# Information retrieval in health

## Indexation, Models, Evaluation

**SJ. Darmoni, MD, PhD & LF. Soualmia, PhD**

- Stéfan Jacques Darmoni = Professor of BioMedical Informatics, Rouen University Hospital & TIBS, LITIS, Normandy University; *Stefan.Darmoni@chu-rouen.fr*

- Lina Fatima Soualmia = Senior Lecturer, TIBS, LITIS, Normandy University; *Lina.Soulmia@chu-rouen.fr*

- Both associated to the LIMICS INSERM U1142, Paris

- Main fields of research
  - Knowledge engineering
  - Terminologies and ontologies, semantic web
  - Information retrieval & automatic indexing

# SIBM in 2015

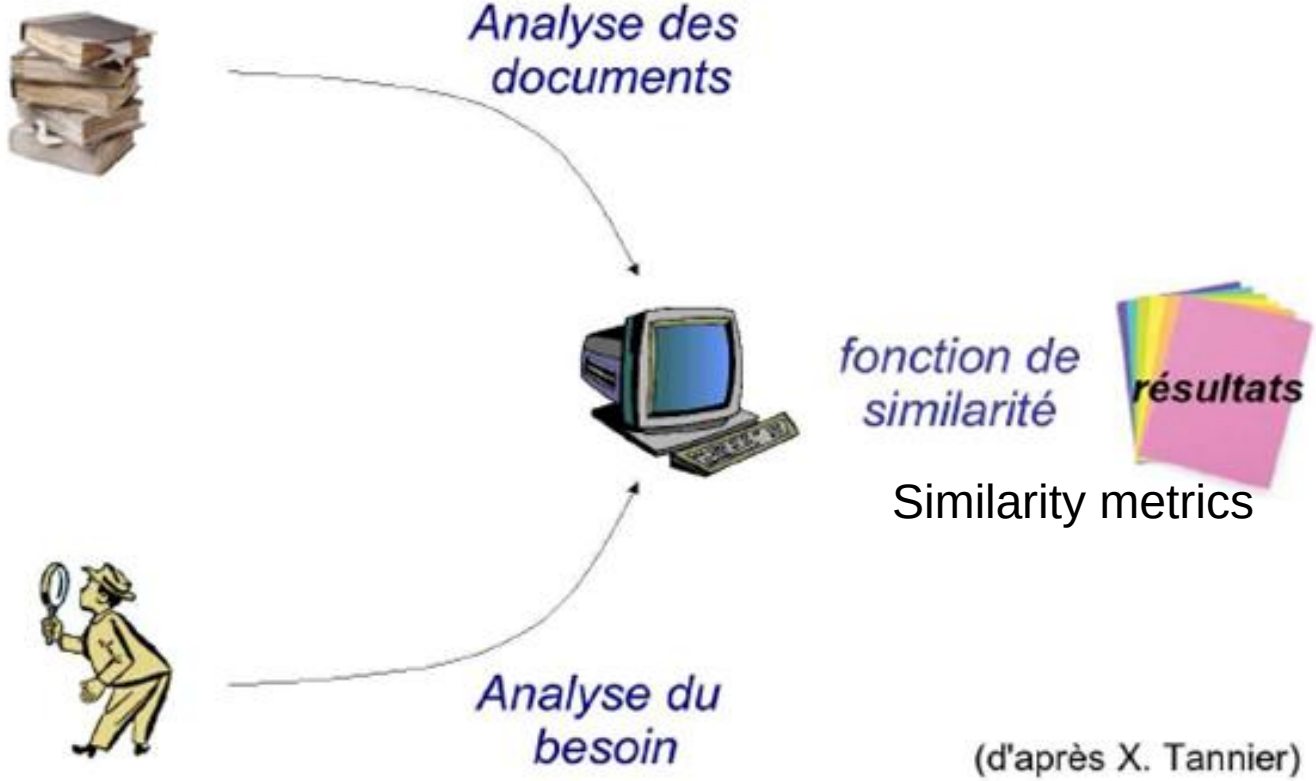
## Department of BioMedical Informatics

| MDs | R&D Engineers | University | Librarians | Assistantes & secrétaires |
|---|---|---|---|---|
| **Stéfan DARMONI** | **Badisse DAHAMNA** | Lina Soualmia (MCF 27ème) | **Benoit THIRION** | **Annie-Claude LANCELEVEE** |
| Philippe MASSARI (retired; 1/4 FT) | Julien GROSJEAN (PostDoc) | Adila MERABTI (PhD student) | **Catherine LETORD (Pharmacist)** | **Angélique MUTEL** |
| **Nicolas GRIFFON** | Ivan KERGOURLAY | Wiem CHEBIL (PhD student) | Gaetan KERDELHUE | **Sandrine VOURIOT (50 % BioStat)** |
| **André GILLIBERT (Public Health Resident)** | Romain LELONG | **Chloé CABOT (**PhD student**)** | Léa SEGAS | |
| Matthieu SCHUERS (GP & PhD student) | Tayeb MERABTI (PostDoc) | Melissa Mary (PhD student BioMérieux) | | |
| Carine PERIGARD (Resident GP) | | | | |
| Joanne PACHECO (Resident GP) | | | | |
| **Jean-Philippe LEROY (1/2 FT)** | | **Grant Hospital** | **Grant Regional Council** | Grant Research projects |

# Information retrieval (IR)

Introduction and main principles of the indexation

# IR: General schema




Analysis of documents

Similarity metrics

Analysis of the need (the query)

# IR: Main difficulties

- *Access, coverage, & response time +++*
  - Large documentary databases (health big data)
- *Relevance*
  - (automatic) metrics to measure relevance (evaluation)
  - Informational need of one give person? Context +++?
- *Exploitation*
  - Relevant documents may not be available in local language
    - huge problem in France;
    - less in Israel, where everyone speaks N languages, with N -> ∞
  - Queried information is difficult to obtain inside the document
    - Q&A vs. documentary information system

# Information retrieval: evolutions

- *Previously*
  - Documentary bases were structured and of small size
  - Access by metadata which describe documents
  - Use of documentary languages by specialists (librarians & information scientists)
- *Nowadays*
  - Most documents in electronic format and multimedia
  - A lot of formats to represent information (sometimes proprietary)
  - Documents dynamically created
  - Document databases with private access (invivisble Web)
  - More and more unstructured data
  - Appearence of semi-structured documents (discharge summaries)

# Performance of IR

- ***Ranking*** of retrieved documents by decreasing scores
  - date of the document (newest first –PubMed, CISMeF-)
  - quality/reputation of the source
  - quality of the indexing vs. the query
  - commercial link (Google)
- ***Evaluation*** by the ***end-user*** depending on:
  - Relevance of documents

