Entrepôts de données

i2b2, intégration et protection des données

Maxime Wack

24 novembre 2020

Historique

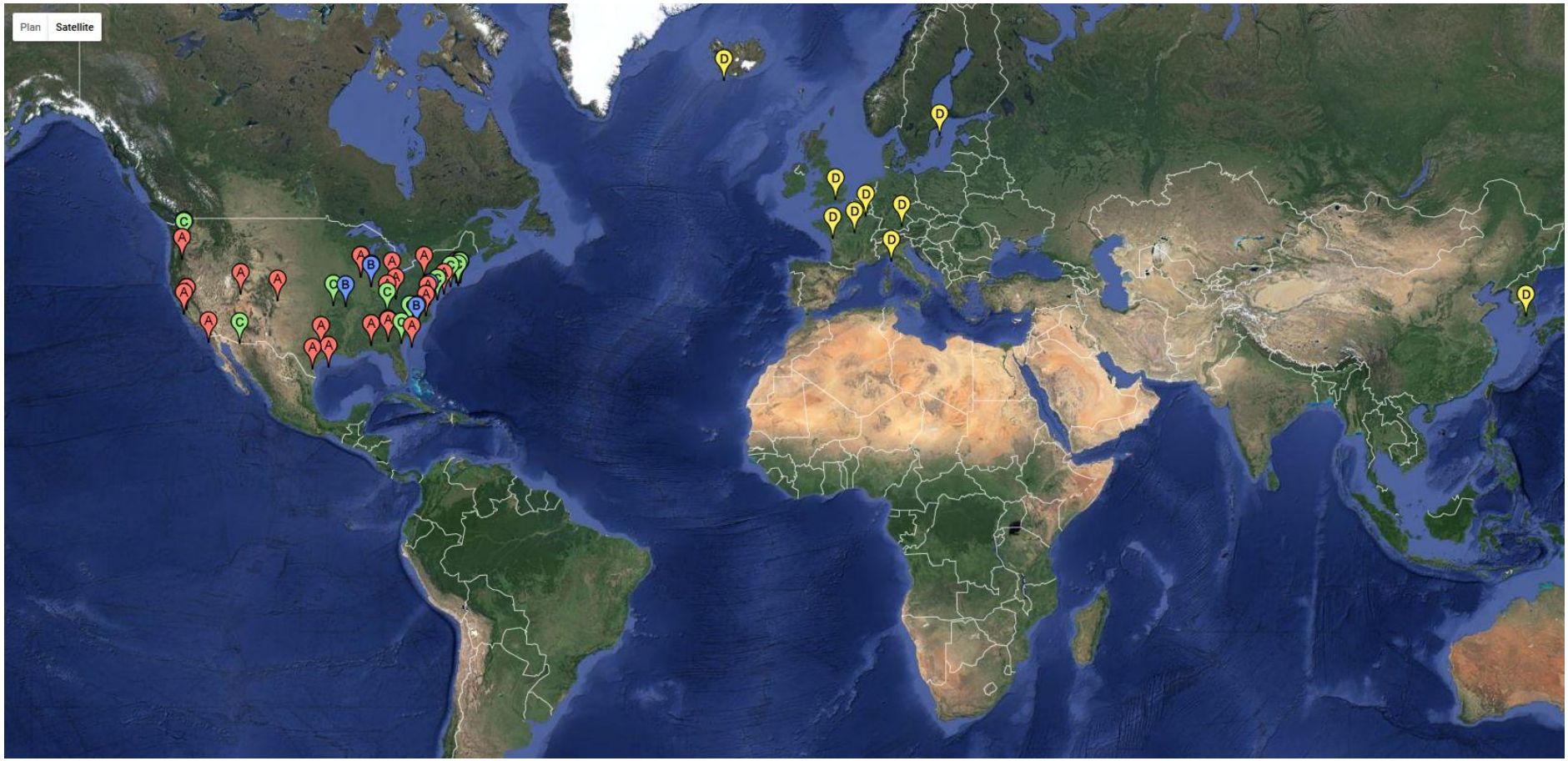
- **Informatics for Integrating Biology and the Bedside**
- Prototype en 2000, Shawn Murphy à Partners HealthCare (MGH, BWH, …), RPDR
- En 2004, projet NCBC (Nation Center for Biomedical Computing) financé par le NIH.
- Partners un des 4 receveurs du financement pour développer i2b2, direction du Pr Isaac Kohane (HMS + BCH), avec participation des hôpitaux affiliés, MIT, HSPH
- Open source, licence custom

Murphy S, Barnett G, Chueh H. Visual query tool for finding patient cohorts from a clinical data warehouse of the partners HealthCare system. Proc AMIA Symp. 2000;1174.

