RONEO N°1 : ENTREPÔTS DE DONNEES, HEBERGEMENT ET DONNEES MASSIVES EN SANTE

Corporation des Carabins Niçois

UFR Médecine
28, av. de Valombrose
06107 Nice Cedex 2

http://carabinsnicois.fr/
roneo.c2n@gmail.com

Date et heure : distanciel
Professeur : Schiappa
Nombre de pages : 14
Ronéiste : Alexandre Mistura et Camilya

SOMMAIRE

Salut je me présente, je suis Alexandre Mistura, je suis en deuxième année de médecine et c'est moi qui vais ronéiser les 3 ronéos de Santé Numérique cette année. Ne lâchez rien, même si c'est difficile, même si vous avez des moments de moins bien (ça arrive à tout le monde), gardez toujours votre objectif en tête, n'oubliez pas pourquoi vous vous battez depuis le début. Pour finir, vos notes aux examens blancs et séances tutorat ne définissent pas vos notes que vous aurez le jour du concours, on a dû souvent vous répéter ça mais c'est la vérité (#vrai2vrai).

PS : je paye un pain au chocolat où vous voulez dans Nice aux 3 premiers ou aux 3 premières qui trouvent LA faute d'orthografe que j'ai dissimulée dans ce paragraphe et qui m'envoient un message sur Messenger (Alexandre Mistura).
Le professeur Schiappa n'ayant pas donné cours en amphi, j'ai fait la ronéo avec le fichier PDF du professeur, si des passages ne sont pas clairs n'hésitez pas à aller voir la fiche de la tutrice sur le forum du tutorat niçois. Bon courage !

Entrepôts de données, Hébergement et Données massives en santé

I. *Définitions*

Les Entrepôts de données cliniques, aussi appelés Integrated Data Repositories (IDR) ou Clinical Data Warehouse (CDW) sont des plateformes utilisées pour l'intégration de plusieurs sources de données autravers d'outils d'analyses spécialisés afin de faciliter le traitement et l'analyse de données massives.

Les données massives en santé appelées **big data** désignent les gros volumes de données qui alimentent l'activité quotidienne d'un hôpital.

Elles sont régies par **3 caractéristiques** ou dimensions :
⇒ **Volume** : Les données proviennent de diverses sources.
⇒ **Vitesse** : Les données sont produites à un rythme de plus en plus soutenu et doivent être traitéesrapidement.
⇒ **Variété** : Les données sont sous des formats différents.

II. *Le problème des Big Data*

→ 90% du volume total des données ont été produites ces 2 dernières années et plus de 80% ne sont toujours pas exploitées.

On compte notamment 8.9 milliards de feuilles de soins (Système national d'information inter-régimes de l'Assurance maladie ou Sniiram) soit 2.3 milliards de Giga-octets en volume.

Pour exemple au Centre Antoine Lacassagne (CAL) on compte :

- 3 millions de comptes rendus médicaux au CAL
- 8 millions d'épisodes
- 300 000 lignes de chimiothérapies
- Des données d'anatomopathologies, de radiothérapie, d'hospitalisations, etc…

On appelle « données structurées » les données représentées ou stockées avec un **format prédéfini** (≈ 20% des données).
→ Malheureusement au CAL, comme ailleurs : 80% des données sont non structurées (texte libre) et donc difficilement exploitables.

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 active
- Traitements/répartition
- Questions cliniques et/ou fondamentales

➥Un entrepôt de données permettra de répondre à toutes ces questions

Qu'est-ce qu'un 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 »

A. Extract – Transform – Load (ETL) +++

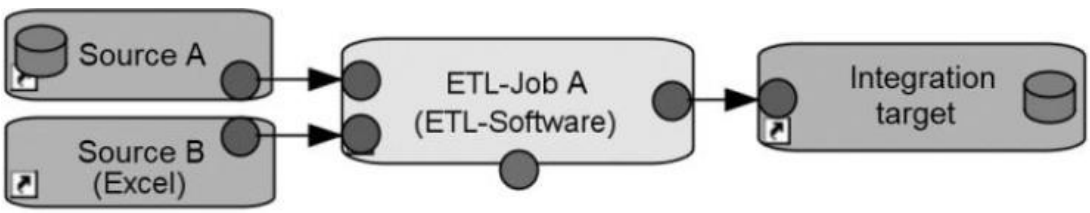
→ Extract : connecter les différentes sources de données et d'extraire les données nécessaires.
↳ Problématique : hétérogénéité des sources de données qui nécessiteront de multiples approchespour la connexion et l'extraction des données.

