

Soutenance de thèse

« Création d'un environnement de gestion de bases de données 'en grille', application à l'échange de données médicales »

Paul De Vlieger
12 Juillet 2011

Sous la direction de:

Jean-Yves Boire – PU-PH
Vincent Breton – DR2 CNRS
Lydia Maigne – MCF



ISIT



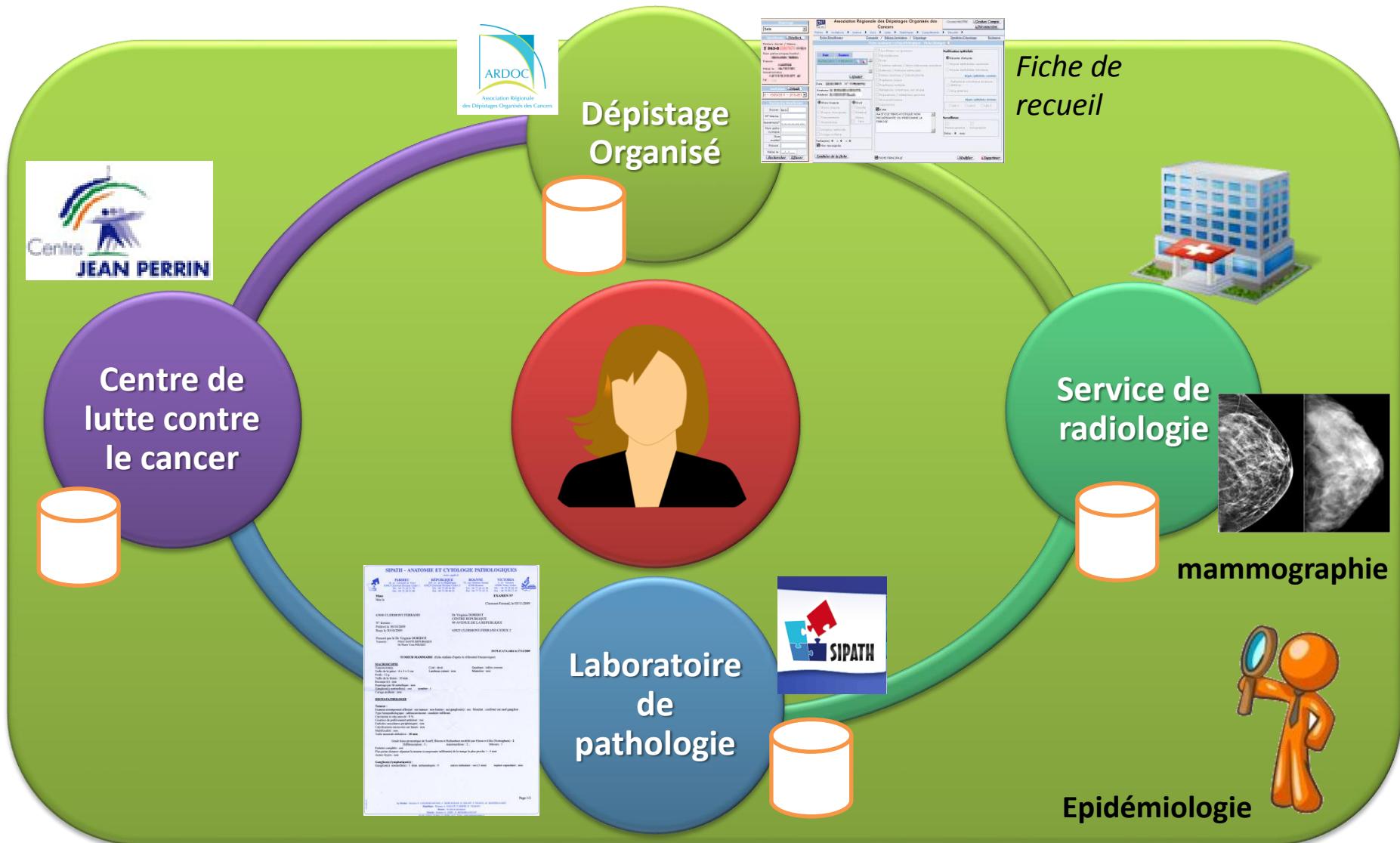


INTRODUCTION

Enjeux de l'échange des données de santé



Parcours médical dans le dépistage organisé du cancer





Objectifs à atteindre



- Créer une infrastructure décentralisée d'échange de données de santé

- Besoins de la communauté médicale

Qualité

- Pour le dépistage organisé des cancers:

- Récupérer les comptes rendus médicaux de façon informatisée

Fiabilité

- Pour les laboratoires d'anatomie pathologique:

- Mettre à disposition les données tout en gardant leur contrôle

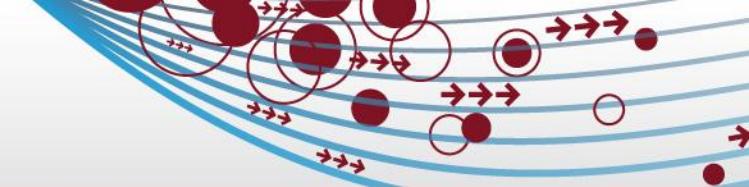
Exhaustivité

- Pour les épidémiologistes:

- Avoir la possibilité d'accéder à l'intégralité des données cancer sur la région

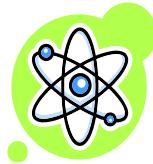
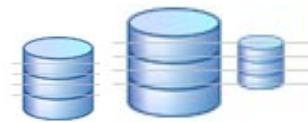


Les grilles informatiques?



● Une infrastructure distribuée

→ Mutualisation des ressources



- Fédération des moyens
- Sécurité
- Interopérabilité



● Création du Réseau Sentinel Cancer Auvergne (RSCA)

- Fondé sur une infrastructure de grille
- Pas de concentration des données
 - ➔ Les sites sont interconnectés

Qualité

● Intégration de multiples sources de données

- De façon distribuée
- Sans intervention manuelle des médecins

Exhaustivité

● Mise en place d'un système d'identification des patients

Fiabilité

- Compatible avec le caractère réparti des grilles
- Permettant de rapprocher les dossiers appartenant à un même patient sous une même identité



Plan

- Partie 1: Les enjeux d'un système distribué de gestion de bases de données pour la santé
 - Systèmes existants, architecture centralisée et ses problèmes
 - Contraintes légales
- Partie 2: Cahier des charges du projet RSCA
 - Acteurs et objectifs fonctionnels
 - Contraintes légales et techniques
- Partie 3: Mise en œuvre de RSCA
 - Architecture
 - Eléments logiciels
- Partie 4: Gestion du patient et des données médicales
 - Modèle d'identification distribué du patient
 - Rapprochement des identités
- Partie 5: Evaluation par rapport au cahier des charges
- Partie 6: Conclusion – perspectives



Partie 1:

LES ENJEUX D'UN SYSTÈME DISTRIBUÉ DE GESTION DE BASES DE DONNÉES POUR LA SANTÉ



Les systèmes existants



- Le traitement des données « cancer » est une problématique ancienne:

- Les registres de l'Institut National de la Statistique et de l'Informatique en Anatomopathologie (INSEE)
- Les registres des cancers (CRISAP) anatomopathologiques
- Le SMSC
- Les données de mortalité

- Point commun de tous ces infrastructures:

- Système déclaratif





Dispositions légales



● La loi informatique et libertés



○ Le traitement des données médicales est soumis à autorisation:

« Il est interdit de collecter ou de traiter des données à caractère personnelle [...] relatives à la santé des personnes »

→ Demande d'autorisation nécessaire à la CNIL pour RSCA

○ Eléments à assurer:

→ Que le patient ait donné son consentement *explicite*

→ Donner la possibilité au patient « *d'accéder, modifier, s'opposer* » au traitement des informations le concernant

→ Fournir des garanties de sécurité et de protection des données par le respect d'un ensemble de recommandations

