Registre National des Cancers

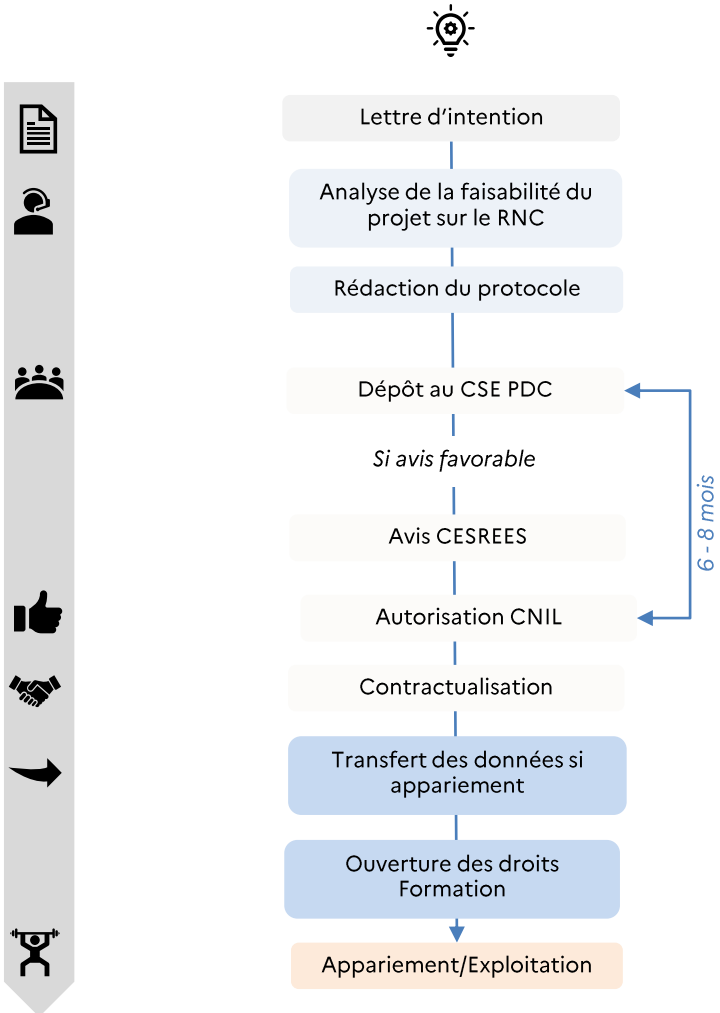
inca Institut National du Cancer

Exploitation des données dans l'EDS-RNC

Mathieu Rocchi, Eliana Wassermann

1ère Journée scientifique du RNC – 03/11/2026

Accéder aux données de l'EDS RNC
Process réglementaire





https://www.cancer.fr/registre-national-des-cancers/

Rôle/responsabilités de chaque acteur, objectifs, données mobilisées ,information des personnes …

Tous les 2 mois

Durée
Liste des personnes autorisées

Mise à disposition dans un espace dédié sécurisé de l'entrepôt de données propre à chaque projet

De données concernant la population de l'étude

Pour les utilisateurs autorisés

Sommaire

01 Démarrage projet et Espace de travail
Présentation des étapes de lancement du projet et de l'espace de travail
PRESENTATION

02 Types des données et Tables INCa
Typologies des données accessibles et focus sur les tables spécifiques
DATA & TABLES

03 Outils d'exploitation et Documentation
Descriptif des outils d'analyse et consultation documentation
OUTILS

04 Support et Prochains Webinaires
Modalités d'assistance technique/métier et événements à venir
ACCOMPAGNEMENT

Processus du lancement d'un nouveau projet

- **Signature convention**
  => Mobilisation différents acteurs au sein de l'INCa

Coordinateur du RNC :
- **Conformité** des aspects administratifs
- **Autoriser le lancement du projet**



Gestionnaire habilitation :
- **Création** accès espace projet



Référent métier INCa du projet :
- Mise en place **méthodologie** de sélection des identifiants du projet