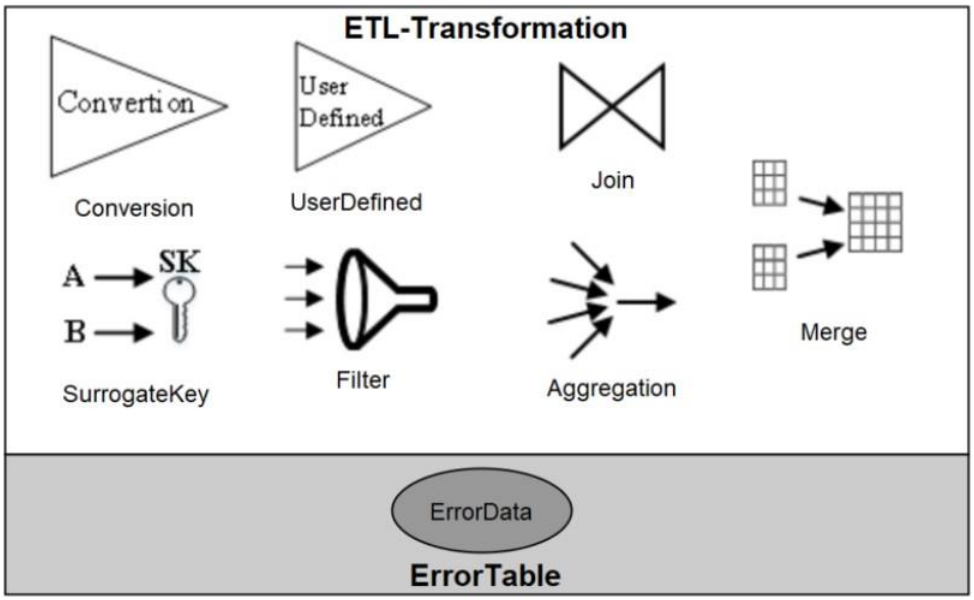
→ Transform : les données extraites sont transformées dans un format spécifique, défini à l'avance. Cette étape facilite l'intégration et la consolidation des données pour l'étape finale.

↳ Problématique : définition et reconnaissance des formats à appliquer, prise en charge des nouvelles données, évolution des formats de données en fonction du temps, interopérabilité des formats.

→ Load : Les données sont transformées dans leurs formes/dimensions finales.
↳ Problématique : la gestion des « anciennes » données versus celles à mettre à jour.

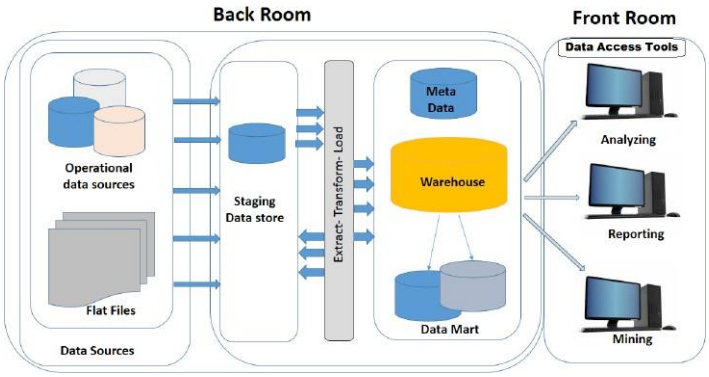
Petit schéma récap que le professeur n'explique pas dans son diapo mais que votre tut de SN explique à merveille dans sa fiche sur le forum (foncez-y vraiment !) :



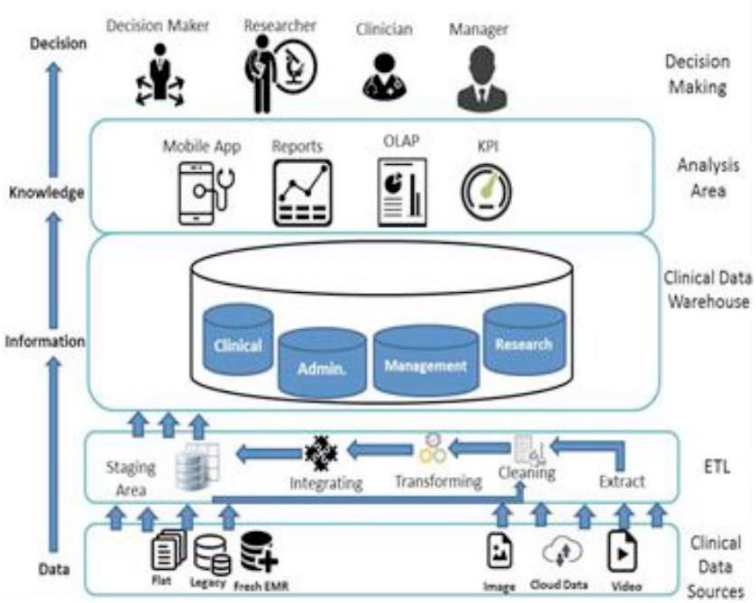


B. Architectures d'un entrepôt de données

Les entrepôts de données contiennent tout ce qui n'est pas « visible » par l'utilisateur, on peut les résumer avec ce schéma :



Le schéma suivant représente quant à lui l'architecture générale d'un entrepôt **de données cliniques** :



