Bonjour à tous, aujourd'hui on va découvrir un nouveau cours qui parle de traitement de données et de big data dans le domaine médical. Je rappelle que ceci est une fiche complète, elle contient la plupart des infos mais sous forme synthétique donc écoutez au moins une fois les diapos sonorisées du prof avant de travailler dessus. C'est parti !

Entrepôts de données Hébergement Données massives en santé

Pour commencer, quelques définitions :

- **Entrepôts de données (cliniques)** : aussi appelés *Integrated Data Repositories (IDR)* ou *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.
- **Données massives en santé (ou big data)** : gros volumes de données qui alimentent l'activité quotidienne d'un hôpital.

Les **big data** sont régis 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 traitées rapidement.
- **Variété** : Les données sont sous des formats différents.

Ces 2 dernières années on a produit 90% du volume total des données et plus de 80% de ces données n'ont pas été exploitées. *(Quelques valeurs : 8.9 milliards de feuilles de soins (Sniiram), 2,3 milliards de Go en volume)*

Au Centre Antoine Lacassagne (CAL) :

- Des millions de comptes rendus médicaux et d'épisodes.
- 300 000 lignes de chimiothérapies.
- Des données d'anatomopathologie, radiothérapie, hospitalisations…
- 80% des données sont non structurées (texte libre) et difficilement exploitables.
- Les données structurées (20%) sont représentées ou stockées avec un format prédéfini.

(Le prof parle du CAL parce qu'il travaille surement là-bas mais en vrai on s'en fou)

Les centres hospitaliers traitent de grands volumes de données tous les jours.
De nombreuses questions sont posées quotidiennement sur :

- La pratique de la médecine
- File active
- Traitements/répartition
- Questions cliniques et/ou fondamentales

Un entrepôt de données permettra de répondre à toutes ces questions

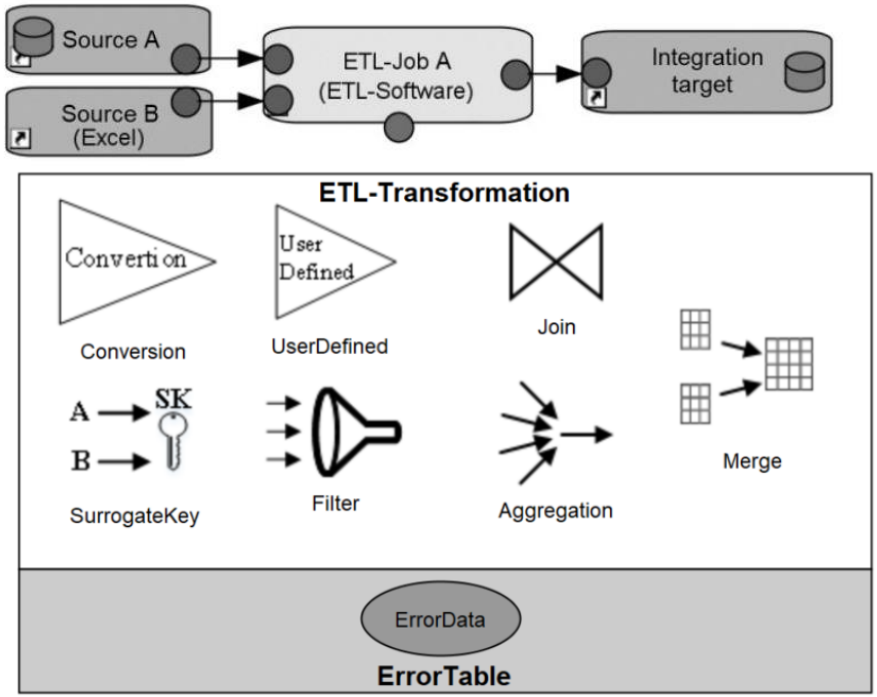
« Un entrepôt de données va recueillir et regrouper les données importantes et les associer aux patients. Les propriétés des variables, des champs, leurs noms, les règles sont définies, idéalement utilisent un standard international. Les données sont solides et ne changeront pas à chaque mise à jour,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