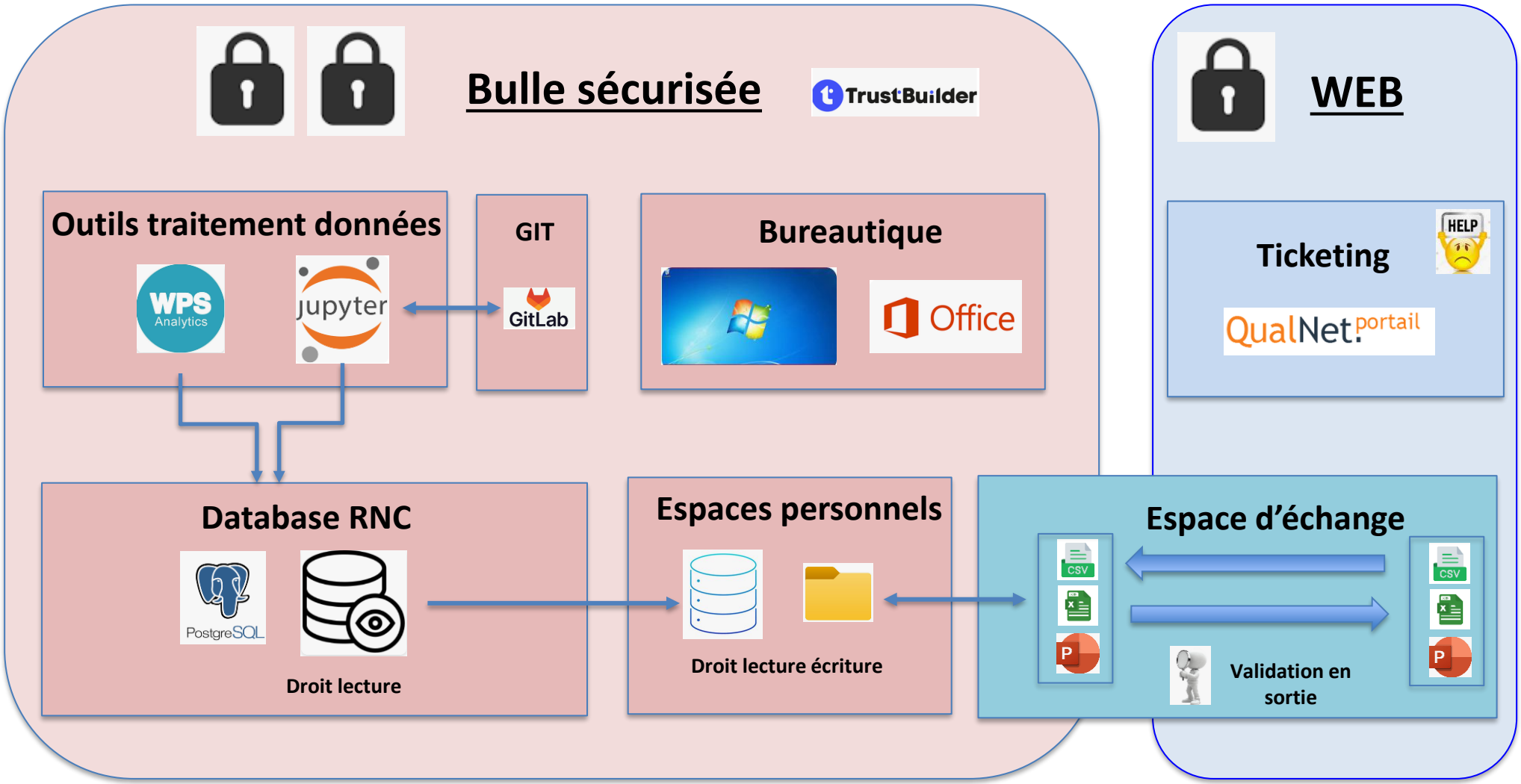
Data engineer / Data manager :
- **Alimentation d'un système de vues** filtrées sur les paramètres du projet



Référent métier INCa du projet :
- **Formation** utilisateurs du projet à l'utilisation de l'EDS-RNC



Présentation de l'espace de travail



Sources de données disponibles dans l'EDS RNC

COHORTE CANCER

SNDS
Système national des données de santé

Extraction annuelle du SNDS de l'ensemble des consommations de soins de personnes atteintes d'un cancer

Tables data management

Institut National du Cancer

REGISTRES DES CANCERS
dans l'EDS RNC



27 registres :

