Bases de connaissances médicales : définition et utilisation

Lorraine Goeuriot
LIG - Université Grenoble Alpes
lorraine.goeuriot@imag.fr
http://mrim.imag.fr/User/lorraine.goeuriot/rimtis4/

Plan du cours

- Qu'est-ce qu'une base de connaissances ?
- Les principales bases de connaissances médicales (BCM)
- Annotation à l'aide de BCM
- Indexation à l'aide de BCM

Plan du cours

- Qu'est-ce qu'une base de connaissances ?
- Les principales bases de connaissances médicales (BCM)
- Annotation à l'aide de BCM
- Indexation à l'aide de BCM

Définitions

- **Une base de connaissance** regroupe des connaissances spécifiques à un domaine spécialisé donné, sous une forme exploitable par un ordinateur [Wikipedia]

- **Un vocabulaire contrôlé** est une liste définie de termes (application informatique)

- **Une terminologie** est l'ensemble des termes, rigoureusement définis, qui sont spécifiques d'une science, d'une technique, d'un domaine particulier de l'activité humaine [wikipedia]

- **Un thesaurus** est un réseau de termes contrôlés, enrichi par des relations associatives pré-définies

- **Une ontologie** est un modèle de description des connaissances basé sur des concepts avec types, propriétés et relations

Les bases de connaissances - utilisation

- Dans le domaine médical, les BC sont utilisées dans différents cadres applicatifs :
  - Etudes épidémiologiques,
  - Codes utilisés dans les rapports médicaux et lors de la facturation des soins,
  - Systèmes experts
  - Recherche d'information
  - …

Plan du cours

- Qu'est-ce qu'une base de connaissances ?
- Les principales bases de connaissances médicales (BCM)
- Annotation à l'aide de BCM
- Indexation à l'aide de BCM

Les bases de connaissances médicales

- Medical Subject Headings (MeSH) : indexation de littérature médicale

- International Classification of Disease (ICD) : code utilisé pour les diagnostics et facturation dans les hôpitaux

- Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT) : représentation des informations spécifiques aux patients

- Foundational Model of Anatomy (FMA)

- Gene Ontology (GO)

- CPT 4 : code des procédures

International Classification of Medicine (ICD)

- Classification statistique internationale des maladies et des problèmes de santé connexes

- Classification médicale codifiée classifiant les maladies et une très vaste variété de signes, symptômes, lésions traumatiques, empoisonnements, circonstances sociales et causes externes de blessures ou de maladies

- Publiée par l'Organisation mondiale de la santé (OMS)

- Mondialement utilisée pour l'enregistrement des causes de morbidité et de mortalité touchant le domaine de la médecine

International Classification of Medicine (ICD)