Vous l'été prochain ->

ETL : Extract – Transform – Load

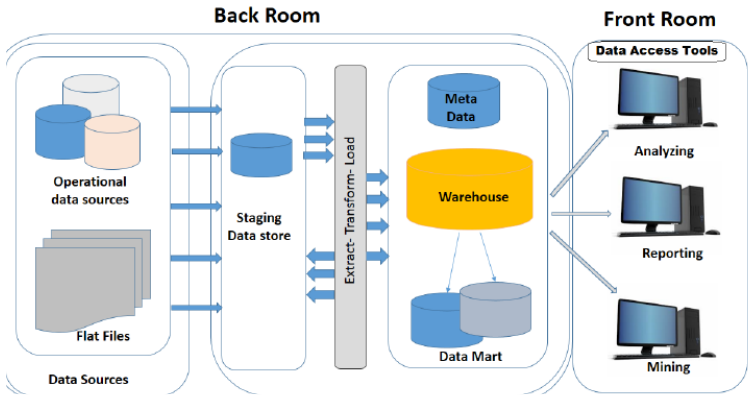
- **Extraction** : connecter les différentes sources de données et d'extraire les données nécessaires. *(Problématique : hétérogénéité des sources de données qui nécessiteront de multiples approches pour la connexion et l'extraction des données)*
- **Transform** : les données extraites sont transformées 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)*

Et voici les schémas qui résument le fonctionnement des bases de données.

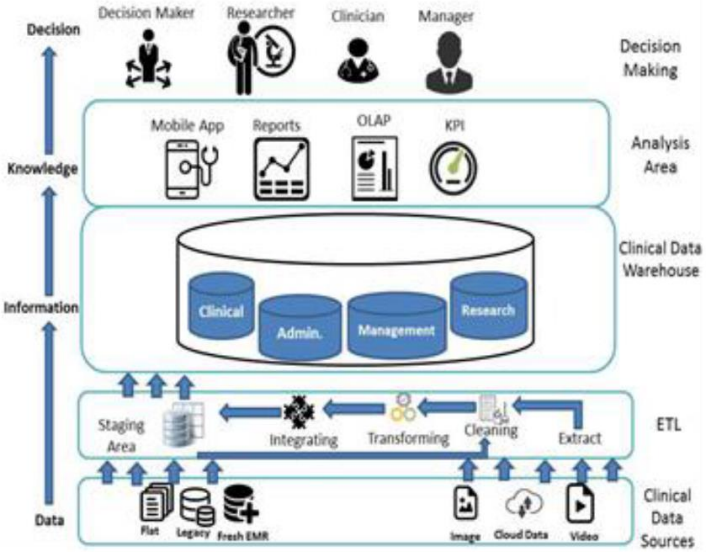


Astuce : Privilégiez la compréhension à l'apprentissage par cœur, pour retenir mieux et à plus long terme.

Architectures d'un entrepôt de données



Architecture générale d'un entrepôt de données cliniques



Les grands types d'architecture :