➤ *4 grands types d'architectures*

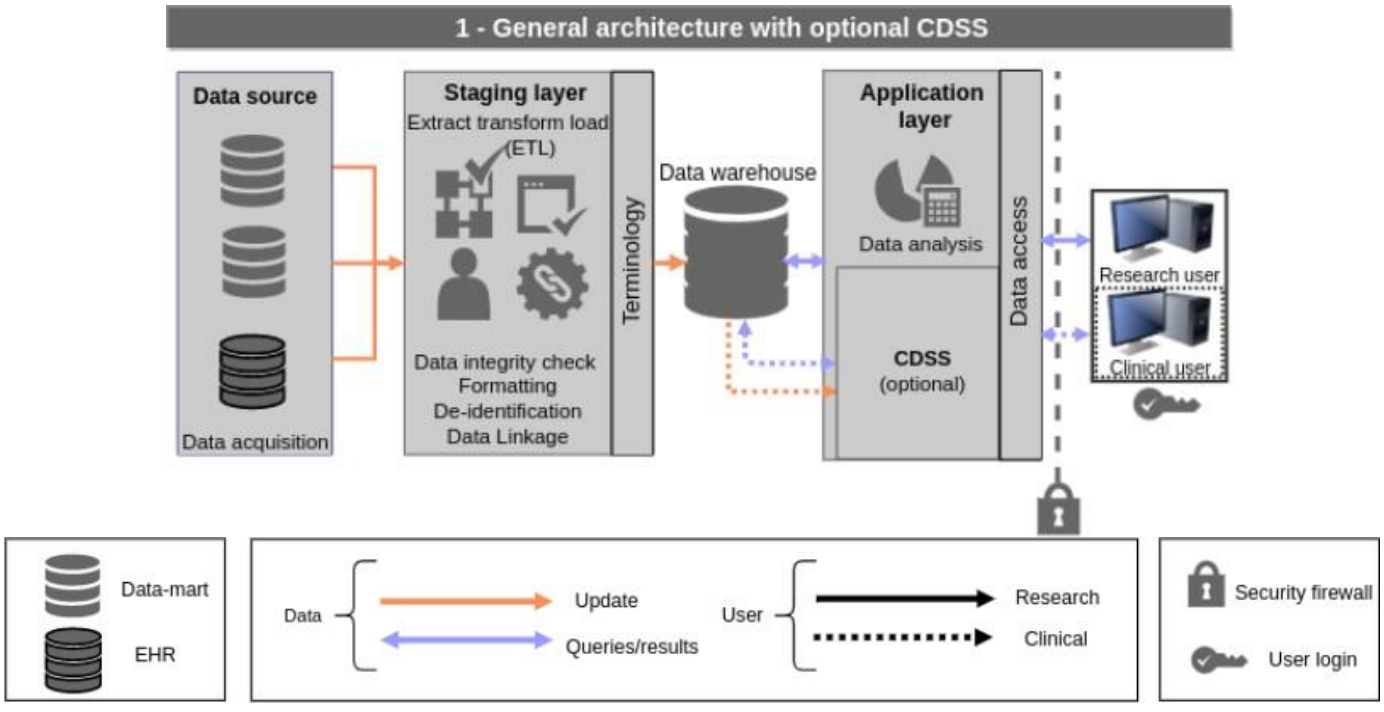
1) General architecture with optional CDSS (Clinical Decision Support System) :

Un data mart ou « magasin de données » est un ensemble de données **ciblées, organisées, regroupées et agrégées** pour répondre à un besoin spécifique, à un métier, ou un domaine donné.

Ici différents data mart sont harmonisés et transférés dans un CDW (= entrepôt), les utilisateurs peuvent interroger directement le CDW au travers d'une interface.

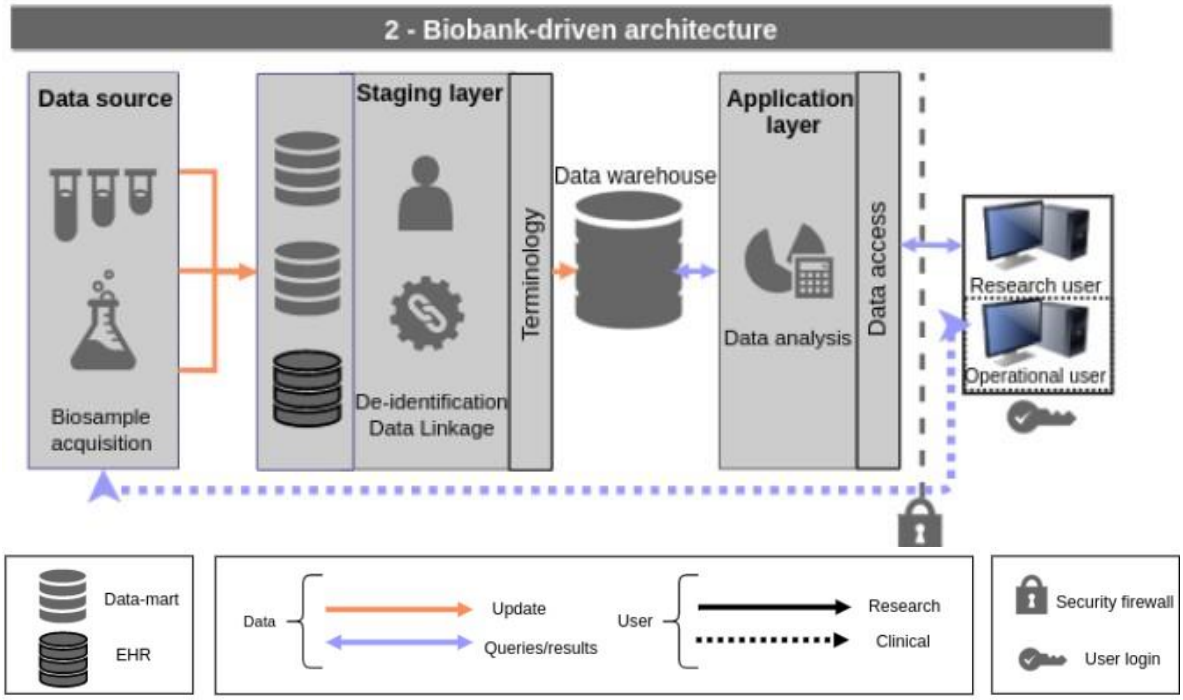
Un CDSS (Clinical Decision Support System) apportera une fonctionnalité de prise de décision en plus (il est optionnel et sert à aider les médecins à prendre des décisions cliniques).

