Entrepôt de données de santé

Stéfan Darmoni, Badisse Dahamna, Romain Lelong, Ivan Kergourlay, Kevin Billey, Julien Grosjean, Clément Massonnaud, Loïc Tanguy

Département d'informatique et d'information médicale, CHU de Rouen
LIMICS U1142 INSERM

Qu’est ce que c’est

- Les Entrepôts de Données de Santé (EDS) sont des outils informatiques permettant la collection, l’intégration puis le traitement des données de santé provenant d’un grand nombre de sources d’information clinique (dossier patient informatisé, système d’information des laboratoires et d’imagerie, prescription informatisée, dossier infirmier…).
- Agrégation d’un **maximum** d’informations disponibles sur les patients
  - Première étape : premier semestre 2017 : entrepôt du DIM (Cora)
- Quelle que soit l’application source (CDP, HEO, Xway, CORA, ICCA…)
  - Reprise des anciennes applications ≠ APHP

➔ Permet de croiser ces informations et donc de sélectionner finement des patients

Objectifs des EDS

- Améliorer le codage PMSI, par la détection semi-automatique d'atypie entre le codage réalisé et les données issues de l'EDS (D2IM) ;
- Améliorer la recherche interventionnelle grâce aux études de faisabilité d'essais cliniques et l'optimisation des inclusions (maison de la recherche) ;
- Détecter des profils de santé particuliers (par exemple, les patients ayant des passages fréquents aux urgences)
- Créer et maintenir de registres et cohortes afin d'optimiser la recherche non interventionnelle sur données épidémiologiques ;
- Créer et maintenir des outils de détection d'événements liés à la vigilance (par exemple, les infections nosocomiales) (Département d'appui à la qualité et à la sécurité des soins (DAQSS))
- Renforcer la prise en compte des indicateurs de qualité ;
- Développer et évaluer les algorithmes d'aide à la décision de demain.

