données et de service « cloud » qui proposent de louer une machine virtuelle que le client peut administrer.

IETF : L'Internet Engineering Task Force, abrégée IETF, littéralement traduit de l'anglais en « Détachement d'ingénierie d'Internet » est un groupe informel, international, ouvert à tout individu, qui participe à l'élaboration de standards Internet. L'IETF produit la plupart des nouveaux standards d'Internet.

IGN : Institut géographique national.

INCa : Institut national du cancer.

Insee : Institut national de la statistique et des études économiques.

InVS : Institut de veille sanitaire.

IRI : International Ressource Identifier, généralise l'URI en acceptant les caractères codés en UTF-8.

J

JAVA : langage de programmation orientée objet.

K

Kernel : se traduit en français par noyau du système d'exploitation, c'est une partie essentielle de l'OS.

KparK : système de sécurité informatique gérant les entrées et les sorties.

L

LIMOS : Laboratoire d'informatique de modélisation et d'optimisation des systèmes

LPC : Laboratoire de physique corpusculaire.

M

MD (4-5) : Comme les fonctions SHA-x se sont des fonctions de hachage.

MOF : Meta-Object Facility standard de l'OMG à destination de l'ingénierie des modèles.

m-santé: Mobile santé (*mHealth ou m-Health* en anglais) se rapporte à la consultation de services connectés à travers des terminaux mobiles comme des tablettes.

N

NAACCR : l'Association des Registres du Cancer Nord Americains.

Notation3 : norme relative à la sérialisation non-XML des modèles RDF, développé pour être lisible par les humains : la notation (N3) est bien plus compacte et lisible que la notation RDF/XML.

O

OMG : Object Management Group, consortium éditant des standards informatiques, tel que UML et XMI.

OMS : Organisation Mondiale de la Santé.

Onduleur : Équipement électrique permettant d'améliorer la qualité du courant fournit aux équipements informatiques et surtout en cas de coupure de l'alimentation électrique du bâtiment les batteries qu'il contient permettront d'assurer la continuité de l'alimentation pour permettre une extension convenable des systèmes.

Onthologie : En philosophie, l'ontologie est l'étude de l'être en tant qu'être, c'est-à-dire l'étude des propriétés générales de ce qui existe.
Par analogie, le terme est repris en informatique comme l'ensemble structuré des termes et concepts représentant le sens d'un champ d'informations, que ce soit par les métadonnées d'un espace de noms, ou les éléments d'un domaine de connaissances. L'ontologie constitue en soi un modèle de données représentatif d'un ensemble de concepts dans un domaine, ainsi que des relations entre ces concepts. Elle est employée pour raisonner à propos des objets du domaine concerné.

OWL : *Web Ontology Language* (OWL) est un langage de représentation des connaissances construit sur le modèle de données de RDF.

P

Paradigme : Modèle représentant le monde considéré.

Q

QoS : Qualité de service (en anglais)

R

Rack : Anglicisme utilisé dans le monde des salles machines pour parler d'un châssis.

Rackable : Capacité d'une ressource informatique à être positionnée dans un rack.

RDF : Resource Description Framework est un modèle de graphe destiné à décrire de façon formelle les ressources Web et leurs métadonnées, de façon à permettre le traitement automatique de telles descriptions.

RDFS : RDF Schema est un langage extensible de représentation des connaissances. Il appartient à la famille des langages du Web sémantique publiés par le W3C.

RFC : Request For Comment, document technique qui décrivent le fonctionnement d'Internet et des systèmes qui y sont reliés, certains RFC deviennent des standards.

S

SaaS : *Software as a Service*, logiciel en tant que service, concept proposé par les hébergeurs de données et de service « cloud » qui proposent de louer un logiciel au mois plutôt que d'acheter une version donnée à un instant 't'.

SAML : Security Assertion Markup Language standard d'échange pour les mécanismes d'authentification comme Schibboleth.

Schibboleth : mécanisme de propagation d'identités au sein d'une fédération, souvent associé avec CAS pour fournir un service de SSO évolué.

Sésame : framework Java de stockage et de requêtage de données RDF.

SFP : Société Française de Pathologie

SGBD : Système de Gestion de Base de Données.

SGDO : Structure de Gestion du dépistage organisé du cancer.

SHA(1-2-256-512-3) : Ce sont différentes fonctions de hachage, seule la fonction SHA-3 peut être considérée

comme fondamentalement différente des autres qui partage une base commune.

SI : **S**ystème d'**I**nformation, ensemble des ressources permettant à l'information de circuler au sein de l'entité considérée.

SKOS : Simple Knowledge Organization System (Système simple d'organisation des connaissances) est une famille de langages formels permettant une représentation standard des thésaurus, classifications ou tout autre type de vocabulaire contrôlé et structuré. Construit sur la base du modèle de données standard RDF, son principal objectif est de permettre la publication facile de vocabulaires structurés pour leur utilisation dans le cadre du Web sémantique. SKOS est, depuis le 18 août 2009, une recommandation du W3C.

SPARQL : SPARQL Protocol and RDF Query Language (prononcé sparkle en anglais : « étincelle ») est un langage de requête et un protocole qui permet de rechercher, d'ajouter, de modifier ou de supprimer des données RDF disponibles à travers Internet.

SSL : Secure Socket Layer

SSO : Single Sign-On système d'authentification unique évitant la saisie de multiple mot de passe pour accéder à différents services.

T

Taxonomie : (ou taxinomie) est la science qui a pour objet de décrire les organismes vivants et de les regrouper en entités appelées taxons afin de les identifier puis les nommer et enfin les classer.

TLS : Transport Layer Security

Triplestore : Un triplestore est une base de données spécialement conçue pour le stockage et la récupération de données RDF.

Triplet RDF : Un triplet RDF est l'unité de données la plus petite contenue dans un graphe RDF au sein d'une base de données de type triplestore (ex. :(sujet, prédicat, objet)).

U

UML : Unified Modeling Language, c'est un langage de modélisation.

URI : l'Uniform Resource Identifier doit permettre d'identifier une ressource de manière permanente, même si la ressource est déplacée ou supprimée (exemple : URL, ISBN, …).

URL : Uniform Resource Locator, est un type d'URI, permettant de localiser une ressource sur un réseau ex : cipiere.fr .

URN : Uniform Resource Name, est un type d'URI, comme le numéro ISBN qui permet de catégoriser une publication.

V

VIP : Virtual Imaging Platform, c'est une plateforme dédiée au transfert de fichiers et aux simulations d'imagerie médicale. VIP a été développé dans l'ANR-09-COSI-03.

