Semantic Clinical Data Warehouse in Rouen University Hospital, Normandy, France

Prof. Stéfan Darmoni, MD, PhD

Badisse Dahamna, M.Sc, Romain Lelong, PhD, Ivan Kergourlay, M.Sc, Flavien Disson, M.Sc, Kevin Eyer, M.Sc, Julien Grosjean, PhD

Department of Digital Health (DDH), Rouen University Hospital, Normandy, France & LIMICS INSERM U1142, Sorbonne University, France

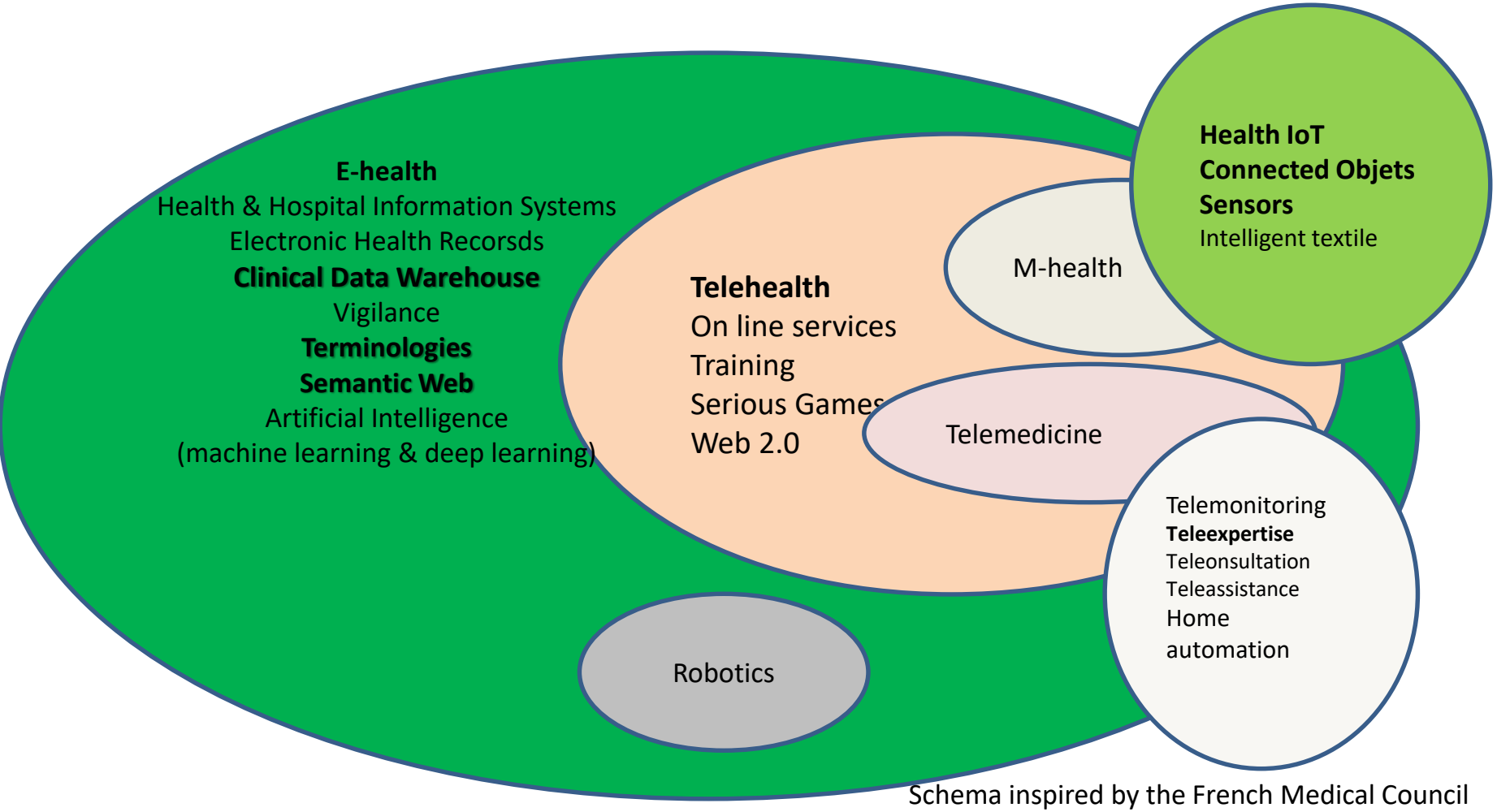
D2IM
Département d'Informatique
et d'Information Médicales

Digital Health

- Digital health is part of the training in medical schools… since 40 years
  - Used to be called Medical Informatics, Medical Computing (US), e-health
- Digital health is subdomain of Public Health in France, which is also part of the training in medical school
  - Epidemiology
  - Occupational medicine
  - Biostatistics
  - Law & Medicine
- Academic associations in digital health:
  - AIM (France) & ILAMI (Israel; URL: https://ilami.org/),
  - EFMI (Europe), AMIA (US)
  - IMIA (Earth and beyond)

LIMICS

Digital health



Schema inspired by the French Medical Council

D2IM
Département d'Informatique et d'Information Médicales

Rouen University Hospital
Department of Digital Health

- SJ. Darmoni, MD, PhD, Prof. of digital health since 2003
- Heterogeneity of competence ; **Key to success +++**
- Double competences +++ (e.g. medicine & computer science; pharmacy & information science; nurse & epidemiology) **Key to success ++**
- N =22 (vs. 100 in Harvard Medical School)
  - Two medical informaticians
  - Seven research engineers in computer science
  - Three medical librarians (CISMeF, LiSSa)
  - Four postresidents (3 public health, 1 pharmacy)
  - Four medical residents (2 public health + 1 GP + 1 Pharmacy)
  - Two PhD applicants
  - To be recruited: two data scientists, one « pedagogical engineer »
- In 2019, RUH DDH became part of the LIMICS (Sorbonne Université) main French lab in biomedical informatics
  - Deputy director of the LIMICS lab, one of the most important in Europe
- 189 scientific papers indexed in PubMed; Hindex = 32; Gindex = 0.86

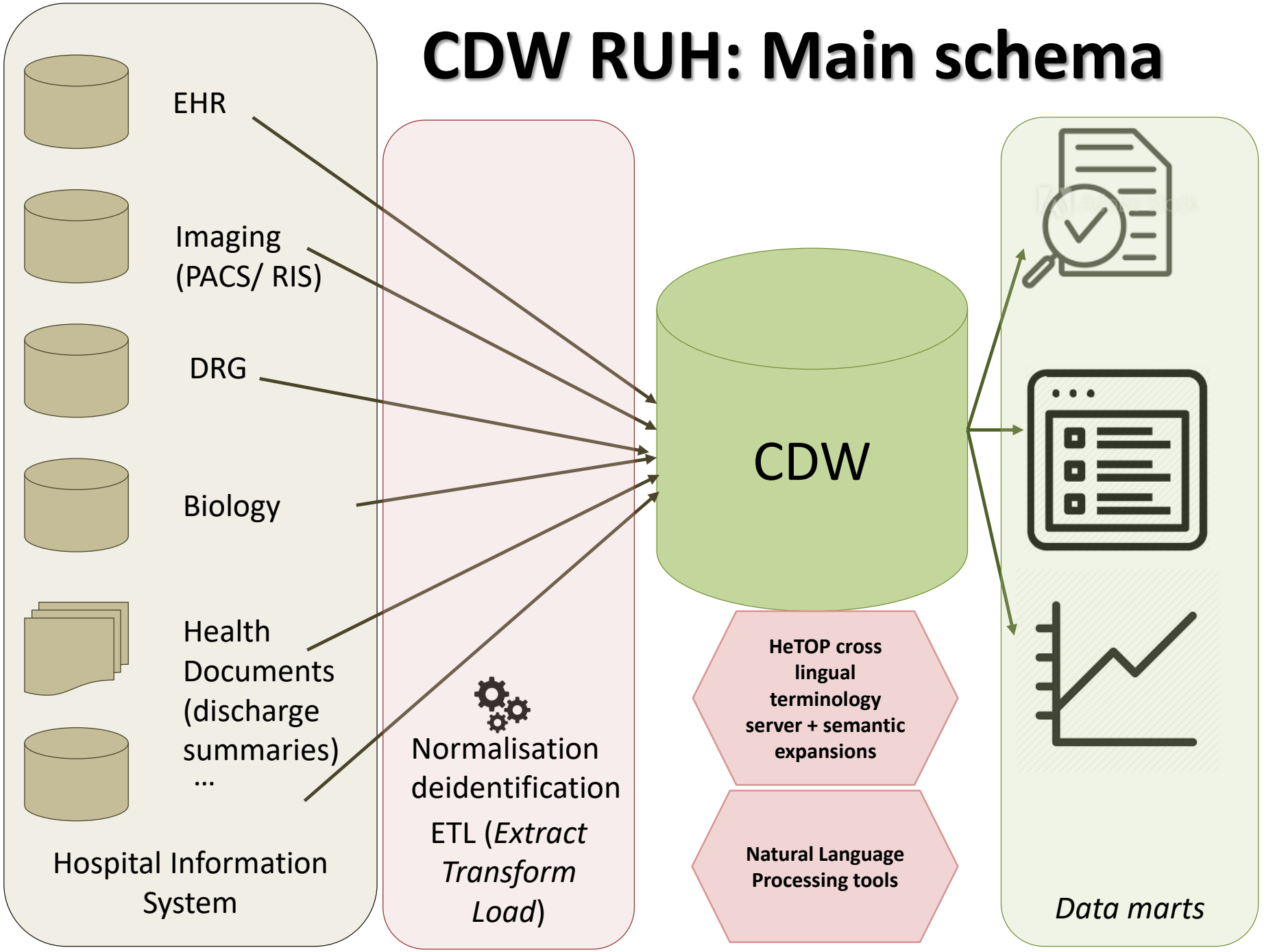
LIMICS

Definition

- Clinical Data Warehouse (CDW) are computer tools allowing collection, integration, then data mining. These data are extracting from various clinical information systems: EHR, LIS, RIS/PACS, CPOE, etc…
- Agregation of a maximum of data available from patients
  - First step, HDW from DRG data
- No matter what is the Clinical Information System (CIS)
  - As far as possible, retrieving all the data from CIS, no matter old they are; going back to 2000 ≠ Paris Hospitals (APHP)

➔ Allowing to link information and to select as precisely as possible the right patients

CDW RUH: Main schema



HDW Objectives in general

- To optimize DRGs by semiautomatic detection of atypical profiles between coding and HDW data (DBI)
- To Improve clinical research thanks to feasibility studies prior to clinical trials & optimization of inclusions
- Detection specific patients profiles
  - e.g. patients frequently admitted to the emergency department
- To create and maintain epidemiological cohorts and registries

HDW Objectives in general

- To detect specific adverse events
  - Vigilances, iatrogenic, cross infections
- To create and follow-up quality indicators
- To Develop and assess computer aided decision support systems (CDASS)
- Access to epidemiology surveillance tools at various level (individual or collective)
- Tools to improve clinical practice
  - Dashboard to provide feedback to individual or collective practice

HDW Objectives for pharmaceutical companies

- To improve clinical research thanks to feasibility studies prior to clinical trials & optimization of inclusions
  - How many patients has the disease A (incidence) in 20xy?
- Detection specific patients profiles
  - e.g. patients frequently admitted to the emergency department
- To create and maintain epidemiological cohorts and registries
- To detect specific adverse events
  - Vigilances, iatrogenic, cross infections
- To create and follow-up quality indicators
- To Develop and assess computer aided decision support systems (CDASS)
- Access to modelization tools for physicians & researchers

CDW State of the art

In the world:

- i2b2  (Harvard MS)
- …

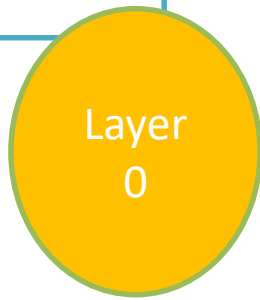
In France :

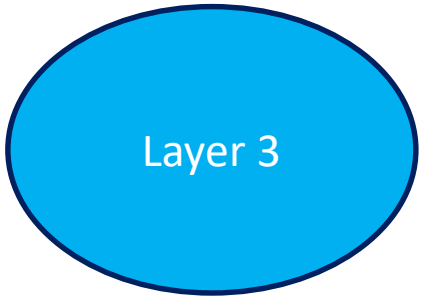
- ehop (Rennes UH)
- Dr Warehouse  (Necker, Foch – Paris)
- ConSoRe
  Continuum Soins Recherche
- EDSaN (Rouen UH)

CDW state of the art

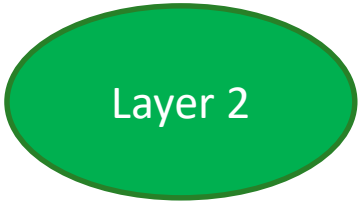
- « CDW » at the national level; in fact, direct use of national EHR, Electronic Health Record
- Existing in some « small » countries (in terms of population)
  - Complexity of digital health not linear with population
  - e.g. Israel (e.g. Clalit), Taiwan, Singapore, Denmark
  - Main publications about COVID-19
- At the HMO level in the US

Semantic HDW based on three independent layers

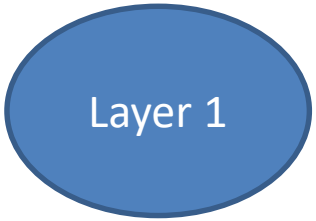




Multiterminology Multilingual Semantic search engine



ECMT Semantic Annotator (NLP & Deep Learning)



HeTOP Terminology server

NoSQL Architecture

HeTOP Terminology server

Layer 1

- Multi Terminology & cross lingual → matrix navigation (among languages & among terminologies
- 100 termino-ontologies included in 55 languages
- 2 M different concepts in English; 0.7 M in French
  - ≈ 165 K concepts in French in UMLS (2018AB) vs. ≈ 630 K in HeTOP (x3.6)
- 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)
- 21.4 M health documents
- Processing time: 30 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

Siefridt C, Grosjean J, Lefebvre T, Rollin L, Darmoni S, Schuers M. Evaluation of automatic annotation by a multi-terminological concepts extractor within a corpus of data from family medicine consultations. Int J Med Inform. 2020 Jan;133:104009. doi: 10.1016/j.ijmedinf.2019.104009. Epub 2019 Nov 1.