Historique

- Depuis : renouvellement du financement à plusieurs reprises
- Installation dans de multiples établissements (200+ : hôpitaux publiques, privés, unités de recherche) à travers le monde
- Extension des fonctionnalités
- Organisation de «challenges» pour l'intégration de nouvelles fonctionnalités
- Plus de 240 articles de recherche sur/utilisant i2b2

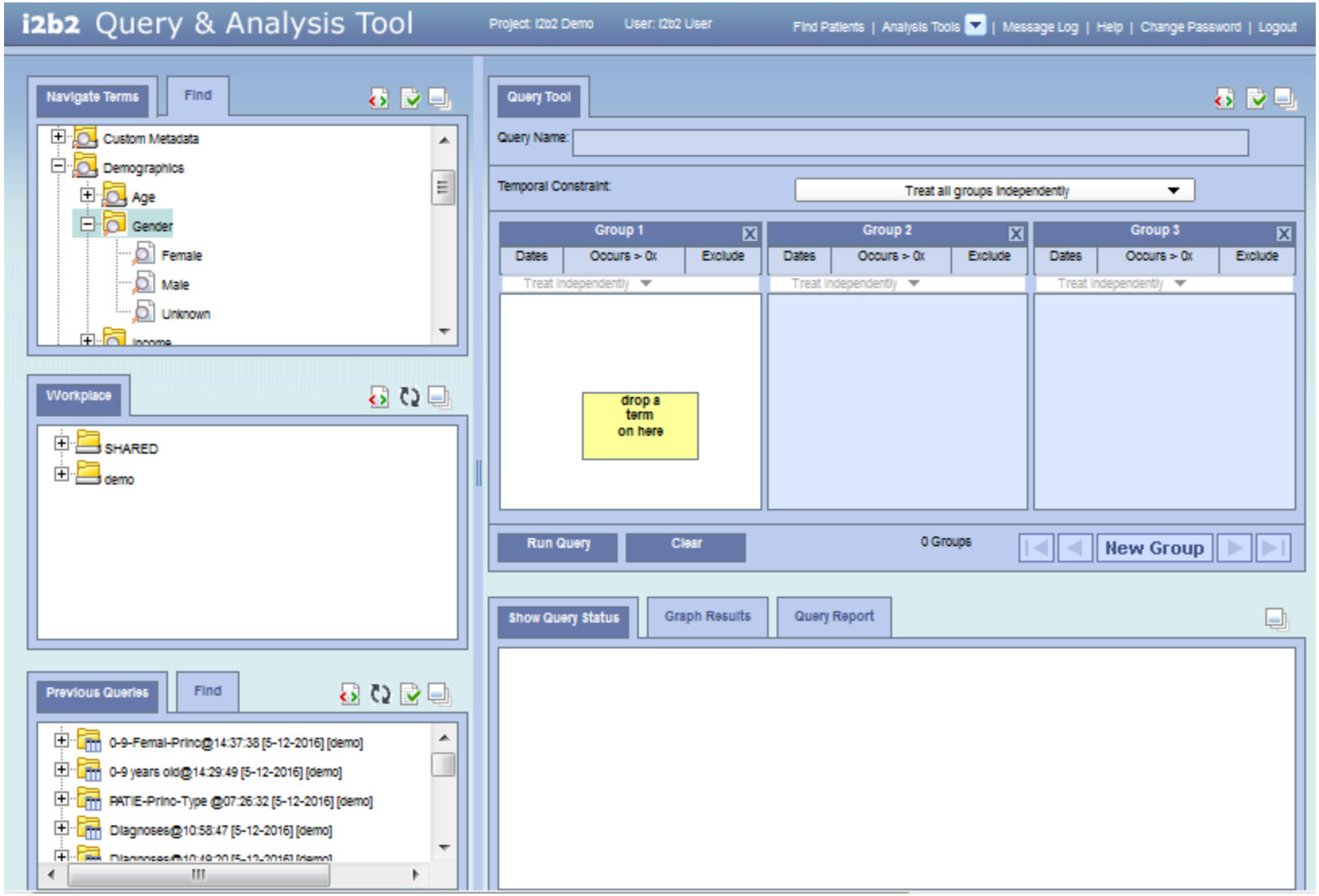
Installations