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

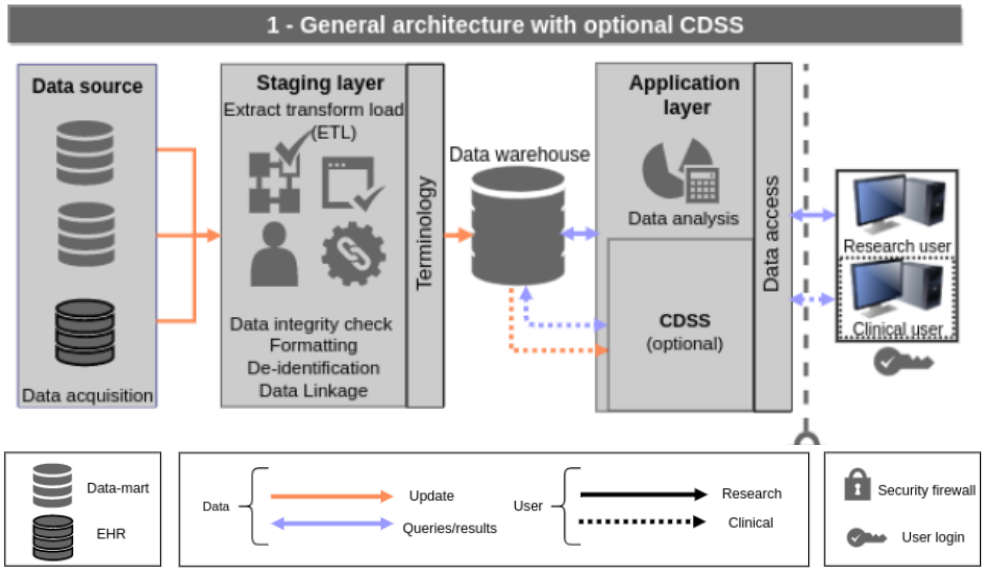
Cette base de données est composée de différents « **data mart** » (« magasin de données » : ensemble de données ciblées, organisées, regroupées et agrégées pour répondre à un besoin spécifique à un métier ou un domaine donné) sont harmonisés et transférés dans un CDW.

Les utilisateurs peuvent interroger directement le CDW au travers d'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 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.



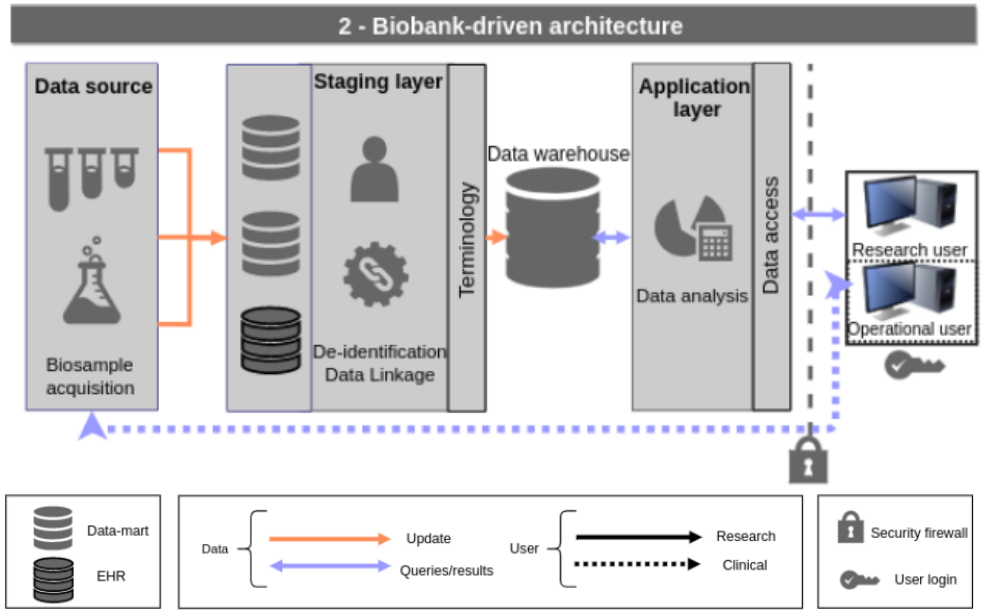
2) Bio-bank driven architecture model

Ce type de base de donnée est construit autour d'un domaine particulier (ici, le biobanking).

Le modèle 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.