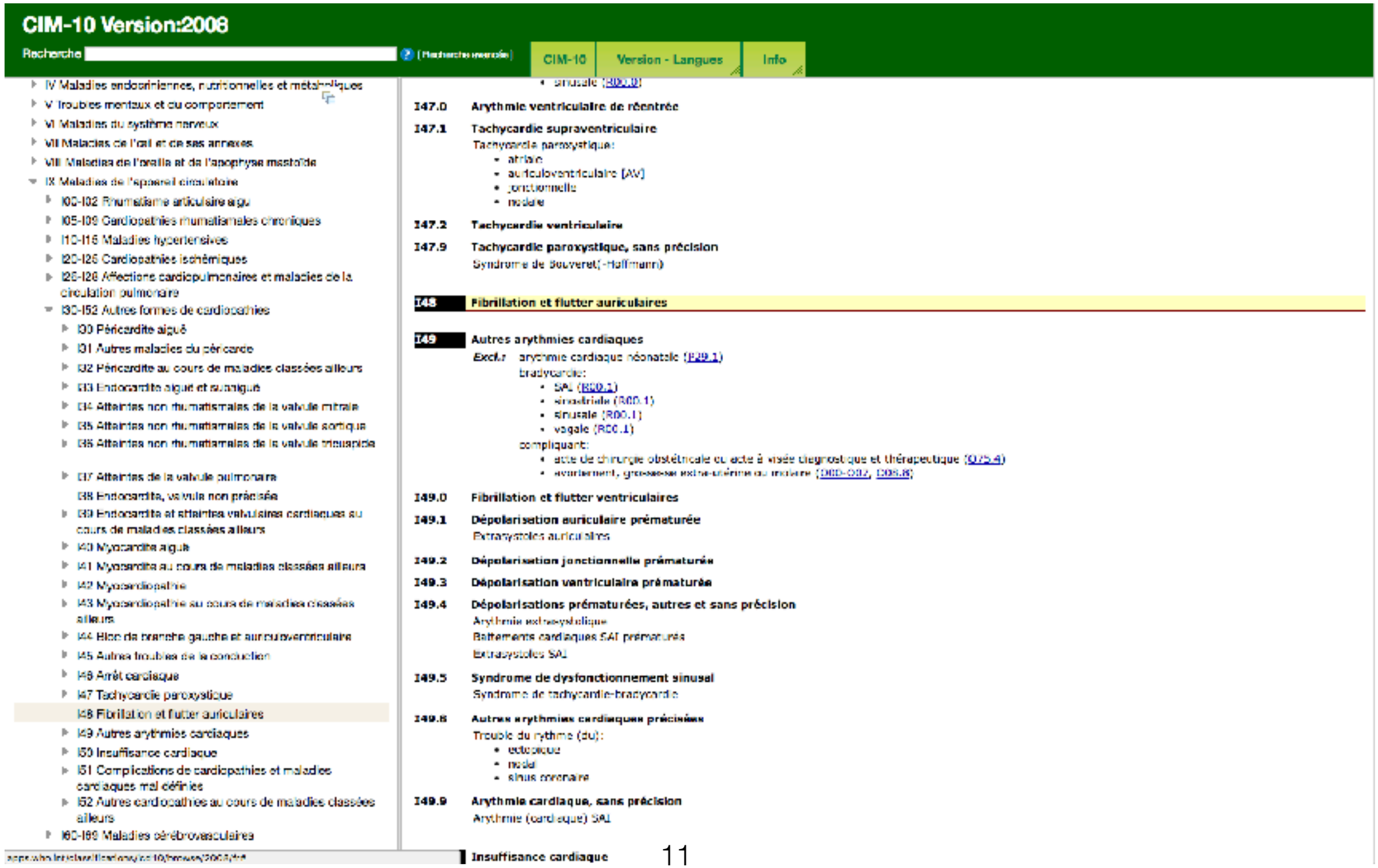
- Version courante : 10 (11 en construction)
- Anglais : ICD-10 http://apps.who.int/classifications/icd10/browse/2016/en
- Français : CIM-10 http://apps.who.int/classifications/icd10/browse/2008/fr

International Classification of Medicine (ICD)

Classification des maladies :

I Certaines maladies infectieuses et parasitaires
II Tumeurs
III Maladies du sang et des organes hématopoïétiques et certains troubles du système immunitaire
IV Maladies endocriniennes, nutritionnelles et métaboliques
V Troubles mentaux et du comportement
VI Maladies du système nerveux
VII Maladies de l'œil et de ses annexes
VIII Maladies de l'oreille et de l'apophyse mastoïde
IX Maladies de l'appareil circulatoire
X Maladies de l'appareil respiratoire
XI Maladies de l'appareil digestif
XII Maladies de la peau et du tissu cellulaire sous-cutané
XIII Maladies du système ostéo-articulaire, des muscles et du tissu conjonctif
XIV Maladies de l'appareil génito-urinaire
XV Grossesse, accouchement et puerpéralité
XVI Certaines affections dont l'origine se situe dans la période périnatale
XVII Malformations congénitales et anomalies chromosomiques
XVIII Symptômes, signes et résultats anormaux d'examens cliniques et de laboratoire, non classés ailleurs
XIX Lésions traumatiques, empoisonnements et certaines autres conséquences de causes externes
XX Causes externes de morbidité et de mortalité
XXI Facteurs influant sur l'état de santé et motifs de recours aux services de santé
XXII Codes d'utilisation particulière

International Classification of Medicine (ICD)



Systematized Nomenclature of Medicine (SNOMED)

- SNOMED Clinical Terms
- Collection de termes organisée en vue d'applications informatiques
- Contient des codes, termes, synonymes et définitions utilisés dans les rapports et documents cliniques
- Une des terminologies cliniques les plus complètes
- Objectif : permettre de représenter et stocker de façon efficace les données cliniques
- Terminologie des rapports médicaux
- http://www.ihtsdo.org/snomed-ct

Systematized Nomenclature of Medicine (SNOMED)

- Couvre : découvertes cliniques, symptômes, diagnostics, procédures, corps humain, organismes, substances, pharmaceutiques, appareil, specimens

- 4 principaux composants :
  - Identifiants des concepts
  - Descriptions
  - Relations
  - Reference Sets : groupement de concepts ou descriptions

Medical Subject Healdings (MeSH)

