

Entrepôts de données, Hébergement, Données massives en santé.

Pr. Renaud SCHIAPPA

Plan du cours

- I. Définitions
- II. Big Data ?
- III. Entrepôt de données cliniques
 - A) ETL
 - B) Architectures
 - C) Données
 - a) Sources et disponibilités
 - b) Formats
 - c) Récupération
 - d) Standardisation et intégration
 - e) I2b2
 - D) Sécurité
 - E) Conseils
 - F) Extraction des données
 - G) Exemples
 - H) Et au CAL ?

I. Définitions

- ❖ **Les entrepôts de données (cliniques) = Integrated Data Repositories (IDR) = Clinical Data Warehouses (CDW)** sont des plateformes utilisées pour l'intégration de plusieurs sources de données au travers d'outils d'analyses spécialisés afin de faciliter le traitement et l'analyse de données massives.
- ❖ **Les Données Massives (en santé) = Big Data**, désignent les gros volumes de données qui alimentent l'activité quotidienne d'un hôpital.

Remarque : Il y en a de + en + de données.

→ On peut les organiser pour :

- ✓ La prise en charge du patient ;
- ✓ Répondre à des questions de recherche.

II. Les 3 caractéristiques des Big Data (en santé) :

- **Volume** : les données proviennent de diverses sources.
- **Vitesse** : les données sont produites à un rythme de + en + soutenu et doivent être traitées vite.
- **Variété** : les données sont sous des formats différents (images, texte...)



Les problèmes avec les Big Data :

- ✓ **90% du volume total** des données ont été produits **ces 2 dernières années** : augmentation de la quantité de données de manière exponentielle et il faut organiser tout ça...
- ✓ **>80%** de ces données ne sont toujours **pas exploitées**.
- ✓ 8.9 Milliards de feuilles de soins informatiques dans la base de la Sécurité Sociale (Sniiram)
- ✓ 2.3 Milliards de Go

⇒ Tout ceci représente beaucoup d'information

Au Centre Antoine Lacassagne : (Artemis : On s'en fout des chiffres du CAL...)

- 3 millions de compte rendus médicaux
- 8 millions d'épisodes (chute, métastase, traitements...)
- 300 000 lignes de chimiothérapies
- Données d'anatomopathologies, radiothérapie, hospitalisations...

→ Problème : **80% de données non structurées** (texte libre) et donc **difficilement exploitables**

→ Seulement **20% de données structurées** : données représentées ou stockées avec un format prédéfini.

III. Entrepôt de données cliniques

Les centres hospitaliers traitent de grands volumes de données tous les jours.

De nombreuses questions sont posées quotidiennement sur :

- La **pratique** de la médecine
- Les **files actives** : elles répondent à des questions très précises.
→ Ex : Combien de patients ont eu un cancer du sein métastatique opéré entre 2010 et 2020, traité par radiothérapie ? quel type de chimiothérapie ?
- Le nombre/répartition des **traitements**
- Questions **cliniques** et/ou **fondamentales**
→ Ex : Est-ce que tous les patients du cancer du sein ont été traités par le même traitement ? Pourquoi certains guérissent et d'autres ont des métastases alors que ces patients semblent être « exactement les mêmes » ? A quel niveau ça se joue ?

⇒ But des entrepôts : répondre à ces questions.

❖ Définition de l'entrepôt de données :

« Un entrepôt de données va recueillir et **regrouper les données** importantes et les **associer** aux patients. Les propriétés des variables, des champs, leurs noms, les règles sont définies, idéalement utilisent un **standard international**. Les données sont solides et ne changeront pas à chaque mise à jour, elles retraceront le parcours du patient et seront à jour »

➔ Explication : si l'on est en communication avec d'autres hôpitaux, il y aura une quantité de données critique. La standardisation va permettre une meilleure communication/efficacité/finesse pour répondre aux questions. Les données sont solides et ne changeront pas à chaque mise à jour. Ces données retracent le parcours du patient et seront à jour.

A) ETL : « Extract – Transform – Load » +++

- Préalable : savoir quelles données on veut extraire.
- Ensuite, il faut les transformer/organiser.
- Enfin, comment les exploiter.

Extraction : connecter les différentes sources de données et **extraire** les données nécessaires.



Problème : **hétérogénéité** des sources de données qui nécessiteront de multiples approches pour la connexion et l'extraction de données.

Ex : Chimiothérapies → on a plusieurs « lignes » par patient.

Alors que le sexe/date de naissance... → on a qu'une seule « ligne » par patient

Transformation : Les données extraites sont transformées dans un format spécifique, défini à l'avance. Cette étape facilite l'intégration et la consolidation des données pour l'étape finale.

→ Est-ce qu'on prend les données telles quelles ou bien est-ce qu'on va les recoder ?

Ex (Chimiothérapie) : Est-ce qu'on prend le nom intégral ? Est-ce qu'il y a des fautes d'orthographe ? → 1 faute d'orthographe donne l'impression qu'il y a 2 traitements différents alors que le traitement est le même. → L'étape de transformation permet de réaliser un contrôle-qualité.



Problème : définition et reconnaissance des **formats** à appliquer, prise en charge des nouvelles données, **évolution des formats** de données en fonction du temps, **interopérabilités** des formats.

Ex (Anatomopathologie) : Avant on disait un « cancer canalaire infiltrant », maintenant on dit un « NST » (No Special Type). Avant on avait toute une base de données avec le nom de « cancer canalaire infiltrant ». D'un coup les « NST » sont arrivées. Du coup ça donnait 2 types de cancer différents sauf qu'il s'agissait de la même pathologie.

Load : Les données sont transformées dans leurs formes/dimensions finales. Il faut charger les données dans l'entrepôt de données de santé.



Problème : La gestion des « anciennes » données VS celles à mettre à jour.