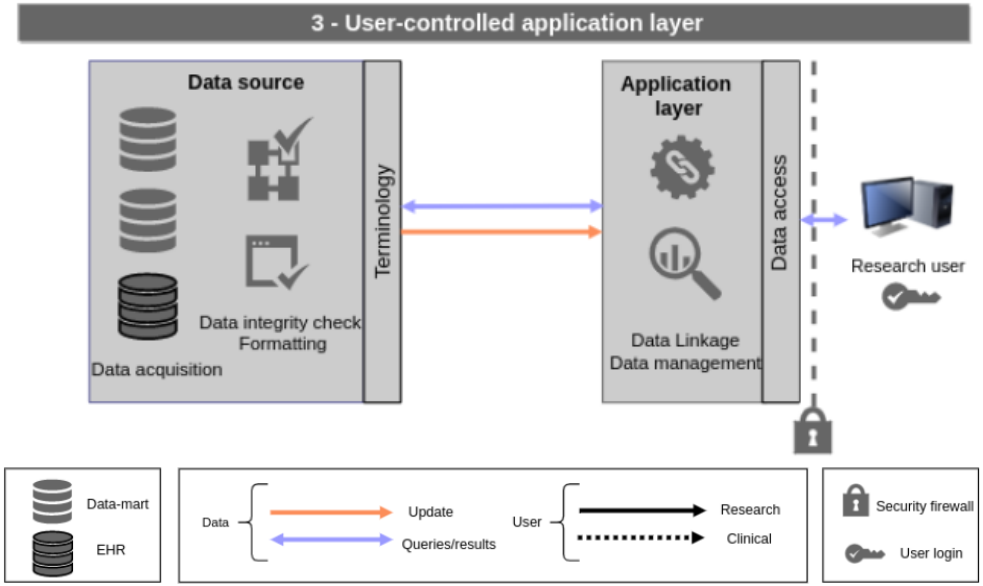


3) User-controlled application layer architecture model

Pas d'étape de transformation particulière.

Pas d'entrepôt de données « central » regroupant les différents data marts.

Les données sont pré-traitées et intégrées directement à partir des données sources seulement quand un/e utilisateur/trice en fait la requête.



4) Federated architecture model

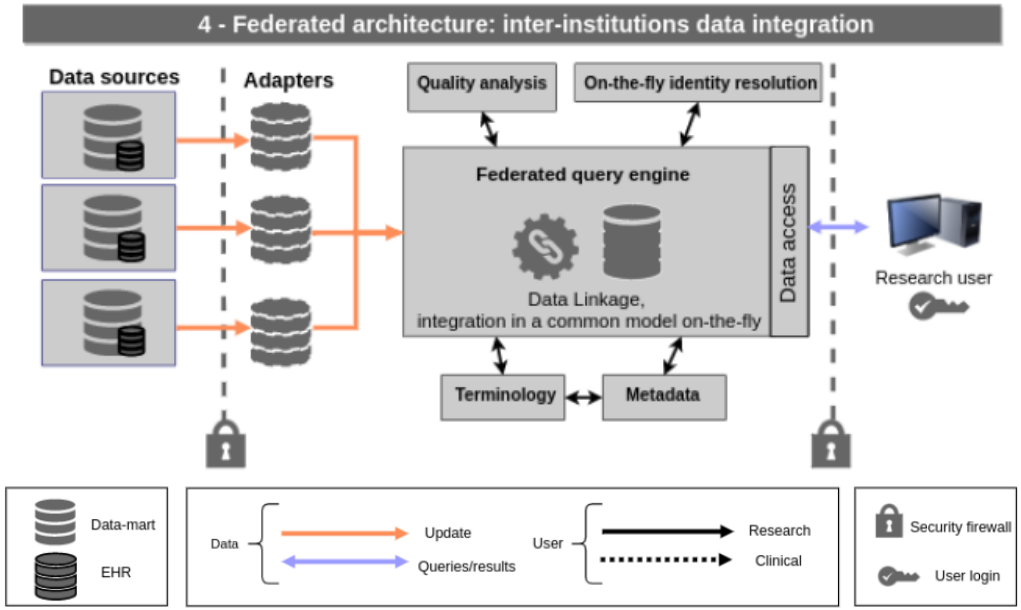
Les données sont récupérées à partir de différents établissements.