W

W3C : Le World Wide Web Consortium, abrégé par le sigle W3C, est un organisme de normalisation à but non-lucratif, fondé en octobre 1994 comme un consortium chargé de promouvoir la compatibilité des technologies du World Wide Web telles que HTML, XHTML, XML, RDF, SPARQL, CSS, PNG, SVG et SOAP.

Web sémantique : Le Web sémantique est un mouvement collaboratif mené par le (W3C) qui favorise des méthodes

communes pour échanger des données.

WHO : Cf. OMS en anglais *World Health Organization*

WSDL : Web Service Description Language.

X

XMI : XML Metadata Interchange est, un standard d'échange de métadonnées UML via XML, défini par l'OMG.

XML : Extensible Markup Language, « langage de balisage extensible » est un langage informatique de balisage générique. L'objectif initial est de faciliter l'échange automatisé de contenus complexes (arbres, texte riche…) entre systèmes d'informations hétérogènes.

Y

YUM : Yellowdog Update Manager gestionnaire de paquets Fedora remplacé par DNF dans Fedora 22.

Z

Zimbra : Société développant une suite collaborative éponyme qui regroupe des services de messagerie et des outils de collaborations à distance.

Annexes

Annexes

A.1 Autorisations CNIL du projets GINSENG

Le : 21/09/2015

Source : http://www.legifrance.gouv.fr/affichCnil.do?id=CNILTEXT000028268603 - date d'accès mai 2015
Date de publication sur legifrance: 05/12/2013

Commission Nationale de l'Informatique et des Libertés

DELIBERATION n°2013-364 du 14 novembre 2013
Délibération n° 2013-364 du 14 novembre 2013 autorisant le Centre national de la recherche scientifique (CNRS) à mettre en œuvre un traitement automatisé de données à caractère personnel ayant pour finalité la mise en place, à titre expérimental, d'une plateforme informatique permettant l'interrogation et le croisement de bases de données mises en œuvre dans le cadre du suivi médical des patients à des fins d'évaluation des pratiques de soins et de recherche épidémiologique.

(Demandes d'autorisation n° 1515344 et 1519026)

La Commission nationale de l'informatique et des libertés,
Saisie par le Centre national de la recherche scientifique (CNRS) d'une demande d'autorisation concernant un traitement automatisé de données à caractère personnel ayant pour finalité la mise en place, à titre expérimental, d'une plateforme informatique permettant l'interrogation et le croisement de bases de données mises en œuvre dans le cadre du suivi médical des patients à des fins d'évaluation des pratiques de soins et de recherche épidémiologique.
Vu la Convention n° 108 du Conseil de l'Europe pour la protection des personnes à l'égard du traitement automatisé des données à caractère personnel ;
Vu la directive 95/46/CE du Parlement européen et du Conseil du 24 octobre 1995 relative à la protection des personnes physiques à l'égard du traitement de données à caractère personnel et à la libre circulation de ces données ;
Vu la loi n° 78-17 du 6 janvier 1978 modifiée relative à l'informatique, aux fichiers et aux libertés, notamment ses articles 8-IV et 25-I-1;
Vu le décret n° 2005-1309 du 20 octobre 2005 modifié pris pour l'application de la loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés
Vu le dossier et ses compléments ;
Sur la proposition de M. Jean MASSOT, commissaire, et après avoir entendu les observations de M. Jean-Alexandre SILVY, commissaire du Gouvernement,
Formule les observations suivantes :

Responsable du traitement	Le Centre national de la recherche scientifique (CNRS) est un établissement public à caractère scientifique et technologique, placé sous la tutelle du Ministère de l'Enseignement supérieur et de la Recherche.
Sur la finalité	Le CNRS envisage de mettre en place, à titre expérimental, une plateforme informatique dénommée GINSENG (Global Initiative for Sentinel E-health Network onGrid), permettant l'interrogation et le croisement de bases de données mises en œuvre dans le cadre du suivi médical des patients, à des fins d'évaluation des pratiques de soins et de recherche épidémiologique. Le traitement projeté vise à tester la faisabilité technique de l'outil et à en mesurer l'intérêt pour la communauté des professionnels à partir de deux études, l'une portant sur le suivi du cancer du sein et l'autre sur la pertinence des césariennes en région Auvergne. Les bases d'associations de dépistage du cancer (ARDOC et ABIDEC), de laboratoires d'anatomocyto-pathologie, de services de santé publique et de biostatistiques et de maternités appartenant à un réseau de périnatalité de la région Auvergne déclarées à la CNIL seront concernées.