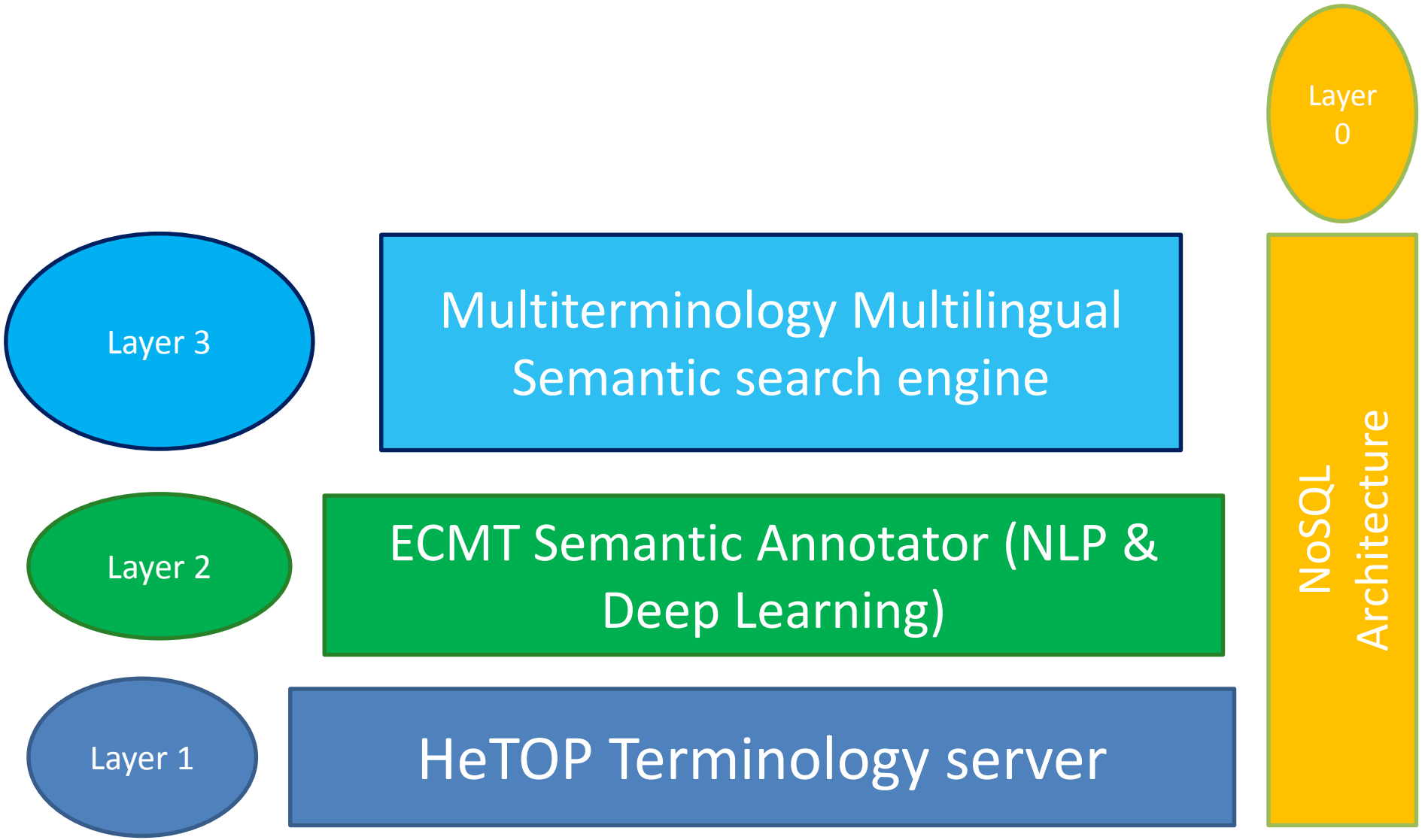
Etat de l’art

- I2B2 : développé à Harvard, gratuit, très utilisé aux États-Unis (80 universités), moins dans le monde (20 institutions dont 3 en France (APHP, Bordeaux et Rennes + Rouen en R/D)
- éHOP : développé à Rennes, 100k€ minimum, 2 installations en France (Rennes et Brest), 4 à venir (Grand Ouest), rien dans le reste du monde.
- Dr WareHouse : développé à Necker ; installé à Saint Anne et Hôpital Foch

Entrepôt de données de santé au CHU de Rouen

- Plusieurs visions stratégiques différentes
  - Aspect multilingue (deux des trois couches)
  - Sémantiques +++
- Conséquence de ces choix
  - Prix +++, notamment du à l'architecture matérielle

Semantic HDW based on three independent layers



HeTOP Terminology server

Layer 1

- Multi Terminology & cross lingual → matrix navigation (among languages & among terminologies
- 70 termino-ontologies included in 32 languages
- 2 M concepts in English; 1.2 M in French
  - 143.762 concepts in French in UMLS (2017AA) vs. 427.912 in HeTOP (x2.98)
- Over 100 million RDF triplets (2014) → big data +++

ECMT Semantic Annotator (NLP & Deep Learning)

Layer 2

- Based on HeTOP; 50 chosen KOS out of 75 (no interface terminologies)
- 11.8 M health documents
- Processing time: 22-24 hours (two servers 1 To; one with 196 cores and the second with 144 cores
- 5.2 G medical concepts extracted; 2.6 G after filtering

Neveol & coll. **Clinical Natural Language Processing in languages other than English: opportunities and challenges.** *Journal of Biomedical Semantics* 2018 **9**:12

Table 7. System performance for ICD10 coding on the **French raw** test corpus in terms of Precision (P), recall (R) and F-measure (F). A horizontal dash line places the frequency baseline performance. The top part of the table displays official runs, while the bottom part displays non-official and baseline runs.