○ Une demande d'autorisation comporte:

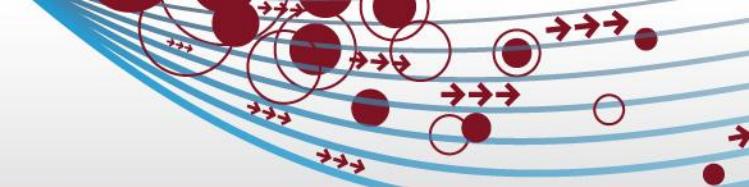
→ Une déclaration « normale »

→ Des annexes: sécurité, interconnexions, type de données échangées

○ Les structures médicales disposent de leur agrément CNIL



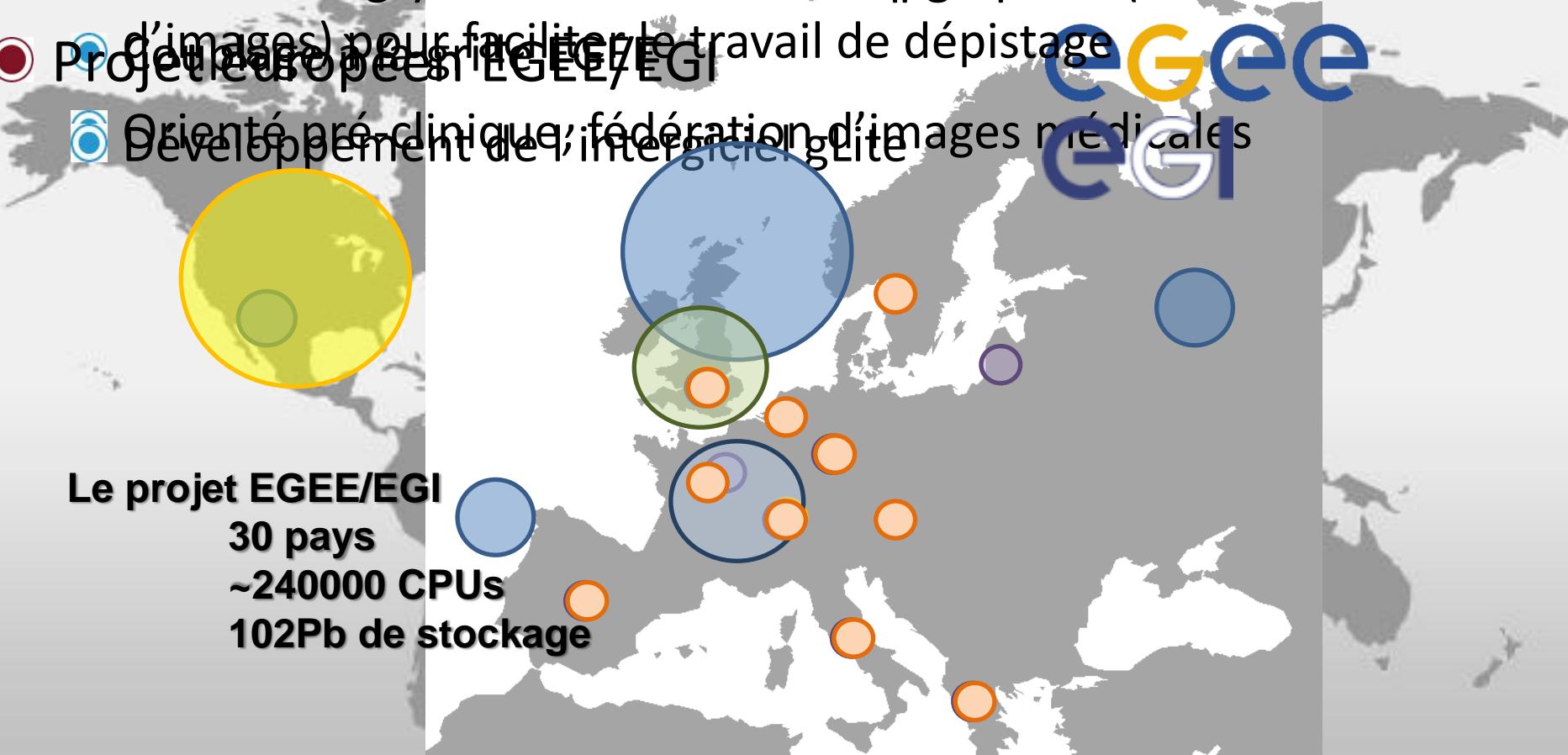
Dispositions légales



- Recommandations de l'ASIP-Santé (e-sante.gouv.fr)
 - Hébergement de données médicales
 - ➔ Soumise à autorisation pour un hébergement externe à une structure médicale
 - ➔ Agrément long à obtenir et investissement coûteux
 - Interopérabilité
 - Traçabilité des systèmes
- Directive Européenne 95/46/CE
 - Harmonisation Européenne de la loi informatique et libertés
 - ➔ Assure une cohérence au sein de l'UE
 - ➔ Création du Contrôleur Européen de la Protection des Données (CEPD)

Les projets de grilles informatiques pour la santé

- ~~Étude de cas : Grille de calcul pour le dépistage pré-clinique des maladies génétiques (2004-2010)~~
- ~~Développement d'outils pour faciliter l'interopérabilité entre les grilles (contenantité d'images) pour faciliter le travail de dépistage~~
- ~~Projets européens EGEE/EGI~~
- ~~Orienté pré-clinique; fédération d'images médicales~~





Focus sur les composants de grilles



● L'intergiciel de grille gLite

● Outils d'exploitation des ressources informatiques

- ➔ Calcul
- ➔ Stockage
- ➔ Bases de données

● Sécurité

- ➔ Infrastructure à clé publique
- ➔ Virtual Organisation Management System (VOMS)

● AMGA

- ➔ Système de gestion de bases de données
- ➔ Compatible avec la sécurité de gLite
- ➔ Permet de fédérer des sources de données distribuées

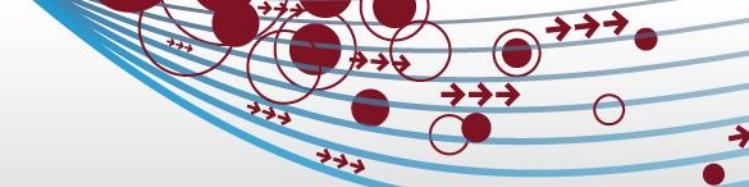


Partie 2:

CAHIER DES CHARGES DU PROJET RÉSEAU SENTINELLE CANCER AUVERGNE



Le consortium RSCA



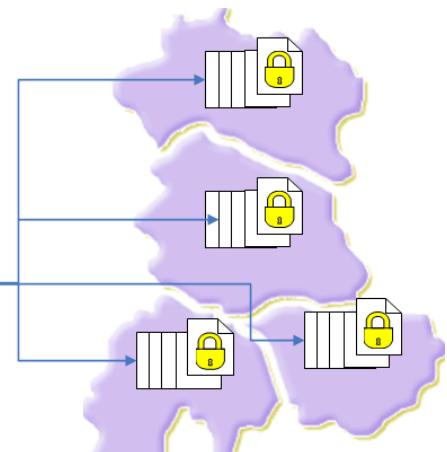
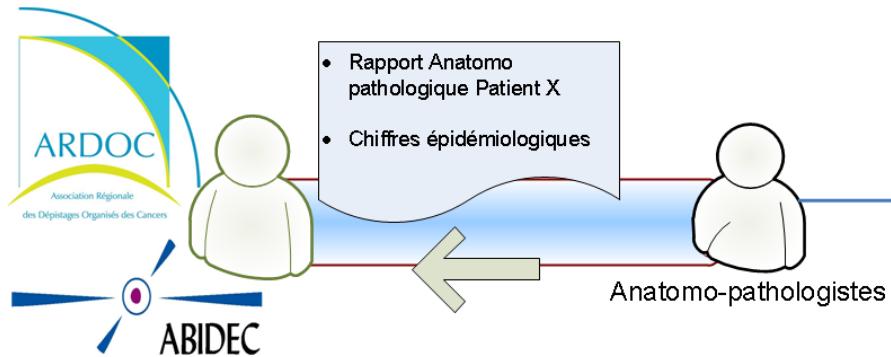
- Association RSCA
- Associations régionales de dépistage organisé des cancers:
 - ARDOC
 - ABIDEC
- Santé publique
 - CHU
- Acteurs techniques:
 - Université Blaise Pascal
 - Université d'Auvergne
 - Société maat-Gknowledge
 - Société OSI-Santé
 - CNRS