	La plateforme GINSENG est un système distribué et non centralisé, qui repose sur la mise en réseau de bases médicales issues de bases existantes mais dont les données d'identité des patients auront été préalablement pseudo-anonymisées comme suit : · pour chaque organisme participant, une base GINSENG qui restera localisée dans l'établissement sera créée à partir de la base de production. Cette nouvelle base restera sous le contrôle de l'établissement, mais elle sera administrée à distance par le CNRS qui est le maitre d'œuvre du projet. · une table de correspondance entre l'identité du patient et un numéro calculé à partir d'un procédé de pseudo-anonymisation sera conservée au sein de chaque site de prise en charge des patients. · les données d''identification des patients seront transmises de manière chiffrée au serveur CNRS qui interrogera, de façon automatisée, l'ensemble des bases afin de savoir si un patient y est référencé et permettra ainsi de relier les informations relatives à un même patient pris en charge sur plusieurs sites. · Les éventuels problèmes d'identito-vigilance seront réglés au sein des organismes chargés du suivi médical des patients. Une fois l'identification validée, un numéro aléatoire sera substitué aux données d'état civil. Dans l'hypothèse où l'expérimentation s'avérerait concluante, cette plateforme pourrait être utilisée dans le cadre d'autres projets de recherche. La Commission considère les finalités poursuivies comme déterminées, explicites et légitimes au sens de l'article 6-2° de la loi du 6 janvier 1978 modifiée. Elle considère qu'il peut être fait application au bénéfice du CNRS des dispositions des articles 8-IV et 25-I, 1° de la loi du 6 janvier 1978 modifiée, qui soumettent à autorisation les traitements comportant des données relatives à la santé et justifiés, comme en l'espèce, par l'intérêt public.
Sur les données traitées	Les catégories de données à caractère personnel traitées sont relatives à l'identification du patient (nom, prénom, date de naissance, sexe, code postal de résidence) à des fins de validation des identités et de chaînage des données dans le cadre de la construction de l'outil. Dans le cadre des deux études pilotes, les données suivantes seront collectées : - données d'identification : sexe, mois et année de naissance - données médicales strictement nécessaires à la mise en œuvre de chacune des deux études envisagées, à l'exclusion de toute date complète. La Commission estime que les données traitées sont adéquates, pertinentes et non excessives au regard de la finalité poursuivie, conformément aux dispositions de l'article 6-3° de la loi du 6 janvier 1978 modifiée.
Sur les destinataires	- Les médecins épidémiologistes et les praticiens médicaux des organismes partenaires du projet, spécialement habilités par ceux-ci, dans la mesure strictement nécessaire à la mise en œuvre d'une étude. - Les personnels du CNRS ne seront jamais en mesure d'associer des données d'identification et des données de santé relatives aux patients. Ces destinataires n'appellent pas d'observation de la part de la Commission.
Sur l'information et le droit d'accès	Les personnes dont les données seront traitées et transmises seront, avant le début du traitement de ces données, individuellement informées de la nature des informations transmises, de la finalité du traitement des données, des personnes physiques ou morales destinataires des données, ainsi que de leurs droits et des modalités pratiques de leur exercice, par la remise d'une notice d'information individuelle. Il appartiendra au médecin responsable de la prise en charge thérapeutique, en contact direct avec les patients, de leur remettre ce document. La note d'information est rédigée de telle sorte que le patient puisse comprendre clairement que le traitement mis en œuvre a vocation à permettre, d'une part, son suivi médical au sein de l'établissement et d'autre part, le recueil et la remontée d'informations à des fins de suivi épidémiologique et d'évaluation des pratiques de soins et qu'il a la possibilité de s'y opposer. Les patients y seront clairement informés du caractère facultatif de leur participation à la plateforme et de la faculté qu'ils ont de retirer leur accord à tout moment et de refuser la transmission de leurs données sans avoir à se justifier et sans conséquence sur leur prise en charge médicale. Pour les patients qui sont d'ores et déjà présents dans la base, une demande de dérogation à l'obligation d'information individuelle a été effectuée par le CNRS

	compte tenu de la difficulté à retrouver les personnes concernées. La Commission prend acte de l'affichage qui sera systématiquement effectué dans l'ensemble des organismes partenaires du projet et de l'information individuelle qui sera effectuée si la personne concernée est à nouveau prise en charge dans l'organisme. La Commission estime que ces modalités d'information et d'exercice des droits n'appellent pas d'observation.
Sur les mesures de sécurité	Les bases GINSENG seront conservées sur chacun des sites de prise en charge des patients dans les conditions d'hébergement de la base de production. Afin de ne pas conserver les données d'état civil dans les bases GINSENG, une table de correspondance sera créée. Le procédé de pseudonymisation associant ces données et un identifiant unique s'appuie sur un algorithme standardisé, ce qui constitue un gage de sécurité théorique et pratique, car il a été développé dans de nombreux langages informatiques. L'interrogation des bases permettant de savoir si un patient est suivi sur un autre site se fera, à l'aide d'un procédé automatique, à partir de données d'identification à l'exclusion de toute donnée médicale. La table de correspondance restera hébergée au sein de chaque organisme et seuls les personnels habilités des sites de prise en charge seront en mesure de décoder les identifiants correspondant aux dossiers des patients qu'ils suivent, si besoin. La présence d'un même patient au sein des différentes bases est vérifiée à la mise en place initiale de la plateforme en requêtant l'ensemble des bases. Une synchronisation a lieu quotidiennement passée cette étape initiale. La transmission des informations est réalisée sur un réseau privé virtuel sous forme chiffrée, ce qui garantit la confidentialité des données échangées. L'exploitation de ces bases par les professionnels de santé s'effectue via un portail web hébergé au CNRS. Des procédures d'habilitation d'accès pour les professionnels de santé sont mises en œuvre par le CNRS sous la haute autorité de praticiens médicaux ayant la responsabilité des données médicales. Une fois habilités, les médecins accèderont au portail web en s'authentifiant au moyen de leur carte de professionnel de santé (CPS). Ils seront alors connectés à une interface web sécurisée à l'aide d'un login/mot de passe, en plus de la CPS. Les seules données médicales des bases GINSENG accessibles à un utilisateur sont celles qui sont strictement nécessaires à l'étude considérée. Cette configuration propre à chaque étude est réalisée par les administrateurs de la plateforme GINSENG du CNRS. Une fois les données transmises à l'utilisateur, elles seront supprimées du serveur. Une traçabilité des accès est mise en œuvre. La CNIL préconise la conservation des traces d'accès pendant une durée de deux ans et un contrôle régulier de ces traces. Les mesures de sécurité décrites par le responsable de traitement sont conformes à l'exigence de sécurité prévue par l'article 34 de la loi du 6 janvier 1978 modifiée. La Commission rappelle toutefois que cette obligation nécessite la mise à jour des mesures de sécurité au regard de la réévaluation régulière des risques.
Sur les autres caractéristiques du traitement	La Commission demande que lui soit communiqué à l'appui de toute demande de généralisation, un bilan portant sur la faisabilité et l'acceptabilité du projet qui sera établi à l'issue de l'expérimentation, prévue le 31 décembre 2014. Elle rappelle que les recherches épidémiologiques ou les études relatives à l'évaluation des pratiques de soins devront, s'il y a lieu, être accomplies conformément aux dispositions des chapitres IX et X de la loi du 6 janvier 1978 modifiée.

Autorise, conformément à la présente délibération, le Centre national de la recherche scientifique (CNRS) à mettre en œuvre, à titre expérimental jusqu'au 31 décembre 2014, le traitement susmentionné.

La Présidente : Isabelle FALQUE-PIERROTIN

Nature de la délibération : Autorisation

Bibliographie

Bibliographie

ADICAP. 2009. “Thesaurus de La Codification ADICAP Index Raisonné Des Lésions V5.” http://medphar.univ-poitiers.fr/registre-cancers-poitou-charentes/documents_registre/adicap_version5_4_1_2009.pdf.

AMA, American Medical Association. 2014. “The Differences between ICD-9 and ICD-10” fact sheet: 4. https://www.unitypoint.org/waterloo/filesimages/for providers/icd9-icd10-differences.pdf.