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

Overall ECMT process

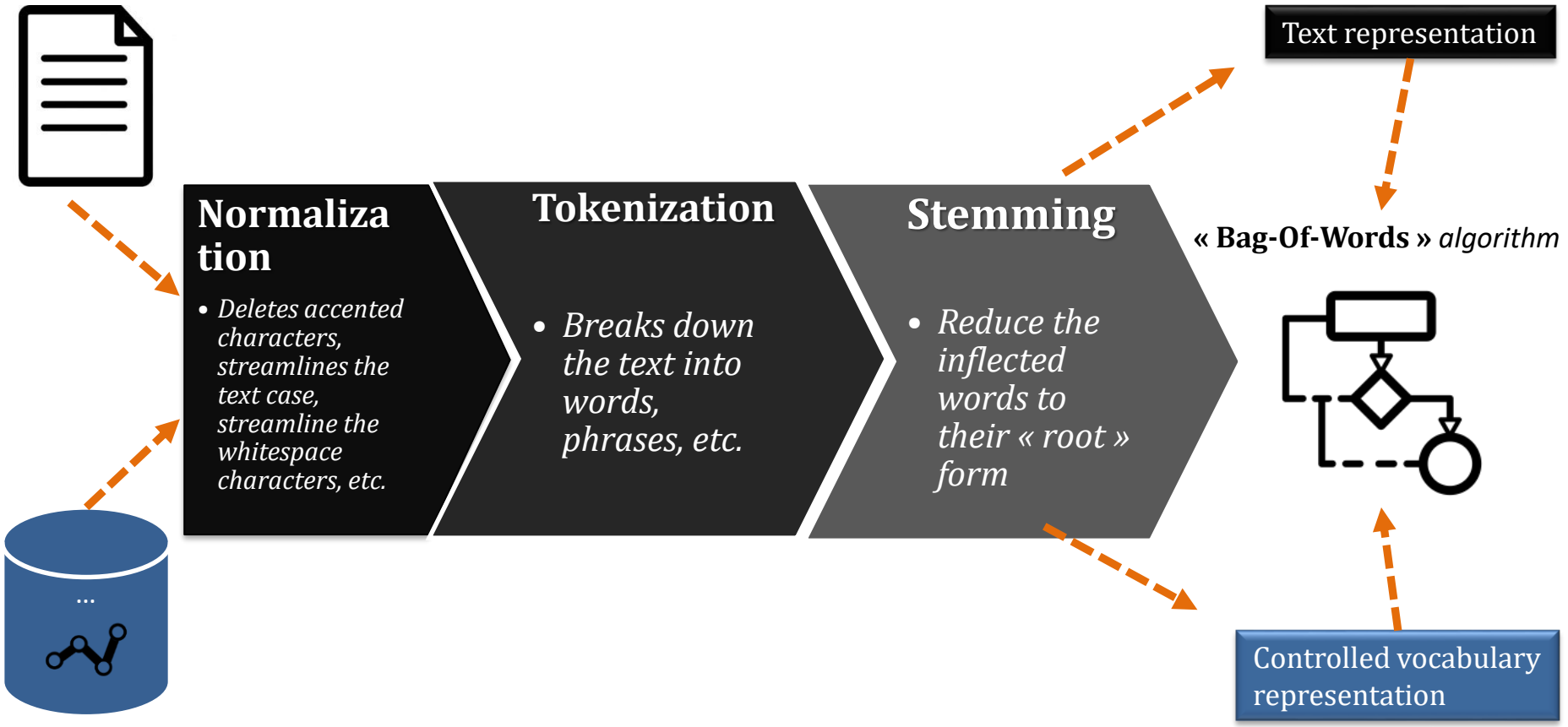


Table 7. System performance for ICD10 coding on the **French raw** test corpus in terms of Precision (P), recall (R) and F-measure (F). A horizontal dash line places the frequency baseline performance. The top part of the table displays official runs, while the bottom part displays non-official and baseline 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
  - Multiterminological queries are possible (≠ PubMed monoterminology MeSH)
  - Explosion (subsumption) based on hierarchy at a multiterminology level not anymore active for EDSaN (active for CISMeF/LiSSa)
  - Semantic expansions (synonyms based on HeTOP and its mappings)
  - Other relations may be used: semantic expansions based on inter-terminology semantic mappings
  - Multilingual +++

Lelong, Romain; Soualmia, Lina F; Grosjean, Julien; Taalba, Mehdi & Darmoni, SJ. Building a Semantic Health Data Warehouse: Evaluation of a search tool in Clinical trials. JMIR Medical Informatics , 2019 Dec 20;7(4):e13917. doi: 10.2196/13917.

Multiterminology Multilingual Semantic search engine

Layer 3

- Very generic tool
- Designed to be used for N patients (regular use of HDW) or for one patient (not politically correct)

→ use of HDW to consult patient data for CARE +++

Two or three generations of software (LIS, RIS, EHR) compiled in one HDW

Show me all the electroretinography reports for this patient

→ one second vs. at least one minute

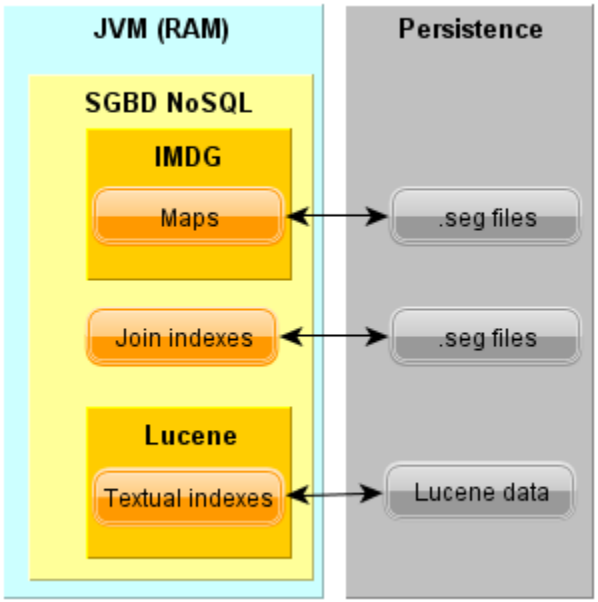
TECHNICAL ASPECTS

NoSQL Architecture IN Rouen CDW

Layer 0

- Java
- Custom NoSQL (In Memory Data Grid – IMDG-): Key-Value store, join indexes (replacing SQL joins) & Lucene (NLP)

- Three powerful servers:
  - 1 To RAM & 196 cores (prod)
  - 1 To RAM & 144 cores (preprod) * 2
  - Semantic Annotator + deep learning



HEALTH DATA WAREHOUSE - SECURE ARCHITECTURE PROPOSAL

HDW PRODUCTION ENVIRONMENT

SECURE HOSTING

HDW APPLICATION SERVER (HTTPS)

- NO NOMINATIVE DATA
- HDW INTERNAL PATIENTS IDS
- ENCRYPTED PARTITION
- ANONYMIZED RECORDS

IMDG

MEDICAL DATA PERSISTENCE SERVER

- NO NOMINATIVE DATA
- HDW INTERNAL PATIENTS IDS
- ENCRYPTED PARTITION
- ANONYMIZED RECORDS

POSTGRESQL 10

NOMINATIVE DATA PERSISTENCE SERVER

- ISOLATED NOMINATIVE INFORMATIONS
- NO MEDICAL DATA
- ENCRYTED JOIN (SALT)
  HDW PATIENTS IDS - HOSPITAL IDS
- ENCRYPTED PARTITION
- ENCRYPTION OF SENSITIVE DATA (NAMES..)
  IN DATABASE

POSTGRESQL 10

RESTRICTION TO AUTHORIZED IP

USERS

REFERENT USER

AUTHENTIFIED SESSION (BROWSER)

D2IM (DIM+SIBM)

SECURE DELIVERY (TEMPORARY)
ENRCYPTED ARCHIVE

HDW INTERNAL NEEDS (APPROVED DEMAND)

SSH

HDW ADMINISTRATORS

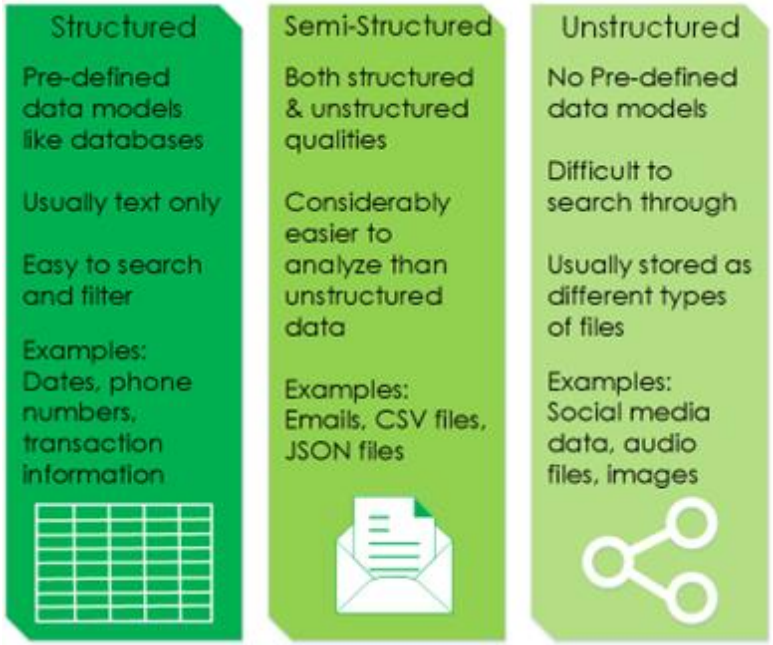
Acess Policy to EDSaN

https://edsan.chu-rouen.fr/edsan/acces-aux-donnees/procedures/

EDSaN has been credited by the French CNIL, respect to European GPDR (nov 2018)

1. Demand sent to Department of Clinical Research (commission of qualification)
2. Transmission to the Scientific and Ethical Committee
3. « Physical » appointment in the DDH to create the EDSaN queries
   - Ethical & Juridical aspects: EDSaN provides access to all RUH health since 2000
   - Scientific aspects : bias of data, bias of tools...
4. Access to queried and specific data in a encrypted environment (nominative access)
5. Potential exports (realized after validation by a specific committee)

Types of Big Data



https://mdaca.io/2021/05/whats-the-big-data/

Structured:

Biology, DRGs data, CPOE data (prescription)

Semi-Structured:

Unstructured:

Text from health documents

Quality of data

1. Quality of data is varying according to the type of data
   - Structured
   - Semi-structured
   - Non structured (80% of the data in France; a little bit less in AngloSaxon countries)
2. The better quality lays in structured data
   - BUT even structured data may generate errors
   - In biology, some errors exists from automats (e.g. kaliemia K+ in blood); supervision by a human biologist, as kaelimia is of utmost importance for the patient +++
3. Least quality for unstructured data
   - Need of a semantic annotator to extract medical concepts
   - Precision (false positive) / Recall (false negative) is varying according to the context

Big Data, definitions

From the 3Vs to 10 Vs (Arriel Benis, PhD)

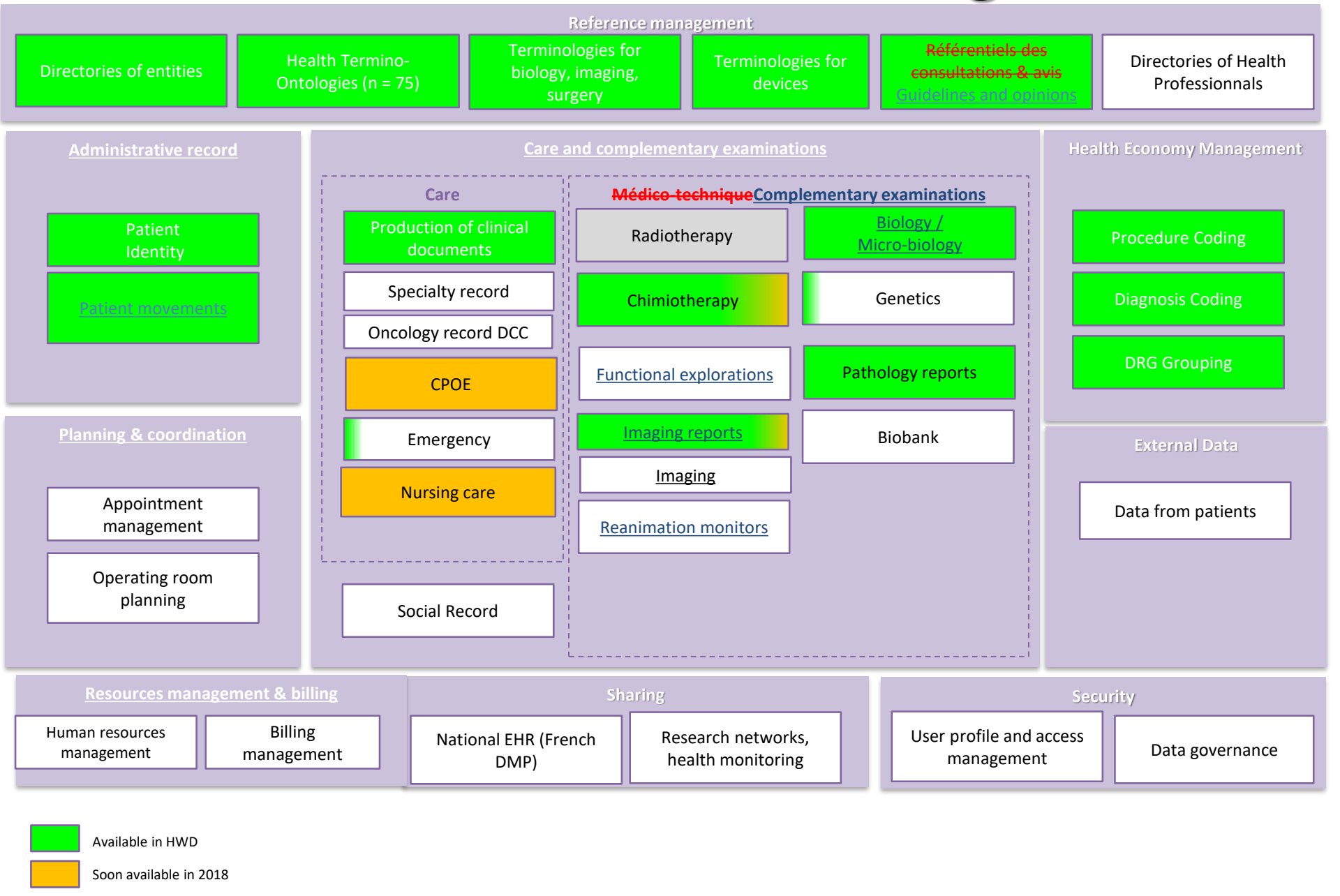
Volume / נפח	•More than 90% of ever generated data where produced these last years
Velocity / מהירות	•Speed at which data is being generated, produced, created, or refreshed
Variety / שוני	•Structured, Semi-Structured, Unstructured data
Variability / השתנות	•Inconsistencies in the data resulting from multiple disparate data types and sources
Veracity / אמינות	•Confidence or trust in the data… dropping when Volume, Velocity, Variety, Variably increase
Validity / תקפות	•Accuracy and correctness of the data is for its intended use
Vulnerability / פגיעות	•Security concerns
Volatility / נדיפות – אי-יציבות	•How old does data need before being considered irrelevant, historic, or not useful any longer?
Visualization /ויזואליזציה	•Challenges due to technical (in-memory, scalability, processing time) and human (perception)
Value	•Driving business value from data