http://www.healthmap.org/i2b2

http://ncats.nih.gov/ctsa/about

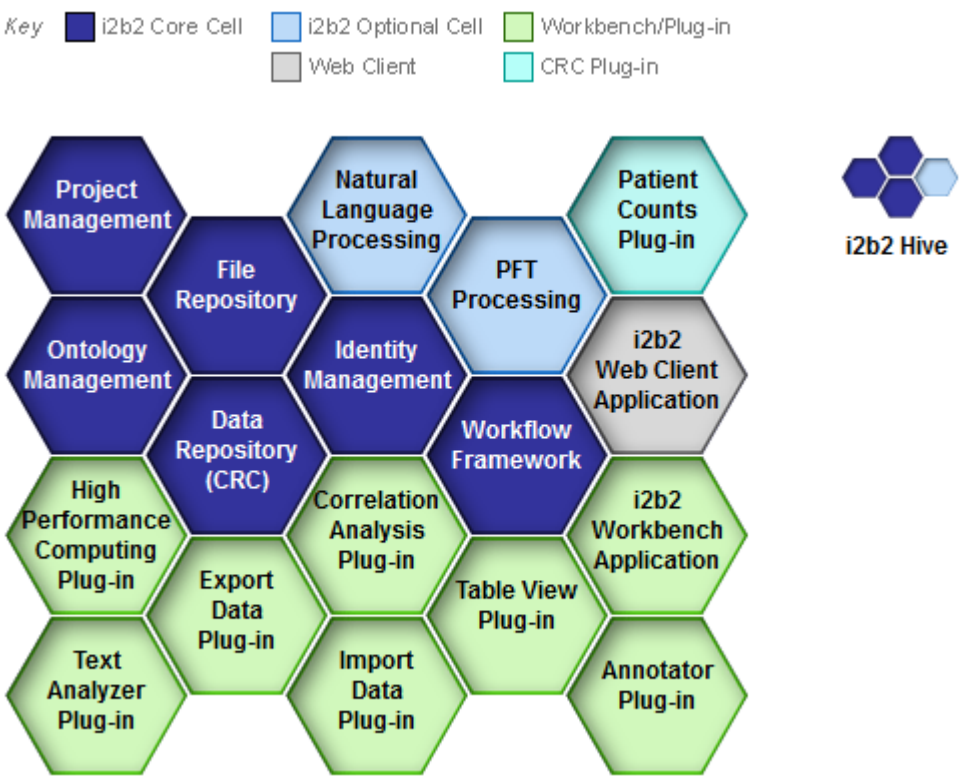
Interface



https://www.i2b2.org/webclient/

Fonctionnement

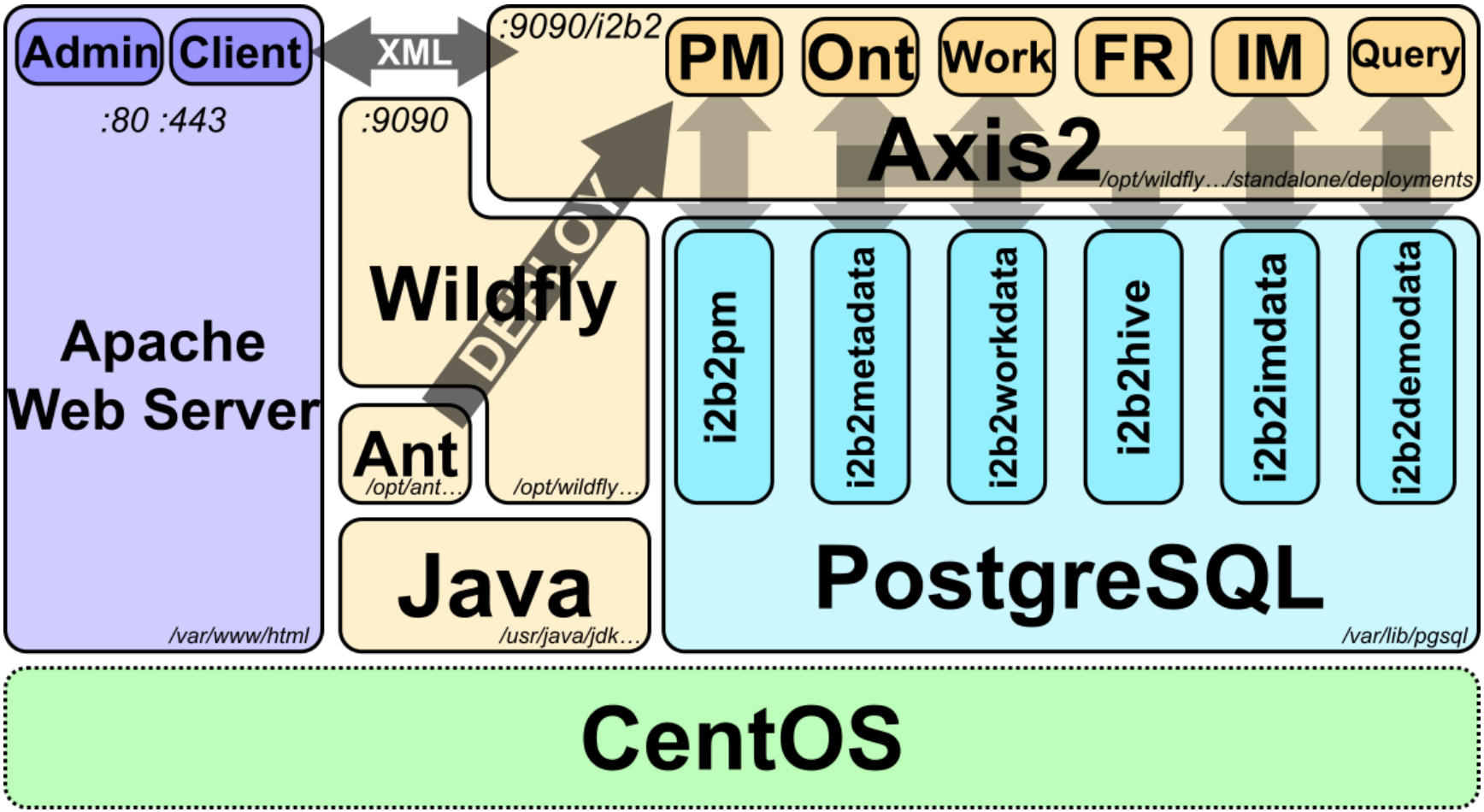
- Ensemble de composants (cells)
- Interagissant entre eux (hive)
- Composants essentiels
  - «ontologie»
  - données
  - fichiers
  - identité
- Composants annexes (NLP)
- Plugins (extensibilité)