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 l'exploitation seulement lors de la session de l'utilisateur/trice et supprimées après.



Conclusion :

Ces différentes architectures offrent différents outils d'analyses, de logiques de présentation et les interfaces de requêtes sont différentes en fonction des types utilisateurs/trices :

- **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 première architecture avec CDSS est la plus appropriée pour les médecins.

Je vais faire mon message de fin ici même si j'ai peu de place : Il reste seulement quelques semaines pour vos examens, c'est une période très importante où il faut se donner à fond pour faire mieux que les autres, parce que ça va forcément être très sélectif, qu'il y ai un concours ou non. Alors courage et rdv l'année pro aux soirées médecine. Dédi aux cotuts, aux tutrices de bioch, à Julia qui mérite bien sa place dans mes dédis et aux ptits potes qui ne liront jamais cette fiche.

Sources

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.

Sources et disponibilité

Chaque source de données cliniques a souvent :

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

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.

Formats

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

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).

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) :

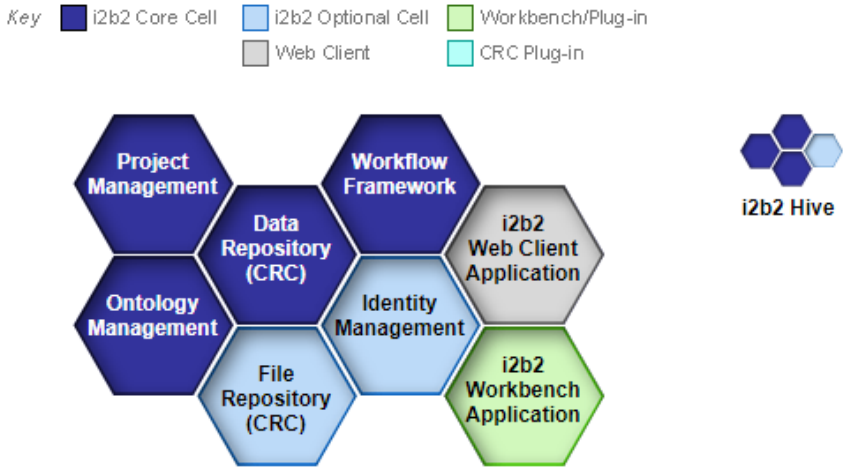
- Un schéma d'organisation permettant l'interopérabilité et le partage des données.
- Une utilisation d'un CDM déjà utilisé par d'autres institution permet de s'affranchir d'une

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.

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.

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 ?

Conseils du prof

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 projet s'élargit.