Amendolia, S Roberto, Michael Brady, Richard McClatchey, Miguel Mulet-Parada, Mohammed Odeh, and Tony Solomonides. 2003. “MammoGrid: Large-Scale Distributed Mammogram Analysis.” *Studies In Health Technology And Informatics* 95: 194–199.

Artmann, Jorg, S Giest, and Jos Dumortier. 2010. “Country Brief: France.” http://ehealth-strategies.eu/database/documents/FRANCE_countrybrief_ehstrategies.pdf.

ASIP-Santé. 2009. “Algorithme de Calcul de l’INS-C.” Paris: ASIP-Santé. http://www.i-med.fr/IMG/pdf/Dossier_de_conception_INS-C_-_Algorithme_de_calcul_v0.0.1.pdf.

———. 2013. “Guide de Mise En Œuvre D’une Authentification Forte Avec Une Carte de Professionnel de Santé (CPS) Dans Une Application Web.” http://integrateurs-cps.asipsante.fr/documents/ASIP-PTS_Guide-de-mise-en-oeuvre-d-une-authentification-forte-avec-une-carte-CPS_20131217_v0.2.4.pdf.

———. 2014a. “Interfaces D’accès Au Système de Messagerie Sécurisées de Santé (MSSanté) - Dossier Des Spécification Fonctionnelles et Techniques - V1.0.0.” http://esante.gouv.fr/sites/default/files/MSS_FON_DSFT_Operateurs_MSSante_v1.0.0.pdf.

———. 2014b. “Dossier de Spéfication Techniques v0.9.5.” Vol. 79. doi:10.1002/cplu.201490022. http://basedaj.aphp.fr/daj/public/file/openfile/id_fiche/11857/id/2722.

Baude, Catherine / Ministère de la jeunesse et des sports. 2007. “Classification Commune Des Actes Médicaux.” Ministère de la santé, de la jeunesse et des sports. http://www.sante.gouv.fr/IMG/pdf/bo0703.pdf.

Beck, F, S Legleye, O Le Nézet, and S Splika. 2005. “Atlas Régional Des Consommations D’alcool 2005 Données INPES/OFDT.” http://www.injep.fr/IMG/pdf/Atlas_alcool_carto_INPES.pdf.

Bhardwaj, Sushil, Leena Jain, and Sandeep Jain. 2010. “Cloud Computing : A Study of Infrastructure As a Service (Iaas).” *International Journal of Engineering* 2 (1): 60–63. http://ijeit.org/index_files/vol2no1/CLOUD COMPUTING A STUDY OF.pdf.

Bisson, Quentin. 2012. “Identification Accélérée Par GP-GPU:” 79.

Bourquard, K. 2007. “Dossier Médical Partagé Ou Personnel: Situation Internationale.” *Pratiques et Organisation Des Soins* 38 (1): 55–67.

Bourret, Christian. 2010. “Electronic Health Record (Dossier Médical Personnel) as a Major Tool to Improve Healthcare in France: An Approach through the Situational Semiotic.” *Networked Digital Technologies*. doi:10.1007/978-3-642-14306-9. http://dx.doi.org/10.1007/978-3-642-14306-9_2.

Breton, Vincent, Kevin Dean, Tony Solomonides, I Blanquer, V Hernandez, E Medico, N Maglaveras, et al. 2005. “The Healthgrid White Paper.” *Studies In Health Technology And Informatics* 112: 249–321. http://www.ncbi.nlm.nih.gov/pubmed/15923733.

“British Medical Journal.” 1986. *British Medical Journal* 293: 659–64.

Cipière, Sébastien. 2016. “Étude Épidémiologique Sur L’impact Du Radon Sur Les Cancers Du Poumon En Auvergne.” Aubière-France.

Cipière, Sébastien, P De Vlieger, D Sarramia, D R C Hill, and L Maigne. 2012. “Development of a Metamodel for Medical Database Management on a Grid Network: Application to Health Watch and Epidemiology for Cancer and Perinatal Health.” In *Cluster, Cloud and Grid Computing (CCGrid), 2012 12th IEEE/ACM International Symposium on*, 892–897. Ottawa, ON: IEEE Computer Society. doi:10.1109/CCGrid.2012.86.

Cipière, Sébastien, Sébastien Gaspard, David Manset, Jérôme Revillard, David Sarramia, Vincent Breton, David Hill, and Lydia Maigne. 2012. “GINSENG (Global Initiative for Sentinel E-Health Network on Grid).” In *Journées Scientifiques Mésocentres et France Grilles 2012*.

Clavier, Carole. 2007. “Le Politique et La Santé Publique: Une Comparaison Transnationale de La Territorialisation Des Politiques de La Santé Publique (France, Danemark).” Rennes 1. http://www.theses.fr/2007REN1G004.

Cohen, W, P Ravikumar, and S Fienberg. 2003. “A Comparison of String Metrics for Matching Names and Records.” In *Communications*, 3:73–78. Citeseer. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.5.9007&rep=rep1&type=pdf.

Cole, Elodia, Etta D Pisano, Mary Brown, Cherie Kuzmiak, M Patricia Braeuning, Hak Hee Kim, Roberta Jong, and Ruth Walsh. 2004. “Diagnostic Accuracy of Fischer Senoscan Digital Mammography versus Screen-Film Mammography in a Diagnostic Mammography Population.” *Academic Radiology*. United States. doi:10.1016/j.acra.2004.04.003.

Cornet, Ronald, and Nicolette de Keizer. 2008. “Forty Years of SNOMED: A Literature Review.” *BMC Medical Informatics and Decision Making* 8 Suppl 1: S2.

Coughlin, Steven S. 2006. “Ethical Issues in Epidemiologic Research and Public Health Practice.” *Emerging Themes in Epidemiology* 3: 16.

cour des comptes. 2009. “Rapports Public Annuel.” http://www.lefigaro.fr/assets/pdf/cour-des-comptes-090204.pdf.

Couvreur, Christophe. 2010. “Projet de Dossier Médical Personnel et Cadre National D’interopérabilité.” *Revue Hospitalière de France*. http://fulltext.bdsp.ehesp.fr/FHF/RHF/2010/536/60_62.pdf.

Cusumano, Michael. 2010. “Cloud Computing and SaaS as New Computing Platforms.” *Communications of the ACM* 53 (4): 27. doi:10.1145/1721654.1721667. http://ebusiness.mit.edu/research/papers/2010.04_Cusumano_Technology Strategy and Management_273.pdf.