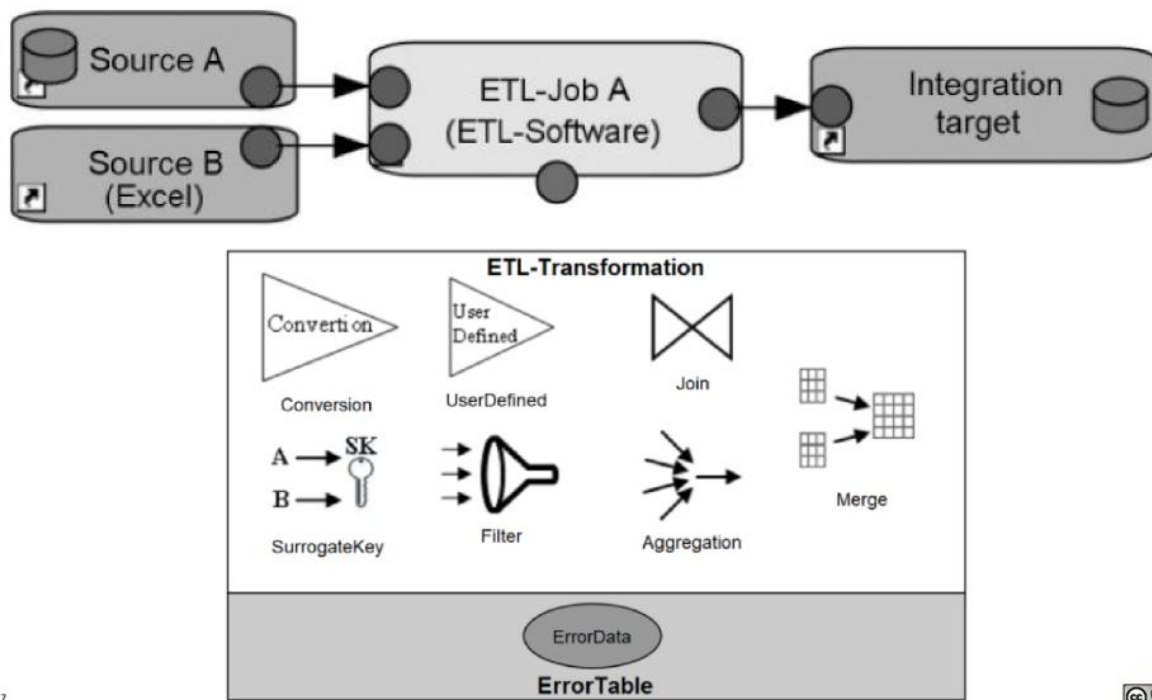
Quand on charge les données, il faut s'assurer que lorsqu'une donnée a changé au cours du temps : est-ce-qu'il faut tout mettre à jour ? Tout supprimer ? Tout recommencer ? Est-ce-que le changement est juste ? Est-ce-que l'ancienne donnée était la donnée correcte ?

Bcp de questions à toutes les étapes...

Le schéma suivant résume tout sur ETL.



ETL : Extract – Transform - Load



Description du schéma

Dans l'ordre, on a :

- 1) Plusieurs sources
- 2) Le travail ETL
- 3) L'envoi des données

Les patients ont un **IPP** (=Identifiant Permanent du Patient càd le nom, prénom...) → on n'a pas le droit de stocker l'identité du patient : on est obligé de pseudonymiser les données.

Ex : Le patient 1 on va l'appeler le patient « A01 ».

Il faut au préalable avoir une **table de correspondance** qui s'applique à toutes les ressources de données. Cette table n'utilise pas forcément toujours la même identification. A chaque fois, il faudra toujours faire attention lorsqu'on veut retrouver son patient.

Ex (chimiothérapie) : il y a un certain nombre de filtres à faire, joindre les données...

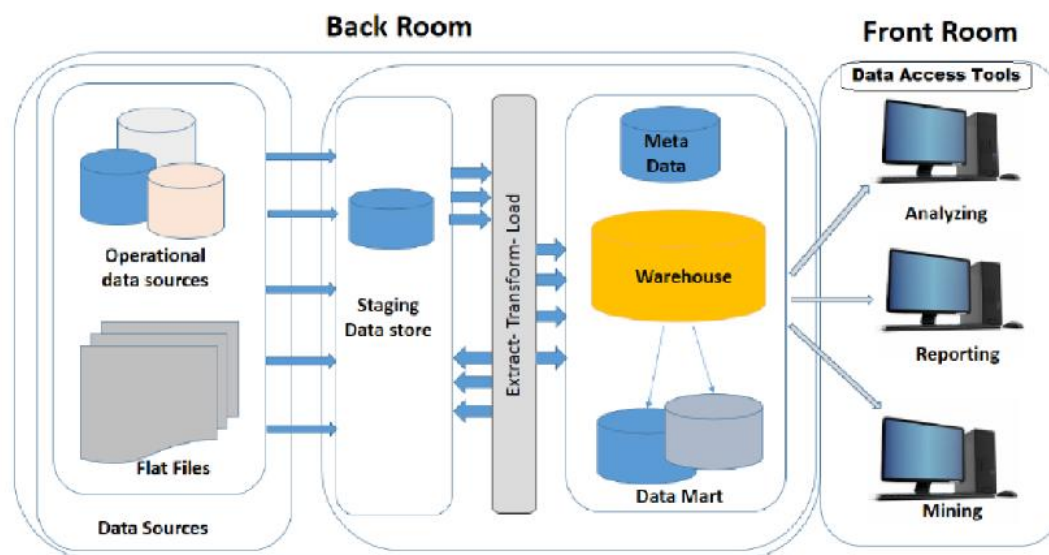
En bas de l'image, vous avez une **table d'erreur**. Elle est très importante car si vous avez 400 000 patients et que l'on en intègre que 50 000 et que les autres ont des erreurs, c'est important de le savoir... Si vous reconnaissez une erreur quelque part, la table d'erreur va vous dire ce qu'il s'est passé.

B) Architectures d'un entrepôt de données

→ Contiennent tout ce qui n'est pas « visible » par l'utilisateur



Architectures d'un entrepôt de données

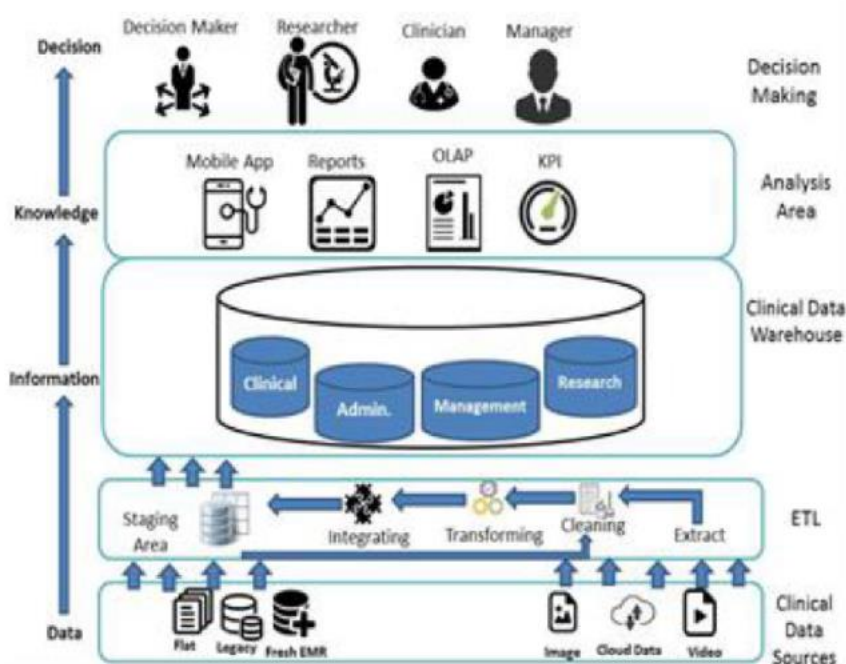


Description du schéma

Dans l'ordre, on a :