- Produit par la National Library of Medicine (NLM) depuis 1960

- Utilisé afin d'indexer et cataloguer les documents biomédicaux

- Cf cours 4

Le projet UMLS

- Origine du projet : proposer un mécanisme permettant de lier différents vocabulaires ainsi que des ressources médicales

- Plusieurs années de mise en place

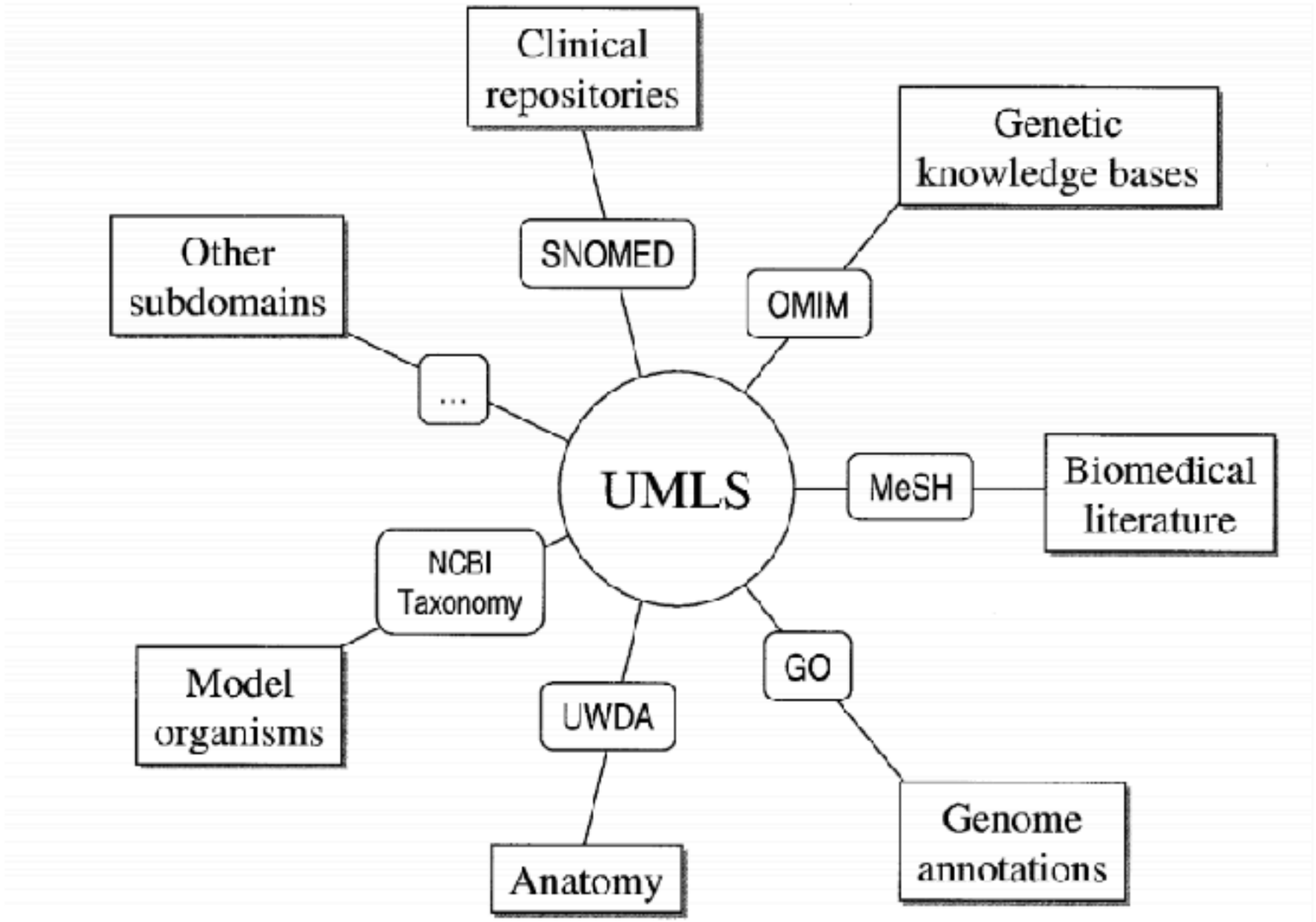
- Contient maintenant :

  - Un métathesaurus

  - Un réseau sémantique (Semantic Network)