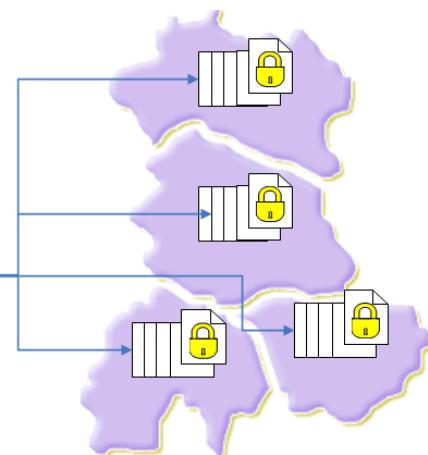
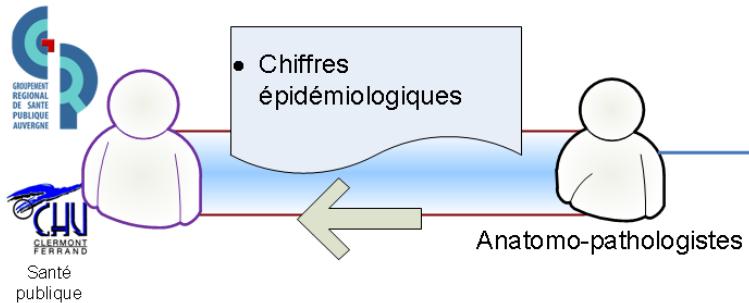


Objectifs fonctionnels

● Pour les associations:



● Pour les structures de santé publique:



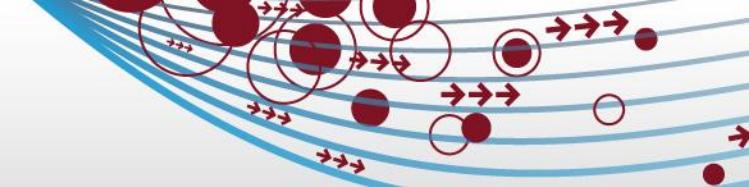


Contraintes du projet

- Laisser les données à l'endroit où elles sont produites
 - Pas d'export, pas de concentration de données
- Garantir un niveau de sécurité suffisant
 - Chiffrement des informations
 - Contrôle des utilisateurs
 - Définition des droits
 - Utilisation de la CPS pour l'authentification
 - Mémoriser tous les accès
- Respecter les contraintes légales
 - Dépôt d'une demande d'autorisation CNIL
 - Suivi des recommandations ASIP-Santé



Définition des données



- A partir des compte rendus de pathologie

- Informations personnelles

- Informations sur l'examen

→ Dates, type

- Données médicales structurées

→ Médecins

→ Codes ADICAP:

OEGSA7B2

- Données médicales non structurées

→ Compte rendu texte

SIPATH - ANATOMIE ET CYTOLOGIE PATHOLOGIQUES www.sipath.fr			
	PARDIEU 18, av. Léonard de Vinci 63063 Clermont-Ferrand Cedex 1 Tél.: 04 73 28 51 70 Fax : 04 73 28 51 80	RÉPUBLIQUE 105, av. de la République 63023 Clermont-Ferrand Cedex 2 Tél.: 04 73 99 46 00 Fax : 04 73 99 46 01	ROANNE 75, rue Général Giraud 42200 Roanne Cedex 2 Tél.: 04 77 44 41 84 Fax : 04 77 33 51
VICTORIA 2, av. Victoria 03220 VICHY Cedex 2 Tél.: 04 70 30 96 10 Fax : 04 70 98 27 42			
EXAMEN N°			
Clermont-Ferrand, le 03/11/2009			
63000 CLERMONT FERRAND			
Dr Virginie DORIDOT CENTRE REPUBLIQUE 99 AVENUE DE LA REPUBLIQUE			
63023 CLERMONT-FERRAND CEDEX 2			
Prescrit par le Dr Virginie DORIDOT Transmis : POLE SANTE REPUBLIQUE Dr Pierre Yves POUGET			
DUPLICATA édité le 27/11/2009			
TUMEUR MAMMAIRE (fiche réalisée d'après le référentiel Oncouvergne)			
MACROSCOPIE			
Tumorectomie. Taille de la pièce : 4 x 3 x 2 cm Poids : 12 g Taille de la lésion : 10 mm Recoupe (s) : non Repérage par fil métallique : non Ganglion(s) sentinel(s) : oui nombre : 1 Curage axillaire : non			
Côté : droit Lambeau cutané : non Quadrant : inféro-externe Mamelon : non			
HISTO-PATHOLOGIE			
Tumeur : Examen extemporané effectué : sur tumeur : non limites : oui ganglion(s) : oui. Résultat : confirmé oui sauf ganglion Type histopathologique : adénocarcinome canalaire infiltrant. Carcinome in situ associé : 0 % Cicatrice de prélèvement antérieur : oui Emboles vasculaires périphériques : non Calcifications retrouvées sur lames : non Multifocalité : non Taille tumorale définitive : 10 mm			
Grade histo-pronostique de Scarff, Bloom et Richardson modifié par Elston et Ellis (Nottingham) : 2 Différenciation : 3 ; Anisonucléose : 2 ; Mitoses : 1			
Exérèse complète : oui Plus petite distance séparant la tumeur (composante infiltrante) de la marge la plus proche > : 4 mm Autres foyers : non			
Ganglion(s) lymphatique(s) : Ganglion(s) sentinelle(s) : 1 dont métastatiques : 0 micro métastase : oui (1 mm) rupture capsulaire : non			
La Pardieu : Docteurs N. CAUCHOIS-GOUJON, C. DESPLECHAIN, H. EGLOFF, F. FRANCK, M. MOSNIER-DAMET République : Docteurs A. GAILLOT, F. MAURY, R. VILMANT Roanne : Secrétariat permanent Victoria : Docteur G. LESSEC, E. RICHARD-COULET			
Site web : www.sipath.fr - Site mail : sipath@wanadoo.fr			

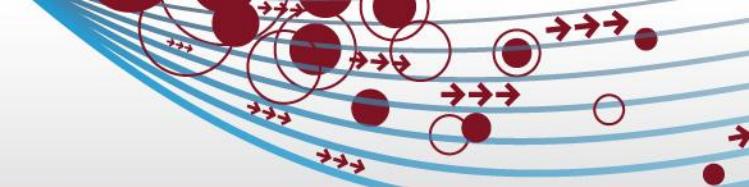


Partie 3:

MISE EN ŒUVRE DU RÉSEAU SENTINELLE CANCER AUVERGNE



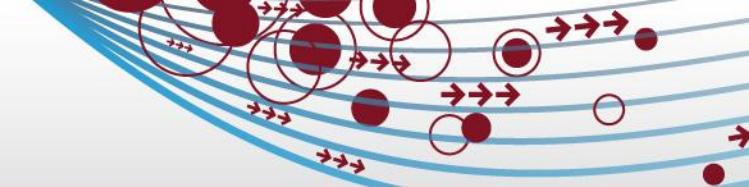
Une grille?