- 1) Les sources de données (flat files = souvent des fichiers textes)
- 2) Etape ETL
- 3) Les données sont transférées dans des entrepôts (Warehouse) :
 - **Data Mart** = « Magasin de données »
 - **Meta Data** = Ce sont des « données sur les données »
 - Ex : au lieu de stocker « masculin » / « féminin », on stocke « 1 » et « 2 ».*
 - Ça prend moins de **place** +++ moins de **temps** mis pour le traitement +++
 - Quand on va décoder les données, les Meta Datas vont nous dire que « 1 » c'est « masculin » ; « 2 » c'est « féminin »
- 4) Les utilisateurs finaux vont avoir accès aux données, qu'il peuvent utiliser de différentes manières :
 - Analyser les données (« Analyzing »)
 - Faire des rapports à la direction (« Reporting »)
 - Chercher du savoir à extraire de ces données, poser des questions plus larges... (« Mining »)

Architecture générale d'un entrepôt de données cliniques



Pour un entrepôt de données cliniques général, à la base il y a les différentes sources de données (texte, bases de données historiques, des bases de données actuelles, images, vidéos, applications ex : cloud, montre connectée....)

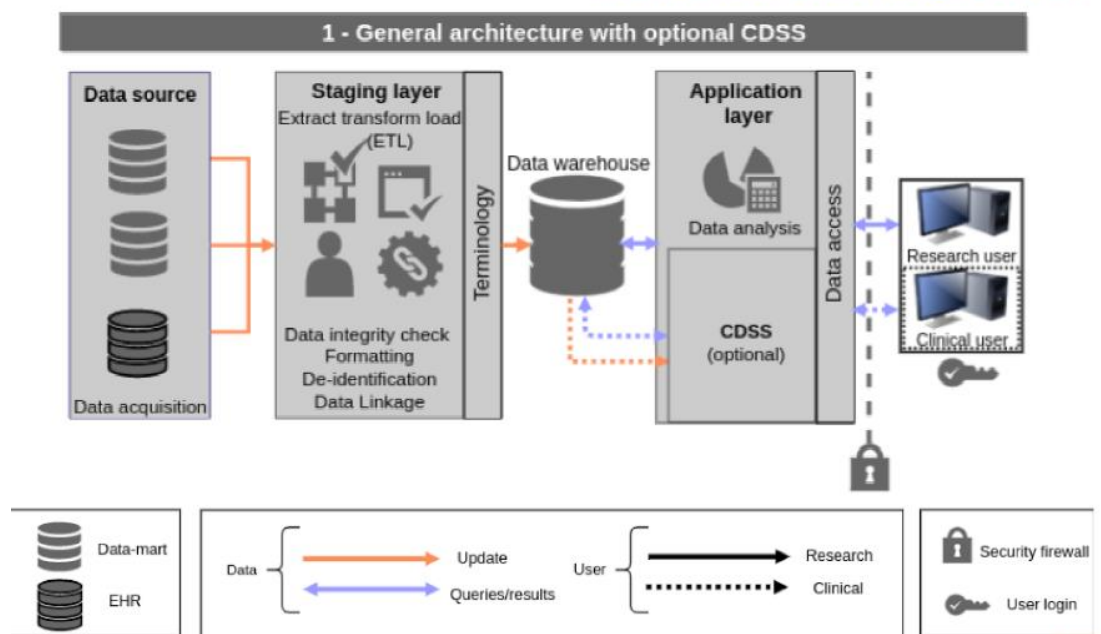
Maintenant on va attaquer les 4 grands types d'architecture...

Les 4 grands types d'architecture :

1) General Architecture with optionnal CDSS (Clinical Decision Support System)

itoine Lacassagne
HLS-DRS, Centre de données
Région Auvergne-Rhône-Alpes

4 grands types d'architecture : General architecture



- Contient la **CDSS** +++ (Clinical Decision Support System) : c'est optionnel, sert à aider les médecins à prendre des décisions cliniques.
- Cette **base de données** est composée de différents **Data Marts** qui sont harmonisés et transférés dans un **CDW** (=Entrepôt).

💡 Data Mart = Magasin de données = ensemble de données cibles, organisées, regroupées et agrégées pour répondre à un besoin spécifique à un métier ou un domaine donné.

→ les utilisateurs peuvent interroger directement le CDW (=Entrepôt) via une interface.

→ Un CDSS apportera une fonctionnalité de prise de décision en plus

- Dans cette organisation, chaque source de données est stockée dans des Data Marts **indépendants**, mais dans le **même établissement**.
- **L'harmonisation** permet de relier et transformer les données.
- L'étape finale de stockage dans une base de données connectée à une interface permet à l'utilisateur/trice **d'accéder aux données** de façon sécurisée.

Explication du schéma

- Tout à gauche : la récupération des données
- ETL
- Entrepôt
- Des applications
- Tout à droite : les utilisateurs, qui en fonction de leurs droits et leur profil, vont avoir accès à un certain nombre de données. Ces données peuvent être :
 - Agrégées
 - Discrètes

On voit sur le schéma des flèches qui indiquent des mises à jour, queries, results qui sont à double sens et peuvent donc revenir. Parfois elles ne sont pas à double sens, elles sont bloquées, càd qu'on ne pourra pas redescendre.