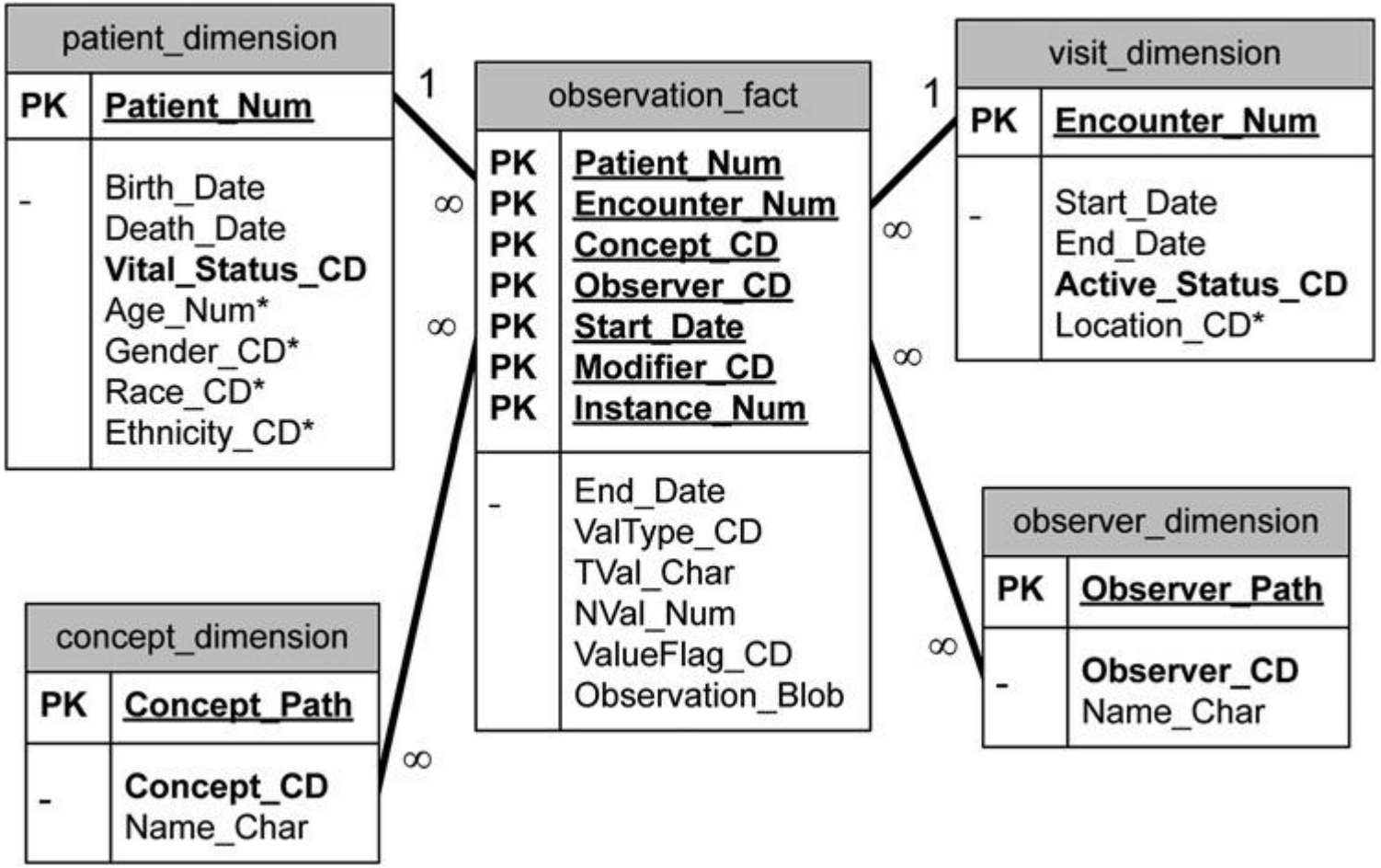
Cells i2b2

- Hive : orchestration des services entre eux
- PM (Project Management) : utilisateurs et projets
- Metadata : représentation des données
- IM : gestion de l'identité des patients
- CRC : hébergement des données
- Work : gestion des requêtes

Cells i2b2


Admin
Client
:80 :443
Apache
Web Server
/var/www/html
XML
:9090/i2b2
PM
Ont
Work
FR
IM
Query
Axis2
/opt/wildfly…/standalone/deployments
:9090
Wildfly
DEPLOY
Ant
/opt/ant…
/opt/wildfly…
Java
/usr/java/jdk…
i2b2pm
i2b2metadata
i2b2workdata
i2b2hive
i2b2imdata
i2b2demodata
PostgreSQL
/var/lib/pgsql
CentOS

Représentation des données



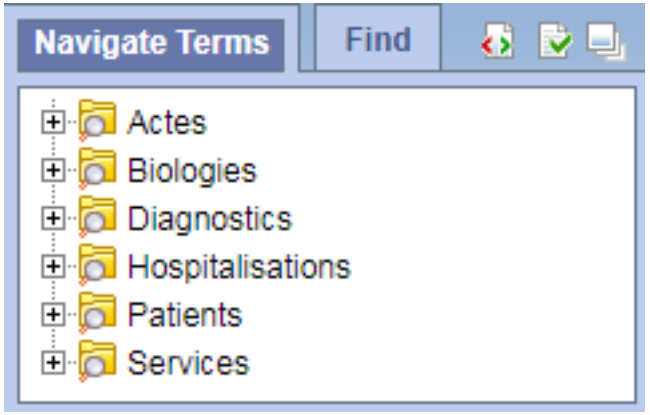
Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc. 2010;17(2):124-30.

Représentation des données

- Structurées, «ontologies»
- Pas de relation sémantique

…ni moteur d'inférence

- Relations
  - hypo/hyperonymie
  - synonymie
  - taxonomie
- Permet l'exploration des données
- représentation de terminologies existantes