Danish-eHealth-Authority. 2013. “Making eHealth Work NATIONAL STRATEGY FOR DIGITALISATION OF THE DANISH HEALTHCARE SECTOR 2013-2017.” Copenhagen. http://www.ssi.dk/~/media/Indhold/DK - dansk/Sundhedsdata og it/NationalSundhedsIt/Om NSI/Strategy2013-17.ashx.

Danish-ministry-of-Health. 2012. “eHealth in Denmark.” Copenhagen. doi:978-87-7601-332-5. http://www.sum.dk/~/media/Filer - Publikationer_i_pdf/2012/Sundheds-IT/Sundheds_IT_juni_web.ashx.

Darby, Sarah, David Hill, Ansi Auvinen, Juan-Miguel Barros-Dios, Hélène Baysson, Francesco Bochicchio, Harz Deo, et al. 2005. “Exposition Au Radon Dans Les Habitations et Risque de Cancer Du Poumon : Analyse Conjointe Des Données Individuelles de 13 Études Cas-Témoins Européennes.” *InVS* BMJ (7485): 6. http://opac.invs.sante.fr/doc_num.php?explnum_id=1722.

De Vlieger, Paul, Jean-Yves Boire, Vincent Breton, Yannick Legre, David Manset, Jérôme Revillard, David Sarramia, and Lydia Maigne. 2010. “Sentinel E-Health Network on Grid: Developments and Challenges.” *Studies In Health Technology And Informatics* 159: 134–145.

De Vlieger, Paul, Jean-Yves Boire, Vincent Breton, Yannick Legré, David Manset, Jérôme Revillard, David Sarramia, and Lydia Maigne. 2009. “Grid-Enabled Sentinel Network for Cancer Surveillance.” *Studies In Health Technology And Informatics* 147: 289–294.

Debin, Marion, Clément Turbelin, Thierry Blanchon, Isabelle Bonmarin, Alessandra Falchi, Thomas Hanslik, Daniel Levy-Bruhl, Chiara Poletto, and Vittoria Colizza. 2013. “Evaluating the Feasibility and Participants’ Representativeness of an Online Nationwide Surveillance System for Influenza in France.” *PloS One* 8 (9): e73675. doi:10.1371/journal.pone.0073675. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3770705&tool=pmcentrez&rendertype=abstract.

DeVlieger, Paul. 2011. “Création D’un Environnement de Gestion de Base de Données ‘en Grille’. Application À L ’ Échange de Données Médicales.” UBP. http://tel.archives-ouvertes.fr/docs/00/71/96/88/PDF/2011CLF1MM11.pdf.

Dolin, Robert H., Liora Alschuler, Calvin Beebe, Paul V. Biron, Sandra Lee Boyer, Daniel Essin, Elliot Kimber, Tom Lincoln, and John E. Mattison. 2001. “The HL7 Clinical Document Architecture.” *Journal of the American Medical Informatics Association* 8 (6): 552–569.

Dolin, Robert H., Liora Alschuler, Sandy Boyer, Calvin Beebe, Fred M. Behlen, Paul V. Biron, and Amnon Shabo. 2006. “HL7 Clinical Document Architecture, Release 2.” *Journal of the American Medical Informatics Association* 13 (1): 30–39.

Doussin, Anne (InVS - DMCT), and Caroline (INSERM – InstitutSanté Publique) RAULT. 2008. “Registres Épidémiologiques et Accès Aux Sources de Données Standardisées : État Des Lieux et Perspectives D’amélioration.” COMITE NATIONAL DES REGISTRES. https://www.google.fr/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=0CCIQFjAA&url=http://www.sante.gouv.fr/IMG/pdf/Contribution_du_Comite_national_des_registres_InVs-Inserm_-propositions.pdf&ei=Y8tdVda8OcG_ygOdhIH4Ag&usg=AFQjCNGu.

Dufrenne, Julien. 2011. “DEMATERIALISATION DES ECHANGES D’INFORMATIONS ENTRE MEDECINS : La Messagerie Sécurisée de Santé Utilisée Par Les Médecins Généralistes.” UNIVERSITE DES ANTILLES ET DE LA GUYANE. http://www.sudoc.fr/153546964.

Dumont, Véronique. 2010. “Controverses Autour de L’échange Électronique de Données de Santé : La Question de L’identifiant Du Patient .” In *Actes Du 15ème Colloque CREIS-Terminal. Les Libertés À L’épreuve de L'informatique. Paris. Juin 2010*. http://www.lecreis.org/colloques creis/2010/IS2010_actes.htm.

Eijo, Juan Francisco Garcia, Marcelo Risk, Francisco Prieto Castrillo, Cesar Suarez Ortega, Maria Boton Fernandez, Alfonso Pardo Diaz, Manuel Rubio del Solar, and Raul Ramos Pollan. 2011. “CardioGRID: A Framework for the Analysis of Cardiological Signals in GRID Computing.” *Journal of Physics: Conference Series* 313 (1): 12010.

European-Commission. 2010a. “Country Brief: Denmark.” http://ehealth-strategies.eu/database/documents/denmark_countrybrief_ehstrategies.pdf.

———. 2010b. “Country Brief: Belgium.” http://ehealth-strategies.eu/database/documents/Belgium_CountryBrief_eHStrategies.pdf.

Fenstermacher, David, Craig Street, Tara McSherry, Vishal Nayak, Casey Overby, and Michael Feldman. 2005. “The Cancer Biomedical Informatics Grid (caBIGTM).” *Conference Proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference* 1: 743–746. doi:10.1109/IEMBS.2005.1616521.

France-Périnat. 2007. “La Santé Périnatale En 2004-2005 - Évaluation Des Pratiques Médicales.” http://www.audipog.net/pdf/cahier_2004_2005.pdf.

Franke, Florian, and Philippe Pirard. 2006. “Le Radon En Corse : Évaluation de L’exposition et Des Risques Associés.” *InVS*: 48. http://opac.invs.sante.fr/doc_num.php?explnum_id=5004.

FRÉDÉRIC BRENON. 2013. “PLUS DE CANCERS EN LOIRE-ATLANTIQUE.” *20 Minutes*. http://www.santepaysdelaloire.com/sites/registre-cancers/files/documents/EPIC/PDF/20_minutes_16-01-2013__p.3_.pdf.

Gagliardi, Fabrizio. 2005. “The EGEE European Grid Infrastructure Project.” In *High Performance Computing for Computational Science - VECPAR 2004 SE - 16*, edited by Michel Daydé, Jack Dongarra, Vicente Hernández, and JoséM.L.M. Palma, 3402:194–203. Springer Berlin Heidelberg. doi:10.1007/11403937_16.