2) Modèle Biobank driven

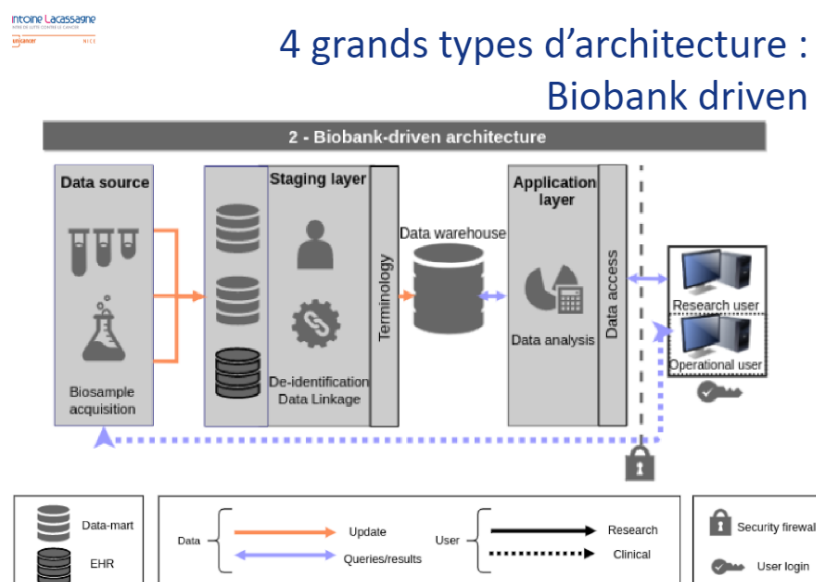
💡 Qu'est-ce qu'une biobanque ? → Dans un hôpital, c'est ce qui va stocker tous les échantillons biologiques.
Ex : prélèvements sanguins, urine, lames d'anatomopathologies...

➔ Des données en découlent.

Le modèle Biobank driven est assez similaire à celui de l'architecture générale. En revanche, cette fois-ci, on s'appuie surtout sur des **échantillons biologiques**.

L'intégration des données cliniques relatives aux échantillons se fait au moment de la partie « **transformation** ».

L'avantage : Accès direct aux **données brutes** des échantillons pour faire des **contrôles qualités** → *permet de comprendre pourquoi certains échantillons sont biaisés.*

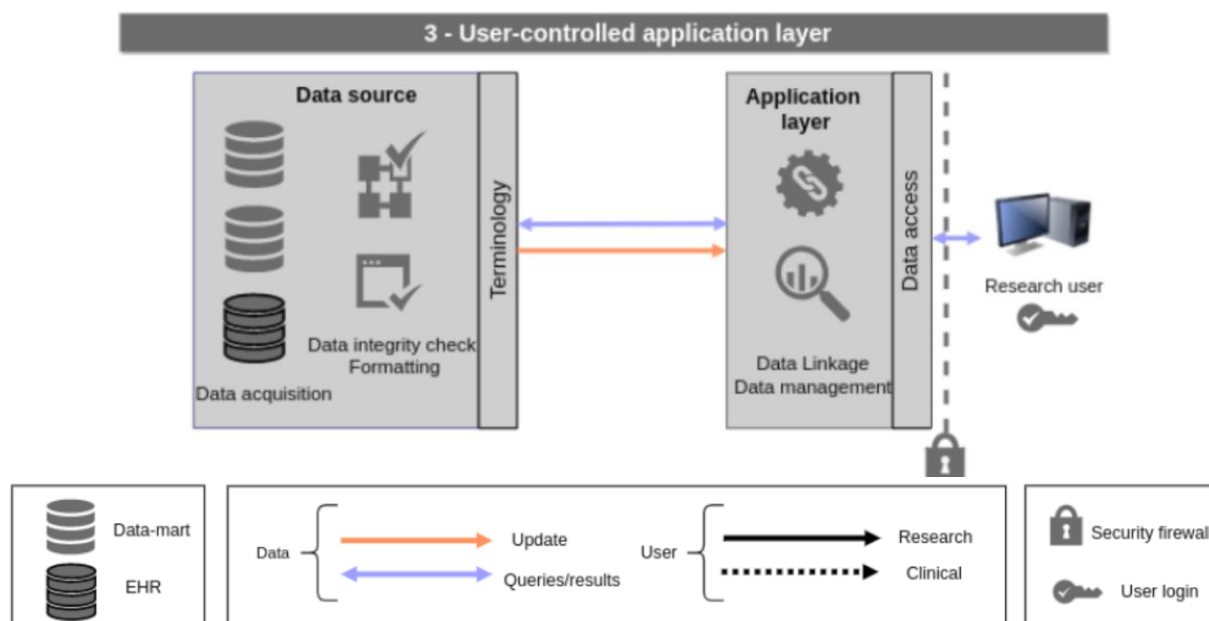


Explication du schéma :

Ici on part de sources de données (échantillons, du vivant qui est récupéré). Ce ne sont pas des bases de données. Le reste est similaire au modèle précédent : ETL, Base de données, les applications, les utilisateurs. Mais cette fois-ci, la flèche est à double sens entre « Operational User » et « Data source »/ « Biosample », ce qui n'était pas le cas pour General Architecture.

3) Modèle User-controlled

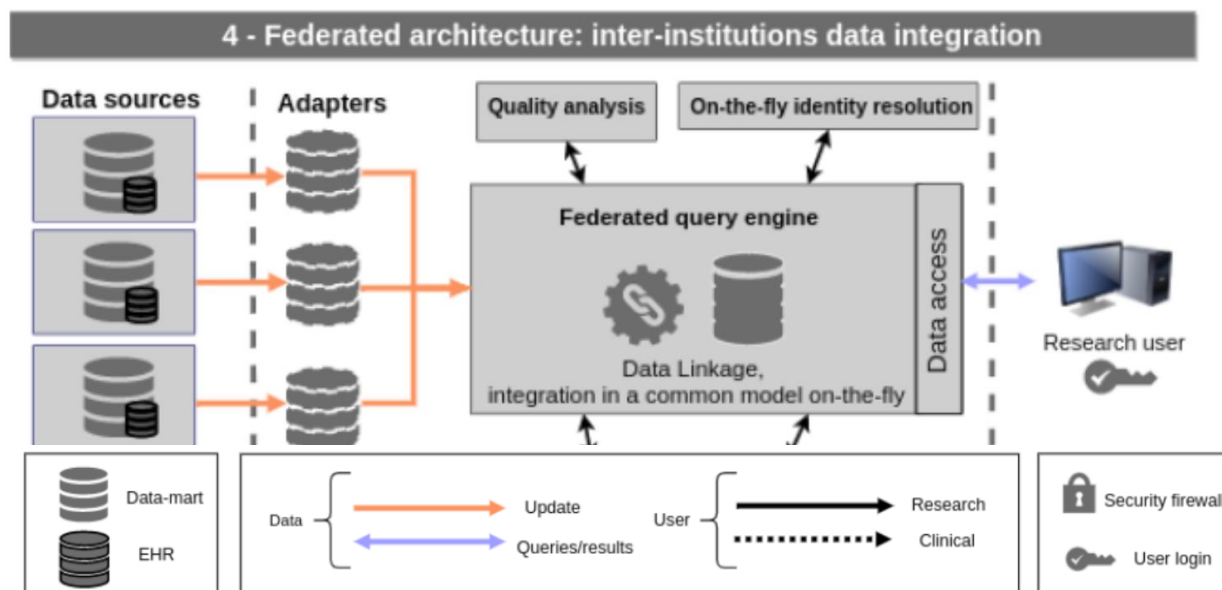
- Il n'y a **pas d'étape de transformation** particulière
 - **Pas d'entrepôt de données « central »** regroupant les Data Marts.
 - Les données sont **pré-traitées** et **intégrées directement** à partir des **données sources** seulement quand un utilisateur(-trice) en fait la requête.
 - Il y a plusieurs bases de données utilisées en même temps l'agrégation des données se fait en même temps.
- Avantage : rapidité
- Inconvénient : exhaustivité des données