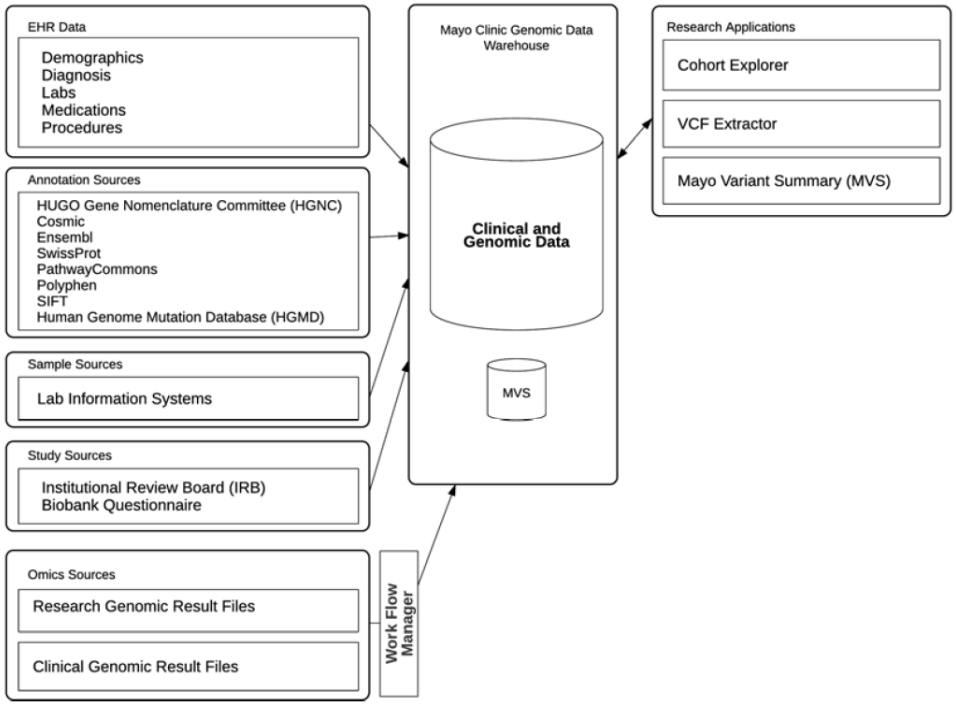
Définir la fréquence des mises à jour, le détail du processus ETL, le niveau d'automatisation.

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

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.

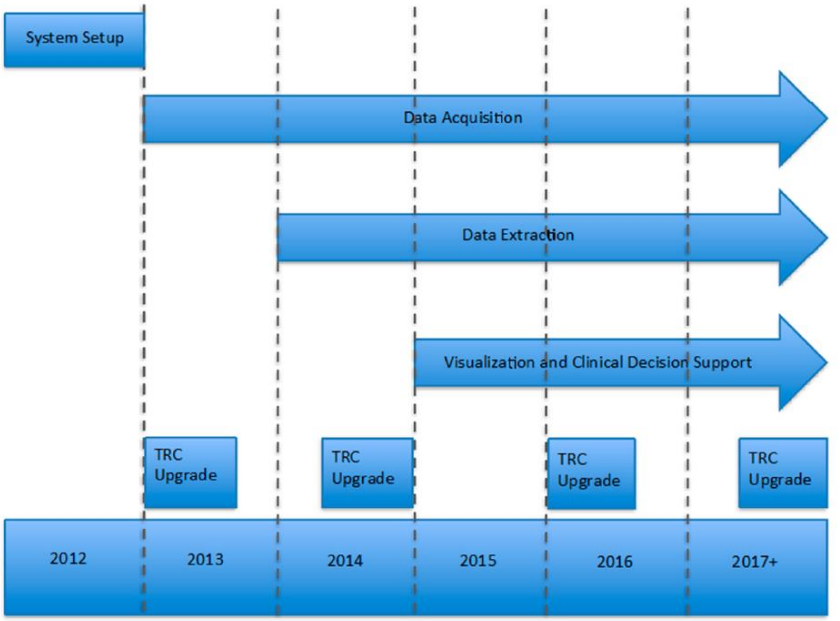
Exemples : Genomic Data Warehousing



A gauche on voit les données, qui sont regroupées dans le warehouse et qui pourront être utilisés 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

Le prof dit : une vingtaine de personnes s'occupent d'implémenter les ressources, volumes de données pas trop importants mais assez honnêtes (ok.)

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

« 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.

	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 :

Année	Nbr de projets	Nbr de départements à l'origine des projets	Projets épidemiologie 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 qui résume l'utilité de l'utilisation des données de santé (je compte sur vous pour comprendre l'anglais).

Et au CAL ? (blc)

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.