	ALL				EXTERNAL			
	Team	P	R	F	**Team**	P	R	F
Official runs	SIBM-run1	**.857**	**.689**	**.764**	SIBM-run1	**.567**	**.431**	**.490**
	LITL-run2	.666	.414	.510	LIRMM-run1	.443	.367	.401
	LIRMM-run1	.541	.480	.509	LIRMM-run2	.443	.367	.401
	LIRMM-run2	.540	.480	.508	LITL-run2	.560	.283	.376
	LITL-run1	.651	.404	.499	LITL-run1	.538	.277	.365
	TUC-MI-run2	.044	.026	.033	TUC-MI-run2	.010	.004	.005
	TUC-MI-run1	.025	.015	.019	TUC-MI-run1	.006	.005	.005
	average	.475	.358	.406	**average**	.367	.247	.292
	median	.541	.414	.508	**median**	.443	.283	.376
Non-official	LIMSI-run2	.872	.784	.825	LIMSI-run2	.700	.594	.643
	LIMSI-run1	.883	.760	.817	LIMSI-run1	.709	.559	.625
	TUC-MI-run1-corrected	.883	.539	.669	TUC-MI-run1-corrected	.780	.290	.423
	TUC-MI-run2-corrected	.882	.536	.667	TUC-MI-run2-corrected	.767	.283	.414
	UNIPD-run1	.629	.468	.537	UNIPD-run2	.350	.381	.365
	UNIPD-run2	.518	.384	.441	UNIPD-run1	.362	.251	.296
	Mondeca-run1	.375	.131	.194	Mondeca-run1	.335	.228	.271
	Frequency baseline	.339	.237	.279	Frequency baseline	.381	.110	.170

Multiterminology Multilingual Semantic search engine

Layer 3

- First step, definition of a model as light and as compact as possible
- Search engine based on Semantic Web
  - Explosion (subsumption) based on hierarchy at a multiterminology level
  - Other relations may be used: semantic expansions based on inter-terminology semantic mappings
  - Multilingual +++

NoSQL Architecture

Layer 0

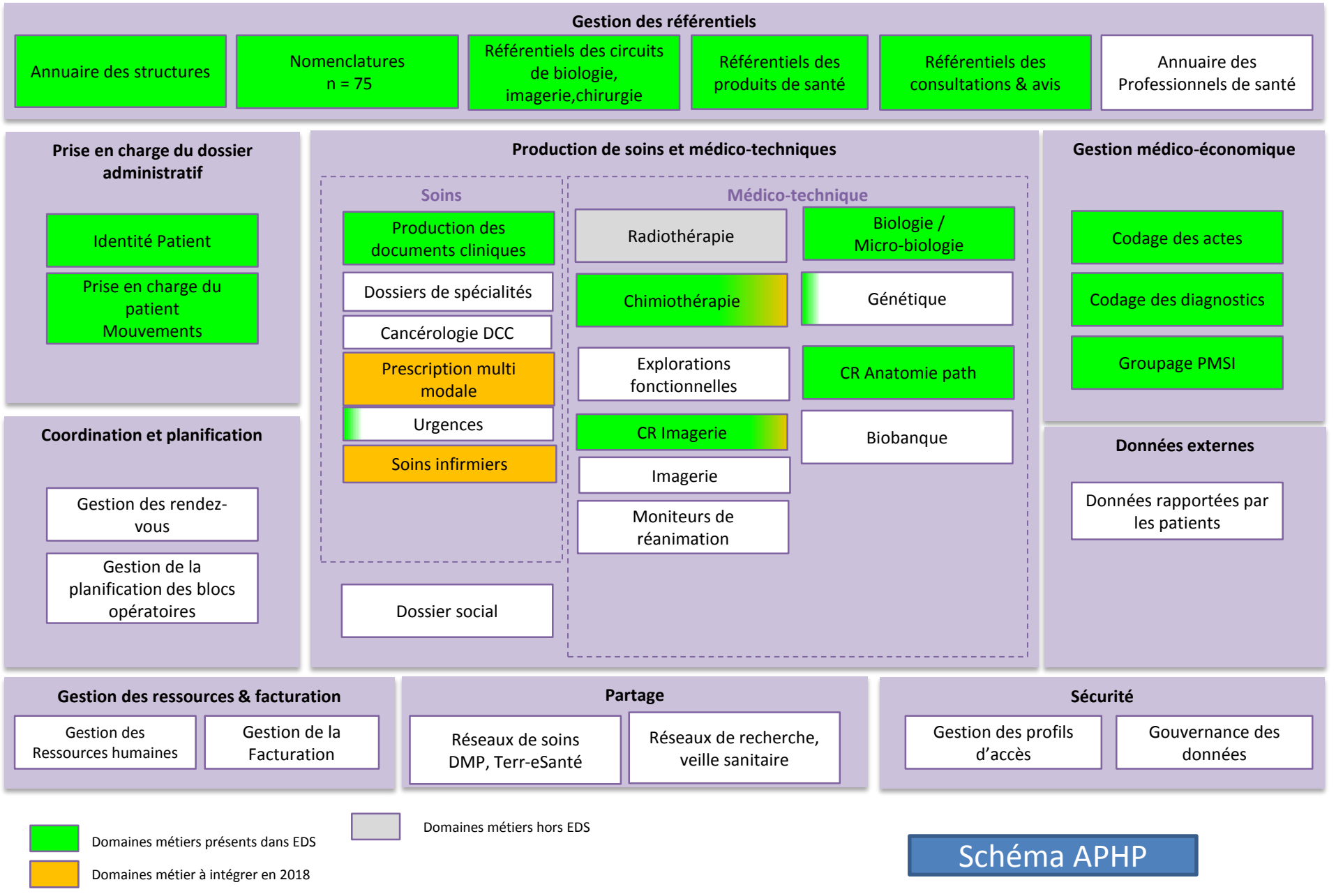
- NoSQL architecture (In Memory Data Grid –IMDG-)
- Two powerful servers:
  - 1 To; one & 196 cores
  - 1 To RAM & 144 cores
- Possibly not sufficient +++