4) Federated architecture

- très « à la mode »

- Les données sont récupérées à partir de **plusieurs établissements différents** qui vont mettre en commun leurs ressources.
- Chaque institution (hôpital) choisit les données qu'elle souhaite partager en utilisant un **adaptateur commun** qui va pré-traiter ces données (cet adaptateur agrège les données en dehors du centre)
- Les données sont intégrées en direct dans un **entrepôt de données « virtuel »** (centralisé en dehors des institutions)
- Les données ne sont présentées pour **l'analyse et l'exploitation** seulement lors de la session de l'utilisateur(-trice) c'est-à-dire qu'à la fin de la session, les données sont supprimées (par souci de sécurité).



Explication du schéma :

- Le *verrou* protège les données ;
 - Les *adaptateurs* pseudonymisent les données : on ne peut pas revenir en arrière ;
 - Au sein du Federated query engine : **toutes les données sont agrégées** ;
- cet Federated query engine contient les données de tous les hôpitaux, il est donc extrêmement **puissant**.

Conclusion sur les 4 types d'architecture

Ces différentes architectures offrent différents outils d'analyse, de logiques de présentation et les interfaces de requêtes sont différentes en fonctions du type d'utilisation qu'on veut et des utilisateurs :

- ❖ **Chercheurs(-ses)** : cherchent des traits cliniques qui permettent d'identifier des cohortes répondant à des questions précises → **toutes** les architectures leurs sont utiles
- ❖ **Médecins** : aide à la prise de décision pour les traitements, interventions, risques pour un(e) patient(e) → La 1ère architecture avec **CDSS** est la plus appropriée.

A retenir +++ : Toutes les architectures ne peuvent pas correspondre à tous les profils.

« Il faut toujours réfléchir en amont à l'utilisation finale de l'entrepôt nécessaire... »

C) Données

Identifier les sources de données constituera le socle du CDW.

Elles varient de format, type, organisation, volume en fonction des départements :

- **Laboratoire** : volume important de résultats biologiques
- **Diagnostic** : souvent non structuré
- **Démographiques** : structurée au début, mais le suivi peut poser des soucis
- **Traitements** : chimiothérapie, radiothérapie (ira, curie, proton, contact etc...), thérapie ciblée, hormonothérapie, immunothérapie : chaque traitement a ses propres caractéristiques.
- **Clinique** : tout « le reste » contenu dans les dossiers médicaux (rechutes, suivi des traitements, habitudes de vie, comorbidités, toxicités, antécédents personnels et familiaux) : pratiquement jamais structuré.

a) Sources et disponibilités

Chaque source de données cliniques a souvent :

- Sa propre organisation
- Son propre standard
- Son propre logiciel d'exploitation
- Son propre « langage »

C'est une étape cruciale d'identification et d'analyse de toutes les sources et spécificités (étape très chronophage). La disponibilité des données en fonction des sources dépend de leur **complétude** et **du design des sources**. Les systèmes « historiques » peuvent ralentir le process car non prévues pour des requêtes fréquentes.

L'augmentation du volume des données cliniques demande la mise en place de nouveaux liens entre les données historiques et les nouveaux systèmes de données.

b) Format

Les types sont très variés :

- Texte (structuré ou non structuré)
- Images
- Vidéo
- Echantillons biologiques
- Réponses
- Puces ADN/ARN
- Données externes (questionnaires, objets de santé connectés)

Les formats le sont également :

- Numérique
- Qualitatif
- Quantitatif
- Séquentiel

c) Récupération

Le traitement des données suivant l'ETL est composé de plusieurs étapes :

1. Extraction (automatique ou manuelle) des données à partir des différentes sources.
2. Anonymisation (optionnel) et attribution d'un identifiant unique.
3. Transformation et standardisation : les données sont d'abord contrôlées à la recherche d'éventuelles erreurs, transformées dans le format cible.
4. Mapping avec la terminologie standard utilisée.
5. Mapping des données entre les différentes sources.
6. Chargement dans la CDW (mise à jour ou ré import total).

d) Standardisation et intégration


Certaines données sont standardisées à la saisie :

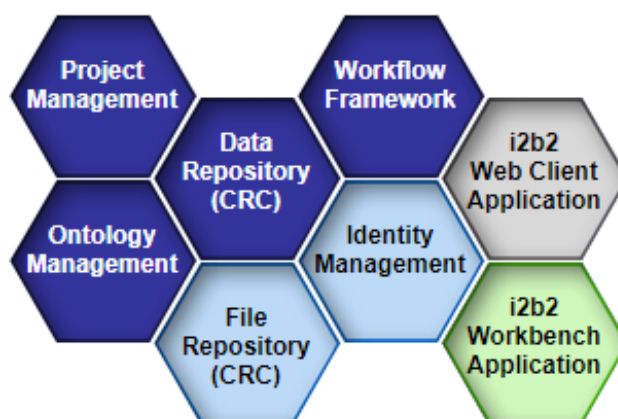
- Utilisations les plus courantes : Classification Internationale des Maladies (CIM-10) et Systematized Nomenclature Of Medicine Clinical Terms (SNOMED-CT)

Utilisation d'un Common Data Model (CDM) :

- Un schéma d'organisation permettant l'interopérabilité et le partage des données.
 - Une utilisation d'un CDM déjà utilisé par d'autres institutions permet de s'affranchir d'une étape importante de sélection des logiciels, plateforme etc...
- ✓ Cela reste une étape cruciale qui peut prendre plus de 90% du temps de construction de l'entrepôt.
 - ✓ **L'Integrating Biology and the Bedside (i2b2)** est un des CDM les plus utilisés.

Key

| | | | | | |
|---|----------------|---|--------------------|---|-------------------|
|  | i2b2 Core Cell |  | i2b2 Optional Cell |  | Workbench/Plug-in |
|  | Web Client |  | CRC Plug-in | | |



Définitions : +++

Project management : sécurité, identification des utilisateurs/trices, rôles.