- **16 registres généraux**
- **9 registres spécialisés**
- **2 Registres nationaux pédiatriques**

Tables data management

Institut National du Cancer

Concrètement dans l'espace projet…

COHORTE CANCER

+

REGISTRES DES CANCERS
dans l'EDS RNC

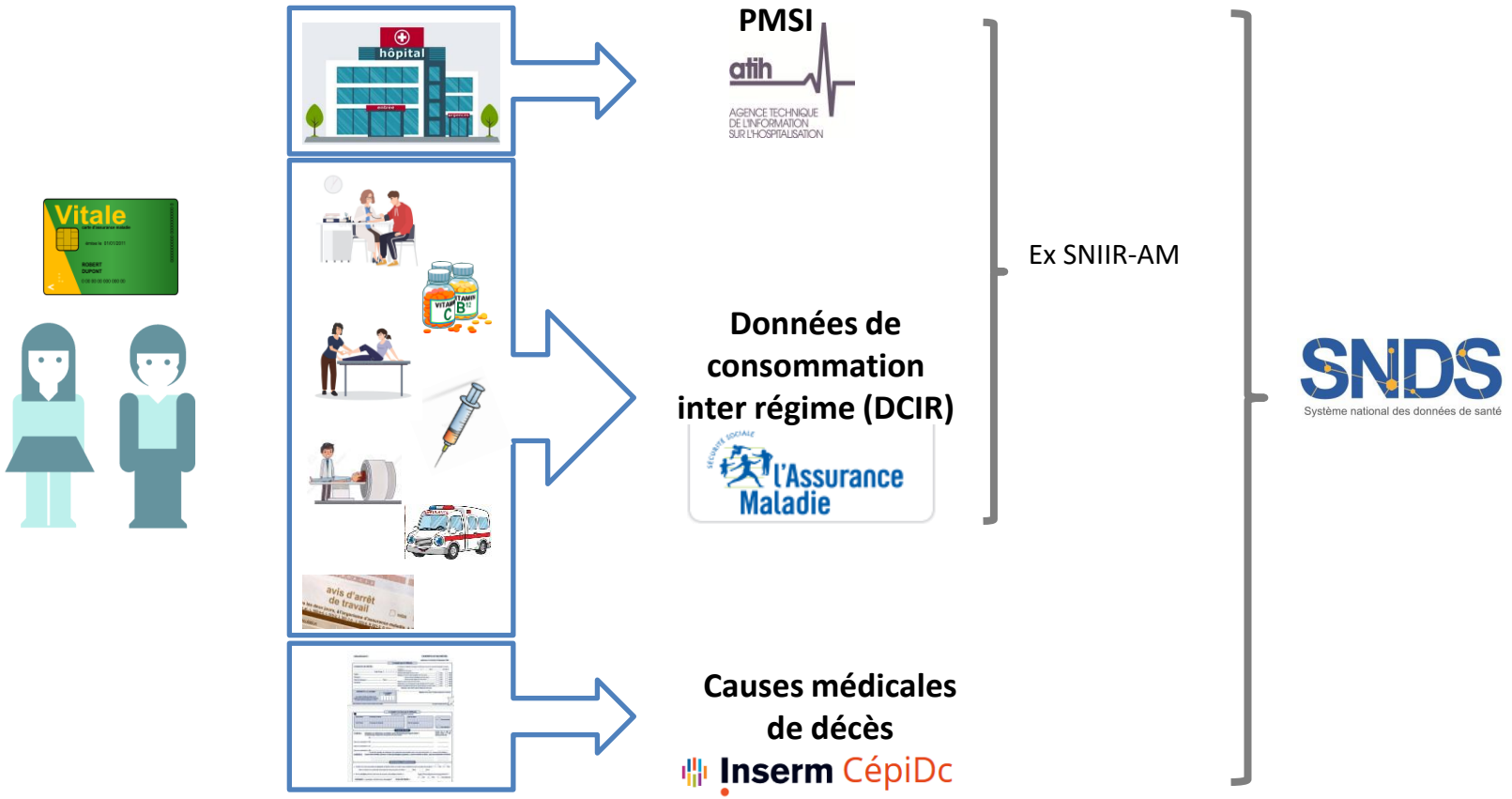
- **FILTRE Liste identifiant**
  Cohorte Cancer en lien avec le projet
- **FILTRE Consommations de soins**
  bornées sur la période autorisée pour le projet