Mise en œuvre – Projet Ruby

Processus

Entrée: données non structurées → **Prétraitement** → **Identification des entités** → **Application des règles métiers** → **Résultat: Données structurées**

1. Préparer les CR pour traitements informatiques
2. Déterminer les segments d'intérêts
3. Segmenter les documents sources
   - Par type de CR
   - Par paragraphe
4. Nettoyer le texte

1. Lister les entités à annoter et celles à extraire via un modèle de règles
2. Annoter à la main un corpus d'entrainement
3. Annoter automatiquement un corpus de test
4. Phase de test et de validation

1. Lister les règles métiers permettant le remplissage du document
2. Appliquer les règles aux résultats du modèle d'annotation
3. Remplir le CSV

	Cancer Primaire						Biopsie				
Patient	Date	Coté	T	N	M	Ménopause	Biopsie	Date	Coté	Type	...
1	6/8/15	GAUCHE	1	0	0	NON	OUI	6/8/15	GAUCHE	CANALAIRE	...
1	6/7/17	GAUCHE	1	1	1	OUI	OUI	6/7/17	GAUCHE	CANALAIRE	...
2	1/2/12	DROITE	1	0	0	OUI	OUI	1/2/12	DROITE	CANALAIRE	...
2	8/6/12	GAUCHE	1	1	1	OUI	OUI	8/6/12	GAUCHE	CANALAIRE	...

Méthodologie

0 – Constituer un jeu d'annotations
- Analyse des comptes rendus
- Sélection des patients valides
- Analyse descriptive du corpus

1 – Créer un pipeline de prétraitement

2 – Annoter manuellement les entités

via un outil d'annotation

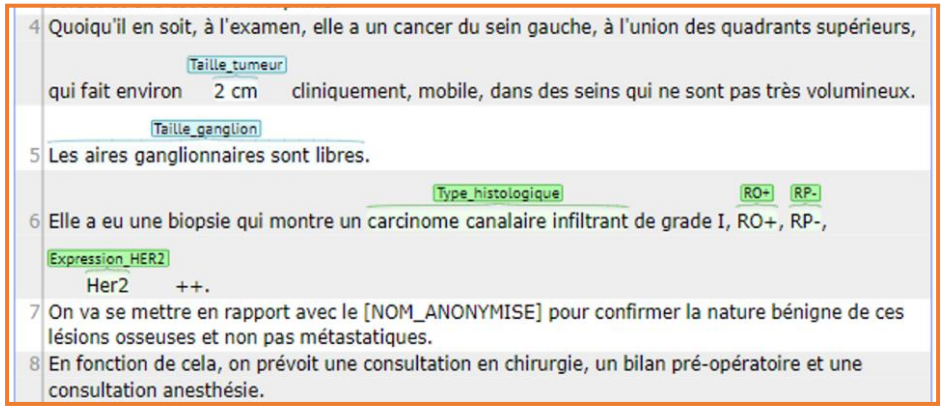
3 – Exécution de l'algorithme d'apprentissage
- Modélisation des annotations automatiques des CRs (*Ex : SpaCy*)
- Test de performance du modèle

4 – Création d'un pipeline de post-traitement

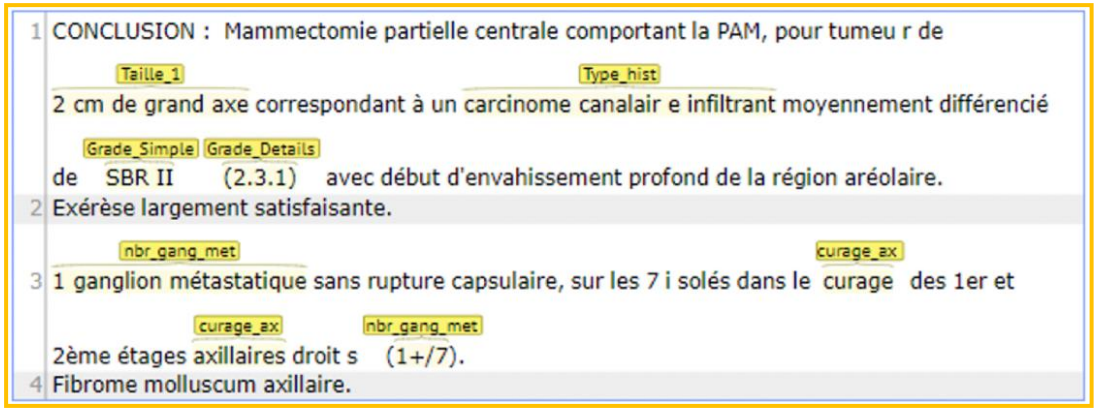
5 – Calculer les KPIs et évaluer les performances de la solution