    Variable from one end-user to an other, his/her knowldege, the context
  - ***response time*** of the system
  - ***Ergonomy of the*** system
    - Perceived ergonomy (SUS questionnaire)

- ***Automatic evaluation***:
  - Boolean comparison of returned documents vs. « ideal » answers (most of the time, manual gold standard)
  - ***Precision*** & ***recall***
  - Evaluation campaigns of IR systems

# Criteria to measure information retrieval & indexing

| | Relevant | Non relevant | |
|---|---|---|---|
| Transmitted documents | A | B | A+B |
| Non transmitted documents | C | D | C+D |
| | A+C | B+D | |

**Recall**= A/A+C= *true positive rate*= *sensitivity* ;
silence = 1 - recall = C/A+C = False negative

**Precision** = A/A+B = *positive predictive value (PPV)*;
noise = 1 – precision = B/A+B = False positive

# Criteria to measure information retrieval & indexing

- F-measure = weighted average of the precision and recall
- General F-measure
  - $F_\beta = (1+ \beta^2)\ PxR\ /\ (\beta^2 \times P) + R$
- $F_1$
  - In most of cases, $\beta = 1$, then $F_1 = 2\ PxR\ /\ P + R$
- According to the context, the developed system will optimize
  - either P or either R
- Other measures
  - MAP (*Mean Average Precision) : area under the curve R/P*
  - P@5, P@10 : precision after 5, 10 found documents => in favor of high/very high précision
  - Error rate = (FP + FN) / relevant

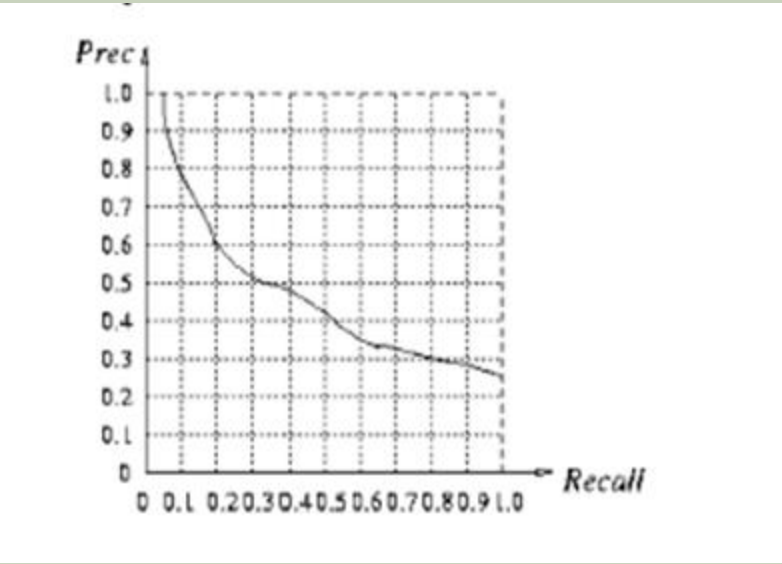
# Automatic evaluation

- ***Recall*** increases with the number of transmitted documents,

whereas the ***precision*** diminishes

- P/R curve to caracterize IR

Systems



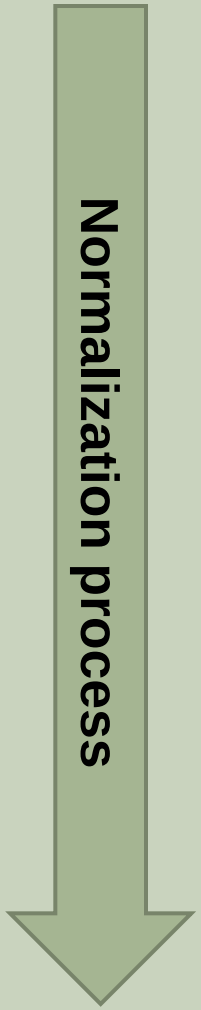

- Specific cases:
  - On very large documentary information systems, the ***relevance*** of the first documents (the first page syndrome) is more important than the ***recall*** => minimization of the ***noise***
  - Secondary objective: miminization of the ***response time***

# Indexing

- Searching the entire corpus of documents to answer a query is impossible:
  - Too many documents
  - Response time much too high
- Therefore, a preliminary phase is mandatory: ***automatic ndexing***
- The goal of the automatic indexation is to « ***transform documents into substitutes able to represent their content*** » [Salton et McGill, 1983]
- ***Manual indexing*** in bibliographic databases (MEDLINE, CISMeF)
- Among difficulties, ***language*** used in documents is a main barrier
  - **Indexing is language dependant +++**