Création Schéma PostGsql spécifique
⇒ **Vues des tables**
de la cohorte cancer filtrées
(sous réserve d'autorisation)

- **FILTRE Liste identifiant**
  Registres Territoriaux en lien avec le projet

Création Schéma PostGsql spécifique
⇒ **Vues des tables**
des registres territoriaux filtrées
(sous réserve d'autorisation)

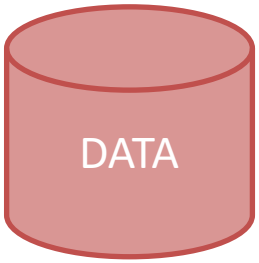
Cohorte Cancer : Les sources de données



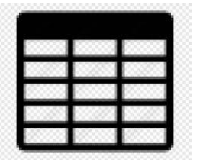
Cohorte Cancer : En quelques chiffres



≃ 14 millions de personnes



≃ 60 milliards d'observations cumulées



≃ 2 600 tables

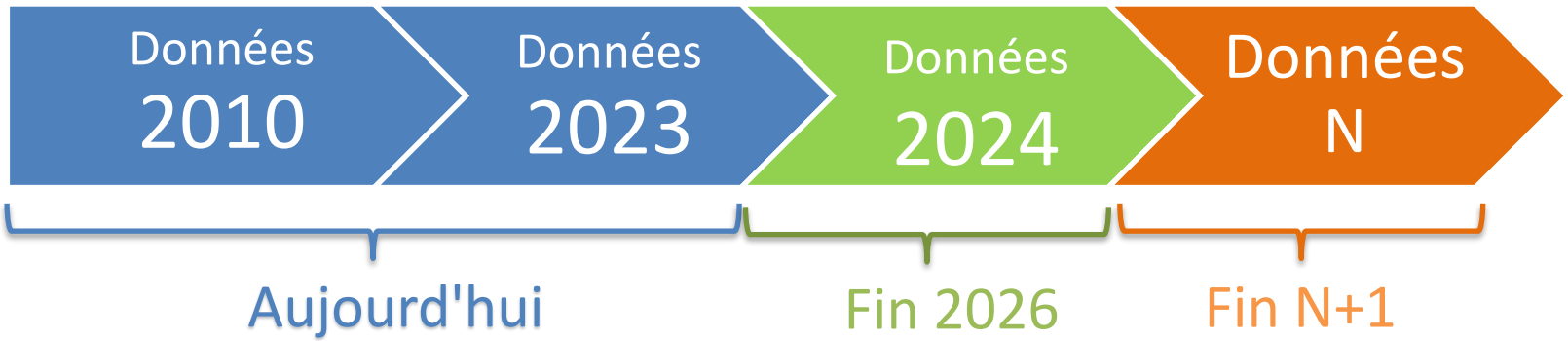
- Données
- Référentiels
- Valeurs

SNDS Système national des données de santé

- Dont ≃ 500 Tables

inca Institut National du Cancer

Données de Consommation de soins disponible



Registres des Cancers :
Les sources de données disponibles dans l'EDS

27 Registres des Cancers

Alimentation annuelle de l'EDS RNC

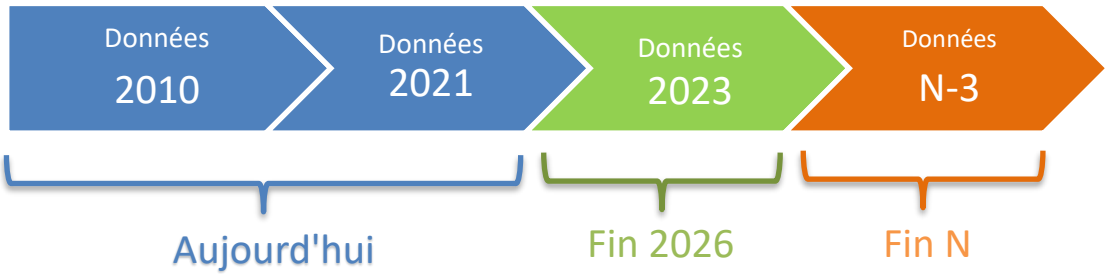
Patients :
- Sexe
- Mois et année de naissance
- Date des dernières nouvelles
- Etat aux dernières nouvelles
- Source d'information de l'état de la personne aux dernières nouvelles