Dans cette organisation, chaque source de données est stockée dans des data mart 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 utilisateur/trice permet d'accéder aux données de façon **sécurisée.**



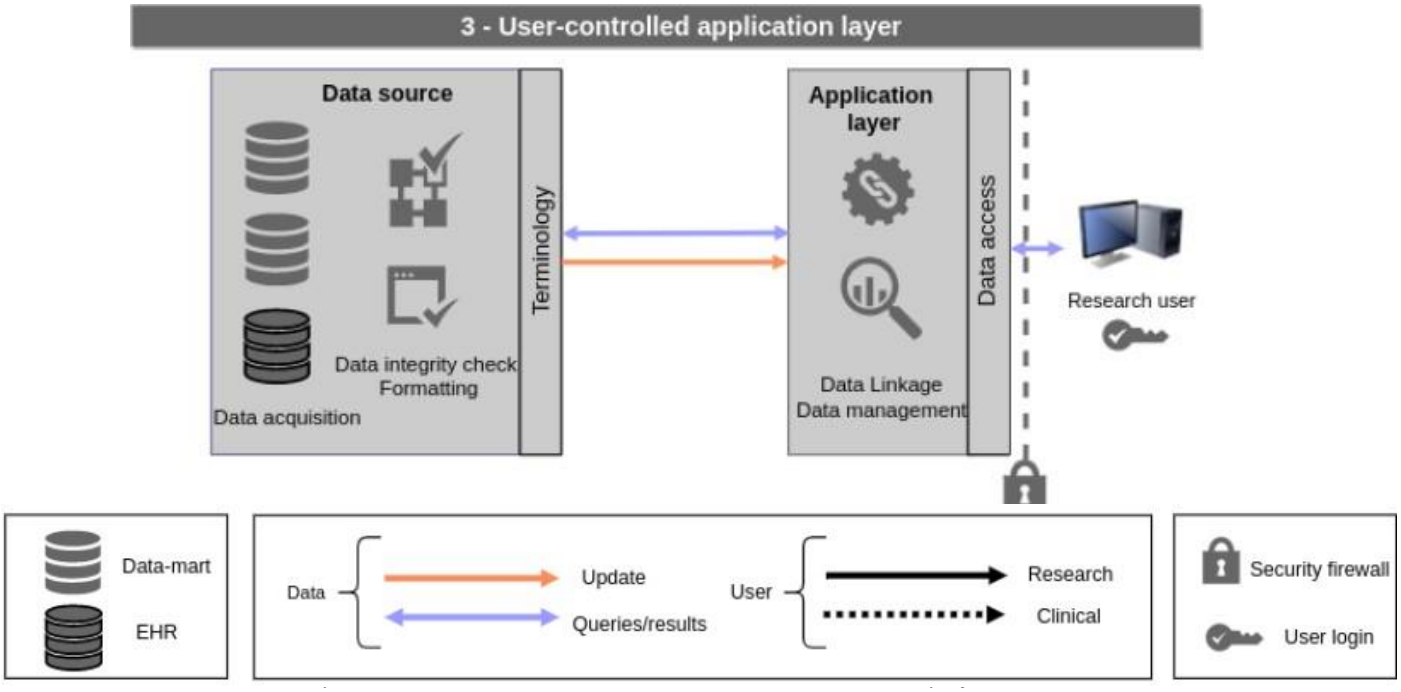
2) Bio-bank driven architecture model :

Il est construit autour d'un domaine particulier (ici, le biobanking), il est similaire au general architecture mais ici tout le modèle s'appuiera sur **la liste des échantillons biologiques disponibles.**
L'intégration des données cliniques relatives aux échantillons se fait au moment de la partie « transformation ».
→ Avantage : permet d'accéder aux données brutes des échantillons, permettant les contrôles qualités.