# Lexical-based Approach

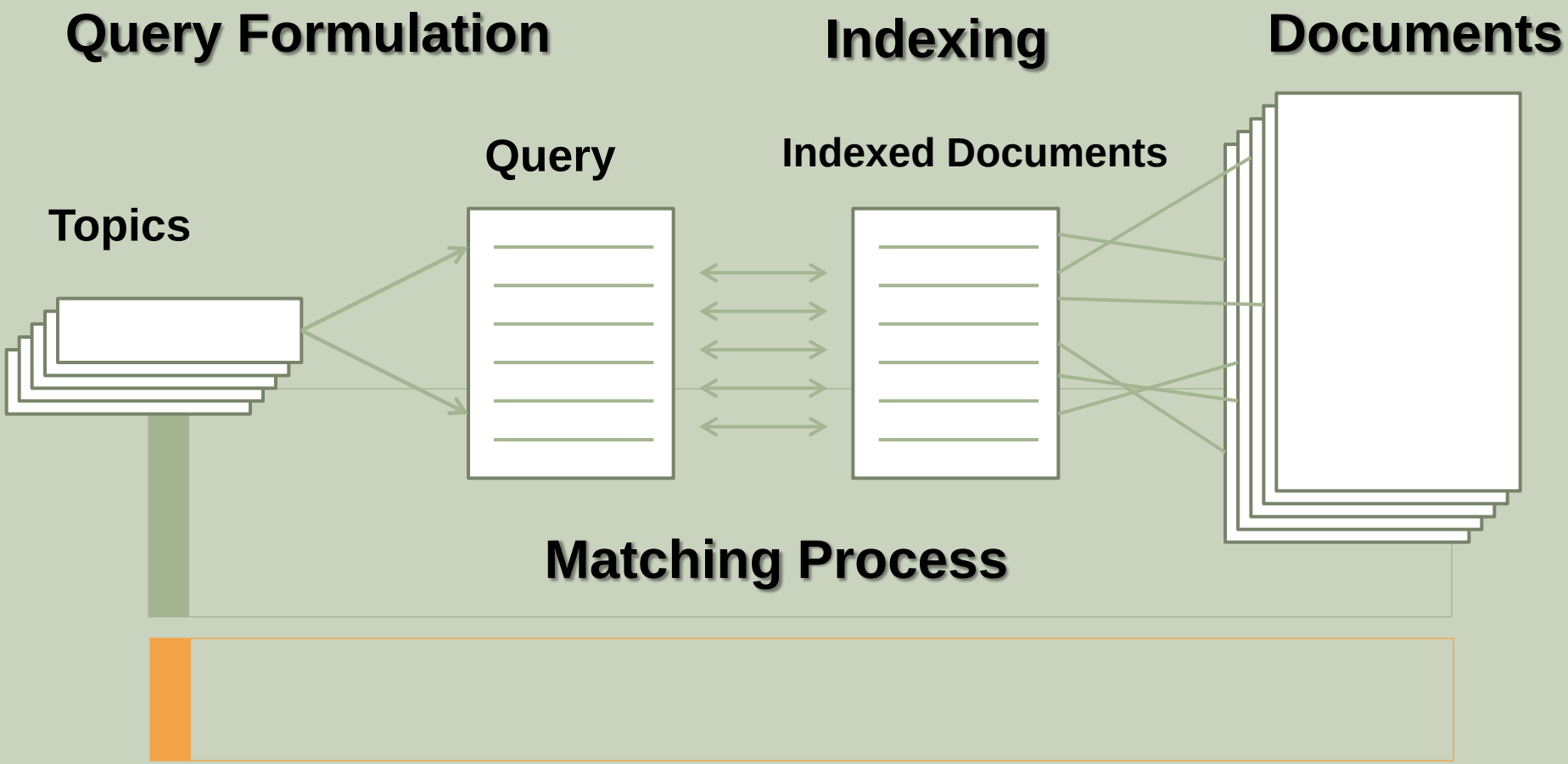
| | |
|---|---|
| Remove genitives | Presence of urinary reducing substances - finding |
| Replace punctuation with spaces | Presence of urinary reducing substances finding |
| Remove Stop words | Presence urinary reducing substances finding |
| Lowercase | presence urinary reducing substances finding |
| Uninflect each word | presence urinary reduce substance find |
| Word order sort | find presence reduce substance urinary |



# Types of Index

- Index may take several forms
  - ***Simple words***: e.g.: *university;  sings → sing*
  - ***Terms***: e.g. reducing Ph, asthma, Aspirin-Induced
  - **Entry (descriptor) of a thesaurus :** e.g. MeSH
  - Concept of a formal ontology: e.g. FMA in anatomy
- Index are more or less easy to exctract
- Index are more or less discriminant
  - ***Good***: *antigen, amyloïd, fiever (in the context of health)*
  - ***Bad***: *enfants, raised, developping*
  - ***Very Bad (stop word)***: *a, for, after…*
- An inverse file associates the index to the documents

# Indexing & Information Retrieval



# Characteristics of the language & IR

- Contrarily to artificial languages, the language is:
  - ***Implicit***: everything is not included in document (e.g. discharge summaries); depending on the ***context +++***
  - ***Redondant***: the language offers many different ways to formulate (more or less) the same content
  - ***Ambiguous***: a same expresion may be interpreted in several ways (e.g. acronyms)
- IR is even more complex:
  - Words may have different meanings in documents
  - Atomic meaning could be either words or expression (combinantion of words) => choice of the « bag of words » algorithm in the SIBM team

# The language is implicit

- Implicit informations are *lacking* to correctly interpret the meaning of a document
- This implicit informations may be correctly « extrapolated » by a human (mainly taking into account *context and knowledge*)
- For automatic indexing, the *context and knowledge* could be (partially) resolved using semantic web technologies (e.g. semantic expansion)

# Redondancy of the language

- *Synonymy* is a strict equivalence of meaning between two expressions (or words):
  - ***total*** : automobile & car (EXACT MATCH –SKOS RELATION)
  - ***partial***:
    - hypernym (generic term): vehicule & bike (BTNT –SKOS RELATION)
    - hyponym (specific term): BMX & bike (NTBT –SKOS RELATION)
    - meronymy (part of): finger & hand
    - holonymy (total of) : upper limb & arm
    - acronyms: as soon as possible & ASAP => generate a lot of ambiguity
    - circumlocutions: lave-vaisselle et machine à laver la vaisselle
    - Partial synonymy is used in UMLS => generate a lot of noise
    - Partial synonymy is used in SIBM => need to be evaluated

# Ambiguity of the language

- *Homonyms* (homographs & polysemy) are words with same characters but a diffrent meaning
  - Acronyms are most of the time homonyms +++
  - In French, IVG (a diagnosis or a procedure)
- *Homographs* are words that belong to different categories but with at least one same inflected form:

- Potential important role of UMLS semantic types to reduce this ambiguity
  - UMLS metathesarus of US National Library of Medicine
  - Over 2 millions medical concepts; each of them with a least of semantic type (e.g. diagnosis, procedure)