Tumeurs :
- Date de diagnostic
- Commune du domicile au moment du diagnostic
- Base du diagnostic
- Topographie (CIM-O)
- Morphologie (CIM-O)
- Latéralité



≃ 1 million de patients

Période disponible



Zoom sur les Tables INCa Institut National du Cancer : Objectifs

- **Synthétiser informations pour utilisation standardisée**

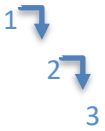
Nettoyage



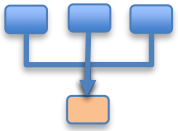
Qualification



Hiérarchisation



Regroupement



Compléter



- **Homogénéiser informations dans le temps :**

2010 2011 2012 ..... 2022 2023 => 2010 2011 2012 ..... 2022 2023

- **Gain de temps**
  - Eviter requête systématique à partir des données sources
- **Possibilité de retour rapide aux données sources**
  - Création d'un numéro technique de consommations unique
- **Simplification utilisation multi-sources**

Zoom sur les Tables inca Institut National du Cancer : Grandes tables

Tables Socio-démographiques

- INC_INFO_GEN_IDT_R — Infos générales
- INC_COM_COD_IDT_R — Communes résidences
- INC_RGM_IDT_R — Régimes
- INC_CMU_IDT_R — CMU et C2S
- INC_IMB_K_R — Référentiel médicalisé ALD historisé et simplifié

Tables localisations

- INC_LOC_IDT_APP_D — Appareil
- INC_LOC_IDT_ORG_D — Organes
- INC_LOC_IDT_ORG_DET_D — Organe détaillé
- INC_LOC_IDT_DGN_D — Tous Diagnostics CIM-10 Cancer

Tables Consommations de soins

- INC_CONS_TOT_D — Toutes consommations de soins en liens avec le cancer
- INC_CONS_MED_K_D — Médicaments anti-cancéreux
- INC_CONS_CCAM_D — Actes CCAM liés au cancer
- INC_CONS_CHIR_D — Chirurgies liées au cancer

Retour Données SOURCES SNDS
Par numéro unique de consommations

Zoom sur les Tables INCa : Consommations

Hospitalisations liées au cancer
- Chirurgies
- Chimiothérapies
- Radiothérapies
- Soins palliatifs
- Séjours médicaux

Médicaments anti-cancéreux
- Ville
- Rétrocession
- Liste en Sus

Radiothérapie
- Cabinet libéral



Tables INCa => Analyse fine des trajectoires de soins par identifiants

Toutes consommations
INC_CONS_TOT_D

Idenitifiant	Type de consommation	Date de début	Date de fin
Ident 1	Séances chimiothérapie 1	23/06/2011	23/06/2011
Ident 1	Séances chimiothérapie 2	15/07/2011	15/07/2011
Ident 1	.....		
Ident 1	Chirurgie	15/09/2011	19/09/2011
Ident 1	.....		
Ident 1	Séances radiothérapie 1	01/11/2011	01/11/2011