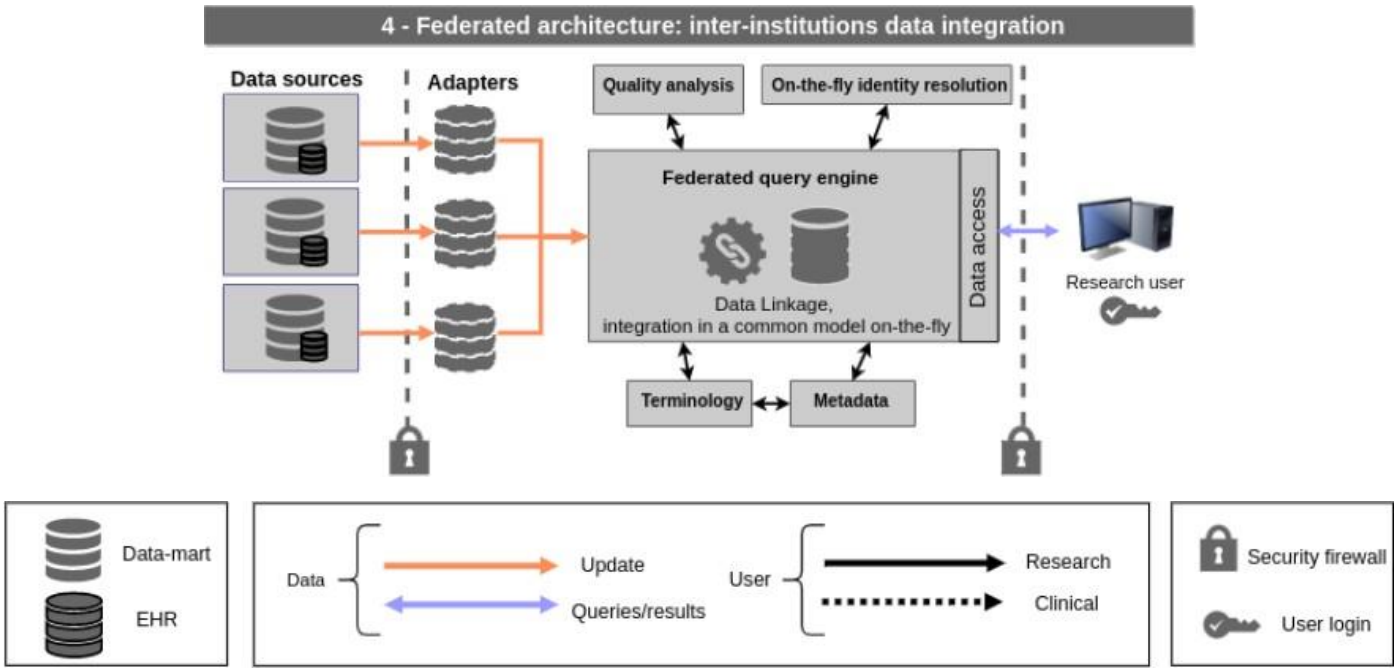
3) User-controlled application layer architecture model :

Il ne demande pas d'étape de transformation particulière, on n'a pas besoin d'entrepôt de données « central » avec les différents data marts.
Les données sont prétraitées et intégrées directement à partir des données sources seulement quandun(e) utilisateur/trice en fait la requête.



4) Federated architecture model :

Les données sont récupérées à partir de différents établissements, chaque institution choisit les données qu'elles souhaitent partager en utilisant un **adaptateur commun** qui va pré-traiter ces données.
Les données sont intégrées en direct dans un entrepôt de données « virtuel » (centralisé en dehors des institutions)
C'est un modèle **flexible** qui permet l'intégration de nombreuses sources.
Les données ne sont présentées pour l'analyse et exploitation seulement lors de la session del'utilisateur/trice et **supprimées.**



Conclusion sur ces 4 types :

→ Ces différentes architectures offrent différents outils d'analyses, de logiques, de présentation

→ Les interfaces de requêtes sont différentes en fonction des types utilisateurs/trices :

- Les chercheurs/ses cherchent des traits cliniques qui permettent d'identifier des cohortes répondant à des questions précises, pour eux toutes les architectures sont utiles

- Les médecins aident à la prise de décision pour les traitements, interventions, risques pour un/e patient/e, pour eux la première architecture avec CDSS est la plus appropriée

C. Les données

1) Sources et disponibilités

Le socle du CDW consiste à identifier les sources de données. 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.

Chaque source de données cliniques a souvent sa propre organisation, son propre standard, son propre logiciel d'exploitation et son propre « language ».

Ainsi l'étape très chronophage d'identification et analyse de toutes les sources et spécificités est très importante !!

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.