Gagliardi, Fabrizio, Bob Jones, Mario Reale, and Stephen Burke. 2002. “European DataGrid Project : Experiences of Deploying a Large Scale Testbed for E-Science Applications.” *Performance Evaluation of Complex Systems Techniques and Tools*: 480–500.

Gjermundrøda, H, M D Dikaiakos, D Zeinalipour-Yazti, G Panayi, and T Kyprianou. 2007. “From Genes to Personalized HealthCare: Grid Solutions for the Life Sciences:ICGrid: Enabling Intensive Care Medical Research on the EGEE Grid.” In *Studies in Health Technology and Informatics*, edited by N Jacq, Y Legré, H Muller, I Blanquer, V Breton, D Hausser, V Hernández, T Solomonides, and M Hofman-Apitius, Volume 126. IOS Press. doi:978-1-58603-738-3.

Gnaegi, Alex, and Cédric Michelet. 2011. “Infomed, Un Projet D’échange Électronique de Données Médicales En Valais.” *Pipette* (4): 10–12.

Gnaegi, Alex, Philippe Wieser, and Georges Dupuis. 2010. “La Stratégie eHealth En Valais.” *Bulletin Des Médecins Suisses |Bollettino Dei Medici Svizzeri* 91 (33): 1247–1250.

HAS. 2006. “Place de La Mammographie Numérique Dans Le Dépistage Organisé Du Cancer Du Sein.” http://www.has-sante.fr/portail/jcms/c_461657/fr/place-de-la-mammographie-numerique-dans-le-depistage-organise-du-cancer-du-sein.

Ihtsdo. 2015. “SNOMED CT Compositional Grammar Specification and Guide.” http://ihtsdo.org/fileadmin/user_upload/doc/download/doc_CompositionalGrammarSpecificationAndGuide_Current-en-US_INT_20150522.pdf?ok.

interop’santé. 2015. “Guide D’interopérabilité Intra-Hospitalier.” http://www.interopsante.org/offres/doc_inline_src/412/WEB-GuideSante2015.pdf.

InVS. 2013. “Dépistage Organisé Du Cancer Colorectal : Guide Du Format Des Données et Définitions Des Indicateurs de L’évaluation Du Programme National.”

Jaro, M A. 1995. “Probabilistic Linkage of Large Public Health Data Files.” *Statistics in Medicine* 14 (5-7): 491–498. doi:10.1002/sim.4780140510. http://doi.wiley.com/10.1002/sim.4780140510.

Jaro, Matthew A. 1989. “Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa , Florida.” *Methodology* 84 (406): 414–420.

Jezewski Serra, D., and E. Salines. 2013. “Évaluation Épidémiologique Du Programme de Dépistage Organisé Du Cancer Colorectal En France.” doi:978-2-11-138318-0.

Kasam, Vinod, Jean Salzemann, Marli Botha, Ana Dacosta, Gianluca Degliesposti, Raul Isea, Doman Kim, et al. 2009. “WISDOM-II: Screening against Multiple Targets Implicated in Malaria Using Computational Grid Infrastructures.” *Malaria Journal* 8 (1): 88. doi:10.1186/1475-2875-8-88.

Katriel, Guy. 2010. “Epidemics with Partial Immunity to Reinfection.” *Mathematical Biosciences* 228 (2): 153–159.

Lehalle, Dominique, and Michèle Sérézat. 2014. “Livre Blanc.” http://www.wobook.com/WBXb7Pf4BI4I/AFHADS_LivreBlanc.html.

Lemoine, Vanessa (INPES). 2013. “Alcool, Tabac et Drogues Illicites : Géographie Des Pratiques Addictives En France.” http://www.inpes.sante.fr/70000/dp/13/dp131107.pdf.

Li, Xinran. 2015. “Évaluation et Amélioration Des Méthodes de Chaînage de Données.” UDA.

Lloyd, S, M Jirotka, A C Simpson, R P Highnam, D J Gavaghan, D Watson, and J M Brady. 2005. “Digital Mammography: A World without Film?” *Methods of Information in Medicine* 44 (2): 168–171.

Manaouil, C. 2009. “Le Dossier Médical Personnel (DMP) : « Autopsie » D’un Projet Ambitieux ?” *Médecine & Droit* 2009 (94): 24–41. doi:10.1016/j.meddro.2009.01.002. http://www.sciencedirect.com/science/article/pii/S1246739109000037.

Marquet, Richard L, Aad IM Bartelds, Sander P Van Noort, Carl E Koppeschaar, John Paget, François G Schellevis, and Jouke Van Der Zee. 2006. “Internet-Based Monitoring of Influenza-like Illness (ILI) in the General Population of the Netherlands during the 2003–2004 Influenza Season.” *BMC Public Health* 6 (Ili): 242. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1609118&tool=pmcentrez&rendertype=abstract.

Mell, Peter, and Timothy Grance. 2011. “The NIST Definition of Cloud Computing.” *NIST Special Publication* 145: 7. http://www.mendeley.com/research/the-nist-definition-about-cloud-computing/.

Merabti, Tayeb, Hocine Abdoune, Thierry Lecroq, Michel Joubert, and Stéfan J. Darmoni. 2009. “Projection Des Relations SNOMED CT Entre Les Termes de Deux Terminologies (CIM10 et SNOMED 3.5).” In *Journées Francophones d’Imagerie Médicale*. Vol. 10.

Molloy, Jennifer C. 2011. “The Open Knowledge Foundation: Open Data Means Better Science.” *PLoS Biology* 9 (12): 1–4. doi:10.1371/journal.pbio.1001195. http://www.plosbiology.org/article/fetchObject.action?uri=info:doi/10.1371/journal.pbio.1001195&representation=PDF.

Nielsen, Claus F, Jens Branebjerg, Casper D Marcussen, Mette A Craggs, Lars Hulbaek, Claus Duedal Pedersen, Editors Fabienne Abadie, et al. 2013. “Strategic Intelligence Monitor on Personal Health Systems , Phase 2 Country Study : Denmark.” doi:10.2791/86918.

ONU. 1946. “Constitution de L’organisation Mondiale de La Santé.” http://www.who.int/governance/eb/who_constitution_fr.pdf.

Paillard, Jean-Christophe, Candice Roudier, Lydéric Aubert, and Blandine Vacquier. 2014. “Les Niveaux de Radon et Leurs Déterminants Dans Les Logements de France Métropolitaine Continentale.” *InVS*. http://opac.invs.sante.fr/doc_num.php?explnum_id=9438.