- Permet de laisser les données où elles se trouvent
 - Satisfait les contraintes des pathologistes
 - Offre un délai d'accès aux données très rapide
 - Contient toujours des informations à jour
- Offre le moyen de sécuriser et de contrôler les accès
 - En utilisant l'infrastructure de sécurité de gLite (GSI)
- Supporte les communautés d'utilisateurs
 - Définition d'organisations virtuelles, de groupes et de droits
 - Permet une extension naturelle à d'autres cas d'utilisation



Création de la VO Sentinelle



● Utilisation du Virtual Organisation Management System (VOMS)

- Gestionnaire d'une Organisation Virtuelle sur une grille
- Utilise les certificats X.509 pour l'authentification des utilisateurs
- Permet de définir les droits
 - ➔ Groupes
 - ➔ Cancer/dépistage
 - ➔ Cancer/épidémio
 - ➔ Rôles
 - ➔ Administration

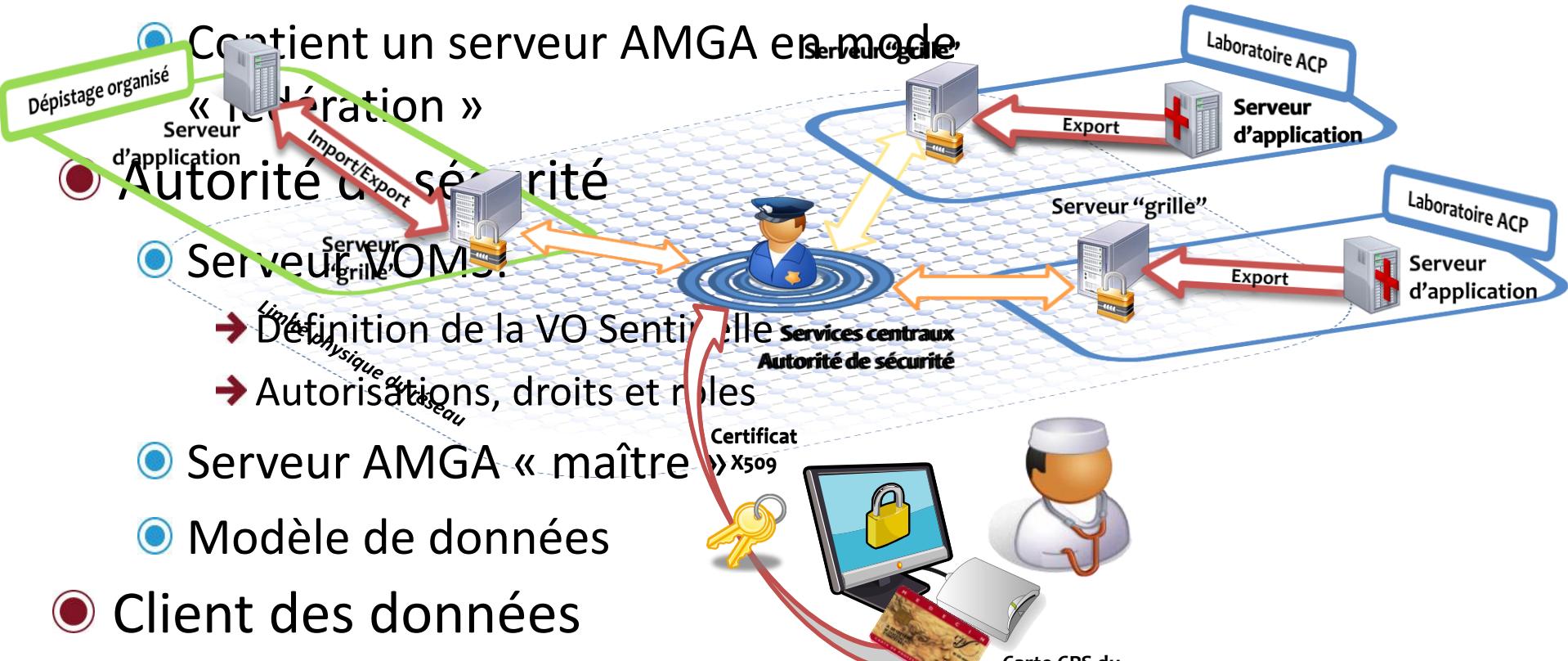
● Utilisation d'AMGA:

- Catalogue de métadonnées compatible avec les grilles
- Récupération des groupes, rôles et droits depuis VOMS
- Fonctionne en mode réPLICATION, fédération
- Permet de définir des Access Control Lists (ACL) sur les données



Architecture globale

● Source de données de pathologie



● Client des données

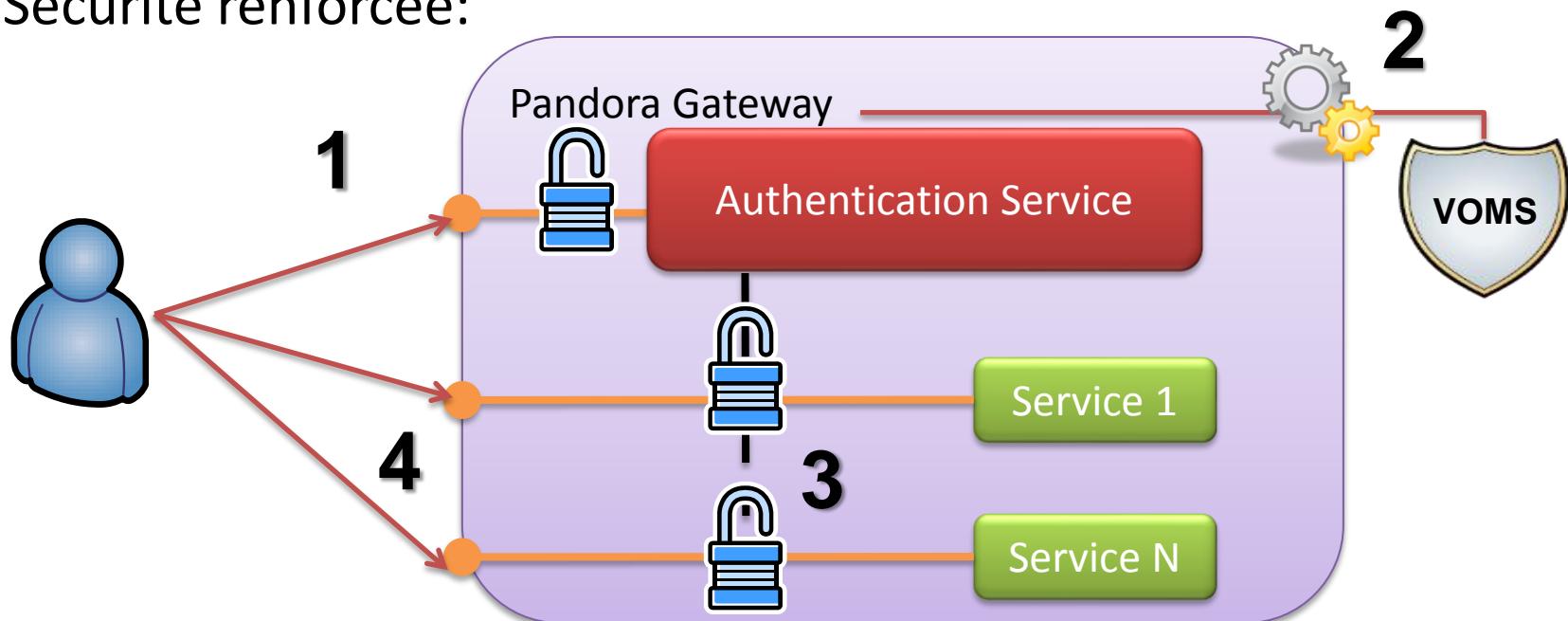
- Se connecte directement au noyau central pour interroger le modèle de données