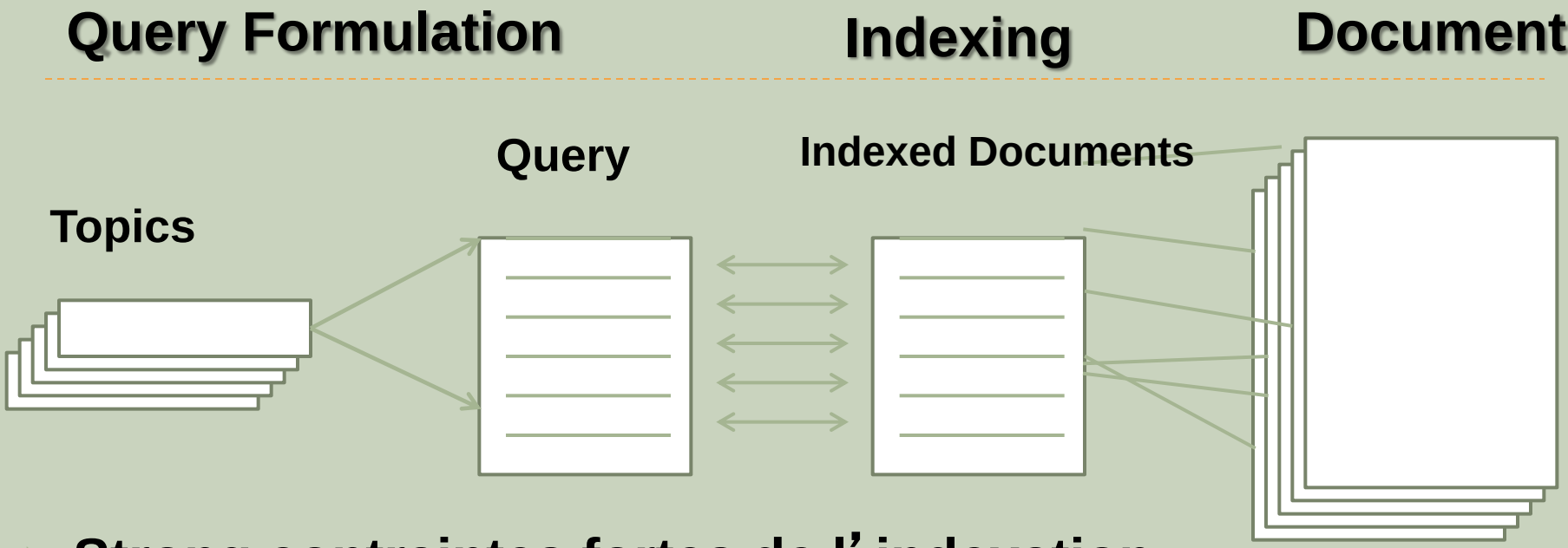
# Ambiguity of the language

- Polysemy are words with several meanings and where all the inflected forms are the same
- More a word is used, more the probability of polysemy is high
- Necessary to define the meaning of a word in the right context
  - One main difficulty in NLP (word sense disambiguation) & Semantic Web
  - Secondary prevention (GP) ≠ secondary prevention (public health)
  - Secondary prevention (GP) ≈ tertiary prevention (public health)
    - Not an Exact Match relation; Use the See Also (or Close) relation

# Properties of the indexing

- *Index* are used to represent the *content* of documents*:*
  - They represent ***only a part*** of the content of the documents
  - They may take ***several forms*** (e.g. simple words, terms, expressions, entries of a thesaurus, etc.)
  - They are ***more ore less difficult to extract***
  - Their storage need more or less ***memory (see later on NoSQL vs. SQL vs. SPARQL)***

# Process of indexing




- **Strong contraintes fortes de l'indexation**
  - Storing large amount of information in a minimal space
  - Extracting all necessary information
  - Allowing efficient access to the index during IR
  - Allowing dynamic indexing

# Chain of indexing

- Segmentation of documents into smaller units (sentences)
  - Drawback: loss of meaning between sentences
- Linguistic normalization
- Production of indexing files

# Segmentation of documents

- Characteristics of documents :
  - formats of file (text, HTML, PDF, etc.)
  - Coding (ASCII, ISO-LATIN-X, Unicode)
  - language(s)
  - non linguistics signes (mathematic formulas, presentation, images, …etc.)
- A collection de documents may contain several languages
  - One index per language or a unique index
  - NLP tools to identify a specific langage
- Level of indexing:
  - Overall documents /subpart of a document / subset of documents (e.g. web site)

# Textual normalization

- Possibility to normalize elements before indexing
  - suppression of *points* into acronyms (***U.S.A.*** in ***USA***)
  - suppression of *accents* (***météo : meteo***)
  - suppression of some *majuscules* (***Et : et***)
- normalization of several data and information
  - *dates* : ***14 juillet 1789 : 14/07/1789 (Fr) : 1789/07/14 (US)***
  - Data about *money*: ***$400 : 400 dollars***
  - *organizations*: IMF: International Monetary Fund
  - Choice of several normalizations may be based on end-users usages
- Possibility not to normalize index:
  - Size of index more important
  - Need of mechanisms to *expand the query* (semantic expansion)

# Linguistic normalization

- Several techniques to **correct**
  - reaccentuation (*meteo météo*)
  - orthographic corrections (*inofrmation information*)
  - grammaticale corrections (*the roses looks! the roses look*)
- Bring several elements to a single form
  - **Stemming** (words of the same stem)

used in linguistic morphology and information retrieval to describe the process for reducing inflected words to their word stem, base or root form—generally a written word form.

- *malade, malades, maladie, maladies, maladive* devient *malad*

  - **Lemmatisation:** (words of the same lemme)

in linguistics, the process of grouping together the different inflected forms of a word so they can be analysed as a single item

- *stomach, gastric,* => canonic form: for a verb, infinitif or for noun, masculin singular: *stomach*

# Linguistic normalization

- **phonetic** normalization (same pronounciation)
  - *Chebyshev : tchebycheff*
  - *Alzeimer : alzheimer*
  - *Very useful in medicine with complex terms (including names of diseases after his/her discoverer*
- Use of other relations (semantic web)
  - Synonyms
  - Hyponyms (explosion in information sciences; subsumption in computer science)
  - Semantic expansion (between terminologies)
    - *kill, assassinate, beat to death, defeat, destroy, do away with, do in, eliminate, eradicate, execute, exterminate, extinguish, finish off, knock off, liquidate, mop up, murder, pip, rack up, shoot dead, slaughter (Source www.sensegate.com)* **=>** *kill*

# Phonetic normalization

- Words with same pronunciation
- Language dependent +++
- Create a set of words (or expressions) with the same pronounciation
- Soundex algorithm (English):
  - Every word is compressed to a reduced form of 4 characters
  - Creation of an index of reduced forms of phonetic equivalents
  - Extract of the algorithm:
    - Keep the first letter of a word
    - Replace the letters a, e, i, o, u, h, w, y by 0
    - Other correspondences: B, F, P, V by 1; C, G, J, K, Q, S, X, Z by 2
    - Suppression of repeated numbers & 0
    - Obtention of a normalized code
      - *Herman* → *H655*

# Normalisation: stemming

- Bring back differents words to their respective stemming
  - Rules are dependent of the language +++
    - E.g. Porter algorithm for English
    - automates, automatic, automation => automat
    - For French, *malade, malades, maladie, maladies, maladive* => *malad*
- Some conventions about reduction phases (for French)
  - Rule examples: *sses => ss; ies => i ; ational => ate ; tional : tion*
- Linguistic normalisation:
  - Significant reducing of index size
  - A lot of pontential errors => P < 1 and R < 1
  - Impossibility to distingish among several forms of the same stemming via the index

# Spelling correction

- Spelling corrections may be due to input errors or wrong OCRs
- Two main approaches
  - Correction in the index
  - Correction in the IR queries (our approach)
- Two main approaches to correct spelling:
  - Correction of words in isolation (ex: inofrmation)
    - Calculation of a distance
    - Possibility to weight operations taking into account frequent errors: input (a→ q), OCR (D→ O)
  - Correction of words in context (flight form Eathrow)
    - Use of large corpora or most frequent queries (log) => our approach