Results

HDW Rouen - volumetry

Patients	2 millions
Stays (hospitalisations, séances, consultations)	22.3 millions
Documents (reports, notes, letters, manual prescriptions...)	21.4 millions > **1G** medical concepts
CPOE	1.8 millions
Lab tests (hematology, biochemistry...)	176 millions
Medical devices	116,000
Diagnostics (ICD-10)	11.6 millions
Procedures (imaging, surgery...) (CCAM)	10.7 millions

HDW Rouen – functional coverage



Main steps are already performed…

- A team of FIVE engineers are devoted to this project from DDH + physicians from the DRG dpt
- Scalability of the French semantic annotator
  - 21 M reports extracted in 30h; possible to rerun again every week if necessary (5ms for one report); in reality, each semester (soberness+++)
- Scalability of the semantic search engine
  - Using a NoSQL architecture and two powerful servers, response time ≈ 2 s
  - Might be used in care context
    - Display the last US of this patient ASAP

EDSaN Platform

EDSaN

D2IM Département d'Informatique et d'Information Médicales

CHU ROUEN NORMANDIE

Consulter vos jeux de données

Cet environnement est consacré à la sélection des données ou des patients afin de créer des cohortes ou des tableaux de données à partir de l'EDSaN. Pour cela, une ou plusieurs requêtes doivent avoir été élaborées et exécutées pour pouvoir rendre accessible les données à explorer.

Accéder à votre espace personnel

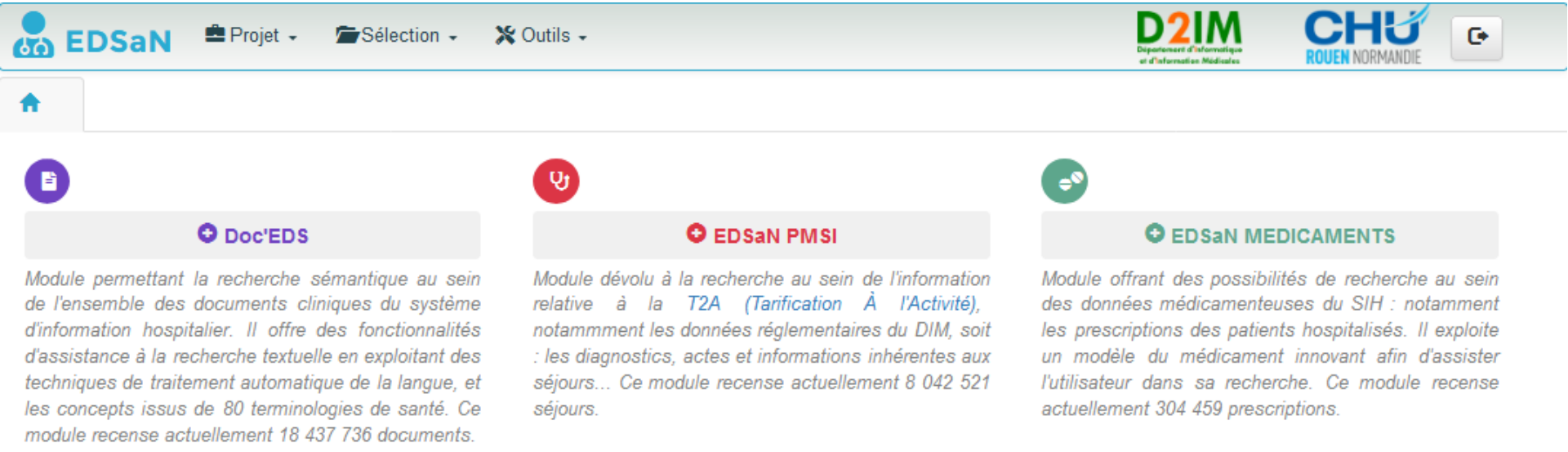
Créer ou ouvrir des requêtes

Uniquement accessible par les administrateurs EDSaN, cet environnement permet de créer des requêtes pour quelles soient notamment disponibles à la consultation par les chercheurs.

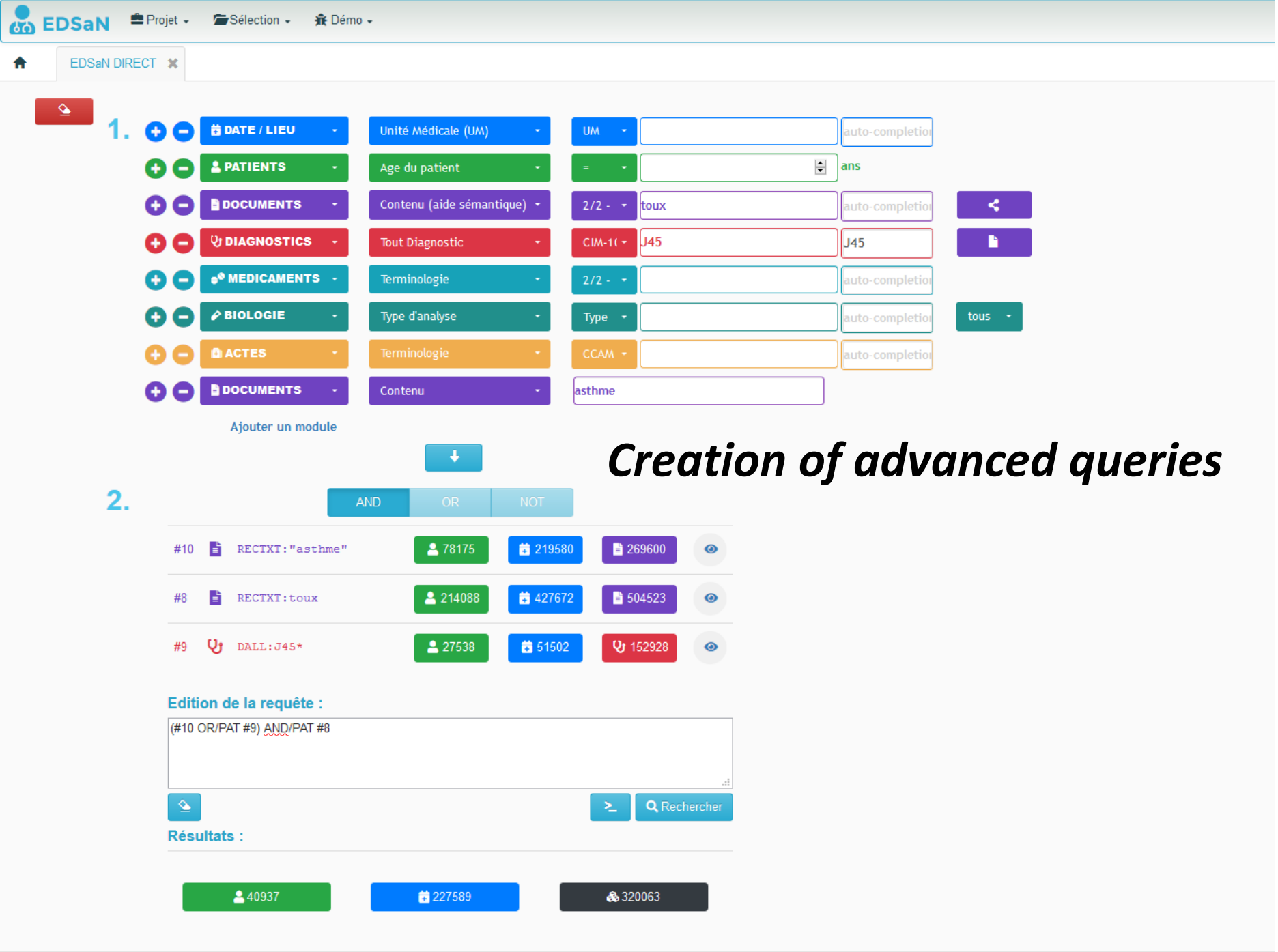
Créer une requête

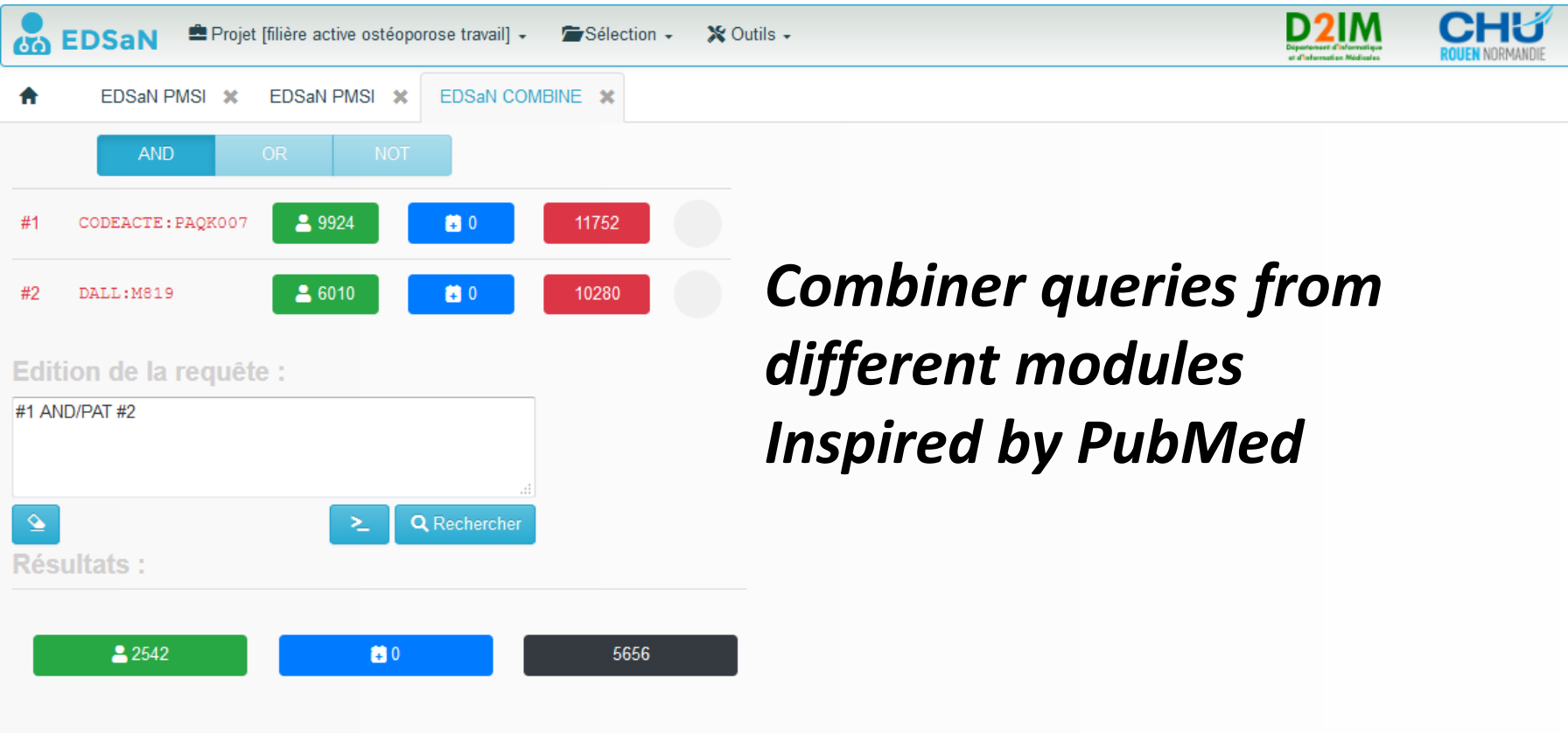
Modular search engine

- 1 module for 1 data type +++
- Currently available:
  - Health documents (80% of valuable information[1])
  - DRGs
  - CPOE
  - Medical devices
  - Lab tests



[1] Raghavan P, Chen JL, Fosler-Lussier E, Lai AM. How essential are unstructured clinical narratives and information fusion to clinical trial recruitment? AMIA Jt Summits Transl Sci Proc. 2014 Apr 7;2014:218-23. PMID: 25717416; PMCID: PMC4333685.