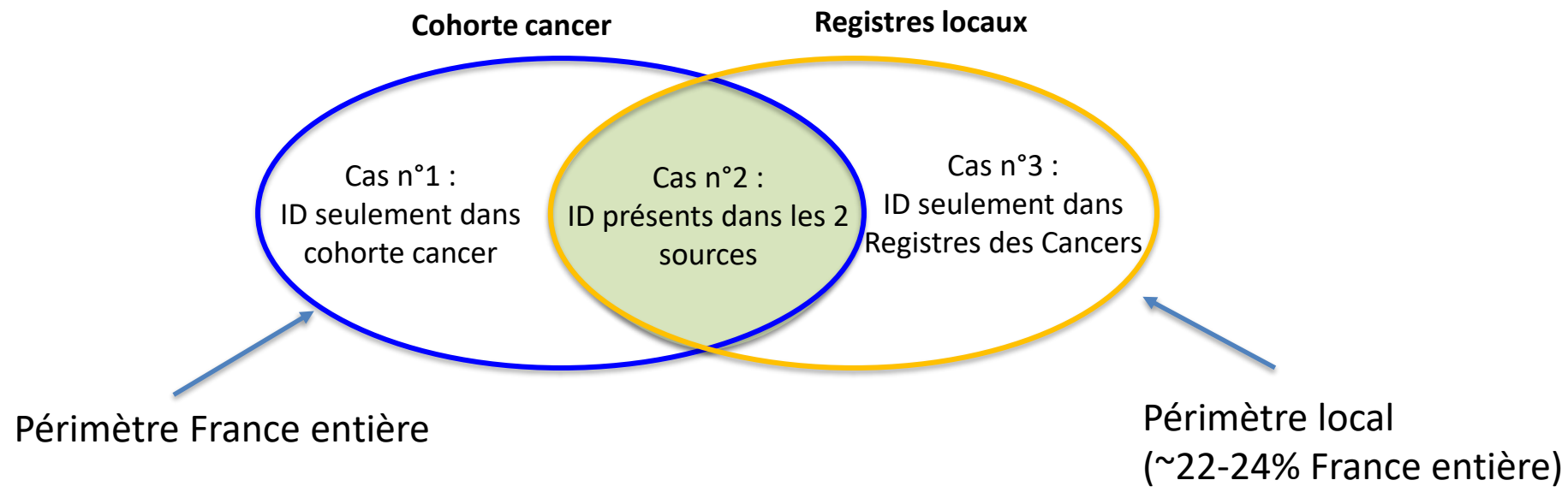
Chirurgies liées au cancer
INC_CONS_CHIR_D
- Diagnostic principal hospitalisation
- Exérèse (oui/non)
- Type du GHM
- Localisation exérèse
- Séjour motivé par un cancer (oui/non)

Actes CCAM liés au cancer
INC_CONS_CCAM_D
- Actes CCAM
- Date précise de l'acte
- Délai depuis l'entrée

Comment croiser les identifiants des 2 sources ?

- **Mise à disposition Vue de correspondance Identifiant :**
  - Identifiant général du RNC **(toujours renseigné)**
  - Identifiant de la Cohorte Cancer **(si accès aux données de la cohorte cancer)**
  - Identifiant des Registres des Cancers **(si accès aux données des registres des cancers)**

- **Cas possibles :**



Tables de Localisations

COHORTE CANCER

INC_LOC_IDT_DGN_D

- Identifiant cohorte cancer
- Informations source localisation
- Date de début et de fin de la consommation de soins ou de l'ALD
- Code CIM-10
- TOP invasif
- TOP métastase
- TOP TEI
- TOP antécédent personnel

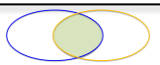
Table de correspondance identifiant

REGISTRES DES CANCERS

TUMEURS

- Identifiant registres locaux
- Date de diagnostic
- Commune du domicile au moment du diagnostic
- Base du diagnostic
- Topographie (CIM-O)
- Morphologie (CIM-O)
- Latéralité

Table de localisation affinée



- Identifiant RNC
- Date et lieu de Consommations de soins ou ALD lors du diagnostic CIM-10
- Ensemble des codes CIM-10 associés au cancer (Codes C et autres que C)
- Date précise du diagnostic
- Description détaillée du diagnostic
  - CIM-O
  - Latéralité

Les données sensibles dans l'EDS RNC



L'INCa a l'autorisation de croiser les données sensibles entre elles.

- **Mois et année de naissance (MM/AAAA)**
- **Commune de résidence**
- **Date de décès (JJ/MM/AAAA)**
- **Date d'exécution des soins**

Pour rappel, l'utilisateur s'engage à

- **Ne pas effectuer de requêtes visant à réidentifier une personne**
- **Ne pas exporter en dehors de l'EDS RNC, un croisement de ces 4 données sensibles par identifiant**
- **Ne pas exporter de données individuelles (avec ou sans identifiant)**

Les outils d'exploitation dans l'EDS RNC

- **Outil de traitements et d'analyses de données**

WPS (World Programming System)