# Similiarity distances: Levenshtein & Stoilos

- Levenshtein distance:
  - Minimal number of elementary operations to go from chain $c_1$ to chain $c_2$

$$LevNorm(c_1, c_2) = \frac{Lev(c_1, c_2)}{Max(length(c_1), length(c_2))}$$

- Stoilos distance:
  - Similarity between two entities depends on their common chains and their differences

$$Sim(s_1, s_2) = Comm(s_1, s_2) - Diff(s_1, s_2) + winkler(s_1, s_2)$$

# Normalisation: variants

- Grouping the term variants (difficult task +++) :
  - Genetic disease
    - Basic terme
  - Disease is genetic
    - syntaxic variant
  - Hereditary disease
    - Semantic variant
  - Genetically determined forms of the disease
    - variante morpho-syntaxique
  - Disease is familial
    - variante syntaxico-sémantique
  - Transmissible neurodegenerative diseases
    - variante syntaxico-sémantique

# Importance of index

- All indexes in documents do not have the same importance
  - Use of **stop words lists, quite complex in health**
  - Stop word most of the time will not be a stop word in a specific context
- ***bag-of-words*** model
  - Number of occurrences of a term in each document
  - **Frequency of a term in a document**
  - Do no take care into account the ordre of the word (in the bag)
  - Indexing the longest bag of words
- How to handle the importance of a term in a document inside a corpus
  - ponderation by ***tf.idf***

# *tf.idf* ponderation

- Calcul of the weight of a term in a document :

$$w_{i,d} = tf_{i,d} * idf_i$$

  - $tf_{i,d}$: frequency of term *i* in the document *d*
  - $idf_i$ : importance of a term *i* into the collection of documents (inverse document frequency)
  - simple metrics: inverse of number of documents in the collection containing the term

$$w_{i,d} = tf_{i,d} * \frac{1}{df_i}$$

  - Most used metrics: log of the ratio between the number of documents in the collection and the number of documents containing the term

$$w_{i,d} = tf_{i,d} * \log(\frac{N}{df_i})$$

  - The weight of a term increases:
    - With frequency of the term in the document
    - With scarcity of the term in documents in the collection

# Information retrieval

Main models and Evaluation

# Information retrieval

- Models of retrieval
  - Three main approaches
- Evaluation
  - Main metrics
  - Pooling
  - Evaluation campaigns

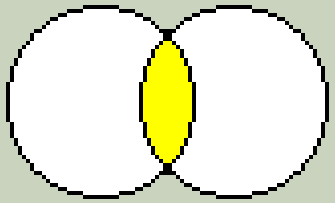
# Three main approaches

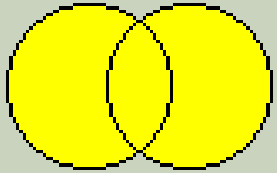
- 1. Models based on set theory
  - Boolean model
- 2. Algebric models
  - vectorial model
- 3. Probabilistic model
- Bayes theorem

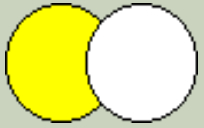
# Boolean model

- First and simplest model
- Based on theory of sets and Boole algebra
- The terms  of the query are either present or absent
  - Binary weight of terms, 0 ou 1
- Therefore, a document is either relevant or not
  - Binary relevance, never partial (exact model)
- The query is built with logic operators
  - AND, OR, NOT
  - (cycling OR swimming) AND NOT doping
- The document is relevant if and only if its content respect the Boolean query

# Boolean querying

AND 

OR 

NOT 

# Searching by proximity

- X NEAR(N) Y
- Searching X AND Y separated by less than N words (excluding stop words)
- Use of position index
- Interesting because searching documents containing X AND Y without limit will generate too many noise
- Need to build the outset of every word in the documents

# Potential extensions

- X NEAR(N) Y
- Weighting of keywords
  - “olympic games AND Beijing AND (swimming:3 OR cycling:4 OR track & fields:2)"
  - Allows a result ranking by best choices performed by the end-user
- => Extended Boolean model

# Boolean model: pro & cons

- Pro:
  - The model is transparent & simple to understand for the user
    - No hidden parameters
    - Reason to select a document is quite clear: this document is relevant for a logic query
  - Adaptated to specialists (information scientists & librarians) & controlled vocabularies
- Cons:
  - Quite difficult to build a complex Boolean query: binary criterion not so efficient
  - Possible to weight terms (extended Boolean model)
  - Impossible to perform a ranking of the results

# Vector model

- Algebraic model :
  - Terms & documents are represented as vectors
  - Similarity measures between a query and a document
    - Ranking list according to this similarity
- Similarity measures: more two documents contain the same terms, more the probability that they represent the same information is high
- Terms & documents are represented as vectors
  - Each dimension corresponds to a separate term
  - The lenghth of the vector is proportional to the weight of the terms
  - Relevance of a document corresponds to similarity between query vector and document vector

# Vector model

query

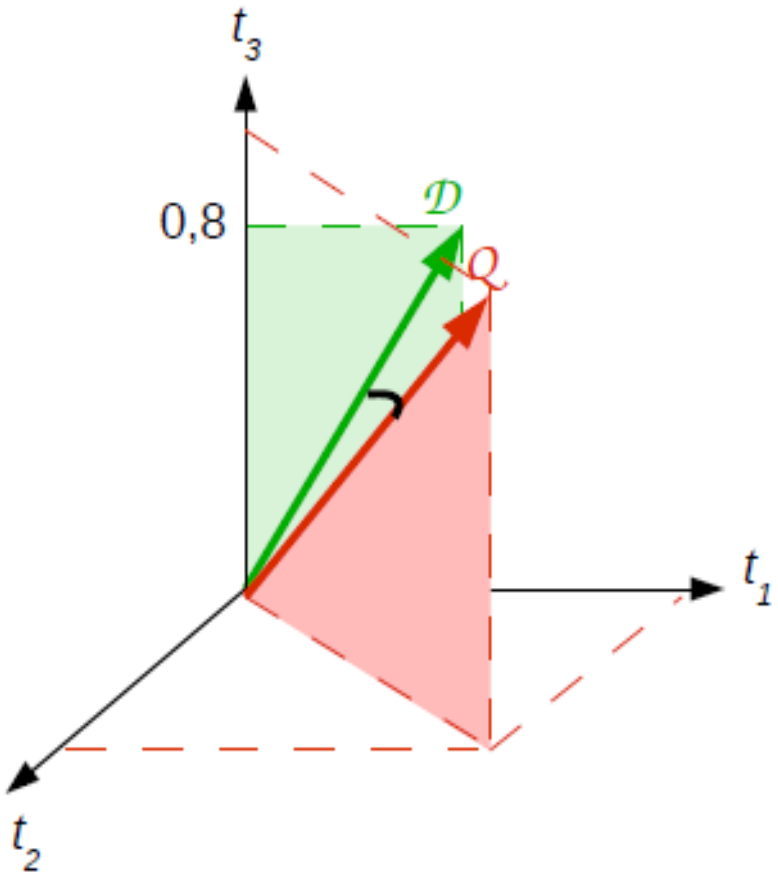
Requête $Q$ : $t_1$ $t_2$ $t_3$

Document $\mathcal{D}$ : ... $t_1$ ... $t_3$ ...