EDSaN
Projet
Sélection
Démo
EDSaN DIRECT
1.
DATE / LIEU
Unité Médicale (UM)
UM
auto-completio
PATIENTS
Age du patient
=
ans
DOCUMENTS
Contenu (aide sémantique)
2/2 -
toux
auto-completio
DIAGNOSTICS
Tout Diagnostic
CIM-1(
J45
J45
MEDICAMENTS
Terminologie
2/2 -
auto-completio
BIOLOGIE
Type d'analyse
Type
auto-completio
tous
ACTES
Terminologie
CCAM
auto-completio
DOCUMENTS
Contenu
asthme
Ajouter un module
Creation of advanced queries
2.
AND
OR
NOT
#10 RECTXT:"asthme" 78175 219580 269600
#8 RECTXT:toux 214088 427672 504523
#9 DALL:J45* 27538 51502 152928
Edition de la requête :
(#10 OR/PAT #9) AND/PAT #8
Rechercher
Résultats :
40937
227589
320063



Combiner queries from different modules
Inspired by PubMed

EDSaN | Liste des projets | Sélections | Options | D2IM | CHU ROUEN NORMANDIE

Projet [my_project]

Sélection :

EDSaN PMSI#2 (17299)

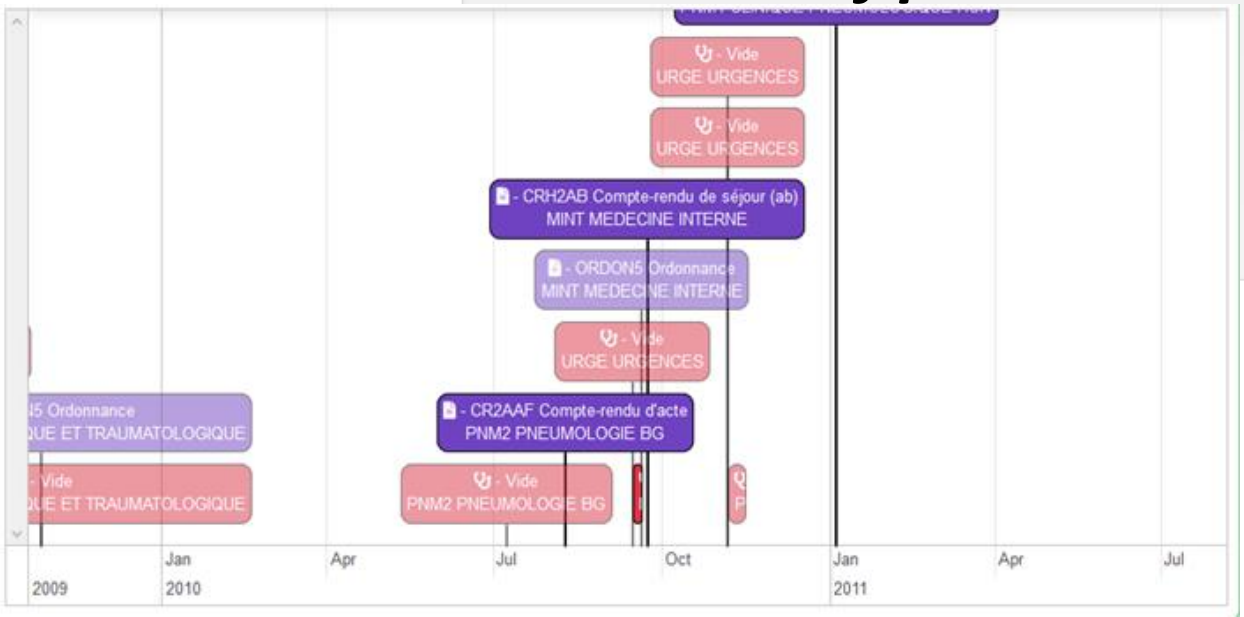
1 de 17299 patients

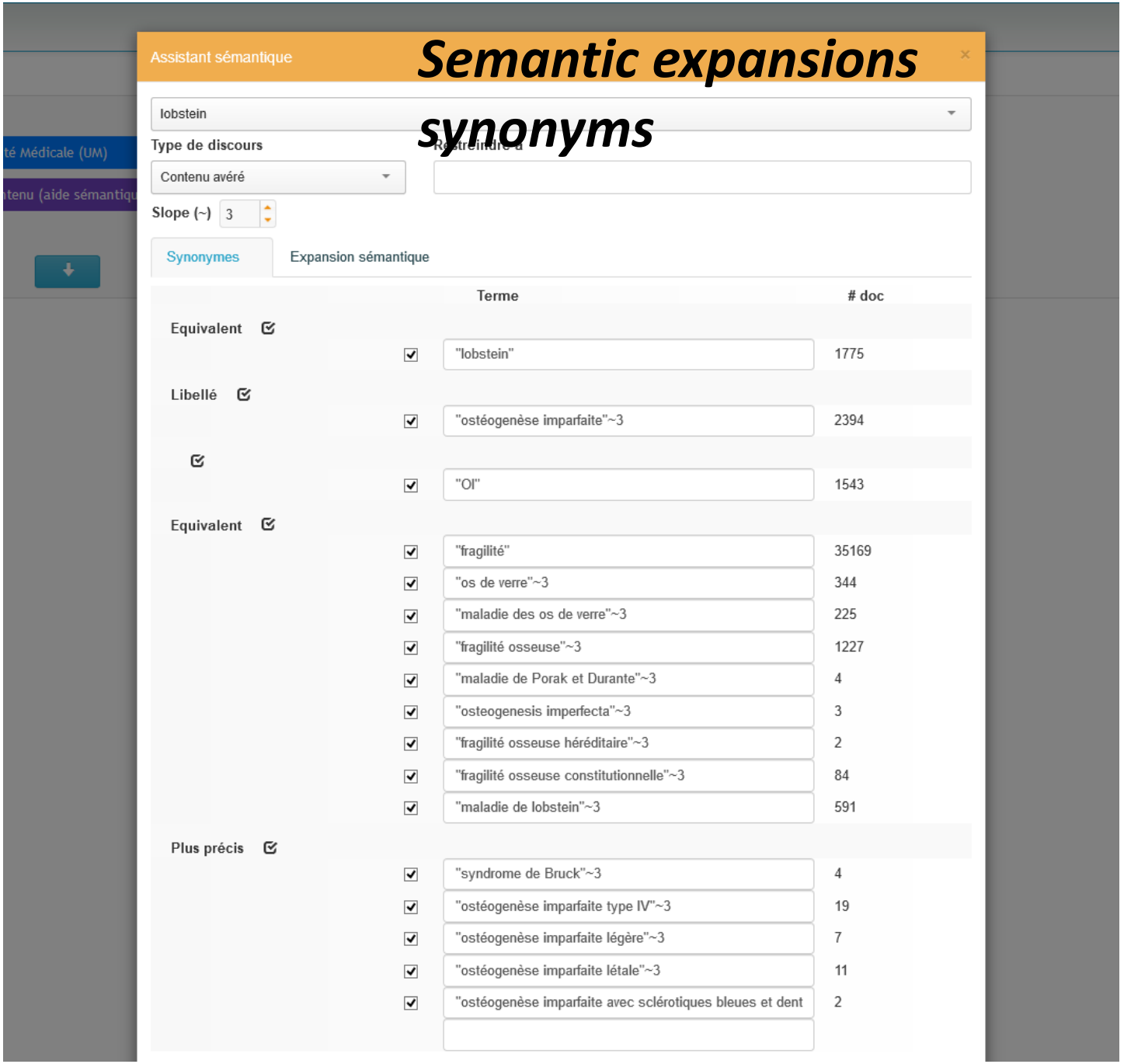
	# Pat.	
Dossiers inclus	33	Supprimer
Dossiers exclus	8	Ajouter
Faux positifs	6	Ajouter
Dossiers à revoir	2	Ajouter

Patient 00578, 1 éléments trouvés; 7 affichés :

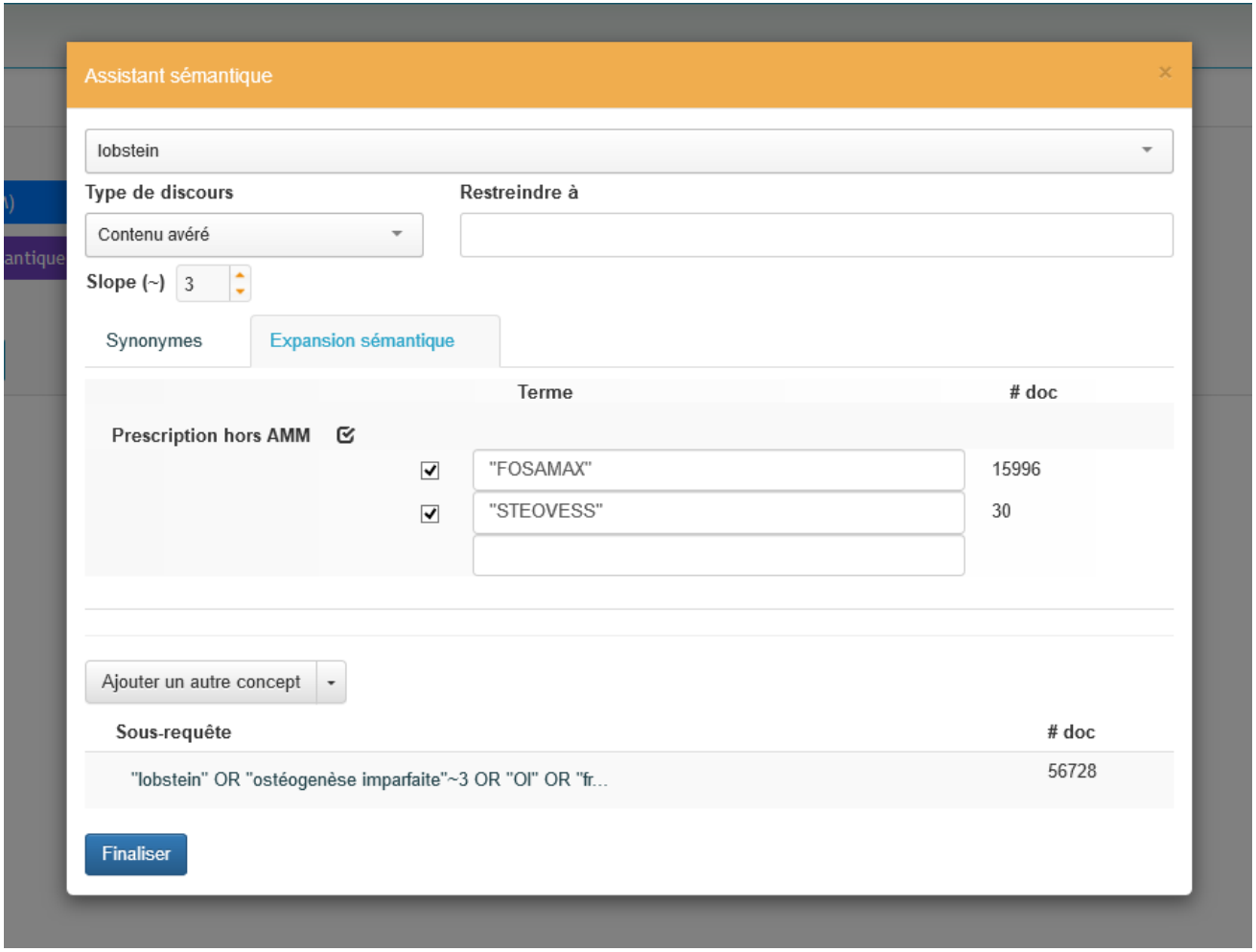
	Date	UF	ID évènement
159653888	2019-11-07	UROL	463647674
708923428	2019-11-05	UROL	463647674
248521848	2019-10-28	UROL	463647674
409120747	2019-10-21	UROL	463647674
180542561	2019-10-18	UROL	463647674
569448439	2019-10-16 00:00:00.0	UROL	463647674
432363505	2019-10-16	UROL	463647674