JUPYTER









- **Moteur SQL pour extraction données**

PostGsql





Moteur de base de données relationnelles



TRINO



Moteur de calcul distribué

Les moteurs SQL Postgsql et TRINO

PostGsql

Moteur de base de données relationnelles

- Actuellement moteur utilisé pour exploiter données EDS RNC

Transfert données vers Datalake

S3 Bucket

- Stockage données en fichier parquet
  - Baisse très importante de la volumétrie
- Temps de traitement beaucoup plus rapide
- Spécifiquement conçu utilisation multi-sources hétérogènes
- Possibilité d'utiliser plusieurs types de moteurs

TRINO

Moteur de calcul distribué

Documentation : métadonnées



Schéma Postgsql spécifique métadonnées

Suivi_maj_tables

- Nom table
- Date événement
- Création/mise à jour
- Motif de mise à jour
  - Correctif
  - Changement méthodo
- Commentaire
- Evolution nom / format colonnes
- Evolution volumétrie table

Tables

- Nom table
- Libellé table
- Première année disponible
- Dernière année disponible
- Type données
- Commentaires

Domaines

- Nom table
- Grand Domaines
- Domaines

Colonnes

- Nom table
- Nom colonne
- Libellé colonne
- Première année disponible
- Dernière année disponible
- Type données
- Table de valeur associé
- Commentaires

Formats

- Nom table
- Nom colonne
- Type de formats
- Première année active
- Dernière année active

Sources : Institut National du Cancer, ATIH, SNDS Système national des données de santé, Plateforme des données de santé French Health Data Hub, Registres des cancers dans l'EDS RNC

Documentation :

Macros programmes SAS et fonction Python

- **Mise à disposition**
  - **Macros programmes** sur / **Fonctions** sur
  - **Objectifs** :
    - **Simplifier utilisation données**
    - **Standardiser pratiques**
    - **Gain de temps**
  - **2 types de macros-programmes / Fonctions :**
    - **Technique:**
      - Initialisation session
      - Modification structure tables
    - **Métier :**
      - Faciliter Requête données SQL
      - Spécifique :
        - » Exemple : Calcul du Score de Charlson
  - **Possibilité de générer nouveaux macros programmes et fonctions selon le projet**

Documentation :

Documentation générale outils, méthodes, formation

- **Espace de documentation**
  - **Aide Outils dans l'espace projet**
    - **Utilisation des langages spécifiques selon l'outil**
      - SAS, R, Python
    - **Utilisation Moteur SQL selon l'outil**
      - PostgreSQL
  - **Documentation générale données sources : Cohorte Cancer et Registres des Cancers**
    - **Fiche méthodes**
      - Exemple scripts sur des sujets thématiques
    - **Méthodologie de création des tables** inca Institut National du Cancer
    - **Descriptif général de constitution des données dans l'EDS RNC**
  - **Supports de Formation**
    - **RNC**
    - **SNDS**

Documentation :

Objectifs de centralisation de l'information

- **A terme => Objectif centraliser documentation**
  - **Connexion de WPS Analytics à GitLab**
  - **Création espace centralisé dans GitLab**
    - **Scripts**
    - **Résultats**
    - **Dictionnaire de données**
    - **Méthodes**
    - **Documentation générale**
    - **Support de formation**

    } PAGES
  - **Faciliter partage des informations dans et en dehors de la bulle sécurisée**
  - **Homogénéiser formats de fichiers (Health-DCAT)**

Support utilisateur

- **Mise en place outil ticketing** QualNet portail
  - **Incident** (problèmes connexion, données, etc…)
  - **Demande** (échange spécifiques fichiers, installation package, etc…)

- **Réponses pluridisciplinaires :**
  - **Support Informatique**
  - **Support Métier :**
    - **Médical : Médecin / Pharmacien**
    - **Data : Data manager / Data scientist / Biostat**

- **Doubles objectifs :**
  - **Apporter aides aux utilisateurs**
  - **Créer interactivité :**
    - **Faire évoluer l'ensemble de l'EDS RNC par rapport aux retours des utilisateurs**



Webinaires réguliers à venir.....

- **Mise en place prochainement de webinaires pour utilisateurs de l'EDS-RNC**



- **Sujets à venir :**
  - Bonnes pratiques générales au sein de l'EDS
  - Documentation
  - Aides techniques **scripts** sur sujets spécifiques
  - Evolutions outils au sein de l'EDS
  - Nouvelles sources de données prochainement intégrées…

MERCI POUR VOTRE ATTENTION

Pour toute question : registrenational@institutcancer.fr