Représentation des données

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

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

Intégration des données

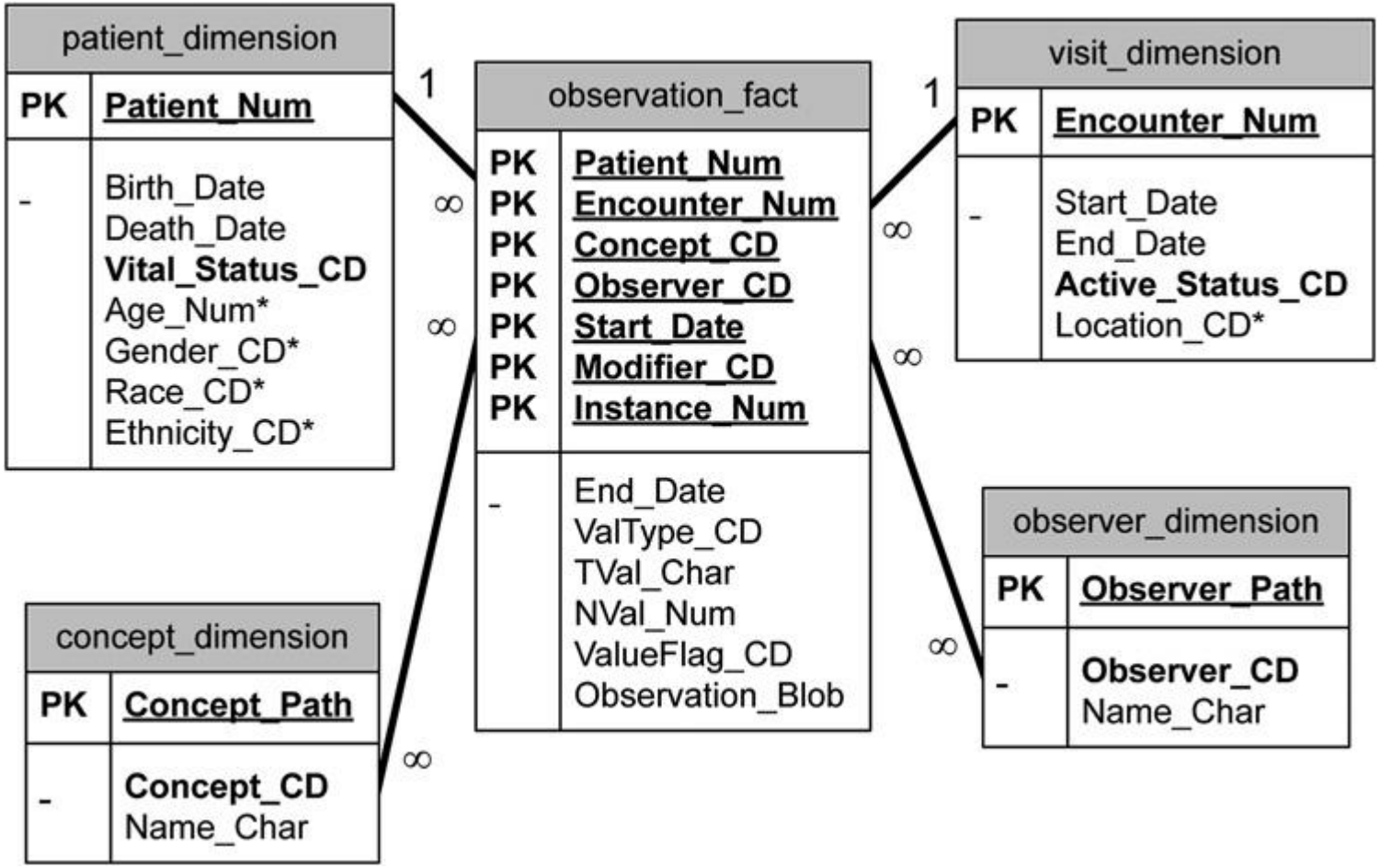
- Intégration des données via une Cell (NLP par exemple)
- Développement d'une Cell custom, communication via des messages XML
- Utilisation de i2b2 workbench
- Utilisation d'outils d'ETL pré-existants (Talend, Kettle)