Paolotti, Daniela, Corrado Gioannini, and Vittoria Colizza. 2010. “Internet-Based Monitoring System for Influenza-like Illness : H1N1 Surveillance in Italy.” In *3rd International ICST Conference on Electronic Healthcare for the 21st Century*, 1–4.

Parshani, Roni, Shai Carmi, and Shlomo Havlin. 2010. “Epidemic Threshold for the Susceptible-Infectious-Susceptible Model on Random Networks.” *Physical Review Letters* 104 (25): 258701. doi:10.1103/PhysRevLett.104.258701.

Passerat-Palmbach, Jonathan. 2009. “Mise En Place D ’ Un Environnement de Sécurité Autour de La Carte de Professionnel de Santé Dans Une Infrastructure de Grille de Données Présenté Par : Remerciements.” Aubière-France.

Pisano, Etta D, Constantine Gatsonis, Edward Hendrick, Martin Yaffe, Janet K Baum, Suddhasatta Acharyya, Emily F Conant, et al. 2005. “Diagnostic Performance of Digital versus Film Mammography for Breast-Cancer Screening.” *The New England Journal of Medicine* 353 (17) (October): 1773–1783. doi:10.1056/NEJMoa052911.

Piwowar, H A, and T J Vision. 2013. “Data Reuse and the Open Data Citation Advantage.” *PeerJ* 1: e175. doi:10.7717/peerj.175. http://www.ncbi.nlm.nih.gov/pubmed/24109559.

Plan cancer. 2013. “Rapport Final Au Président de La République.” Paris. https://www.google.fr/url?sa=t&rct=j&q=&esrc=s&source=web&cd=2&ved=0CC4QFjAB&url=http://www.e-cancer.fr/component/docman/doc_download/10625-rapport-final-du-plan-cancer-2009-2013&ei=Uj5BVP6SHMToaJ7AgOAO&usg=AFQjCNFFCRdlfMkVn1mFCUD9DJt4oUwzFA&sig2=ZaAcPMJr.

Poletto, Chiara, Sandro Meloni, Vittoria Colizza, Yamir Moreno, and Alessandro Vespignani. 2013. “Host Mobility Drives Pathogen Competition in Spatially Structured Populations.” *PLoS Computational Biology* 9 (8).

Quantin, C, C Binquet, K Bourquard, F Allaert, B Gouyon, C Ferdynus, R Pattisina, G Harmenil, S Pequignot, and J Gouyon. 2004. “Estimation de La Valeur Discriminante Des Traits D’identification Utilisés Pour Le Rapprochement Des Données D’un Patient.” *Revue d’Épidémiologie et de Santé Publique* 52 (5) (October): 431–440. doi:10.1016/S0398-7620(04)99079-7. http://linkinghub.elsevier.com/retrieve/pii/S0398762004990797.

Quantin, Catherine, Gouenou Coatrieux, François André Allaert, Maniane Fassa, Karima Bourquard, Jean-Yves Boire, Paul De Vlieger, Lydia Maigne, and Vincent Breton. 2009. “New Advanced Technologies to Provide Decentralised and Secure Access to Medical Records: Case Studies in Oncology.” *Cancer Informatics* 7: 217–229. http://www.ncbi.nlm.nih.gov/pubmed/19718446.

Riviere, Jean-Philippe;(CNOM/VIDAL). 2013. “2ème Baromètre Sur Les Médecins Ayant Un Smartphone : L’utilisation En Consultation Se Banalise.” http://www.vidal.fr/actualites/13131/2eme_barometre_sur_les_medecins_ayant_un_smartphone_l_utilisation_en_consultation_se_banalise/.

RSPA. 2001. *CHARTE CONSTITUTIVE DU RESEAU DE SOINS PERINATALS D ’ AUVERGNE*. rspa.

———. 2004. *Avenants À La Charte Du Réseau de Santé Périnatale D ’ Auvergne Formant Convention Constitutive En Date Du 30 Mars 2004*. France: RSPA.

Sauleau, Erik A, Jean-Philippe Paumier, and Antoine Buemi. 2005. “Medical Record Linkage in Health Information Systems by Approximate String Matching and Clustering.” *BMC Medical Informatics and Decision Making* 5 (1): 32. http://www.ncbi.nlm.nih.gov/pubmed/16219102.

Skaane, Per, and Arnulf Skjennald. 2004. “Screen-Film Mammography versus Full-Field Digital Mammography with Soft-Copy Reading: Randomized Trial in a Population-Based Screening Program--the Oslo II Study.” *Radiology* 232 (1) (July): 197–204. doi:10.1148/radiol.2321031624.

Tilston, Natasha L, Ken T D Eames, Daniela Paolotti, Toby Ealden, and W John Edmunds. 2010. “Internet-Based Surveillance of Influenza-like-Illness in the UK during the 2009 H1N1 Influenza Pandemic.” *BMC Public Health* 10 (1): 650. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2988734&tool=pmcentrez&rendertype=abstract.

Tizzoni, Michele, Paolo Bajardi, Chiara Poletto, José J Ramasco, Duygu Balcan, Bruno Gonçalves, Nicola Perra, Vittoria Colizza, and Alessandro Vespignani. 2012. “Real-Time Numerical Forecast of Global Epidemic Spreading: Case Study of 2009 A/H1N1pdm.” *BMC Medicine* 10 (1): 165. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3585792&tool=pmcentrez&rendertype=abstract.

Tourlière, B., D. Rouzaire, and C. Bertin. 2007. “Cartographie Du Potentiel D’émanation Du Radon En Auvergne.” *BRGM* 07POLE03: 61.

VAN DER AALST, Wil, and Kees Max VAN HEE. 2008. *Workflow Management: Models, Methods, and Systems*. Edited by MIT Press. MIT Press. MIT Press. http://wwwis.win.tue.nl/~wvdaalst/publications/p120.pdf.

Vendittelli, Françoise, Catherine Crenn-Hébert, and Véronique Tessier. 2007. “Abécédaire de L’évaluation Des Pratiques Professionnelles.” http://www.audipog.net/pdf/epp_intro.pdf.

Vulliet-Tavernier, Sophie. 2002. “La CNIL et La E-Santé.” *Médecine & Droit* 2002 (52): 3–4. doi:http://dx.doi.org/10.1016/S1246-7391(02)83002-0. http://www.sciencedirect.com/science/article/pii/S1246739102830020.