weight

poids $w_{\mathcal{D},t1} = 0.45$

poids $w_{\mathcal{D},t3} = 0.80$



# Vector model: similarity mesures

- Dice

$$RSV(\vec{Q},\vec{D}) = \frac{2\sum w_{iQ} \times w_{iD}}{\sum w_{iQ} + \sum w_{iD}} \qquad \frac{2|A \cap B|}{|A| + |B|}$$

- Jaccard

$$RSV(\vec{Q},\vec{D}) = \frac{\sum w_{iQ} \times w_{iD}}{\sum w_{iQ} + \sum w_{iD} - \sum w_{iQ} \times w_{iD}} \qquad \frac{|A \cap B|}{|A \cup B|}$$

- Overlap

$$RSV(\vec{Q},\vec{D}) = \frac{\sum w_{iQ} \times w_{iD}}{min(\sum w_{iD}, \sum w_{iQ})} \qquad \frac{|A \cap B|}{min(|A|, |B|)}$$

# Vector model: similarity mesures

- Produit scalaire

Scalar product

$$RSV(\vec{Q},\vec{D}Q)=\vec{Q}.\vec{D}=\sum_{i=1}^{n} w_{iQ}\times w_{i\mathcal{D}}$$

- Cosinus de l'angle

Cosine of the angle

$$RSV(\vec{Q},\vec{D})=\frac{\vec{Q}.\vec{D}}{|\vec{Q}|\times|\vec{D}|}=\frac{\sum w_{iQ}\times w_{i\mathcal{D}}}{\sqrt{\sum w_{iQ}^{2}}\times\sqrt{\sum w_{i\mathcal{D}}^{2}}}$$

- Distance euclidienne

Euclidian distance

# Vector model: pro & cons

- Pro:
  - Query language is more simple (liste de mots clés)
  - Better results thanks to the weights
  - Selection of documents with partial relevance is possible
  - Ranking possible based on the matching documents/query
- Cons:
  - In this model, all the terms independant (main problem +++)
  - « black box » syndrome => the end user does not understand why a document is selected according to the query

- Overall, vector model is the most used model in IR (not true in health)

# Evaluation campaigns

- TREC (Text REtrieval Conference) :
- Every year since 1992
- Sponsored by DARPA
- Several research axes:
  - Multimedia: image, vidéo, Web
  - Specific query types: Q&A, interactive, cross-lingual
- Specific domains: genomics, law
- Specific ways of expression: blogs, spams...
- CLEF (CrossLanguage Evaluation Forum), for European languages
- NTCIR, for Asian languages

# Evolutions in the indexing

- **Several modalities of web indexation that may be intricated**
  - **Documentary Indexation**: thesaurus, description of ressources
  - **Automated Indexation**: based on NLP (Natural Language Processing)
  - **Social Indexation**: tags of web 2.0
  - **Semantic Indexation**: metadata (XML, RDF) and ontologies (OWL)

# Automatic indexation: semantic search engines

- Appearence of new search engines:
  - Hakia:
    - Born in 2006
    - Natural langage
    - Mixte of semantic analysis, ontology, fuzzy logic, and artificial intellingence
  - Powerset :
    - Born in May 2008; bought by Microsoft in July 2008
    - Semantic search on Wikipedia
    - Analysis of sentence containging the words of the query
    - Proposal of new key-words

# Evolutions in indexation:

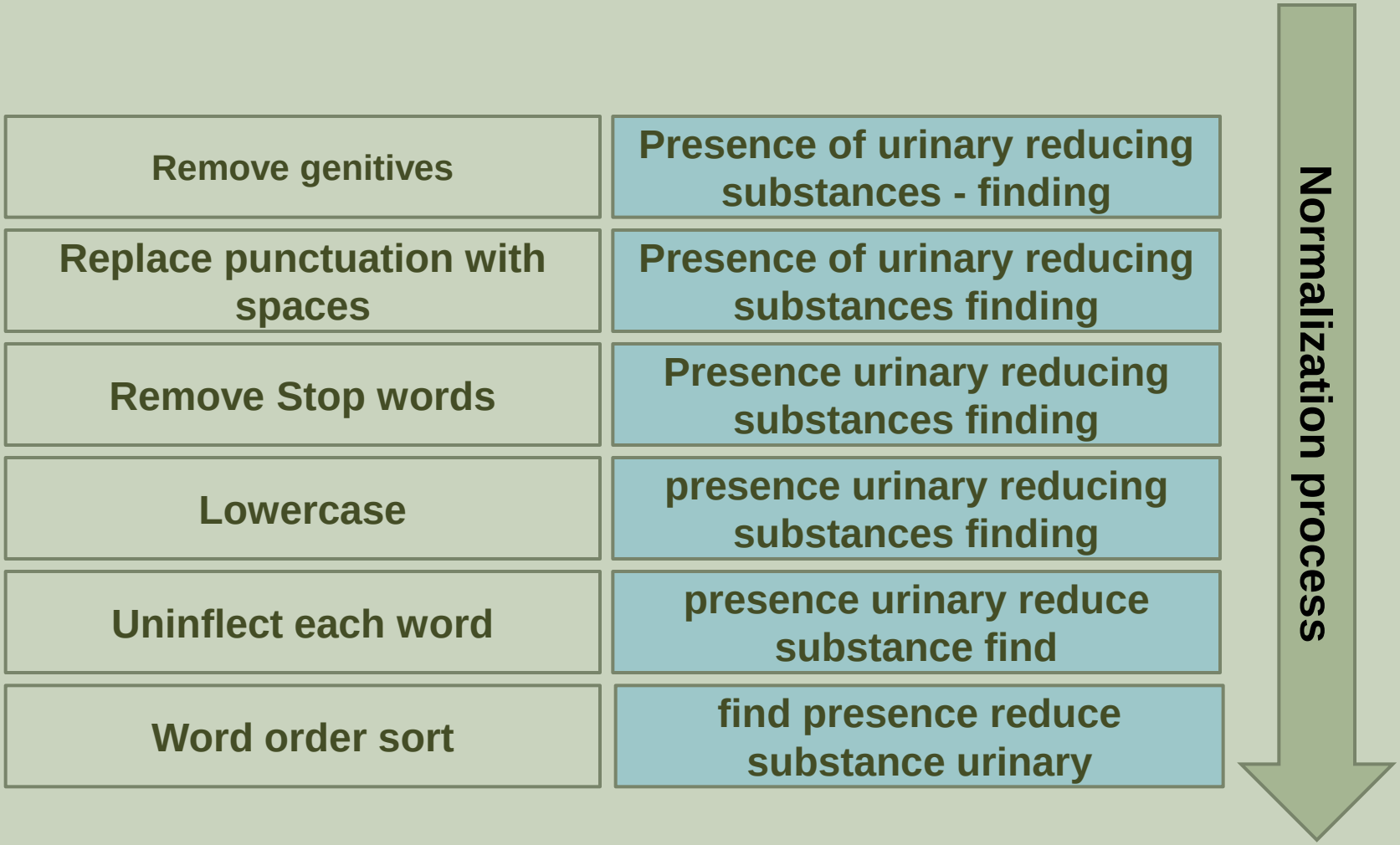
## Social indexation: tags & folksonomies