Pandora Gateway

● Infrastructure développée par maat-Gknowledge

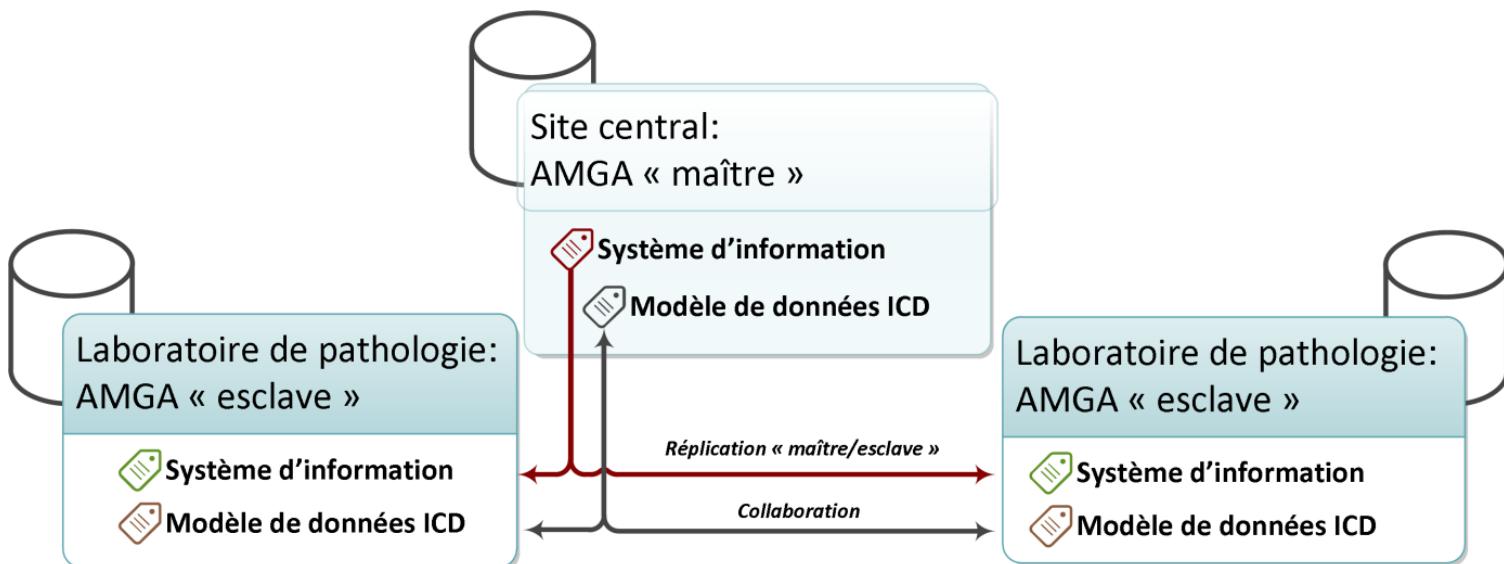
- Projet Health-e-Child
- Couche d'abstraction de la complexité des grilles
- Architecture orientée service (Services Web)
- Sécurité renforcée:





Le modèle de données

- Utilisation d'« Integrated Case Data (ICD) » dans AMGA
 - Représentation générique des données médicales
 - Utilisation de concepts
 - Patient, Visit, Exam, Clinical Variable
 - Restitution de l'information au moyen d'un service web





Modèle de données anatomopathologique



● Concepts ICD liés à la base anatomopathologique

■ Patient

■ Visit / Medical Event

■ Clinical Variable

NURES ID	Identifiant dossier interne
NUDDEXT	Identifiant de l'examen (public)
DATPREL	Date ou le prélèvement a été effectué
DATENREG	Date ou le prélèvement a été reçu au labo
NUPAT	Numéro du patient
NOMPAT	Nom patronymique
PRENOM	Prénom
ADRESSE1	Adresse du patient
ADRESSE2	Complément d'adresse 2
ADRESSE3	Complément d'adresse 3
CODPOSTAL	Code postal du patient
VILLE	Ville
PAYS	Pays
SEXЕ	Sexe
DATNAISSANCE	Date de naissance

NOMLEC	Médecin lecteur
INSEELEC	Code INSEE du médecin lecteur
NOMLEC2	Médecin lecteur 2
INSEELEC2	Code INSEE du 2ème médecin lecteur
NOMMED	Nom de médecin (prescripteur)
INSEEMED	Code INSEE du médecin prescripteur
NURES	Numéro dossier, interne à la BDD
DATVALIDATION	Date de validation
MODPREL	INFORMATIONS ADICAP
TYPTECH	
ORGANE	
LESION	
DATCODAGE	Date de finalisation
NOMORIG	Origine du prélèvement
RESULTATCCL	Lien vers le fichier
NOMLEC	Médecin lecteur
INSEELEC	Code INSEE du médecin lecteur



Interfaçage avec les Systèmes d'information

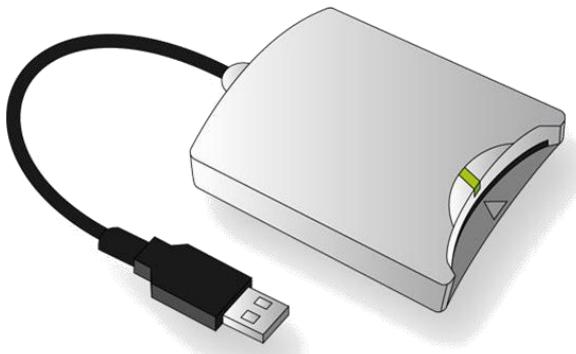


- Pour les fournisseurs de données de pathologie
 - Solution pour le cabinet SIPATH:
 - ➔ Les données sont « poussées » au format XML depuis les bases métier vers le serveur « grille »
 - ➔ L'intégration est ensuite effectuée dans ICD
- Pour les associations de dépistage
 - ICD permet d'exposer aux travers de services web les données
 - ➔ La gateway se charge de l'authentification des utilisateurs
 - ➔ Amga offre un accès transparent aux sites distribués en mode « fédération »
- Pour la santé publique
 - Les clients se connectent directement au serveur central
 - ➔ Ils peuvent interroger le réseau comme une base de données classique



Eléments de sécurité: CPS

- Mise en place d'un système d'authentification des utilisateurs par Carte de Professionnel de Santé [1]



- Contenu d'une CPS:

- Informations sur le porteur de la carte (personne, métier, ...)
- Deux certificats X509: un pour l'authentification et un pour la signature

[1] J. PASSERAT-PALMBACH, *Mise en place d'un environnement de sécurité autour de la Carte de Professionnel de Santé dans une infrastructure de grille de données*. 2009, Rapport d'ingénieur, Filière : Informatique des Systèmes Embarqués, Université Blaise Pascal - Clermont-Ferrand.



Eléments de sécurité: CPS



● Utilisation de PKCS#11

- Modèle standard d'accès aux périphériques de sécurité
- Fonctionnement multiplateforme