- Insertion directe dans la base de données (postgreSQL ou Oracle)
- **R2b2**

R2b2

- Package R
- https://github.com/maximewack/R2b2

- Administration de la plateforme
  - «projets» (mini datamarts)
  - utilisateurs
- Gestion des données
  - représentation
  - intégration

Représentation des données

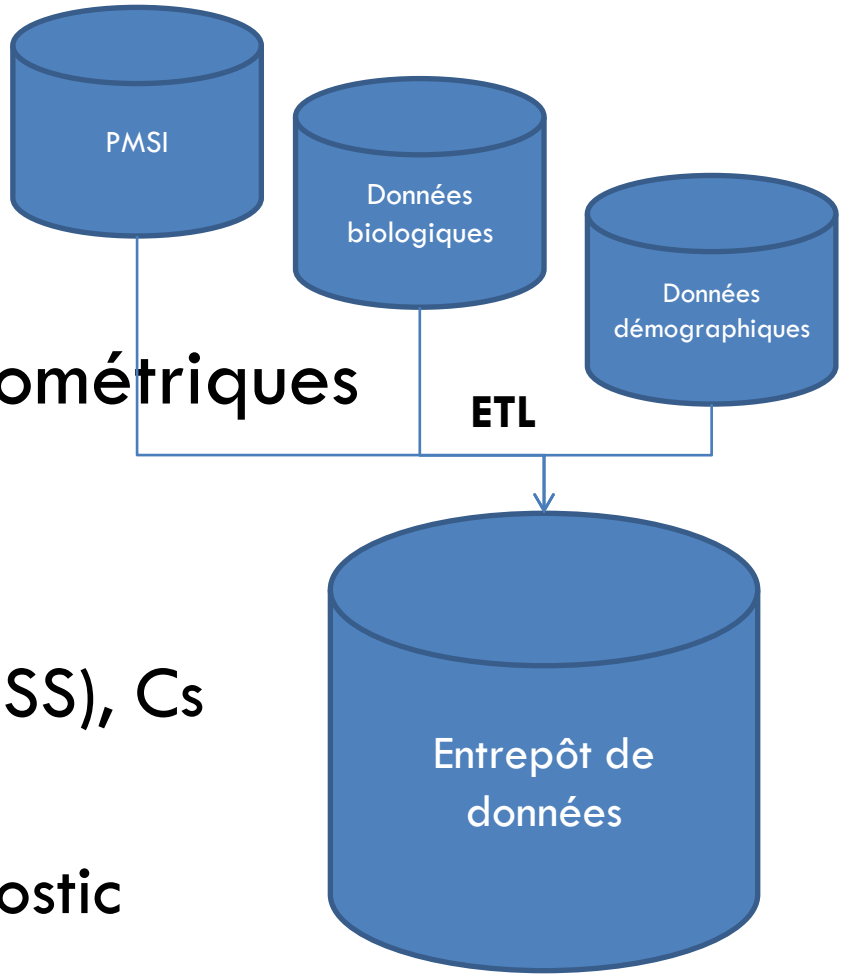


Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc. 2010;17(2):124-30.