Ontology management : gère la terminologie.

Data repository : gère les données structurées, permet l'interrogation et la visualisation des données.

File repository : stocke les « gros » fichiers (images, puces)

Workflow Framework : gère les interactions entre les différentes « hives ».

Identity management : anonymisation des patients.

Web client application : permet aux utilisateurs/trices d'interroger le CDW.

Workbench : application permettant d'analyser les données de façon plus précise.

D) Sécurité

Il est crucial de fixer les règles de sécurité des données dès la conception de l'entrepôt :

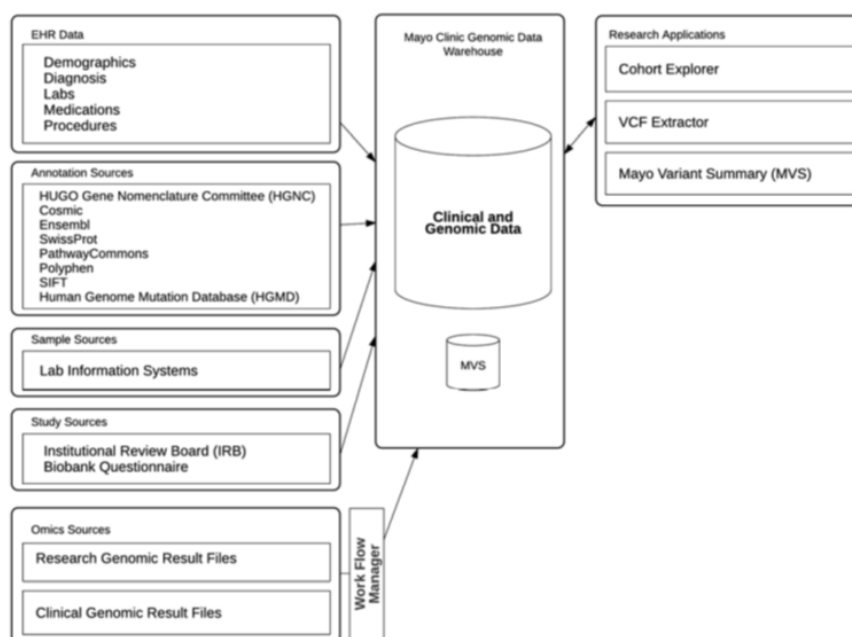
- Comment sont stockées les données ? Physiquement sur site ? Prestataire externe dans le « cloud » ? Est-il certifié Hébergeur de données de santé ?
- Quelle est la politique de sauvegarde ? Sites multiples ? Protection vol physique ou électronique ?
- Comment est contrôlé l'accès aux données ? Qui a les droits ? Qui décide des types d'accès ?
- Est-ce que les données des patients sont anonymisées ? Pseudonymisées ? En clair ?
- Est-ce que chaque accès aux données est tracé ? Des audits de sécurité réalisés ?

E) Conseils du Professeur

- 1) Penser sur le long terme pour assurer la longévité du projet : s'affranchir de contraintes de formats propriétaires permettra la réutilisation du système.
- 2) Commencer par choisir l'architecture souhaitée basée sur les besoins des utilisateurs/trices.
- 3) Sélectionner un CDM déjà utilisé par d'autres institutions afin de bénéficier de l'aide et l'expérience d'une plus grande communauté.
- 4) A chaque fois que cela est possible : adopter une terminologie. Essayer de l'appliquer dès le début du traitement des données et rajouter des terminologies plus spécifiques lorsque le scope du projet s'élargit.
- 5) Définir la fréquence des mises à jour, le détail du processus ETL, le niveau d'automatisation.
- 6) Communiquer à chaque étape tout en consultant régulièrement les utilisateurs/trices.

F) Extraction des données

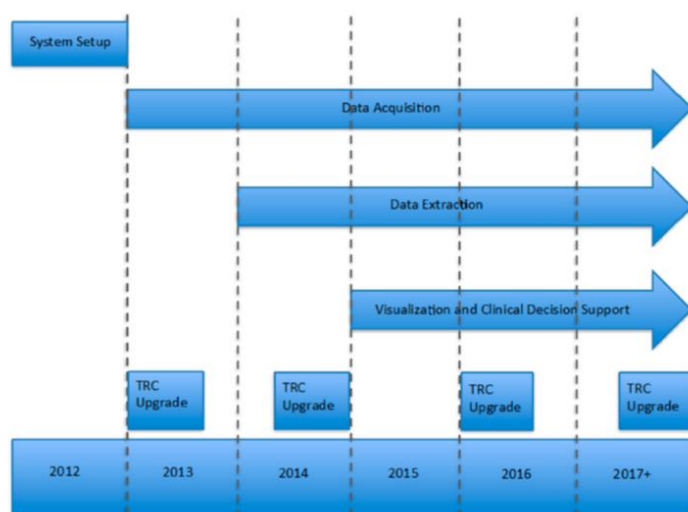
- **« Trop d'attributs »** : les données structurées des patients peuvent être reliées à un grand nombre de variables, il faudra sélectionner précisément les variables d'intérêt.
- **« Plusieurs valeurs »** : certaines variables ont des valeurs répétées par patient (toxicités, comorbidités), ce qui pose des soucis de taille variable de dimensions (un patient pourra avoir une ligne ou plusieurs : comment traiter un patient avec une seule chimiothérapie vs un patient avec 5 lignes ?).
- **« Données temporelles »** : comment placer la rechute au bon moment (et pas avant le diagnostic principal) ? Importance de mettre en place des règles et de regarder ce qui a pu être fait par ailleurs.
- **Effectuer des évaluations de la qualité des données** : permet d'identifier les problèmes à la source des données plutôt que de les régler dans l'entrepôt final.



A gauche on voit les données, qui sont regroupées dans le Warehouse et qui pourront être utilisées par les chercheurs.

EHR: Electronic Health Records

VCF: Variant Call Format



Ceci est la timeline de l'entrepôt

TRC=logiciel qui est mis à jour

Il a fallu 3 ans entre le début de l'étude et la première visualisation)

Table 1. Mayo Oracle Translational Research Center (TRC) implementation resources.

| Area | Role | Number of Members |
|--------------------|-------------------------|-------------------|
| IT | Database Administrator | 2 |
| IT | Data Pipeline Architect | 2 |
| IT | Architect | 2 |
| IT | Programmer | 6 |
| IT | Support Analyst | 2 |
| Bioinformatics | Bioinformatician | 2 |
| Biostatistics | Data Scientist | 2 |
| Project Management | Project Manager | 2 |
| Executive | IT Executive | 2 |
| Executive | Clinician | 1 |

IT: Information Technology.