● Développement d'un module d'accès aux certificats CPS

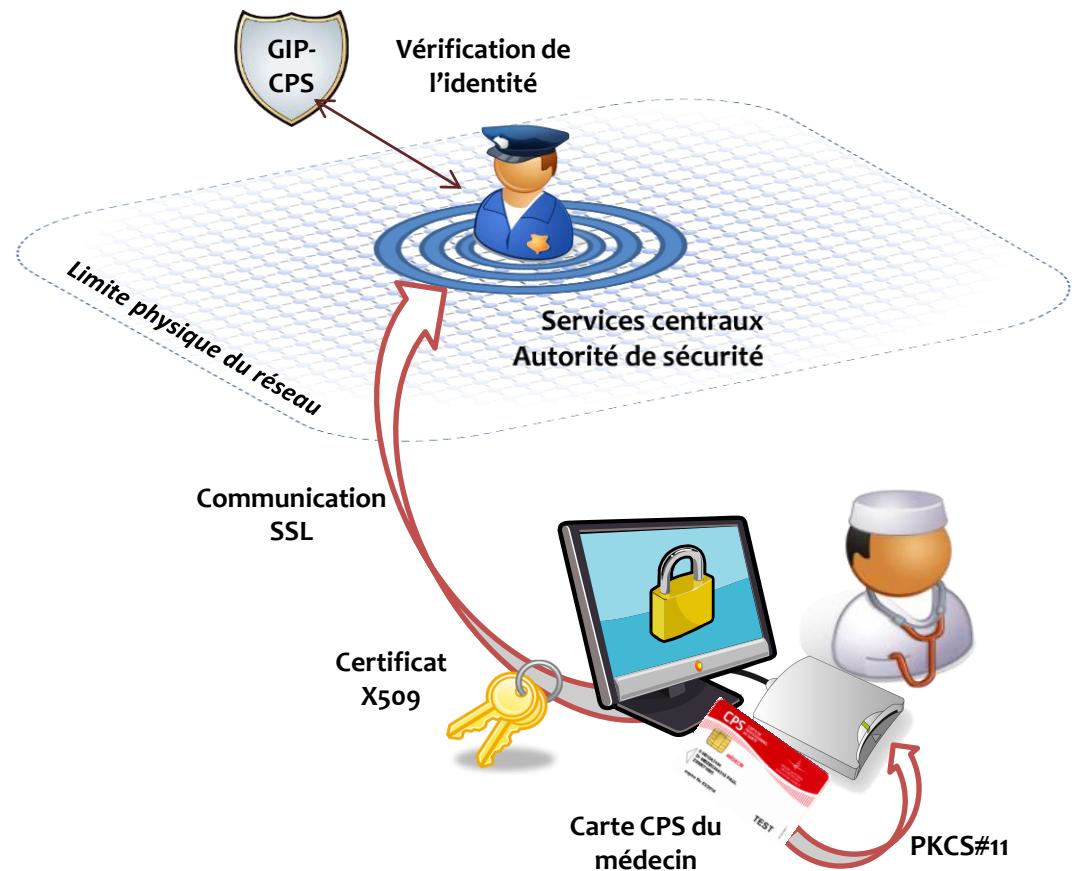
- Intégration au service d'authentification de la Gateway





Eléments de sécurité: CPS

- Inscription de la chaîne de certification GIP-CPS dans VOMS





Partie 4:

GESTION DU PATIENT ET DES DONNÉES MÉDICALES POUR LE RÉSEAU SENTINELLE CANCER AUVERGNE



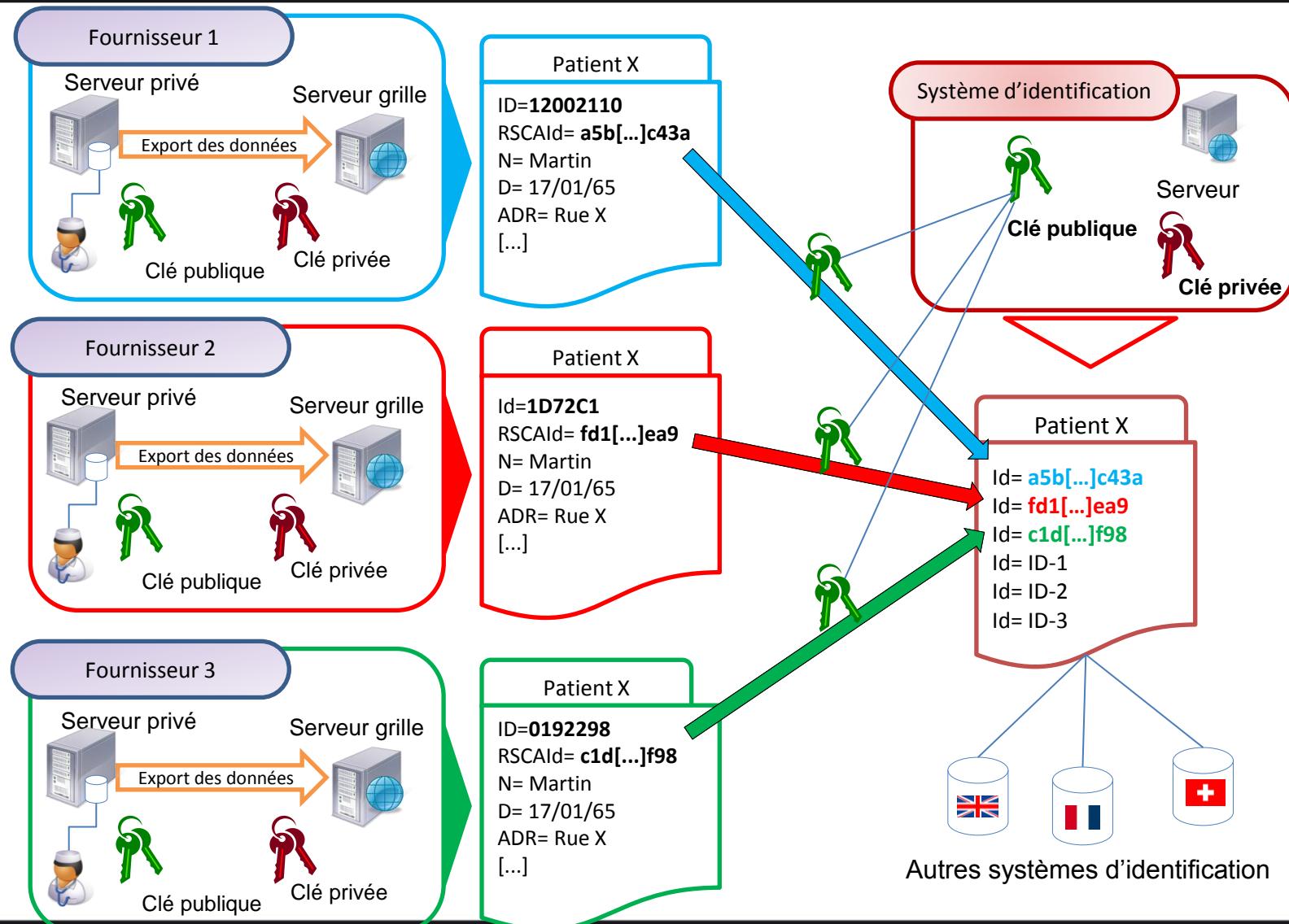
Identification du patient

- Problème « ouvert » dans le système de soin français
 - Utilisation du numéro de sécurité sociale prohibé par la CNIL
1 79 03 63 877 122 56
 - Adoption de l'Identifiant National de Santé marginal
13 196 952 729 666 105 357
- Besoin de définir un nouveau modèle
 - Respect des contraintes de l'ASIP
- Création d'un modèle dynamique et distribué du patient
 - Définition d'un identifiant anonyme de type 'uuid' propre à chaque patient et à chaque base:
162bb7e6-8fe4-4c69-854d-48c184e29825
 - ➔ Statistiquement unique: 10^{36} possibilités
 - Utilisation systématique du chiffrement de l'identifiant
 - ➔ Fédération des identifiants au niveau du serveur central

[1] P. LEACH, M. MEALLING, ET R. SALZ, A UNIVERSALLY UNIQUE IDENTIFIER (UUID) URN NAMESPACE. 2005, RFC 4122.