Intégration des données

Sources de données

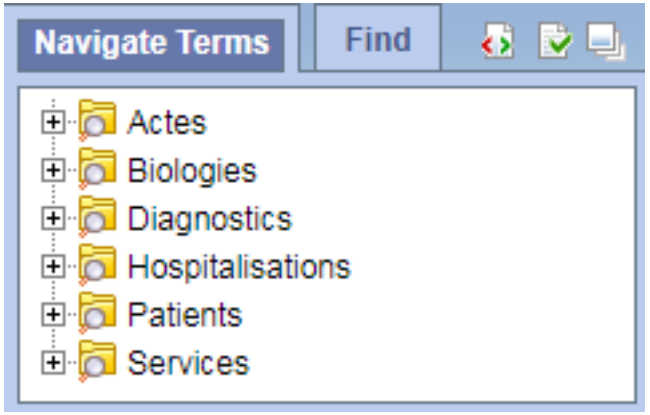
- PMSI : diagnostics et actes
- Biologies via le DPI
- Démographiques et morphométriques



Intégration des données

- visit = venue (hospitalisation (RSS), Cs
- provider = unité médicale
- RUM = période pour un diagnostic

Représentation des données



- Diagnostics PMSI : **CIM-10**
- Actes PMSI : **CCAM**
- Structure des services
- Données démographiques
- Biologies : définition d'une terminologie locale
  - extraction de 2302 concepts
  - alignement de 633 synonymes vers 241 termes
  - 1320 termes après exclusions et alignement
  - Organisation hiérarchique : milieu → bilan → mesure

Accès aux données

- Niveaux d'accès (inspirés HIPAA)
  - OBFSC : données agrégées floutées
  - AGG : données agrégées
  - LDS : accès limité aux données identifiantes
  - DEID : LDS + notes complètes
  - PROT : accès aux données complètes

Accès aux données

- Niveaux d'accès (inspirés HIPAA)
  - **OBFSC** : données agrégées floutées
  - **AGG** : données agrégées
  - LDS : accès limité aux données identifiantes
  - DEID : LDS + notes complètes
  - **PROT** : accès aux données complètes

Rôles utilisateurs

- USER : utilisateur «simple»
- MANAGER : gère les utilisateurs et leurs requêtes
- ADMIN : administrateur pour la plateforme

Anonymisation des données

- Tables *patient_mapping* et *encounter_mapping*
- Associent un identifiant i2b2 aux patients

→ pseudonymisation

- Possibilité de choisir la méthode de pseudonymisation

Gouvernance

- Pas de cadre légal spécifique aux entrepôts
- Cadre législatif encadrant :
  - le partage des données de soin (*CSP L1110-4, L1110-12*)
  - la recherche sur la personne humaine (*Loi Jardé*)
  - les bases de données nominatives (*Loi Informatique et Libertés*)

Gouvernance

Exemple de l'HEGP :

- Niveau 1 : AGG pour tous les praticiens
- Niveau 2 : subset anonymisé (validation scientifique)
- Niveau 3 : subset identifiant (validation scientifique + déclaration CNIL)

Gouvernance

Fonctionnement des requêtes sur le SIH au CHRU de Nancy

- confiées au DIM
- accord du chef de service pour l'extraction de données produites
- accord du chef de pole pour les données d'un pole
- examen par la CIM pour des données couvrant tout l'établissement
- on confie la liste nominative au responsable médical ayant signé la demande

Gouvernance

Au CHRU de Nancy :

- données complètes (extraction de concepts)
- données agrégées (comptes de patients)
- données agrégées masquées (comptes approximatifs)

Stratégie d'accès mimant la politique d'accès aux données du SIH

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

Gouvernance

- Déclaration CNIL simple par la CIL de l'établissement
- Permise par les périmètres d'accès accordés aux données
- Deux options offertes pour l'identification des patients par les responsables médicaux

Consentement

- Non-opposition
- Filtre à la source (extraction des données)

→ problème avec l'utilisation épidémiologique/surveillance

- Modélisation du consentement

→ nécessité de former les utilisateurs

À l'HEGP

- Déploiement depuis 2008
- Données depuis 2000
- 1.5 To de données (dont ~80% d'index)
- Base de données Oracle

➢ optimisations : indexes, tables virtuelles, vues

- 30Go de texte (CRs)