Consultation and selection of patients

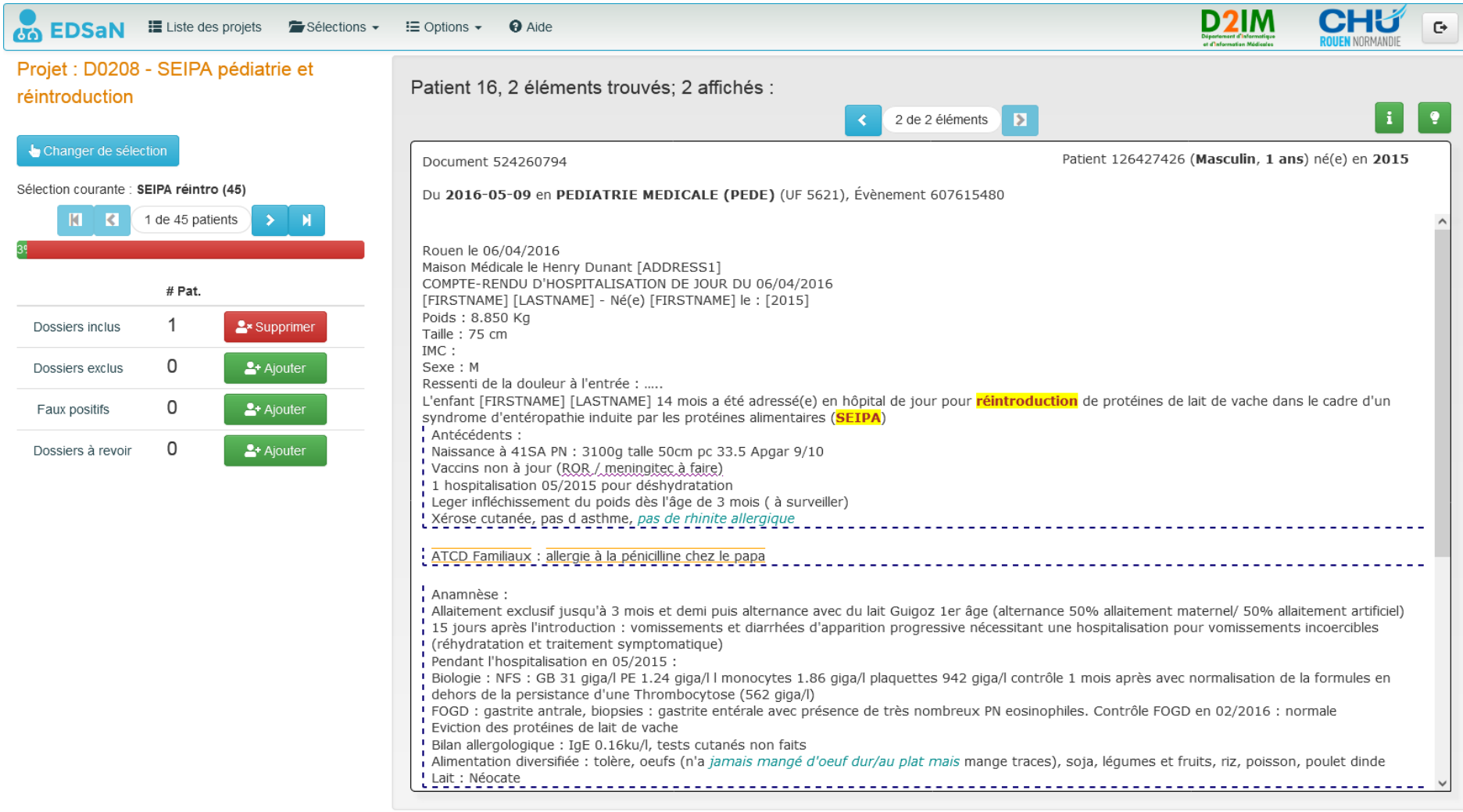



Assistant sémantique
Semantic expansions
synonyms
lobstein
Type de discours
Contenu avéré
Slope (~) 3
Synonymes
Expansion sémantique
Terme
doc
Equivalent
"lobstein" 1775
Libellé
"ostéogenèse imparfaite"~3 2394
"OI" 1543
Equivalent
"fragilité" 35169
"os de verre"~3 344
"maladie des os de verre"~3 225
"fragilité osseuse"~3 1227
"maladie de Porak et Durante"~3 4
"osteogenesis imperfecta"~3 3
"fragilité osseuse héréditaire"~3 2
"fragilité osseuse constitutionnelle"~3 84
"maladie de lobstein"~3 591
Plus précis
"syndrome de Bruck"~3 4
"ostéogenèse imparfaite type IV"~3 19
"ostéogenèse imparfaite légère"~3 7
"ostéogenèse imparfaite létale"~3 11
"ostéogenèse imparfaite avec sclérotiques bleues et dent 2

Semantic expansions
Exploitation of HeTOP semantic relations



Selection tool

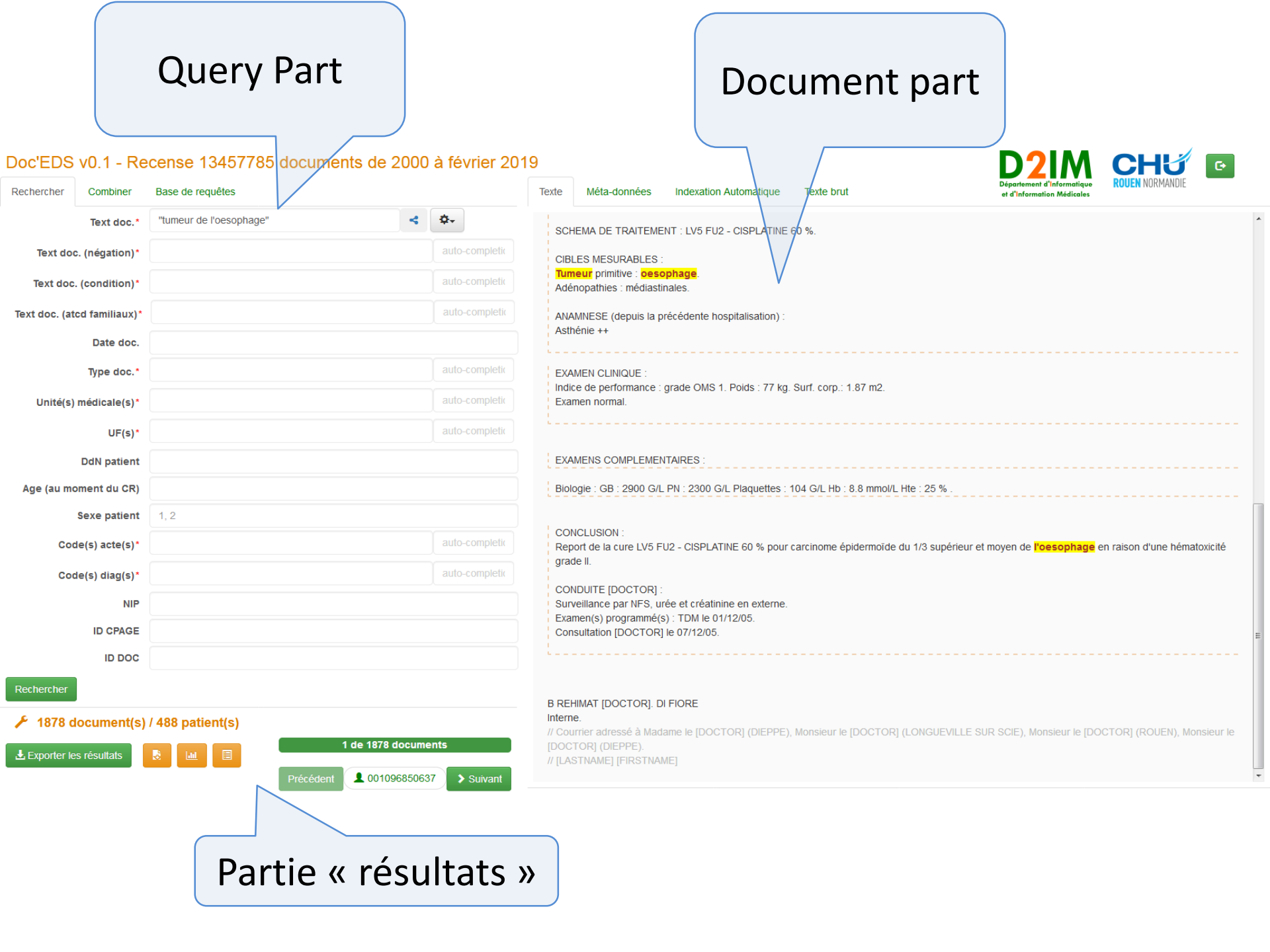


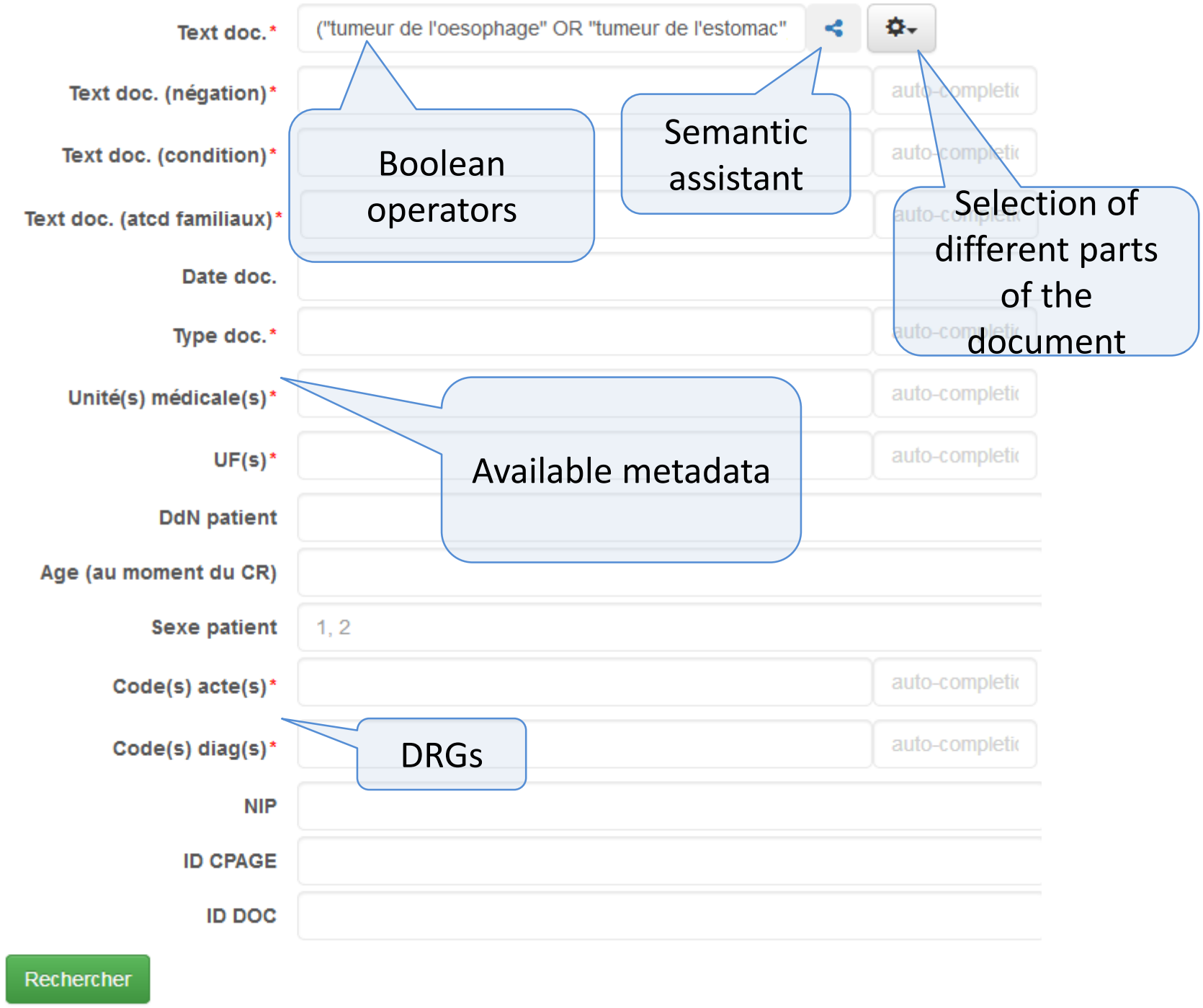
Doc'EDS

- New tool developed in 2019, based on the first 20 use cases since March 2018
- No semantics… for now => better cost-effectiveness
- Based on Lucene
- To better understand the deeper structure of the document (80% of the health information in France)
  - Negation
  - Hypothesis
  - Family history
  - Segmentation of the document

Doc'EDS

- Allow « manual dataminig »: oxymoron!
- Tool to validate the quality of the queries: continuous improvment of the search engine
- Based on rules, regular expressions and frequency analyses


Query Part
Document part
Doc'EDS v0.1 - Recense 13457785 documents de 2000 à février 2019
D2IM
Département d'Informatique et d'Information Médicales
CHU ROUEN NORMANDIE
Rechercher
Combiner
Base de requêtes
Text doc. *
"tumeur de l'oesophage"
Text doc. (négation) *
Text doc. (condition) *
Text doc. (atcd familiaux) *
Date doc.
Type doc. *
Unité(s) médicale(s) *
UF(s) *
DdN patient
Age (au moment du CR)
Sexe patient
1, 2
Code(s) acte(s) *
Code(s) diag(s) *
NIP
ID CPAGE
ID DOC
auto-completi
Rechercher
1878 document(s) / 488 patient(s)
Exporter les résultats
1 de 1878 documents
Précédent
001096850637
Suivant
Texte
Méta-données
Indexation Automatique
Texte brut
SCHEMA DE TRAITEMENT : LV5 FU2 - CISPLATINE 60 %.
CIBLES MESURABLES :
Tumeur primitive : oesophage.
Adénopathies : médiastinales.
ANAMNESE (depuis la précédente hospitalisation) :
Asthénie ++
EXAMEN CLINIQUE :
Indice de performance : grade OMS 1. Poids : 77 kg. Surf. corp.: 1.87 m2.
Examen normal.
EXAMENS COMPLEMENTAIRES :
Biologie : GB : 2900 G/L PN : 2300 G/L Plaquettes : 104 G/L Hb : 8.8 mmol/L Hte : 25 % .
CONCLUSION :
Report de la cure LV5 FU2 - CISPLATINE 60 % pour carcinome épidermoïde du 1/3 supérieur et moyen de l'oesophage en raison d'une hématoxicité grade II.
CONDUITE [DOCTOR] :
Surveillance par NFS, urée et créatinine en externe.
Examen(s) programmé(s) : TDM le 01/12/05.
Consultation [DOCTOR] le 07/12/05.
B REHIMAT [DOCTOR]. DI FIORE
Interne.
// Courrier adressé à Madame le [DOCTOR] (DIEPPE), Monsieur le [DOCTOR] (LONGUEVILLE SUR SCIE), Monsieur le [DOCTOR] (ROUEN), Monsieur le [DOCTOR] (DIEPPE).
// [LASTNAME] [FIRSTNAME]
Partie « résultats »


Text doc.*
("tumeur de l'oesophage" OR "tumeur de l'estomac"
Text doc. (négation)*
Text doc. (condition)*
Text doc. (atcd familiaux)*
Date doc.
Type doc.*
Unité(s) médicale(s)*
UF(s)*
DdN patient
Age (au moment du CR)
Sexe patient
1, 2
Code(s) acte(s)*
Code(s) diag(s)*
NIP
ID CPAGE
ID DOC
Rechercher
Boolean operators
Semantic assistant
Selection of different parts of the document
Available metadata
DRGs

RAMI[DOCTOR] 5 mg, un par jour
FUROSEMIDE 20 mg, un par jour
SELOKEN 200 mg, 1 par jour
SINVASTATINE 40 mg, un par jour
INNOHEP 14000, une injection par jour
IMOVANE 7.5 mg, un au coucher
VESICARE 5 mg, un par jour
UROREC 8 mg, un par jour

MODE DE VIE
Marié (femme aidante)

HISTOIRE DE LA MALADIE :
Patient de [redacted] ans hospitalisé le matin [redacted]/04 à Becquerel pour une cure de radiothérapie dans le cadre de la prise en charge d'un carcinome épidermoide de **l'oesophage** suivit au CHU en gastro-entérologie (médecin référent : [DOCTOR]).
Le patient décrit des expectorations mêlées de sang survenues en deuxième moitié de nuit (nuit du [redacted] au [redacted]/04). Un second épisode de crachats sanglants après effort de toux au cours de l'évaluation pré radiothérapie. Dans le contexte de **tumeur** de **l'oesophage**, le patient est transféré en unité de soins intensifs de gastro-entérologie au CHU pour une probable hématémèse.