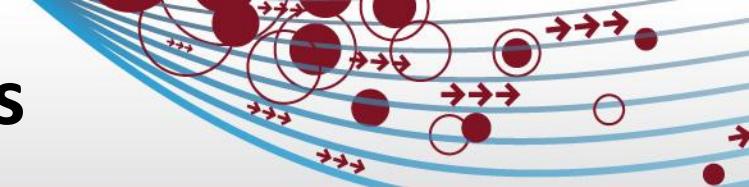


Mécanisme d'identification



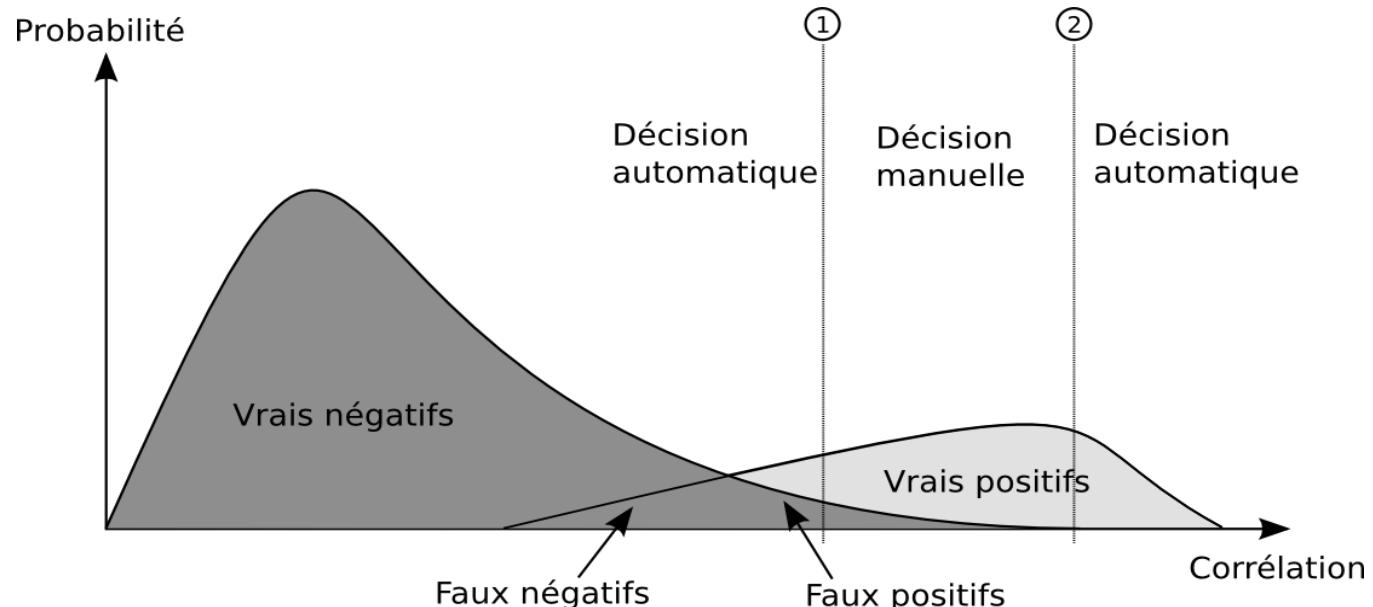


Rapprochement des identités



● But

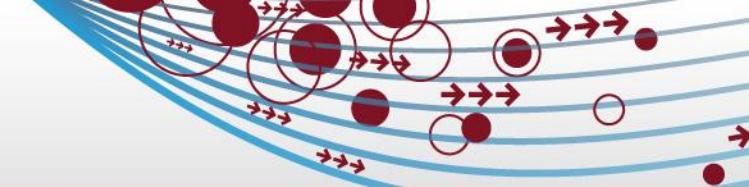
- Comparer les traits d'identification des patients
 - ➔ Jusque 27% d'erreur d'identification sur 3 bases de 100000 patients [1]
- Calcul d'un « score » de similarité entre deux patients
- Définir les seuils de décision ① et ② permettant de rapprocher ou non deux patients



[1] C. FRIEDMAN ET R. SIDELI, *Tolerating spelling errors during patient validation*. Computers and Biomedical Research, 1992. **25**(5): p. 486-509.



Comparateur de patients



- Définition de règles en fonction du type de données:
 - Entier, Date et Chaîne de caractères
- Pondération de ces règles en fonction de leur pertinence

	Nom	Prénom	Date de naissance	Adresse
Type	Chaîne	Chaîne	Date	Chaîne
Précision	•••	•••	•••	•
Elément bloquant	X		X	
Poids (similaire)	•••	•••	•••	•
Poids (différent)	•••	•	••	•



Algorithmes de comparaison de chaînes de caractères

Algorithmes analytiques

- Comparaison mathématique des chaînes de caractères

- Compte un nombre de différence entre les chaînes

- Meilleur algorithme [1]:

- Distance de Jaro-Winkler

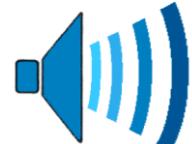
- Exemple:

$\text{JaroWinkler("Philippe", "Mhilippe") = 77\%}$

$\text{JaroWinkler("Philippe", "Filip") = 0\%}$



Algorithmes phonétiques



- Calcul d'un « code » phonétique

- Comparaison des codes de manière analytique

- Meilleur algorithme [1]:

- Phonex (ou Phonex-fr)

- Exemple:

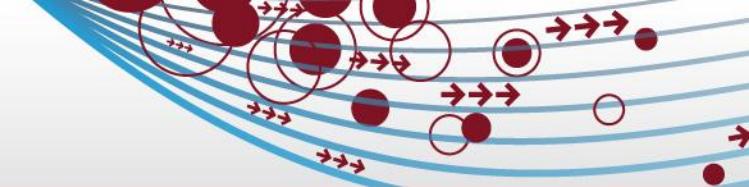
$\text{Phonex("Philippe", "Mhilippe") = 0\%}$

$\text{Phonex("Philippe", "Filip") = 100\%}$

[1] P. CHRISTEN, *A comparison of personal name matching: Techniques and practical issues*. Data Mining Workshops, 2006: p. 290-294.