Résultats

EDS Rouen - volumétrie

- 1,86 M patients
- 13,28 M séjours (H – C – séance)
- 11,9 M documents (depuis 2000)
  - 5 G concepts médicaux extraits (ECMT)
- 121,1 M analyses biologiques unitaires (Na, K) (depuis 2004)
- PMSI : 9,4 M diagnostics (CIM10) ; 8,6 M actes (CCAM)

EDS Rouen – couverture fonctionnelle



Prototype Mars 2018

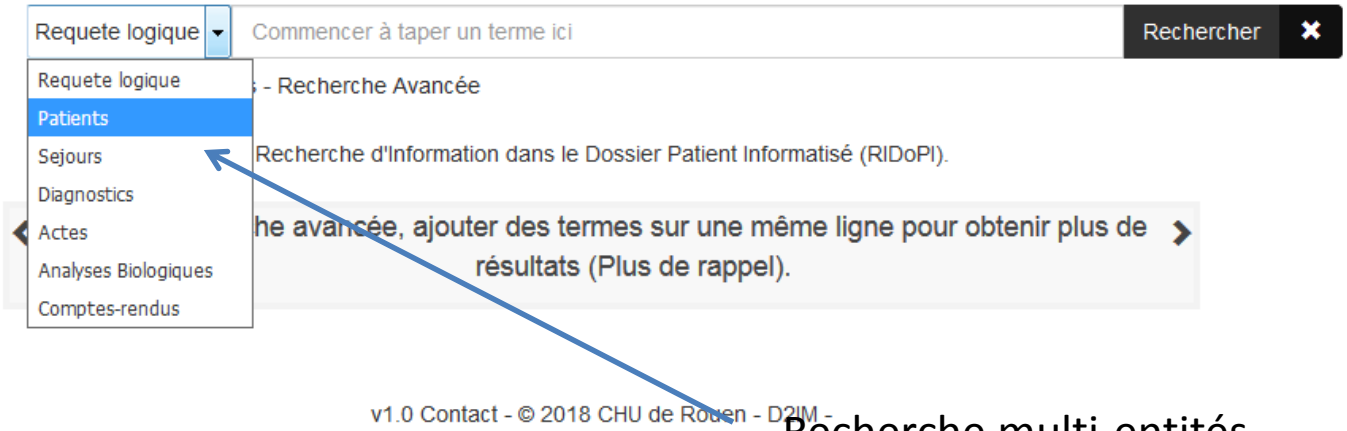
207 357 patients (12% de la taille totale estimée au CHU de Rouen)

- 1 656 275 séjours
- 1 154 371 diagnostics PMSI
- 1 456 697 actes PMSI
- 14 241 431 résultats d'analyses biologiques
- 671 442 documents textuels

RIDoPI

Recherche d'Information dans le Dossier Patient Informatisé

CHU ROUEN NORMANDIE



Recherche multi-entités

ASIS

Accès Sémantique à l'Information de Santé

Type d'entité recherchée :