2) Formats

Les types de données sont très variés :	Et les formats le sont également :
- Texte (structuré ou non structuré) - Images - Vidéo - Échantillons biologiques - Réponses - Puces ADN/ARN - Données externes (questionnaire, objets de santé connectés)	- Numérique - Qualitatif - Quantitatif - Séquentiel

3) Récupération

Le traitement des données suivant l'ETL comporte 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).

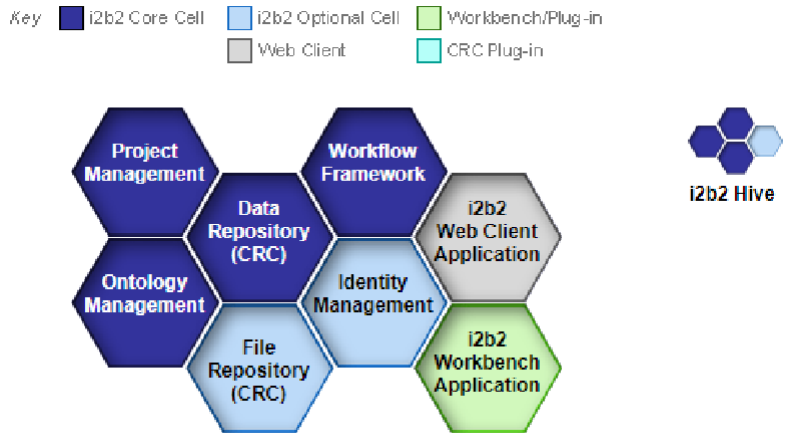
4) Standardisation et intégration

→ Certaines données sont standardisées à la saisie, les plus courantes sont la Classification Internationale des Maladies (CIM-10) et le Systematized Nomenclature Of MEDicine-Clinical Terms (SNOMED-CT)

→ Cependant on peut aussi utiliser un Common Data Model (CDM) :

C'est un schéma d'organisation permettant l'interopérabilité et le partage des données.
Une utilisation d'un CDM déjà utilisé par d'autres institution permet de s'affranchir d'une étape importante de sélection des logiciels, plateforme etc… et cela reste une **étape cruciale** qui peut prendre plus de 90% du temps de construction de l'entrepôt.
Le Common Data Model le plus utilisé est **Integrating Biology and the Bedside (i2b2).**

5) i2b2



- 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 intéractions entre les différentes « hives ».
- Identity management : anonymisation des patients.
- Web client application : permet aux utilisateurs/trices d'interroger le CDW.
- Workbench : application permettant d'analyse 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 stockée 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

→ Penser sur le long terme pour assurer la longévité du projet : s'affranchir de contraintes de formats propriétaires permettra la réutilisation du système.
→ Commencer par choisir l'architecture souhaitée basée sur les besoins des utilisateurs/trices.
→ Sélectionner un CDM déjà utilisé par d'autres institution afin de bénéficier de l'aide et l'expérience d'une plus grande communauté.
→ 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 projets'élargit.
→ Définir la fréquence des mises à jour, le détail du processus ETL, le niveau d'automatisation.

→ Communiquer à chaque étape tout en consultant régulièrement les utilisateurs/trices.

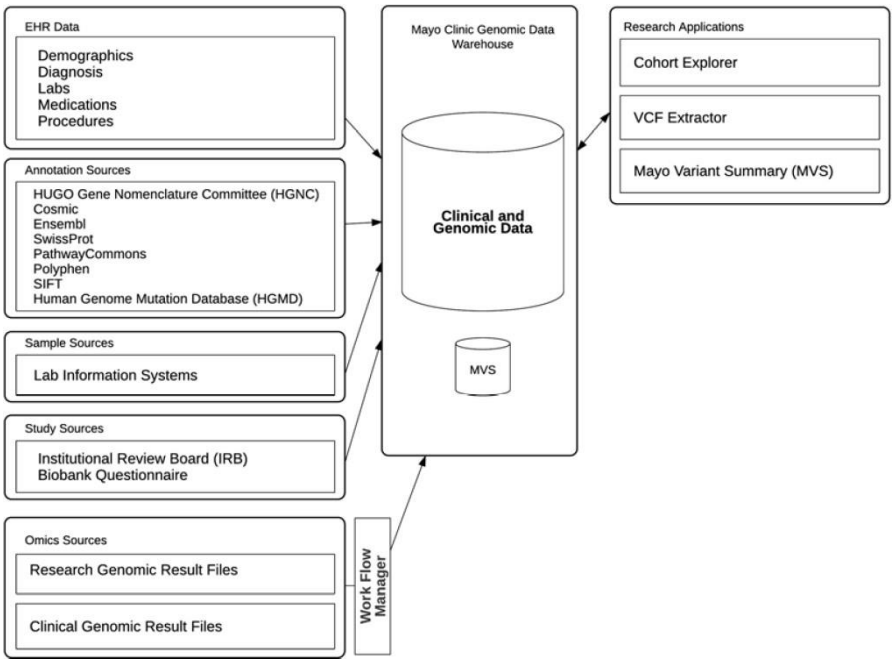
Pour l'extraction des données il faut faire attention :

- « 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 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.