Expérimentation

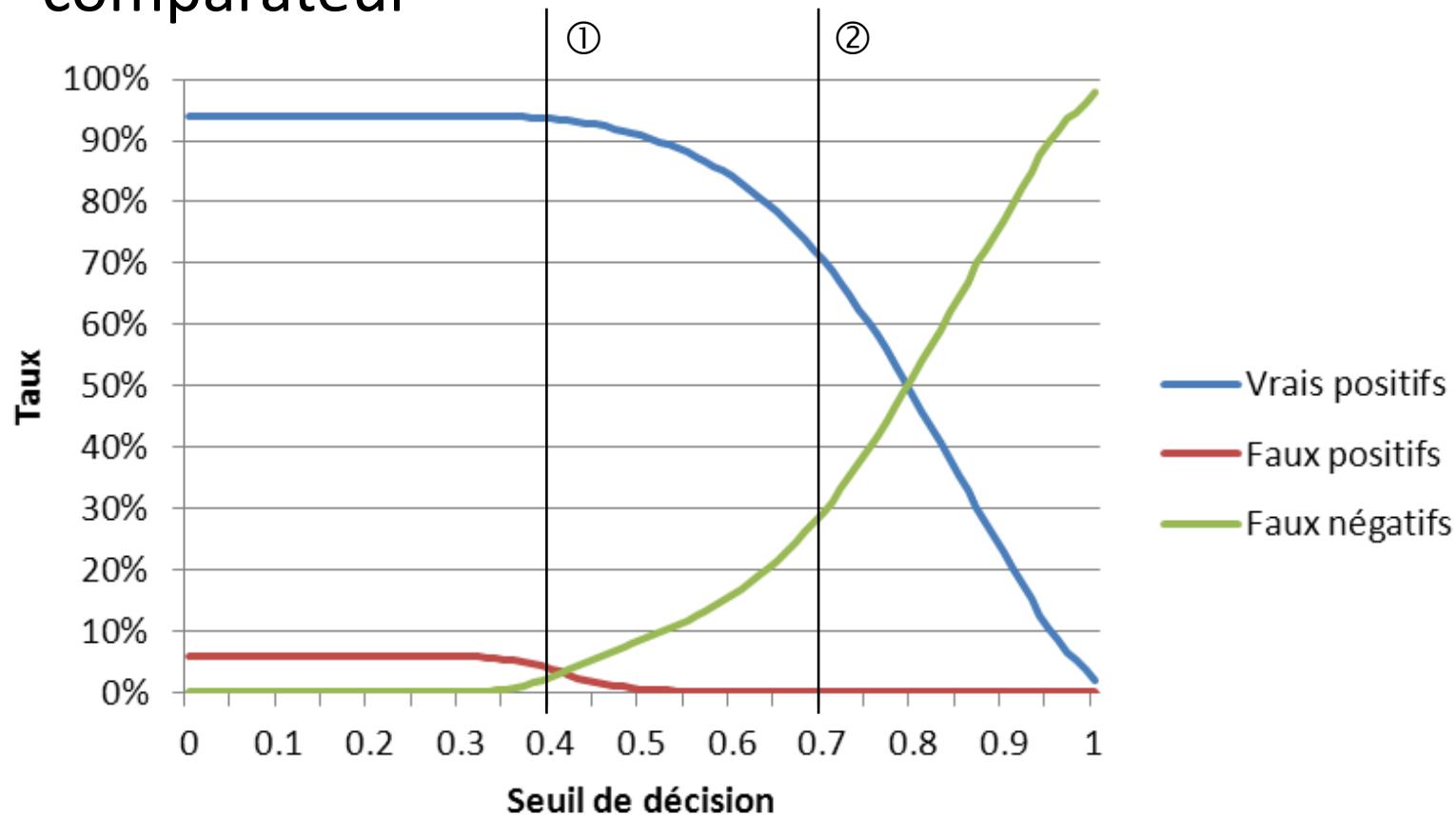


- Utilisation de données « vraisemblables » issues du réseau sentinelle
 - 70000 individus
 - Utilisation du **nom** et du **prénom** des patients
- Insertion manuelle de biais dans une seconde base
 - Loi de poisson, $\lambda=1$
- Utilisation d'une méthode combinant à la fois Jaro-Winkler et Soundex pour la comparaison des chaînes



Expérimentation : résultats

- Ajout de l'adresse et de la date de naissance au comparateur



→ Seuil de décision valable pour ① = 0.4 et ② = 0.7



Analyse du comparateur



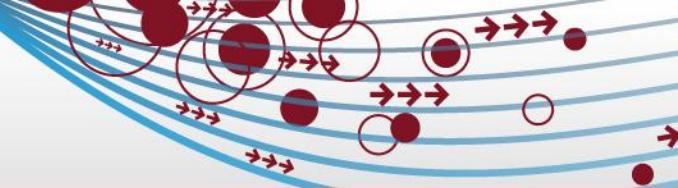
- Bon taux de rapprochement automatique
 - Nécessite un maximum d'informations sur le patient
- Complexité quadratique de l'algorithme
 - 70000 enregistrements de 4 champs = 19,6 milliards de comparaisons
 - ➔ 10heures d'exécution sur un processeur actuel
- Perspectives d'amélioration
 - Arrêt de la comparaison si le score déjà inférieur au seuil ①
 - Lancement parallèle de l'algorithme
 - Amélioration des algorithmes de comparaison
 - ➔ Utilisation des ressources GP-GPU



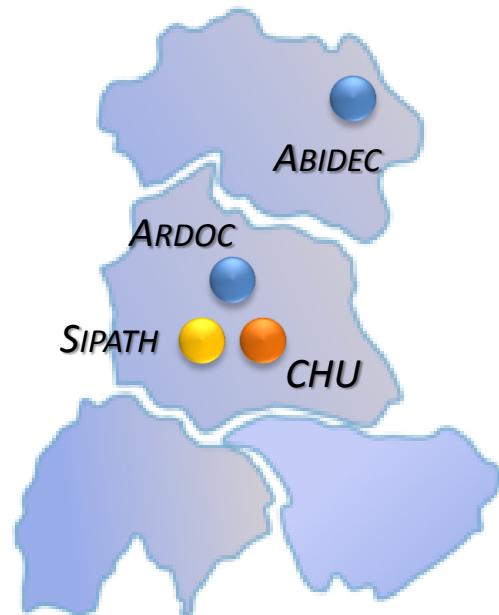
Partie 5:

EVALUATION

Etat du Réseau Sentinelles Cancer Auvergne



- L'infrastructure de grille en place
 - Services centraux (AMGA, VOMS), VO Sentinelles
→ Hébergés par maat-G France
- 4 sites sont équipés
 - Les associations de dépistage (ARDOC, ABIDEC)
→ Accès au réseau difficile (lignes ADSL)
 - La SIPATH (Pôle santé république)
→ 90% des cancers du sein
 - La faculté de médecine /CHU
→ Développements, tests
- Obstacles
 - Juridiques (accord CNIL)
 - Installation sur site (physique et réseau)





Réponse au cahier des charges



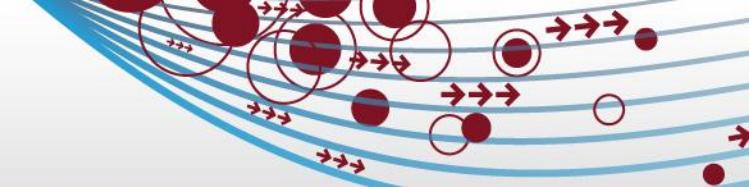
- Ne pas exporter les données 
- Contraintes de sécurité
 - CPS 
 - Système d'autorisation central 
 - Traçabilité 
- Intégration
 - Pathologistes 
 - Modèle de données
 - Système d'information
- Intégration 
- Dépistage 
- Epidémiologistes
 - Modèle de données 
 - Interface 
- Identification des patients 
- Rapprochement d'identités 
- Imagerie médicale 
- Accord CNIL 

Comparaison avec les grilles pour la santé

PROJETS	DATES	INFRASTRUCTURE OPÉRATIONNELLE	GESTION DES IMAGES MÉDICALES	GESTION DE L'IDENTITÉ	ANALYSES ÉPIDÉMIOLOGIQUES
 eDiaMoND	2002 - 2005				
 caBIG®	2004 - ?				
 ACGT <small>Advancing Clinico Genomic Trials on Cancer</small>	2006-2010				
 Health-e-Child	2006-2010				
 NEUROLOG	2006-2010				
 SENTINELLE GRID NETWORK	2008 - ?				



Conclusion



● Le Réseau Sentinelles Cancer Auvergne

- Apporte une réponse technique à la contrainte du respect de la propriété des données au moyen des technologies des grilles informatiques
- Offre un cadre sécurisé pour accéder aux dossiers anatomopathologiques
- Permet un accès rapide au données « cancer » pour des enquêtes épidémiologiques

● Il comprend

- La mise en place des éléments de sécurité requis par la CNIL et l'ASIP-Santé
- Le déploiement d'un modèle dynamique et distribué pour l'identification des patients
- Une méthode de rapprochement des identités nécessitant peu d'intervention manuelle

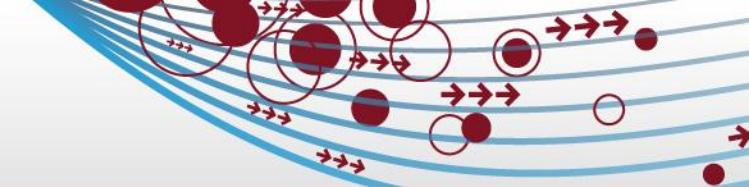


Perspectives de RSCA

- Extension du réseau aux cabinets de radiologie pour les mammographies numériques
 - Nécessaire pour les relectures
- Amélioration de l'outil d'identification et de rapprochement des patients
 - Taux de rapprochement automatique
 - Vitesse d'exécution
- Intégration des données dans les SI des structures de dépistage



Perspectives de l'outil



- Modèle de grille générique pour l'accès décentralisé aux données de santé
 - Enjeu majeur de société
 - Possibilité d'extension à d'autres cas d'utilisation
 - ➔ Autres applications médicales
 - ➔ Autres types de données (imagerie, puce à ADN, ...)
 - Possibilité d'extension géographique
 - ➔ A l'échelle française
 - ➔ A l'international
- Le projet ANR GINSENG (2011-2014)
 - Poursuit les objectifs de RSCA
 - Cadre d'utilisation cancer, périnatalité et urgences
 - Plus grande échelle