Walker, Jan, Eric Pan, Douglas Johnston, Julia Adler-Milstein, David W. Bates, and Blackford Middleton. 2005. “The Value of Health Care Information Exchange and Interoperability.” *Health Affairs (Project Hope)* Suppl Web: 10–18. doi:10.1377/hlthaff.w5.10. http://www.providersedge.com/ehdocs/ehr_articles/The_Value_Of_Health_Care_Information_Exchange_And_Interoperability.pdf.

Warden, R. 2011. “Impact of caBIG on the European Cancer Community.” *Ecancermedicalscience* 5: 1–7. doi:10.3332/ecancer.2011.225. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3223955&tool=pmcentrez&rendertype=abstract.

Warren, R, A E Solomonides, C Del Frate, I Warsi, J Ding, M Odeh, R McClatchey, et al. 2007. “MammoGrid--a Prototype Distributed Mammographic Database for Europe.” *Clinical Radiology* 62 (11): 1044–1051.

WHO, World Health Organization. 1992. “ICD-10: International Statistical Classification of Diseases and Related Health Problems: 10th Revision.”

Winkler, W E. 2007. “Automatically Estimating Record Linkage False Match Rates.” *Methods* Statistics (#2007-05): 5. http://www.amstat.org/sections/srms/proceedings/y2006/Files/JSM2006-000318.pdf.

Winkler, WE. 1990. “String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage.” *Proceedings of the ASA Section on Survey Research Methods*: 1184–1187. http://www.eric.ed.gov/ERICWebPortal/recordDetail?accno=ED325505.

Winkler, William E. 1999. “The State of Record Linkage and Current Research Problems.” *Statistical Research Division US Census Bureau*: 1–15. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.39.4336.

———. 2005. “Overview of Record Linkage and Current Research Directions.” *U.S. Bureau of the Census* (2005nov15). Research Report Series: 6. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.79.1519.

World Health Organization. 2011. “Global Observatory for eHealth: Atlas: eHealth Country Profiles.” *Observatory*. Vol. 1. World Health Organization. http://whqlibdoc.who.int/publications/2011/9789241564168_eng.pdf.

140

[140] Bibliographie réalisée à l’aide de https://www.mendeley.com/ et du composant intégré dans MS Word.

“

Ce sera quelque chose d'admirable, s'il fait d'aussi belles cures qu'il fait de beaux discours.

- Jean-Baptiste Poquelin, *le malade imaginaire*

Publications & Conférences

Sébastien CIPIÈRE — Publications & Conférences

2015

P. Schweitzer, S. Cipière, A. Dufaure, H. Payno, Y. Perrot, D. R. C. Hill, L. Maigne
"Performance Evaluation of Multithreaded Geant4 Simulations Using an Intel Xeon Phi Cluster,"
Scientific Programming, vol. 2015, Article ID 980752, 10 pages, 2015. doi:10.1155/2015/980752.

2014

X Li, A Guttmann, S Cipière, J Demongeot, JY Boire, L Ouchchane
Utilisation de l'algorithme EM pour estimer les paramètres du chaînage probabiliste d'enregistrements
Revue d'Épidémiologie et de Santé Publique 62, S196

X Li, A Guttmann, S Cipiere, L Maigne, J Demongeot, JY Boire, L Ouchchane
Implementation of an extended Fellegi-Sunter probabilistic record linkage method
using the Jaro-Winkler string comparator
Biomedical and Health Informatics (BHI), 2014 IEEE-EMBS International Conference

S Cipiere, G Ereteo, A Gaignard, N Boujelben, S Gaspard, D Manset, V Breton, DRC Hill, T Glatard, F Cervenansky, J Montagnat, J Revillard, L Maigne
Global Initiative for Sentinel e-Health Network on Grid (GINSENG): Medical Data Integration and Semantic Developments for Epidemiology
Cluster, Cloud and Grid Computing (CCGrid), 2014 14th IEEE/ACM International Conference

X Li, A Guttman, S Cipière, L Maigne, JY Boire, L Ouchchane
Comparaison de performance des algorithmes de rapprochement de patients
Revue d'Épidémiologie et de Santé Publique 62, S76-S77

F Jaziri, E Peyretaillade, M Missaoui, N Parisot, S Cipière, J Denonfoux, A Mahul, P Peyret, DRC Hill
Large Scale Explorative Oligonucleotide Probe Selection for Thousands of Genetic Groups on a Computing Grid: Application to Phylogenetic Probe Design Using a Curated Small Subunit Ribosomal RNA Gene Database
The Scientific World Journal 2014

2013

X. Li, A. Guttmann, S. Cipière, L. Maigne, J-Y. Boire, L. Ouchchane
Évaluation des algorithmes de rapprochement de patients par traits d'identification nominatifs,
Revue d'Épidémiologie et de Santé Publique, 61, S322, Elsevier Masson, 31/10/2013

2012

S. Cipière, P. De Vlieger, D. Sarramia, D. R. C. Hill, L. Maigne,
"Development of a metamodel for medical database management on a grid network",
2012 12th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (ccgrid 2012),
ISBN: 978-0-7695-4691-9, Ottawa, Canada.

S. Cipière, P. De Vlieger, D. Sarramia, D. R. C. Hill, L. Maigne,
GINSENG (Global Initiative for Sentinel E-health Network on Grid),
EGI community Forum 2012, Munich, March 2012.

S. Cipière, P. De Vlieger, S. Gaspard, D. Manset, J. Revillard, D. Sarramia, D. R.C. Hill, L. Maigne,
GINSENG, une infrastructure de grille au service de l'e-santé et de l'épidémiologie,
conférence GISEH (Gestion et Ingénierie des SystèmeEs Hospitaliers) 2012, Québec, Canada, septembre 2012.

2011

F. Jaziri, M. Missaoui , S. Cipière, P. Peyret , D. R.C. Hill ,
Large Scale Parallelization Method of 16S rRNA Probe Design Algorithm on Distributed Architecture: Application to Grid Computing,
IEEE ICI2011, International Conference on Informatics and Computational Intelligence,
Bandung, Indonesia, 12-14 December 2011

F. Jaziri, S. Cipière, M. Missaoui, J. Denonfoux, E. Dugat-Bony, N. Parisot, A. Mahul, S. Rimour, E. Peyretaillade, P. Peyret, D. R.C. Hill,
Détermination de sondes oligonucléotidiques pour biopuces phylogénétiques en environnement grille de calcul.,
Rencontres Scientifiques France Grilles, Lyon 2011, 4 p.

NOTES

NOTES