Table 2. Mayo Oracle TRC production hardware.

| Component | Quantity | CPU | Memory | Disk Space | Manufacturer |
|------------------------------|----------|--------------------------|--------|------------|-------------------------------------|
| Oracle Exadata Database | 2 | Intel Xeon X5675 24 Core | 192 GB | 19 TB | Oracle, Redwood City, CA, USA |
| Application Server | 2 | Intel Xeon X5687 16 Core | 24 GB | 500 GB | Hewlett-Packard, Palo Alto, CA, USA |
| Oracle ZFS Storage Appliance | 1 | N/A | N/A | 2.5 TB | Oracle, Redwood City, CA, USA |

« Une vingtaine de personnes s'occupent d'implémenter les ressources, volumes de données pas trop importants mais assez honnêtes »

Table 4. Mayo Oracle TRC post-implementation resources.

| Area | Role | Number of Members |
|--------------------|------------------------|-------------------|
| IT | Database Administrator | 1 |
| IT | Architect | 1 |
| IT | Programmer | 2 |
| IT | Support Analyst | 2 |
| Bioinformatics | Bioinformatician | As-needed |
| Project Management | Project Manager | 1 |

« Une dizaine de personnes pour la post-implémentation et des millions de lignes de données pour l'entrepôt de données. »

On arrive à la fin de cet exemple sans aucune autre remarque dessus.

On arrive également bientôt à la fin de ce cours alors courage.

Un autre exemple : George Pompidou University Hospital Clinical Data Warehouse

Ils utilisent i2b2.

Ils ont défini 3 niveaux d'accès aux données :

- Premier niveau : Seulement accès aux données agrégées répondant aux critères de sélection (e.g. : combien de patientes triples négatives opérées entre 2010 et 2020).
- Deuxième niveau : cohortes anonymes avec les données détaillées.
- Troisième niveau : Cohorte avec toutes les données, non anonyme.

Artemis : Les images qui suivent sont là pour illustrer, n'apprenez pas ça SVPPPP

Table 5. Mayo Clinic genomic data warehouse data statistics.

| Data Type | Total |
|---------------------------------------|----------------|
| Samples with Genomic Results | 11,734 |
| Research Samples | 9712 |
| Clinical Samples | 2022 |
| Research Studies with Genomic Results | 71 |
| Total Variant Count | 8,612,759,579 |
| Total Omics Results (Rows) | 68,431,547,534 |
| Total Patient Count | 9,283,510 |
| Total Subject Count | 149,714 |

| | September 2009 | December 2013 | July 2016 |
|--------------------------------------|-------------------|------------------|-----------|
| Concepts | | | |
| Biology (thousands) | 7.29 | 9.1 | 11.2 |
| Diagnostic codes (ICD10) (thousands) | 21.36 | 39.91 | 40.25 |
| Drugs (thousands) | 31.36 | 33.67 | 41.6 |
| Data facts | | | |
| ICD Diagnosis (millions) | 1.87 | 2.94 | 7.67 |
| Clinical items (millions) | 20.8 | 61.1 | 122.2 |
| Laboratory results (millions) | 62.8 | 98.0 | 124.3 |
| Drug orders (millions) | 0.95 | 3.2 | 6.4 |
| Text reports (millions) | 0.16 | 2.36 | 3.7 |

Entre 2009 et 2016, on est passé de 21 millions à 41 millions.

L'entrepôt permet de réaliser de nombreux projets : (*Artemis* : j'aurais dit qu'on s'en fout...)

| Année | Nbr de projets | Nbr de départements à l'origine des projets | Projets épidémiologie clinique | Projet département de santé | Recherche clinique |
|-----------|----------------|---|--------------------------------|-----------------------------|--------------------|
| 2011 | 13 | 5 | 8 | 5 | 0 |
| 2012 | 4 | 4 | 1 | 3 | 0 |
| 2013 | 13 | 10 | 8 | 4 | 1 |
| 2014 | 22 | 11 | 14 | 5 | 3 |
| 2015 | 22 | 10 | 9 | 13 | 0 |
| Total (%) | 74 (100%) | 17 (71%) | 40 (54%) | 30 (41%) | 4 (5%) |

Summary table**What was already known on the topic**

- Reuse of health data is a major issue for better patient care management and facilitates clinical and epidemiological researches
- Hospital have deployed clinical data warehouses to facilitate reuse of health data
- Reuse procedures have to guarantee both easy access for clinicians and patient privacy

What this study added to our knowledge

- Deployment of a CDW is a long-term process from conception to end-user CDW adoption.
- Clinicians are not prepared to formulate complex queries and navigate through the different nomenclatures that populate a CDW.
- Strong collaboration between clinicians, biomedical informatics, biostatistics and epidemiology specialists is needed to complete successfully research project using a CDW.

« Ceci est un article en Anglais qui résume l'utilité de l'utilisation des données de santé »

Et au CAL ?

(Artemis : j'aurais dit qu'on s'en ble...)

C'est le début du projet de lancement de la plateforme de données.

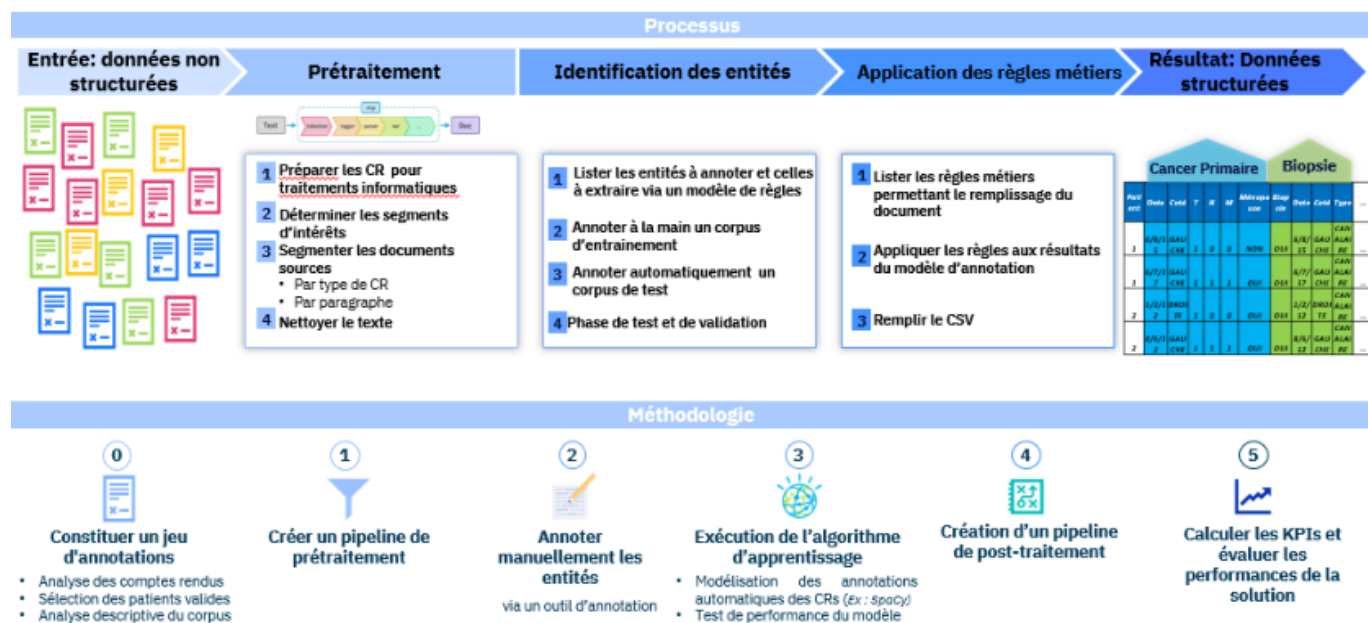
Analyse des sources de données disponibles.