Patient(s) | Hospitalisation(s) | Sejour(s) | Prise(s) en charge en UF | Analyse(s) biologique(s) | Diagnostic(s) | Acte(s) | Compte(s)-rendu(s)

	Patient(s)	Sexe	Homme / Femme / Autre	
ET	Diagnostic(s)	Terminologie(s)	...	Commencer à saisir
ET	Analyse(s) biologique(s)	Type de l'analyse	...	Commencer à saisir
ET	Sejour(s)	Date d'entrée dans l'épisc	=	2018-04-30

Ajouter une ligne

Construction de la requete :

Exemple de Requetes - Requête en language moteur

ASIS

Accès Sémantique à l'Information de Santé

Searched entity type :

Patient | Hospitalization | Stay | Patient managment | Biological test | Diagnosis | Procedure | Record

	Patient	Gender	Male / Female / Other	
ET	Diagnosis	Terminology(ies)	...	Enter terms
ET	Biological test	Type of biological test	...	Enter terms
ET	Stay	undefined	=	2018-04-30

Ajouter une ligne

Construction de la requete :

(@1 PATIENTS Male 105696)

Exemple de Requetes - Requête en language moteur

of concept provides Information Retrieval capabilities among a data subset of the health data warehouse of the university hospital of Rouen. Access to these data is regulated and an official request must be made at the

Health Data warehouse devoted to University Hospital of Rouen.

ASIS

Accès Sémantique à l'Information de Santé

Type d'entité recherchée :

Patient(s) | Hospitalisation(s) | Sejour(s) | Prise(s) en charge en UF | Analyse(s) biologique(s) | Diagnostic(s) | Acte(s) | Compte(s)-rendu(s)

	Patient(s)	Sexe	Homme / Femme / Autre				
ET	Diagnostic(s)	Terminologie(s)	...	J45 asthme	Expl	Source de données	Type
ET	Analyse(s) biologique(s)	Type de l'analyse	...	Commencer à saisir			
ET	Sejour(s)	Date d'entrée dans l'épisc	=	2018-04-30			

Ajouter une ligne

Construction de la requete :

(@1 PATIENTS Homme 105696) ET (@2 DIAGNOSTICS J45 asthme... 56)

Exemple de Requetes - Requête en language moteur

ASIS

Accès Sémantique à l'Information de Santé

Type d'entité recherchée :

Patient(s) | Hospitalisation(s) | Sejour(s) | Prise(s) en charge en UF | Analyse(s) biologique(s) | Diagnostic(s) | Acte(s) | Compte(s)-rendu(s)

	Patient(s)	Sexe	Homme / Femme / Autre				
ET	Diagnostic(s)	Terminologie(s)	...	J45 asthme	Expl	Source de données	Type de diagnostic
ET	Analyse(s) biologique(s)	Type de l'analyse	...	Commencer à saisir			
ET	Sejour(s)	Date d'entrée dans l'épisc	=	2018-04-30			

Source de données : Données structurées ✔ / Documents médicaux

Ajouter une ligne

Construction de la requete :

(@1 PATIENTS Homme 105696) ET (@2 DIAGNOSTICS J45 asthme... 625)

Exemple de Requetes - Requête en language moteur

Cas d'usage

- En mars 2018, demande de récupération de résultats d'analyses biologiques de nourrissons sur une période donnée et pour une liste de patients donnés

- En juin 2018, demande d'identification de patients atteints d'endocardite après une pose de TAVI (valve aortique)
  - Travail en complément du DIM : identifier dans les documents textuels la mention de TAVI ; exploitation de l'annotateur automatique (ECMT)
  - Nombre de cas retrouvés par l'entrepôt DIM = 30
  - Nombre de cas retrouvés par l'entrepôt Rouen étendu = 53 (2 non retrouvés, présents dans l'entrepôt DIM)
  - 23 ont pu l'être grâce à l'exploitation des données non structurées présentes dans documents de santé (et traitement par l'ECMT)

Questions ?

- Couriel : Stefan.Darmoni@chu-rouen.fr

Ce document a été réalisé par la Cellule TICE Médecine de la Faculté de Médecine de Rouen (Courriel : Florence.Charles@univ-rouen.fr).