F. Exemples

1) Genomic Data Warehousing @Mayo Clinic

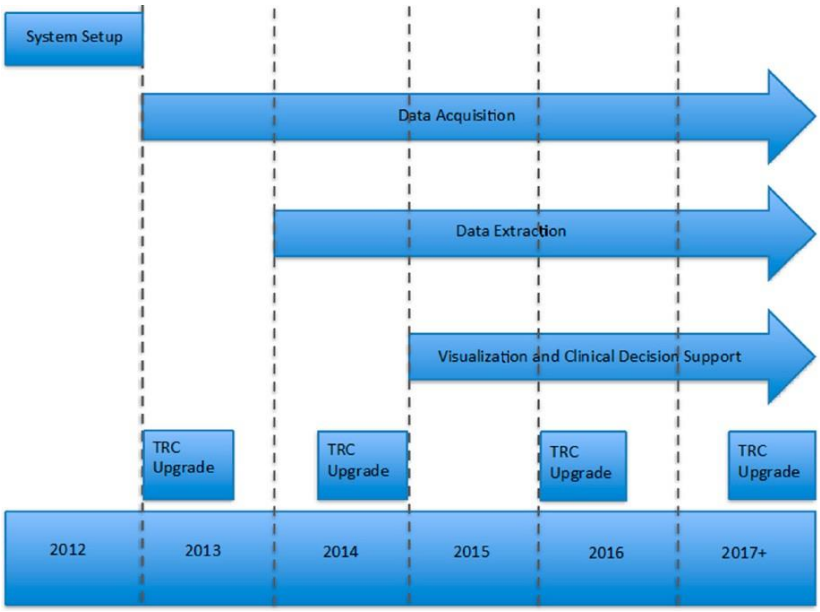
→ Le prof fait peu (voir pas mdr) de remarque sur cet exemple, essayez de lire un peu les docs et faire des liens avec les points du cours juste avant <3



EHR: Electronic

Health RecordsVCF:

Variant Call Format



Ci-contre c'est la timeline de l'entrepôt. Le logiciel qui est mis à jour est TRC.

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

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

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

2) George Pompidou University Hospital Clinical Data Warehouse

Dans ce centre ils utilisent un i2b2 et 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 : Cohorte anonyme avec les données détaillées.
- Troisième niveau : Cohorte avec toutes les données, non anonyme.

L'entrepôt permet de réaliser de nombreux projets :

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%)*

	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

Un petit article en Anglais qui résume l'utilité de l'utilisation des données de santé pour finir cet exemple en beauté.

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.

« Tableau récapitulatif

Ce qui était déjà connu sur le sujet

- *La réutilisation des données de santé est un enjeu majeur pour améliorer la gestion des soins despatients et faciliter les recherches cliniques et épidémiologiques.*
- L'hôpital a déployé des entrepôts de données cliniques pour faciliter la réutilisation des donnéesde santé
- *Les procédures de réutilisation doivent garantir à la fois un accès facile pour les cliniciens et la vieprivée des patients*

Ce que cette étude a ajouté à nos connaissances

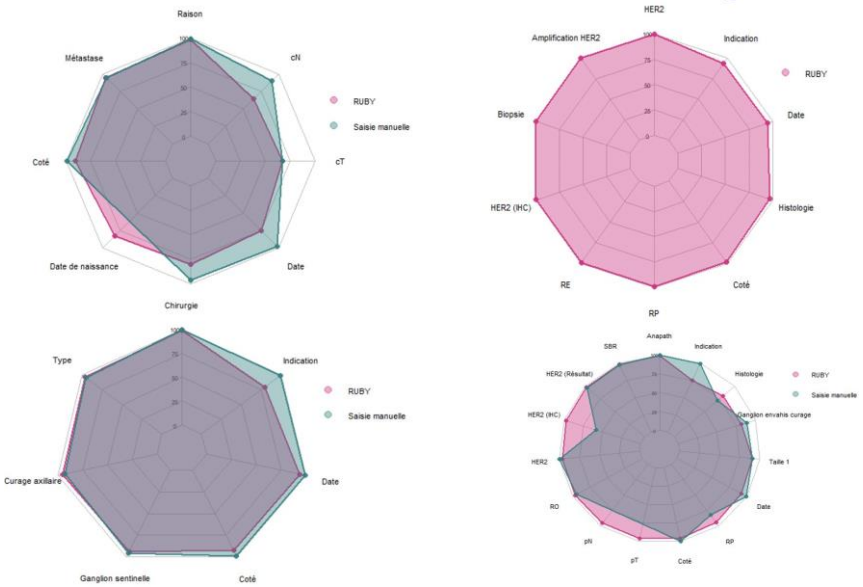
- *Le déploiement d'un CDW est un processus à long terme à partir de la conception jusqu'au « end-user CDW adoption » (oui j'ai pas compris…)*
- Les cliniciens ne sont pas prêts à formuler des questions complexes et naviguer à travers les différentes nomenclatures d'un CDW.
- *Une solide collaboration entre les cliniciens, et les spécialistes en information biomédicale, en biostatistiques et en épidémiologie est nécessaire pour mener à bien un projet de recherche à l'aided'un CDW. »*