Mise en place d'une structuration automatique des données grâce à des algorithmes d'intelligence artificielle (projet RUBY) : au lieu de requêter les données textuelles, des données structurées seront présentées directement pour intégration à la plateforme de données de santé.

Projet RUBY

Projet RUBY

Mise en œuvre



Confidentiel Centre Antoine Lacassagne & IBM France

Projet RUBY

Mise en œuvre

Annoter manuellement des entités: des exemples d'annotation

Patient

4 Quoiqu'il en soit, à l'examen, elle a un cancer du sein gauche, à l'union des quadrants supérieurs, qui fait environ 2 cm cliniquement, mobile, dans des seins qui ne sont pas très volumineux.

5 Les aires ganglionnaires sont libres.

6 Elle a eu une biopsie qui montre un carcinome canalaire infiltrant de grade I, RO+, RP-, Expression_HER2 Her2 ++.

7 On va se mettre en rapport avec le [NOM_ANONYMISE] pour confirmer la nature bénigne de ces lésions osseuses et non pas métastatiques.

8 En fonction de cela, on prévoit une consultation en chirurgie, un bilan pré-opératoire et une consultation anesthésie.

Anapath: Segment Conclusion

1 CONCLUSION : Mammectomie partielle centrale comportant la PAM, pour tumeur de 2 cm de grand axe correspondant à un carcinome canalaire infiltrant moyennement différencié de SBR II (2.3.1) avec début d'envahissement profond de la région aréolaire.

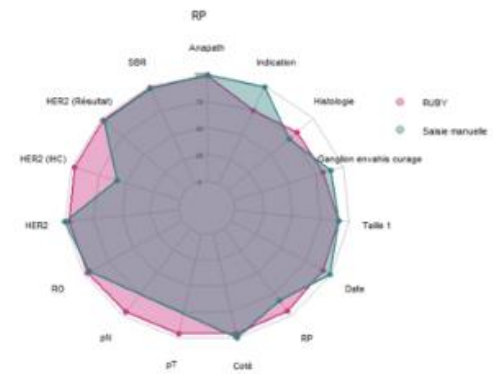
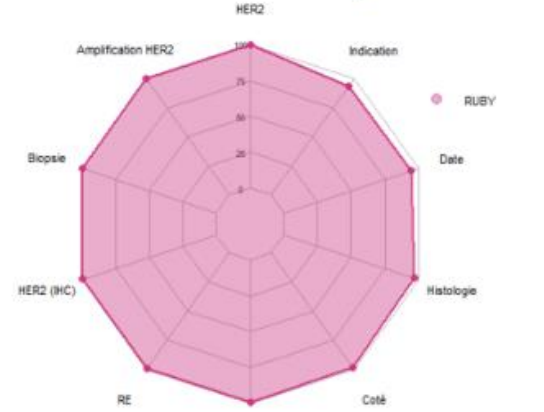
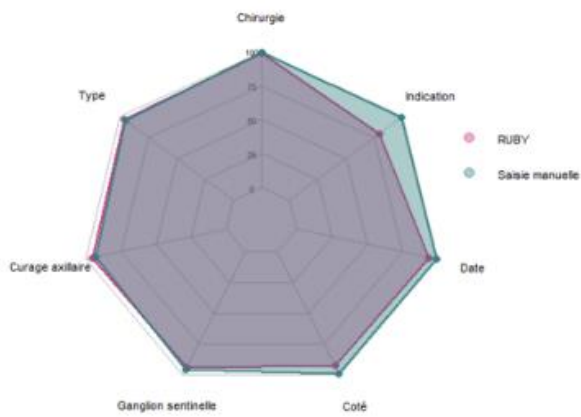
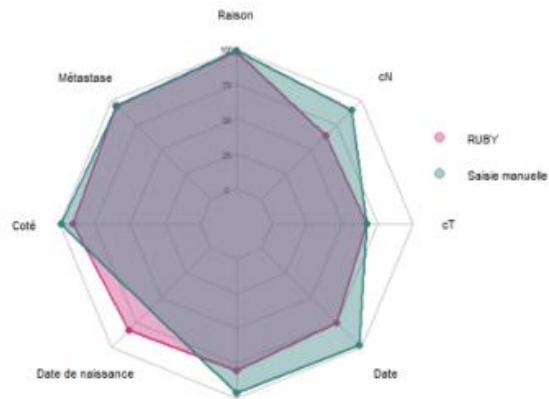
2 Exérèse largement satisfaisante.

3 1 ganglion métastatique sans rupture capsulaire, sur les 7 îlots solés dans le curage des 1er et 2ème étages axillaires droits (1+/7).

4 Fibrome molluscum axillaire.

- Après avoir choisi le corpus, l'annotation se fait fichier par fichier
- En sélectionnant le ou les termes à annoter, l'annotation est réalisée en choisissant la catégorie relative au(x) terme(s) sélectionnée dans un menu déroulant
- Les carrés colorés au-dessus des mots ou phrases correspondent aux entités identifiées dans le CR
- Les entités identifiées ne peuvent pas se chevaucher: un mot fera partie d'une seule entité sur un CR
- Chaque CR a ses propres entités à identifier, mais un CR peut être utilisé pour rechercher de l'information sur d'autres CR.
- Dans l'exemple sur **Consultation**, les entités en vert correspondent aux entités de **Biopsie**

Projet RUBY



FIN DU COURS !!!!