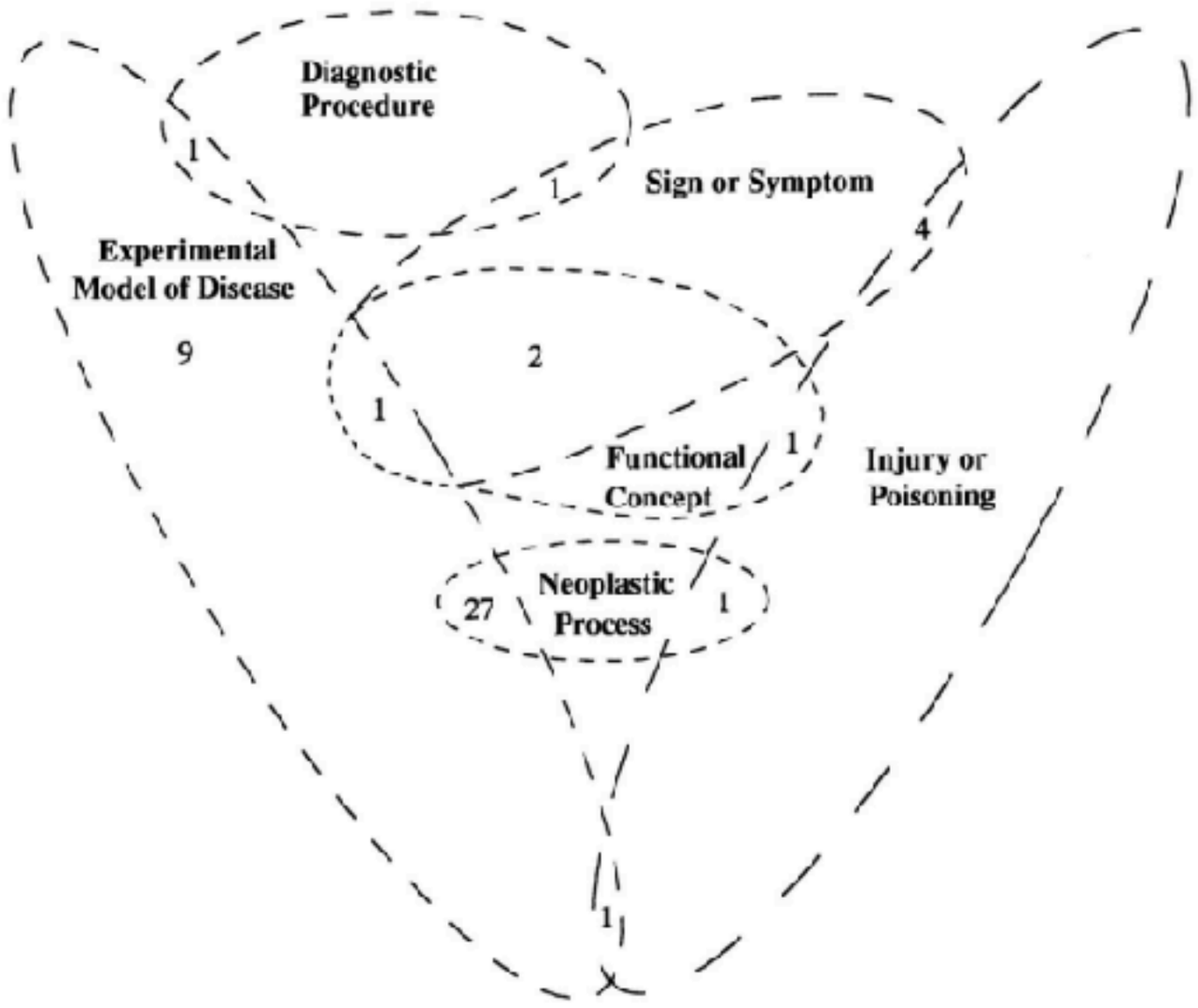
  - Un ensemble d'outils de traitement de la langue (the Specialist Lexicon)

Le projet UMLS



Bodenreider, O. (2004) The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research, 32, D267-D270.*

Le projet UMLS



Gu, H., Perl, Y., Geller, J., Halper, M., Liu, L.-m. & Cimino, J. J. (2000) Representing the UMLS as an Object-oriented Database: Modeling Issues and Advantages. *Journal of the American Medical Informatics Association, 7, 1, 66-80.*

Le métathesaurus UMLS

- Rassemble plus de 100 vocabulaires
- Toutes les entrées désignant la même chose sont rassemblées sous la forme d'un *concept*
- Chaque concept correspond à plusieurs *termes*, qui représentent une expression du concept (pas des variations lexicales)
- Pour chaque *terme* sont listées les *chaînes* correspondant aux variations lexicales, et à chaque *chaîne* sont associés des *atomes* correspondant aux entrées des vocabulaires
- Pour chaque *concept*, une *chaîne préférée* et un *terme préféré* sont indiqués, ils correspondent à la *forme canonique* du concept
- Chaque concept, terme, chaîne et atome a un identifiant unique : CUI, LUI, SUI, AUI