EXAMEN CLINIQUE :
Taille : 176 cm : 76 kg. IMC : 24.5 kg/m². Température : 36.4°C Pouls : 77 par mn. TA : 109/70. *EVA : 0.* Hb : 12.6 g/Dl. Bonne tolérance hémodynamique de l'hémorragie, *pas de déglobulisation*
Dénutrition. OMS 2-3.
Conjonctives colorées. *Absence d'ictère.*
Cardiovasculaire : BDC irrégulières. *Pas de souffle. Pas d'OMI Pas de Turgescence des jugulaires.*
Respiratoire : Râles crépitants à la base pulmonaire gauche. Toux importantes suivies d'expectorations teintées de sang.
Digestif : présence d'une sonde naso-gastrique d'alimentation entérale. Abdomen souple, indolore. BHA présents et normaux. *Absence d'argument en faveur d'une ascite clinique.* Toucher rectal : selles dures de couleur normale (pas de méléna *ni de rectorragies*). Prostate de taille augmentée avec disparition du sillon médian.
Neurologique : bonne orientation temporo-spatiale. *Pas de déficit sensitivo-moteur.*

BIOLOGIE :
Hémoglobine à 12,3 g/dl. .Sérologie et antigénurie aspergillaire en cours.

RADIOLOGIE :
Radio de thorax :
TDM du [redacted]02/20[redacted] : Stabilité de l'épaississement circonférentiel oesophagien avec multiples adénomégalies médiastinales.Apparition de plusieurs nodules lobaires inférieures gauches :

Anonymization (patients and physicians)

Detection of suspiscion/hypotheses/ doubts/future

Détection of negations

SCHEMA DE TRAITEMENT : Cacis (Campo + Cisplatine) + radiothérapie.

CIBLES MESURABLES :
Tumeur primitive : oesophage.
Adénopathies : oui.

ANAMNESE (depuis la précédente hospitalisation) :

Antécédents : chirurgie sur le ménisque droit. Chirurgie sur varices superficielles des membres inférieurs. Terrain atopique : eczéma.
Frère décédé d'un cancer du poumon il y a 30 ans.
Oncle décédé d'un cancer de l'oesophage il y a 16 ans.

EXAMEN CLINIQUE :
Indice de performance : grade OMS 0. Poids : 89 kg. Surf. corp.: 2 m2.

EXAMENS COMPLEMENTAIRES :

Biologie : GB : 6600 G/L PN : 3890 G/L Plaquettes : 272000 G/L Hb : 13,5 g/dl. TP : 99 %.
Urée : 0,24. Créat : 7,3 mg/l. TCA : 1,18.

TRAITEMENT RECU :
Chimiothérapie :
- Campo 60 mg/m2/j soit 120 mg/j (soit 100 % de la dose standard).
- Cisplatine 30 mg/m2/j soit 60 mg/j (soit 100 % de la dose standard).
- Traitement anti-émétique préventif par Zophren.
- Traitement anti-émétique curatif par Solumédrol.

Radiothérapie : à prévoir.

« Segmentation » du document

Détection des antécédents familiaux

Analyses

Différents bilans

Statistiques & bilans

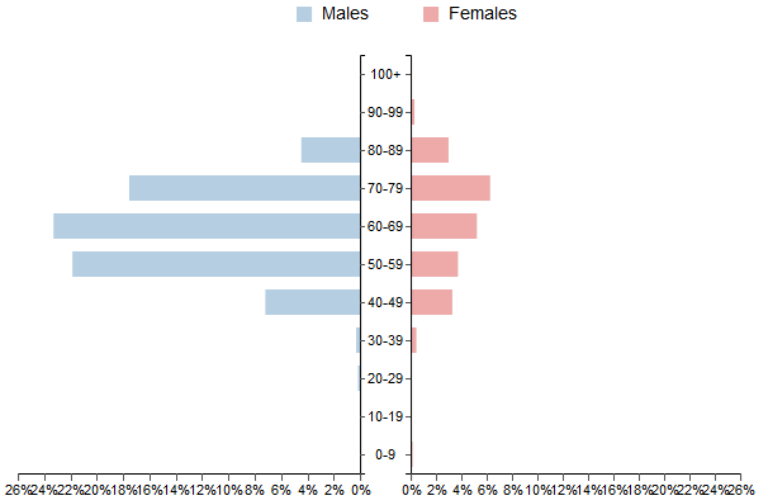
- Stats démographiques
- Date doc.
- Type doc.
- Unité(s) médicale(s)
- UF(s)
- DdN patient
- Age (au moment du CR)
- Sexe patient
- Code(s) acte(s)
- Code(s) diag(s)

150 femmes (22,4%)

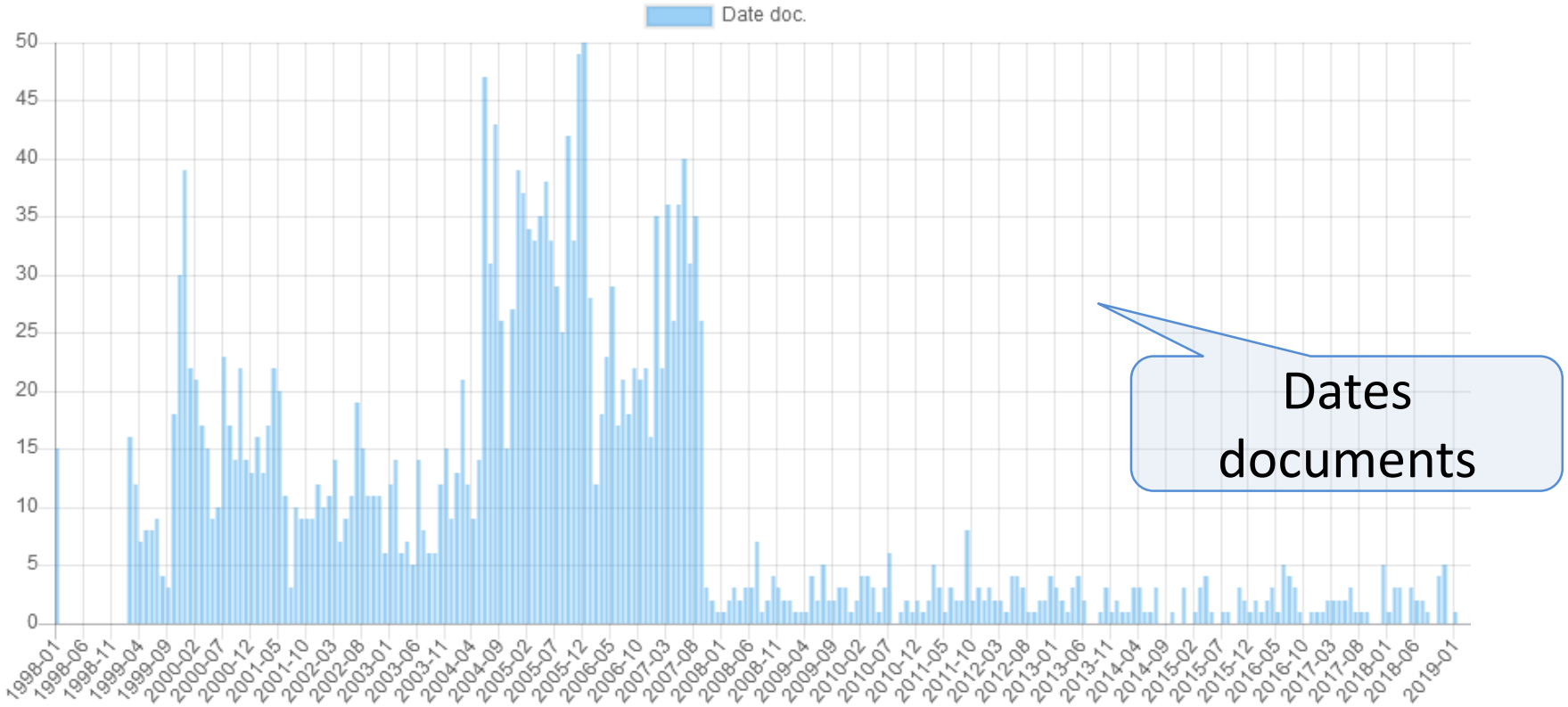
520 hommes (77,6%)

Statistiques des âges (au moment du CR)

Moyenne	63,3
Écart type	11,5
Minimum	7
Maximum	93
Q1	54
Q2	63
Q3	72



Population



SEJUM	#	%
DIGE HEPATO GASTRO ENTEROLOGIE NUTRITION	1685	61,9%
CGCD CHIRURGIE GENERALE ET DIGESTIVE	235	8,6%
RADI IMAGERIE CENTRALE	149	5,5%
URGE URGENCES	128	4,7%
PHIE PHARMACIE	93	3,4%
ORLO O.R.L ADULTES	61	2,2%
PNM1 CLINIQUE PNEUMOLOGIQUE HCN	59	2,2%
PHYS PHYSIOLOGIE DIGESTIVE	34	1,2%
CARD CARDIOLOGIE	25	0,9%
REAC REANIMATION CHIRURGICALE	23	0,8%

Répartition dans les unités médicales / services

Diagnostics
PMSI, etc.

Code(s) diag(s) | DP | DA | DR

Afficher 10 **éléments** **Filtrer :**

155 valeurs distinctes, 2369 occurrences au total

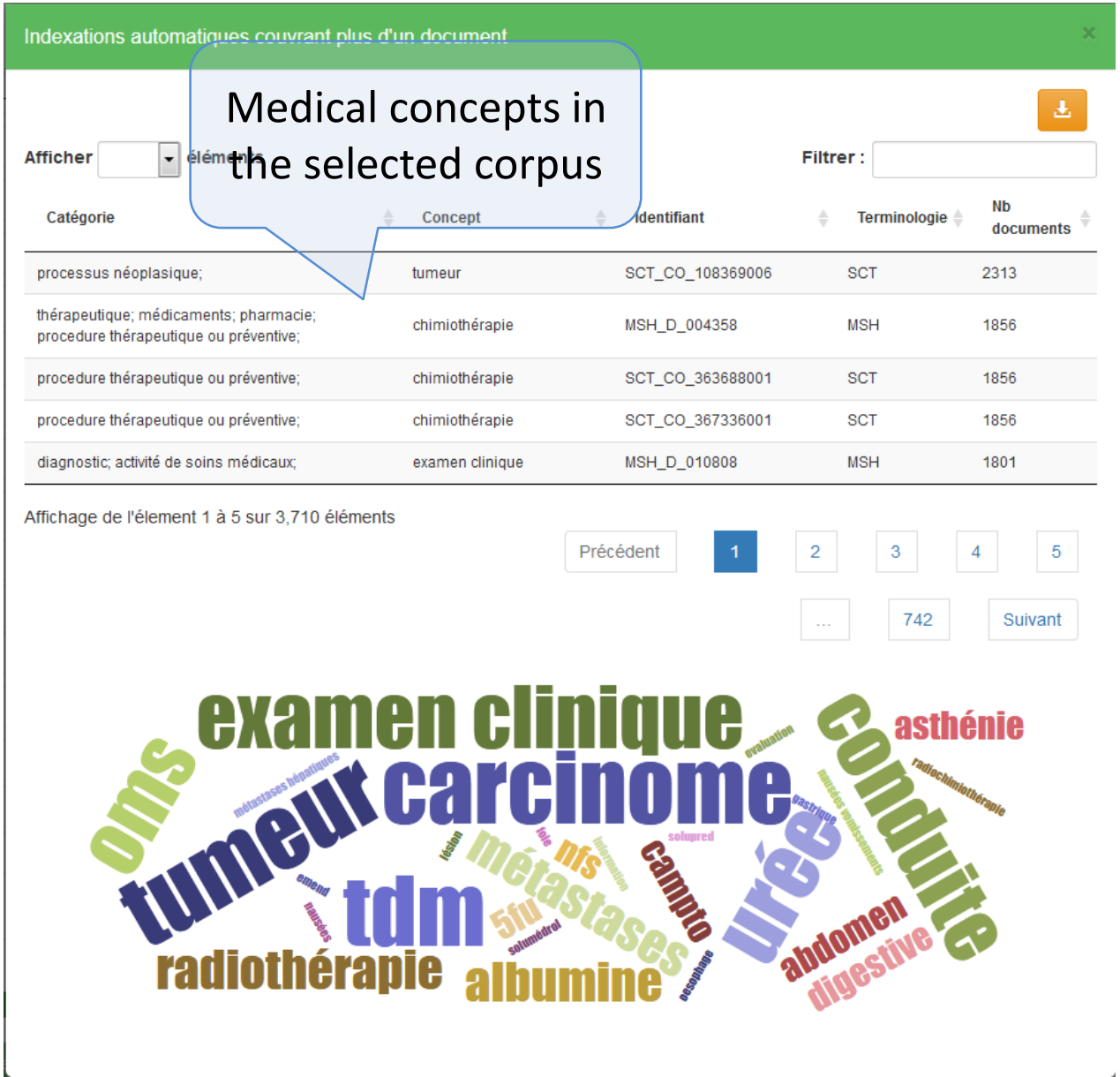
DIAGCODES	#	%
Z511 séance de chimiothérapie pour tumeur	1292	54,5%
Z530 acte non effectué en raison de contre-indication	302	12,7%
C155 tumeur maligne du tiers inférieur de l'oesophage	86	3,6%
C151 tumeur maligne de l'oesophage thoracique	66	2,8%
Z087 examen de contrôle après traitements combinés pour tumeur maligne	63	2,7%
C154 tumeur maligne du tiers moyen de l'oesophage	50	2,1%
C150 tumeur maligne de l'oesophage cervical	39	1,6%
C153 tumeur maligne du tiers supérieur de l'oesophage	39	1,6%
Z452 ajustement et entretien d'un dispositif d'accès vasculaire	34	1,4%
C159 tumeur maligne de l'oesophage, sans précision	24	1,0%