- **The principle of folksonomy:**
  - Form of « **collaborative decentralized spontaneous classification** », based on terms chosen s'appuyant sur les termes choisis par les utilisateurs
  - **Objectif :** facilitate the indexing of contents and IR
  - Tags may be applied to **web signets**, **photos**, **vidéos, or blogs (tag clouds)**
  - Creation of a community of **« specialists »** among Internet end-users

# CISMeF semantic search engine

- Born in 2000
- Then, based on one single terminology (MeSH), identical with PubMed ATM & bilinguism (Fr En)
- Based on « bag of words » algorithm & Boolean querying
  - match on MeSH thesaurus & title of documents (first step)
  - If no answer, query on CISMeF metadata (Dublin Core)
  - If no answer, query of full text (Oracle tool, Google CSE)
- 2006: multiterminology
- 2012: multilinguism
- 2015: NoSQL

# Lexical-based Approach (NLP)



# 2014: shift from Oracle SQL to NoSQL

- NoSQL vs. SPARQL vs. SQL (two junior engineers)
- Best response time with NoSQL
- Choice of Infinispan datagrid
  - stand-alone into one hospital +++ confidentiality of health data
- For optimal perfomance, nearly all the data are placed in RAM => 128 Go RAM
- Serialisation of the data on file systems
- Use of Lucent indexes for textual query (previously SQL)
- SOA (service oriented) architecture
- Powerful server(s)
  - Xeon 2690 v3 biprocessor; each proc with 12 cores (17501 CPU mark); 128 Go RAM
  - Efficient on parrallel and sequential processing

# Automatic indexing in SIBM

- ECMT v3 MultiTerminology Concept Extractor
  - Based on crosslingual multiterminology portal www.hetop.eu
    - ≃500,000 concepts in French (>327,000 different CUIs)
  - Language dependent
  - Integrated in a software suite (Alicante)

- To be compared to several tools existing for health English
  - Based on UMLS (more than 2 million concepts)

CiSMeF
Catalogue et Index des Sites Médicaux de langue Française

# Extracteur de Concepts Multi-Terminologique (ECMT v3)

Examens paracliniques : * Biologie : globule plaquettes 245 000 mm3, TP 88 , ionogramm 3,8 mmol l, ASAT 23, ALAT 25, phosphatase Radiographie pulmonaire : récidive d'un pne régulier et sinusal sans trouble de la repolar supra aortiques et trans- crânien (23 04): t interne gauche dès son origine, sténose de irrégulière du bulbe ancienne calcifiée, évalu artères vertébrales droite et gauche. Pas d' au niveau du polygone de Willis et notamme postérieures. * Angioscanner des TSA (12 04) : présence d'une infiltration athéromateuse au niveau de la crosse aortique. Infiltration athéromateuse calcifiée au niveau de l' artère carotide commune. Au niveau du bulbe, infiltration athéromateuse plus importante avec présence d'une plaque calcifiée quasi circonférentielle. Thrombose complète de la carotide interne sur tout son trajet depuis le bulbe jusqu'à sa portion terminale. A droite, infiltration athéromateuse prédominant également au niveau du bulbe avec présence de plaque calcifiée sans sténose significative retrouvée. Calcifications au niveau de l' artère carotide interne dans sa portion intra caverneuse. Sténose moyennement serrée de l' origine de la vertébrale droite. Plusieurs boucles vasculaires sur le trajet de la vertébrale droite. Artère vertébrale gauche est le siège de boucle vasculaire. Pas de sténose significative retrouvée. Au niveau du polygone de Willis, celui-ci est bien opacifié avec présence d'une artère communicante antérieure et des deux artères communicantes postérieures. * Scanner thoracique sans et avec injection : plusieurs adénopathies centimétriques connues stables, épanchement péricardique gauche, dont l' épaisseur maximale mesure 16,4 mm. Bonne opacification des gros vaisseaux médiastinaux. On retrouve la lésion bourgeonnante au sein de la lumière de la bronche souche gauche mesurée ce jour à 22 x 33 mm, responsable d'une atélectasie complète du poumon gauche. Présence d'un emphysème sous cutané thoracique antérieur gauche et axillaire gauche en rapport avec le drainage thoracique. * Echographie de stress : non réalisable en raison du pneumothorax. Evolution dans le service : Sur le plan respiratoire : le patient bénéficie de la pose d'un drain thoracique à gauche pour évacuation de l' épanchement aérique. La radiographie pulmonaire après mise en place montre un décollement persistant ainsi qu'une atélectasie du lobe inférieur gauche, avec ascension de la coupole diaphragmatique. Le drain est donc mis en aspiration sans amélioration de la radiographie thoracique. Une fibroscopie bronchique est réalisée et montre une obstruction par un caillot frais au niveau du tronc souche gauche expliquant la non réexpension du poumon. Compte tenu de la taille du caillot, nous programmons une bronchoscopie au tube rigide sous anesthésie générale le 16 04 2009 qui permet la thermocoagulation de la lésion qui apparaît comme un caillot, puis extraction de celui-ci. Après extraction, il existe un saignement régulier issu de B6 qui inonde l' arbre bronchique gauche. On procède à un nettoyage au fibroscope souple de tout le lobe supérieur qui apparaît sain, sans anomalie. La lobaire inférieure est totalement obstruée par un caillot. La lobaire inférieure est cathétérisable mais l'orifice d' entrée est totalement déformé, refoulé par une tumeur issue de B6 saignant facilement au contact. On termine le geste en thermocoagulant B6. Des biopsies de carène, éperon lobaire supérieur et inférieur gauche sont réalisées et sont négatives.