À l'HEGP

- Démographie (ddn, ddc, sexe, visites)
- PMSI (diag, actes, GHM)
- CR (imagerie, opération, hospit, observ)
- Bio
- Anapath
- Prescriptions
- Radiothérapie/Chimiothérapie

À l'HEGP

- 1M patients uniques
- 4M visites
- 500M observations
- «vues» excluant les patients ayant exprimé leur opposition, mais présents dans la base de données

Biologies - Qualité

- Qualité au cours du temps
- Nécessité de méta-données sur les biologies

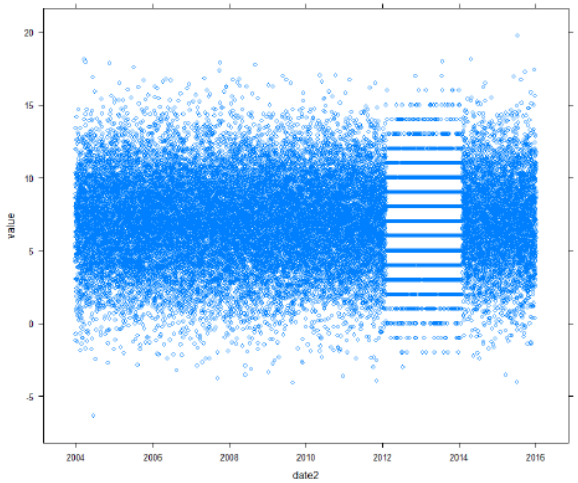


Figure 5 - Discrétisation

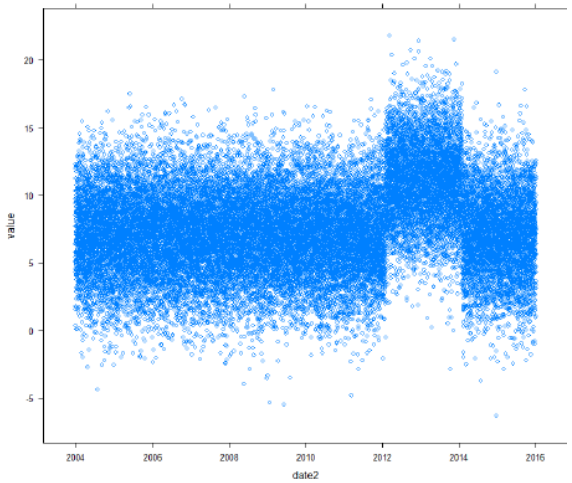


Figure 6 - *Shift*

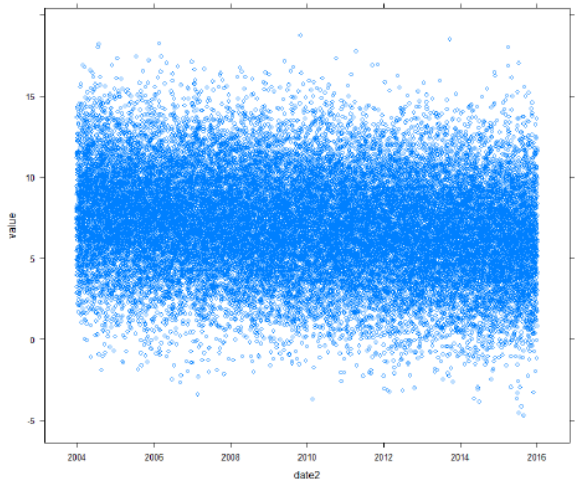


Figure 7 - Tendance

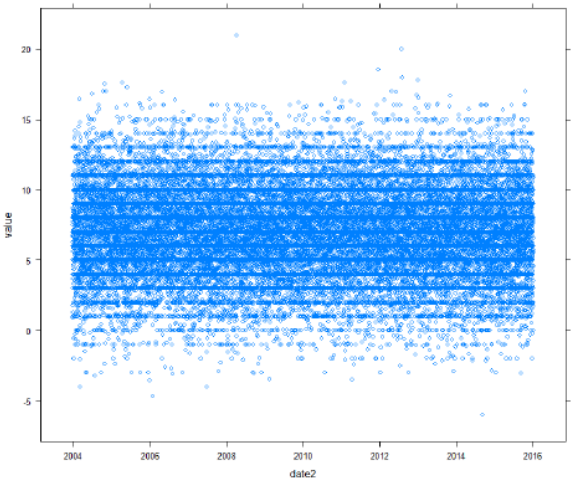


Figure 8 - Discrétisation mélangée

SHRINE

- Réseau d'i2b2
- Accès limité aux données
- Comptes par datamart
- Permet l'échange entre plusieurs hopitaux
- Couche d'inter-opérabilité (Common Data Model)
- Data Stewart