Annoter manuellement des entités : exemples d'annotation

Patient

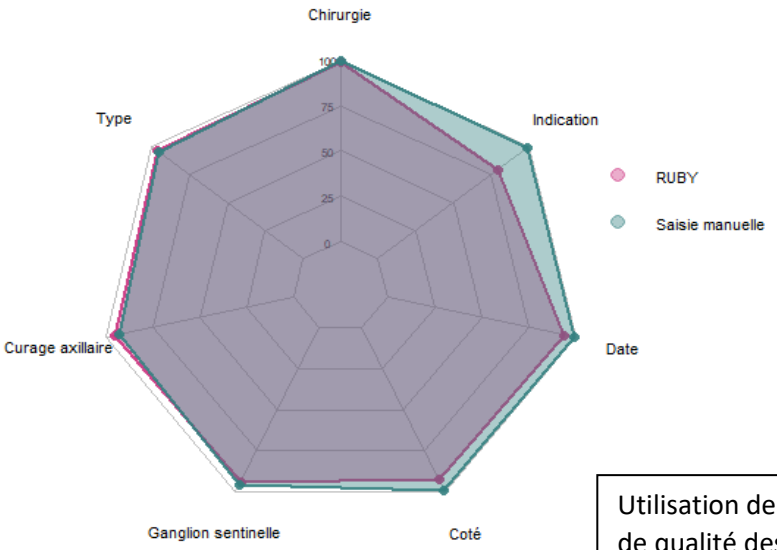


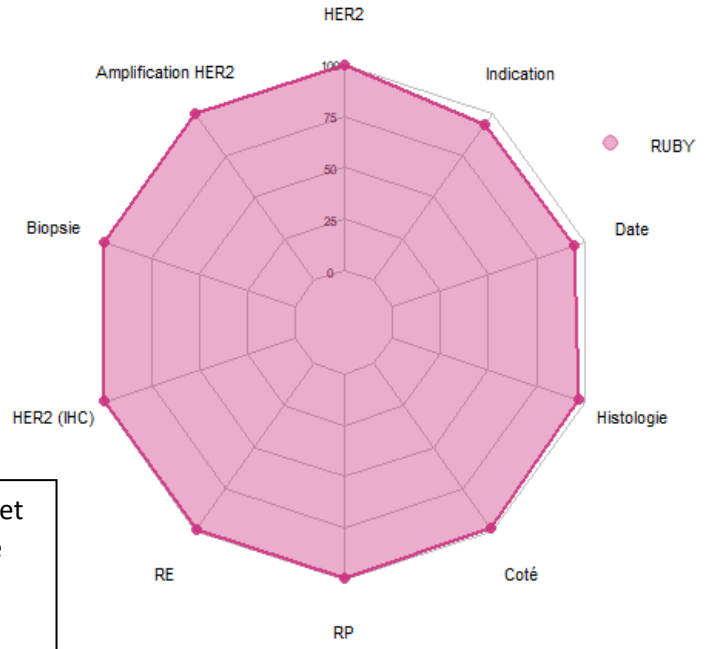
Anapath: Segment Conclusion



- 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.





Utilisation de Ruby = Gain de temps et de qualité des données considérable par rapport à la saisie manuelle des données.