G. Et au Centre Antoine Lacassagne ?

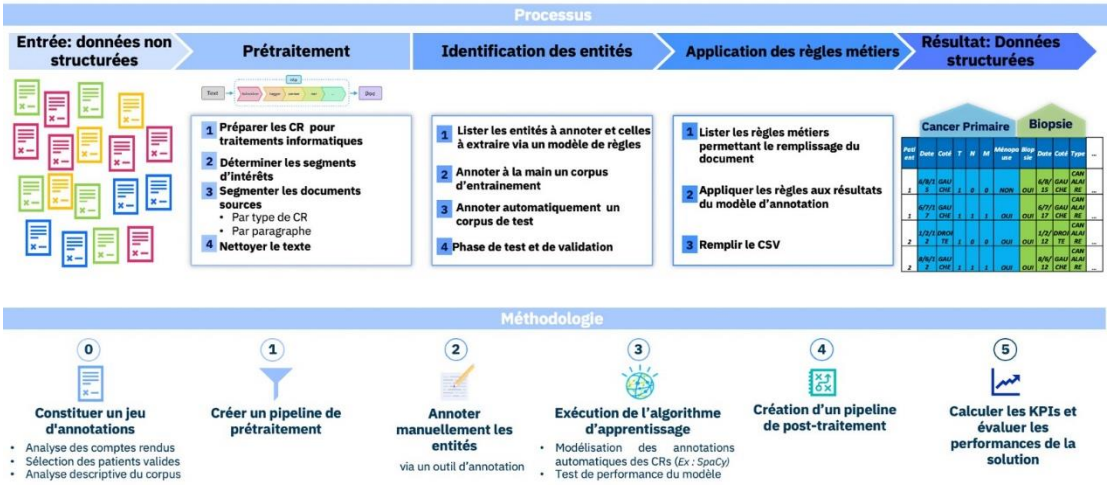
- Au CAL, c'est le début du projet de lancement de la plateforme de données.
- On analyse des sources de données disponibles.

Il y a surtout la mise en place d'une structuration automatique des données grâce à des algorithmes d'intelligence artificielle (**projet RUBY** publié dans JCO Clinical Cancer Informatics) : au lieu de requêter les données textuelles, des données structurées seront présentées directement pour intégration à la plateforme de données de santé.

→ Je vous remets les 3 diapo du prof sur ce projet :



Mise en œuvre



Mise en œuvre

Annoter manuellement des entités : des exemples d'annotation

Patient

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.
Les aires ganglionnaires sont libres.
Elle a eu une biopsie qui montre un carcinome canalaire infiltrant de grade I, RO+, RP-, Her2 ++.
On va se mettre en rapport avec le [NOM_ANONYMISE] pour confirmer la nature bénigne de ces lésions osseuses et non pas métastatiques.
En fonction de cela, on prévoit une consultation en chirurgie, un bilan pré-opératoire et une consultation anesthésie.

Anapath: Segment Conclusion

CONCLUSION : Mammectomie partielle centrale comportant la PAM, pour tumeu r de
2 cm de grand axe correspondant à un carcinome canalair e infiltrant moyennement différencié de SBR II (2.3.1) avec début d'envahissement profond de la région aréolaire.
Exérèse largement satisfaisante.
1 ganglion métastatique sans rupture capsulaire, sur les 7 i solés dans le curage des 1er et 2ème étages axillaires droit s (1+/7).
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**

Les Dédis :

♥ *Dédi à vous parce que cette ronéo n'est vraiment pas simple.*

♥ *Dédi à toute ma famille (en particulier à mon génie de frère et à mes parents qui me faisaient à manger tous les jours et à qui j'ai dû demander peut-être 10 000 fois de se taire parce que je révisais).*

♥ *Dédi au pomo de douche (le futur dentaire qui règnera sur Villefranche), au stylo et au putois.*

♥ *Dédi à Maribosome (je suis sûr que vous êtes déjà tombés sur au moins un de ses beaux messages sur le forum…elle en a envoyé tellement).* 😊

♥ *Dédi à la meilleure matière (#biochdivine).*

♥ *Dédi à mes fillots, Matteo et Johana, vous allez tout déchirer, je crois très fort en vous !!!!!*

♥ *Dédi à Narine ma Co-marraine qui a du mal à se lever le matin snif.* 😕

♥ *Dédi à Lucas (on s'est vu 2-3 fois à Vaugrenier avec l'Asptt athlé, j'espère que tout se passe bien pour toi, envoie moi un message si tu vois cette ronéo).*

♥ *Dédi à Balenciaga, l'homme qui révise moins vite que son ombre.*

♥ *Dédi à Quillan du Ping-Pong de Valrose, courage mec tu vas tout casser !!!*

♥ *Et enfin dédi à moi d'avoir réussi toudemêm.*