Affichage de l'élement 1 à 10 sur 155 éléments

Précédent 1 2 3 4 5 … 16

Suivant

Data mining (ECMT) => word cloud



Filtering (diseases, procedures, …) : here drugs

Afficher éléments

Filtrer : PHA_RAC

Catégorie	Concept	Identifiant	Terminologie	Nb documents
Substance pharmacologique; produit chimique inorganique;	cisplatine	PHA_RAC_1826	PHA	1150
chimie organique; Substance pharmacologique;	zophren	PHA_RAC_8895	PHA	1015
chimie organique; Substance pharmacologique;	campto	PHA_RAC_1470	PHA	517
chimie organique; Substance pharmacologique;	solumédrol	PHA_RAC_7444	PHA	401
médicament clinique;	solupred	PHA_RAC_7445	PHA	306

Affichage de l'élement 1 à 5 sur 332 éléments (filtré de 3,710 éléments au total)

Précédent 1 2 3 4 5 … 67 Suivant

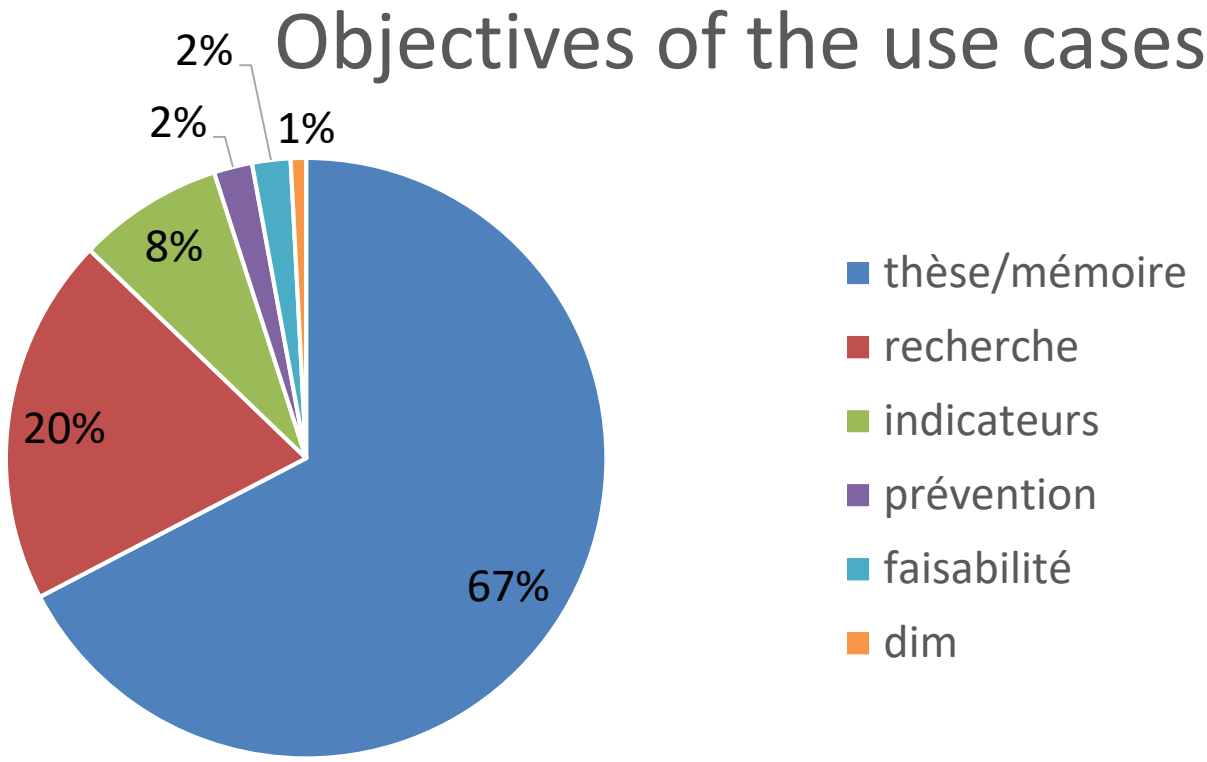
Where EDSaN stands in August 2022

State of the play

- Operational since the beginning 2018 for test cases and prototypes
- CNIL agreement (GPDR +++) obtained in October 2020
- August 2022: **273 use cases already treated** (integrated in a specific registry, according to GPDR requirements)

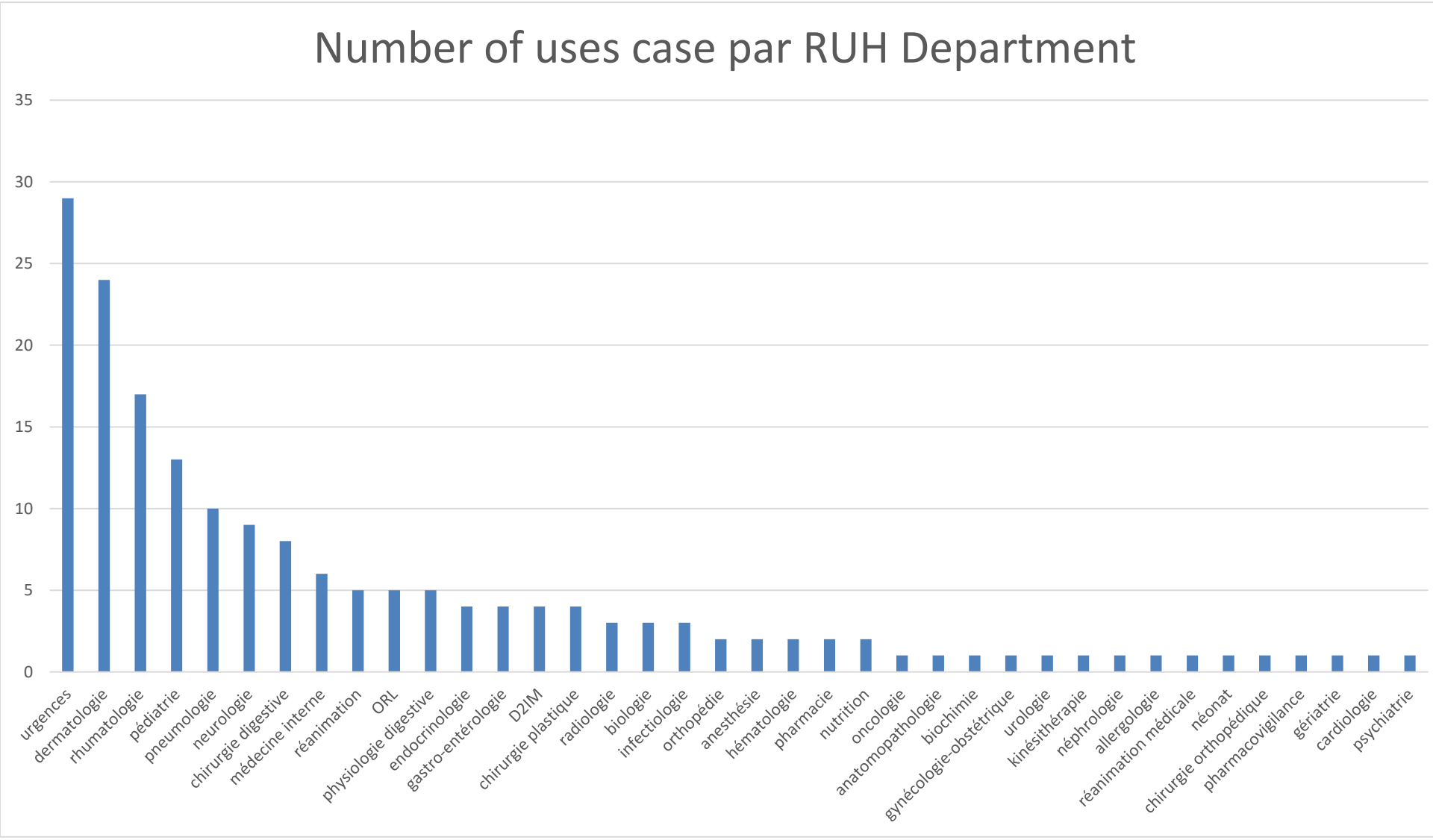
Use cases of the SCDW

- Since March 2018, over 263 use cases treated from SCDW
  - huge feedback to modify our top priorities to be developed
  - One example: to heavily invest on the semantic annotator as 80% of the questions were directly linked to health documents vs. lab tests or drugs
- One example: all lab tests during several years for specific infants
- In June 2018, to identify all the patients with endocarditis AFTE TAVI Transcatheter Aortic Valve Implantation
- To identify the TAVI acronym in the documents, thanks to the semantic annotator (ECMT)
  - Number of cases retrieved based on DRGs data warehouse = 30
  - Number of cases retrieved based on current RUH data warehouse = 53 (2 false negatives)
  - 23 new cases retrieved thanks to EDSaN!!!



COVID-19 uses cases:

- Steering committe: each day, how many health documents integrating « COVID 19 »… weeks before any ICD10 codes…
- Follow-up using the same tool of chronic disease (stroke, MI): decrease of 1/3 patients in RUH
- In 2021, based on HAS guideline, list of 6,000 patients to be vaccinated in « emergency »


Number of uses case par RUH Department
35
30
25
20
15
10
5
0
urgences
dermatologie
rhumatologie
pédiatrie
pneumologie
neurologie
chirurgie digestive
médecine interne
réanimation
ORL
physiologie digestive
endocrinologie
gastro-entérologie
D2IM
chirurgie plastique
radiologie
biologie
infectiologie
orthopédie
anesthésie
hématologie
pharmacie
nutrition
oncologie
anatomopathologie
biochimie
gynécologie-obstétrique
urologie
kinésithérapie
néphrologie
allergologie
réanimation médicale
néonat
chirurgie orthopédique
pharmacovigilance
gériatrie
cardiologie
psychiatrie

Keys to « success »

- Strong cooperation forte DDH – Department of Clinical Research (DCR) – Department of Computer Science (DCS)
- With DCR
  - Implementation of the governance
  - Managment of the CNIL relation (GPDR) +++
- With DCS
  - Partnership about ETL/ELT
  - Creation of a data lake upstream EDSaN (and also for specific uses cases besides EDSaN)
- Large involvment of the DPO (Data Protection Officer) & the person in charge of the cybersecurity of the Hospital Information System (French acronym RSSI)

Still in development

- Computer aided decision tool (exemple to semi-automatically calculate several quality indicators for interdisciplinary medical team conference; production of documents during the COVID outbreak)
- Dashbord creation from EDSaN, important for hospital & furthermore, for general practice
- Automatic extraction of specific items from unstructured data (*feature extraction*)

Perspectives

- Allowing mapping and comparison to national database (HDH/SNDS)
- Extending EDSaN to real world data to general practice (P4DP project 2022-5; CNGE)
- **EDSaN implementation in other Normandy hospitals?**
- And beyond… cooperation with several public partners (other French hospitals –Rennes, APHP Paris, Lille-), private partners (Dedalus)

Perspectives
Towards « One Health »

1. Clinical
   - Hospital
   - General Pratice
   - National (EHR Israël, Denmark, Singapore, Taiwan; France & US: DRGs)
2. Omics
   - Genomics, Transcriptomics, Metabolomics, Nutriomics…
3. Environment
4. IoT & medical devices
5. Social Media, email…
6. Veterinary medicine
7. …

Deep learning in RUH, France

- Medical word embeddings (Word2Vec, FastText,Glove)
  - PhD Emeric Dynomant (OMICX)
  - Two different corpus
    - 12 M health documents from HSDW
    - 180 K abstracts from LiSSa, French bibliographic database
- Doc2Vec, Patient2Vec (in progress)
  - PhD Mikaël Dusenne, MD
    - Hybrid semantic annotator (“old” NLP + deep NLP)
    - Doc2Vec2DRGs, using an other tool !!! ELMO?