Effacer

33 phrases annotées en 270 ms. 194 codes distincts identifiés.

**artère carotide commune** (MSH_D_017536)

- **catégorie(s) :** médecine et chirurgie vasculaire; cardiologie; partie du corps, organe ou composant d'un organe; système d'organisme;
- **ascendant(s) :** arborescence MeSH (MSH_D_ARBO); artères carotides (MSH_D_002339); système cardiovasculaire (MSH_D_002319); artères (MSH_D_001158); Anatomie (MSH_D_A); vaisseaux sanguins (MSH_D_001808);
- **descendant(s) :** artère carotide interne (MSH_D_002343); artère carotide externe (MSH_D_002342);
- **relié(s) :** artère carotide primitive, sai (SNO_NO_T-45100);

| | Ter. | Code | CUI | Cond. | Ctxt. |
|---|---|---|---|---|---|
| | SNO | G-C410 | C0332310 | | |
| | SNO | G-C470 | C1331475 | | |
| | SNO | F-06305 | C0231301 | | |
| | MSH | D000124 | C0001072 | non | |
| | SNO | T-D8100 | C0004454 | | |
| | MSH | D000410 | C0001899 | | 25, |
| ancien | SNO | G-A463 | C0580836 | | |
| anesthésie | MED | T140 | C0002903 | | 16 |
| anesthésie générale | SNO | P1-C0010 | C0002915 | | 16 |
| anesthésie générale | MSH | D000768 | C0002915 | | 16 |
| anesthésie pour bronchoscopie | SNO | P1-C2A04 | C0198953 | | 16 |
| anomalie | SNO | M-01130 | C1457869 | pas | |
| antérieur | SNO | G-A105 | C1704448 | | |
| aorte, sai | SNO | T-42000 | C0003483 | | |
| appareil respiratoire, sai | SNO | T-20000 | C0521346 | | |
| arbre | SNO | L-D0012 | C0040811 | | |
| artère carotide commune | MSH | D017536 | C0162859 | | |
| artère carotide interne | MSH | D002343 | C0007276 | | |
| artère cérébrale postérieure | MSH | D020769 | C0149576 | | |
| artère communicante antérieure | SNO | T-45530 | C0149562 | | |
| artère communicante postérieure | SNO | T-45320 | C0149559 | | |
| artère vertébrale | MSH | D014711 | C0042559 | | |
| artère vertébrale droite | SNO | T-45710 | C0226230 | | |
| artère vertébrale gauche | SNO | T-45720 | C0226231 | | |
| artères | MSH | D001158 | C0003842 | | |
| aspartate aminotransferases | MSH | D001219 | C0004002 | | 23, |
| aspiration | SNO | F-29200 | C0700198 | | |
| aspiration (technique) | MSH | D013396 | C0038638 | | |
| atélectasie | SNO | D2-60300 | C0004144 | | |
| atélectasie complète | SNO | D2-60306 | C0264497 | | |
| atélectasie pulmonaire | MSH | D001261 | C0004144 | | |
| athérome | SNO | M-52100 | C0264956 | | |
| B05BC02 - urée | ATC | B05BC02 | C0041942 | | 3,8 |
| biopsie | SNO | P1-03100 | C0005558 | | |
| biopsie | MSH | D001706 | C0005558 | | |
| bon | SNO | G-A223 | C0205170 | | |
| bronche souche gauche | SNO | T-26500 | C0225630 | | |
| bronche, sai | SNO | T-26000 | C0205039 | | |
| bronches | MSH | D001980 | C0006255 | | |

# Metrics of ECMT (April 2015 CLEF eHealth in French)

| | TP | FP | FN | P | R | F1 |
|---|---|---|---|---|---|---|
| Anatomy | 142 | 149 | 54 | 0.4880 | 0.7245 | 0.5832 |
| Chemistry | 153 | 38 | 108 | 0.8010 | 0.5862 | 0.6670 |
| Devices | 13 | 12 | 6 | 0.5200 | 0.6842 | 0.5909 |
| Disorders | 375 | 96 | 209 | 0.7962 | 0.6421 | 0.7109 |
| Geography | 14 | 4 | 7 | 0.7778 | 0.6667 | 0.7179 |
| Live Beings | 125 | 38 | 31 | 0.7669 | 0.8013 | 0.7837 |
| Objects | 3 | 16 | 28 | 0.1579 | 0.0968 | 0.1200 |
| Phenotype | 14 | 35 | 17 | 0.2857 | 0.4516 | 0.3500 |
| Physiology | 60 | 33 | 74 | 0.6452 | 0.4478 | 0.5286 |
| Procedure | 195 | 105 | 109 | 0.6500 | 0.6414 | 0.6457 |
| | | | | | | |
| **Overall** | **1094** | **526** | **643** | **0.6753** | **0.6298** | **0.6518** |

# Similiarity distances: Stoilos

- communality function:

$$Comm(s_1, s_2) = \frac{2 * \sum_i length\,(\,MaxComSubString_i\,)}{length(\,s_1\,) + length(\,s_2\,)}$$

- Example : $S_1$= « Trigonocepahlie » et $S_2$ = « Trigonocephalie »

Comm(Trigonocepahlie,Trigonocephalie)= 0.866.

length($MaxComSubString_1$)=length(Trigonocep)=10
length($MaxComSubString_2$)=length(lie)=3

# Similiarity distances: Stoilos

- Difference function:
- Based on length of chains, which where not matched at the previous step

$$Diff\ (s_1, s_2) = \frac{uLen_{s_1} * uLen_{s_2}}{p + (1-p) * (uLen_{s_1} + uLen_{s_2} - uLen_{s_1} * uLen_{s_2})}$$

*$S_1$= « Trigonocepahlie »* et *$S_2$= « Trigonocephalie »  p=0.6 ;*

*$uLenS_1$= 2/15 et $uLenS_2$ =2/15;*

*Then:  Diff($S_1$,$S_2$)=0.0254.*

- Winkler parameter:

$$Winkler\ (s_1, s_2) = L * P * (1 - Comm(s_1, s_2))$$