Medical word embeddings querying page

12M ▾ endocardite Search

GloVe	infectieuse, myocardite, eto, native, streptocoque, bovis, bactériémie, faecalis, _endocardite
FastText (CBOW)	proprio_septive, myopericardite, septo_optique, endo_aortique, endocartite, endoculaire, acrodermite, rhino_septale, salmonellose, épidermolyse
FastText (Skip-Gram)	endocartite, endocardique, proctologique, extancilline, septo_basale, prolongements, recanalisée, précentrale, dantrolene, podoscopique
Word2Vec (Skip-Gram)	bovis, sanguinis, eto, gordonii, gallolyticus, aorto_mitrale, mutans, infectieuse, streptoccoque, salivarius
Word2Vec (CBOW)	endocartite, _endocardite, native, bovis, médiastinite, myocardite, mutans, gallolyticus, myopéricardite, tamponnade

Home

Medical word embeddings querying page

a ▼ endocardite Search

ord2Vec (Skip-Gram) endocardites, bactériémie, infectieuse, valvulopathie, septicémie, mycotique, spondylodiscite, médiastinite, valvulaire, bioprothèse

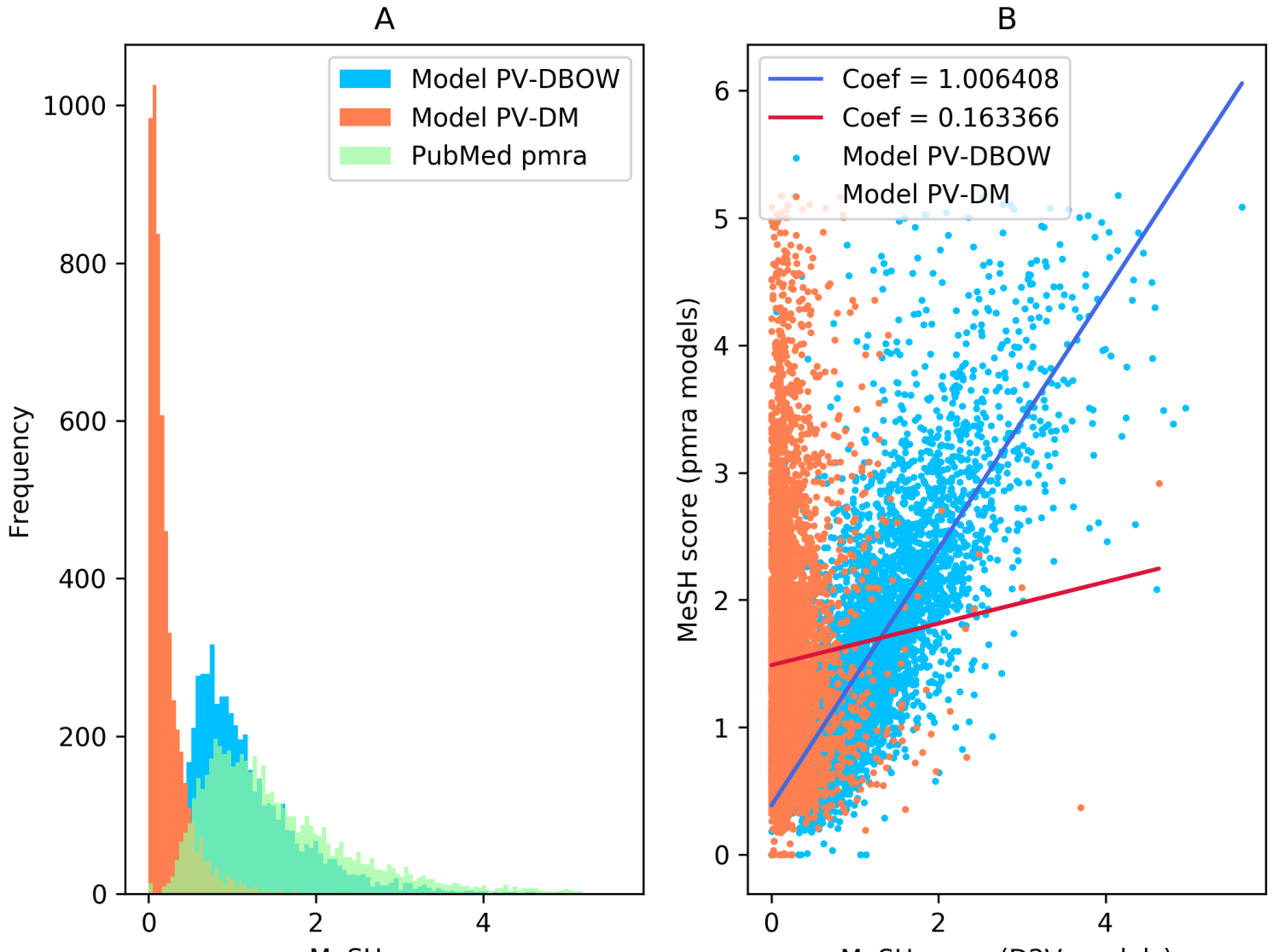
Home

Wordembeddings in two different contexts

QUERY: “facebook”

LiSSa corpus (300k)	internet, twitter, web, blog, e_learning, blogs, internautes, tic, game, …
RUH documents (12M)	reproches, injures, messages, insultes, rumeurs, ex_conjointe, menaces, insultant, …

Doc2Vec2PubMed



Valorization

- **Alicante**
  - SME from the North of France
  - Exclusive sales of RUH DBI tools: HeTOP & ECMT
    => two French hospitals + Dedalus
  - Integration of the semantic search engine in 2019
- **MedaïS Scientific**
  - Counseling on digital health

Publications https://edsan.chu-rouen.fr/edsan/communication/publications/

About EDSaN

- Romain Lelong; Badisse Dahamna; Romain Leguillon; julien Grosjean; Catherine Letord; SJ Darmoni & Lina F. Soualmia. Assisting Data Retrieval with a Drug Knowledge Graph. **ICIMTH2021**, 2021.
- Stéfan J. Darmoni. IA au sein d'un entrepôt de données de santé à Rouen. **Bulletin de l'AfIA** , 04, Number 112, Pages 18-20, 2021.
- Pressat-Laffouilhère T, Balayé P, Dahamna B, Lelong R, Billey K, Darmoni SJ, Grosjean J. Evaluation of Doc'EDS: A French Semantic Search Tool to Query Health Documents from A Clinical Data Warehouse. BMC Med Inform Decis Mak. 2020 Sep. DOI : 10.21203/rs.3.rs-59497/v1
- Dynomant E, Lelong R, Dahamna B, Massonaud C, Kerdelhué G, Grosjean J, Canu S, Darmoni SJ. Word embedding for French natural language in healthcare: a comparative study. JMIR Med Inform. 2019 Jul;7(3):e12310. DOI : 10.2196/12310
- Lelong R, Soualmia LF, Grosjean J, Taalba M, Darmoni SJ. Building a Semantic Health Data Warehouse: Evaluation of a search tool in Clinical trials. JMIR Med Inform. 2019;30. DOI : 10.2196/13917
- Siefridt C, Grosjean J, Lefebvre T, Rollin L, Darmoni SJ, Schuers M. Evaluation of automatic annotation by a multi-terminological concepts extractor within a corpus of data from family medicine consultations. Int J Med Inform. 2019. DOI : 10.1016/j.ijmedinf.2019.104009
- Dynomant E, Lelong R, Dahamna B, Massonnaud C, Kerdelhué G, Grosjean J, Canu S, Darmoni SJ. Word Embedding for French Natural Language in Healthcare: A Comparative Study. MEDINFO 2019: Health and Wellbeing e-Networks for All; 2019 Aug 26-30; Lyon, France. Stud Health Technol Inform. 2019;264:18-2. DOI : 10.3233/SHTI190195
- Grosjean J, Letord C, Charlet J, Aimé X, Danès L, Rio J, Zana I, Darmoni SJ, Duclos C. Un modèle sémantique d'identification du médicament en France. Atelier IA & Santé; 2019 Juil 1; Toulouse, France.
- Dynomant E, Darmoni SJ, Lejeune E, Kerdelhué G, Leroy JP, Lequertier V, Canu S, Grosjean J. Doc2Vec on the PubMed corpus: study of a new approach to generate related articles. 2019 Nov.
- Lelong R. Accès sémantique aux données massives et hétérogènes en santé. Normandie Université. 2019 Juin.
- Ndangang M, Grosjean J, Lelong R, Dahamna B, Kergourlay I, Griffon N, Darmoni SJ. Terminology Coverage from Semantic Annotated Health Documents. Stud Health Technol Inform. 2018;255:20-4. DOI : 10.3233/978-1-61499-921-8-20
- Lelong R, Soualmia LF, Sakji S, Dahamna B, Darmoni SJ. NoSQL technology in order to support Semantic Health Search Engine. MIE 2018: Medial Informatics Europe; 2018 Apr 24-26; Gothenburg, Sweden.
- Lelong R, Soualmia LF, Dahamna B, Griffon N, Darmoni SJ. Querying EHRs with a Semantic and Entity-Oriented Query Language. Stud Health Technol Inform. 2017;235:121-5. DOI : 10.3233/978-1-61499-753-5-121
- Cabot C. Recherche d'information clinomique au sein du Dossier Patient Informatisé : modélisation, implantation et évaluation. Normandie Université, 2017.

Publications based on EDSaN data

- Julien Grosjean; Thibaut Pressat-Laffouilhère; Marie Ndangang; Jean-Philippe Leroy & Stéfan J. Darmoni. Using Clinical Data Warehouse to optimize the vaccination strategy against COVID-19: a use case in France. **MedInfo**, 2021
- Deroualle T, Dominique S, Darmoni SJ, Grosjean J, Monti F, Lequerré T, Vittecoq O. Les formes rhumatologiques de la sarcoïdose sont associées à un risque accru de développer une spondyloarthrite : résultats d'une étude cas-témoins monocentrique rétrospective. Rev Rhum. 2020 Déc;87 Suppl 1:SA98-9. DOI :10.1016/j.rhum.2020.10.170
- Flouriot C, Avinee G, Joulakian M, Alarcon C, Marchand C, Savoure A, Durand E, Tron C, Frebourg N, Litzler PY, Chapuzet C, Eltchaninoff H. Infective endocarditis after transcatheter aortic valve implantation, a comparison with endocarditis occurring in surgical aortic prosthesis and native aortic valve patients. Eur Heart J. 2019 Oct;40 Suppl 1:S3664. DOI : 10.1016/j.acvdsp.2019.02.139

DONNÉES DE SANTÉ

ENTREPÔT DE DONNÉES DE SANTÉ NORMAND (EDSaN)

DONNÉES DE SANTÉ DES PATIENTS

NORMALISATION DES INFORMATIONS

INFORMATICIENS

TRAITEMENT DES DONNÉES

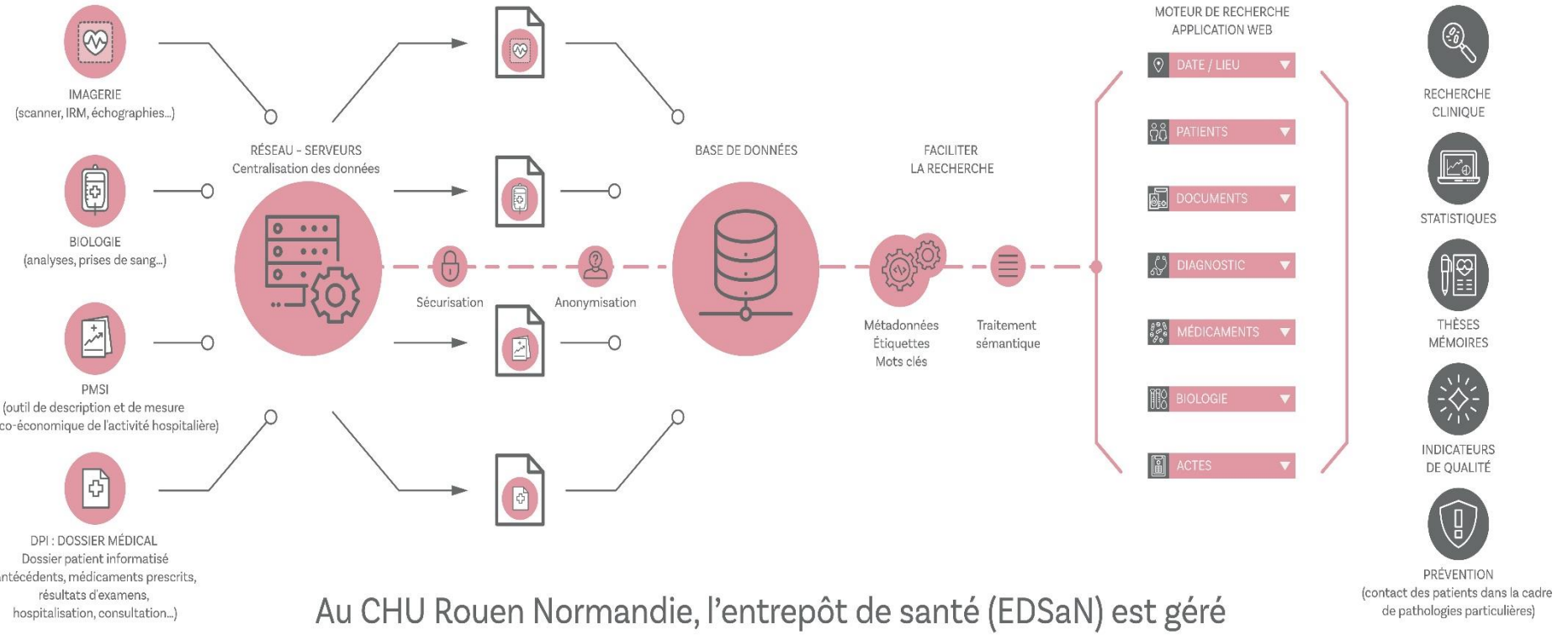
INFORMATICIENS

CONTRÔLE DES DONNÉES

COMITÉ SCIENTIFIQUE ET ÉTHIQUE

UTILISATEURS

CHERCHEURS - MÉDECINS
SOIGNANTS - ÉTUDIANTS



Au CHU Rouen Normandie, l'entrepôt de santé (EDSaN) est géré par le département de l'informatique et de l'information médicale.

Réalisation XG d'après D2IM / CHU Rouen

Thank you for listening!

Questions?

Email:
Stefan.Darmoni@chu-rouen.fr
Julien.Grosjean@chu-rouen.fr