Le métathesaurus UMLS

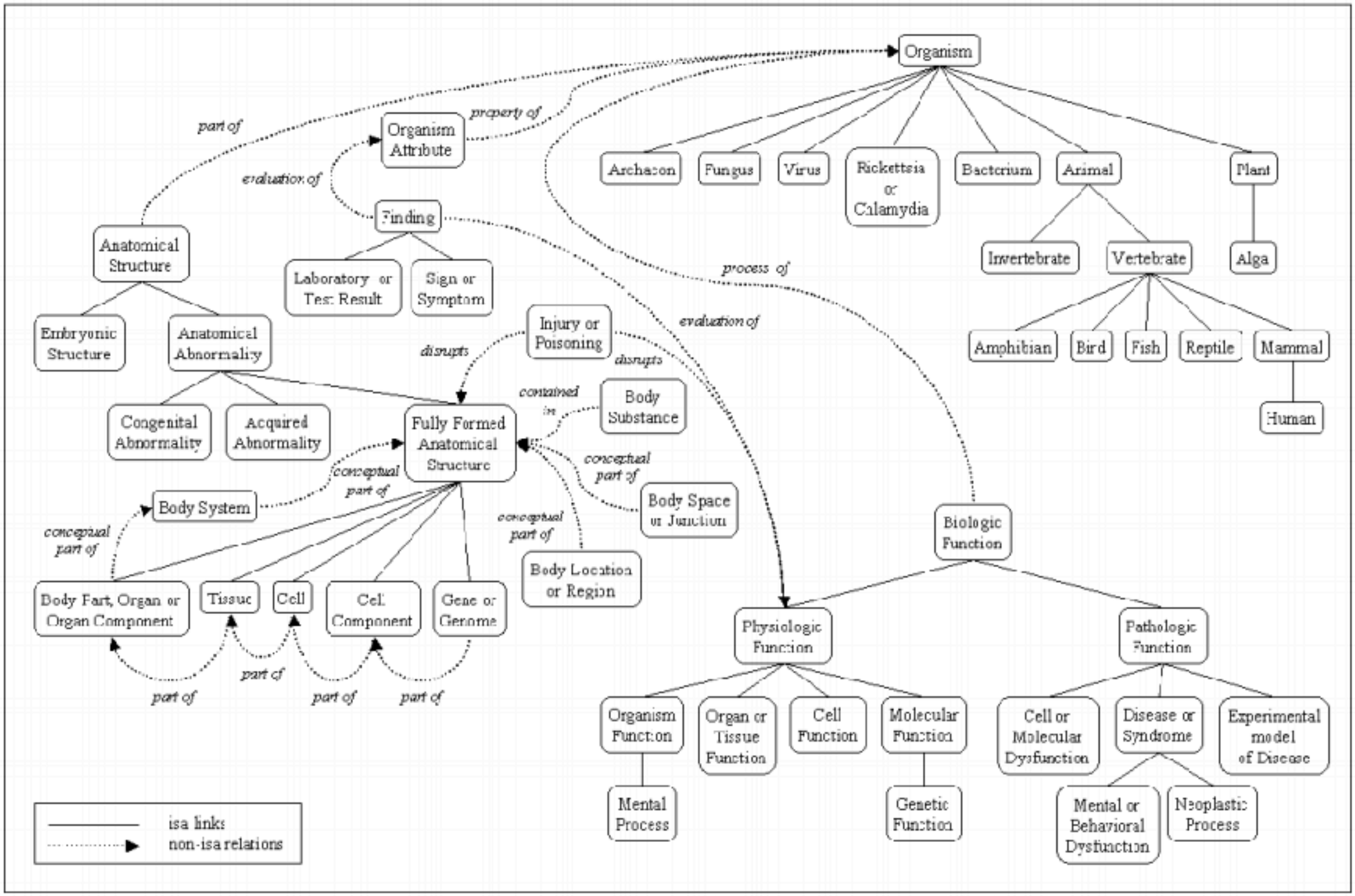
Concept (CUI)	Term (LUI)	String (SUI)	Atom (AUI)
C0004238 Atrial fibrillation (preferred) Atrial fibrillations Auricular fibrillation Auricular fibrillations	**L0004238** Atrial fibrillation (preferred) Atrial fibrillations	**S0016668** Atrial fibrillation (preferred)	**A0027665** Atrial fibrillation (from MSH)
			A0027667 Atrial fibrillation (from PSY)
		S0016669 Atrial fibrillations (plural variant)	**A0027668** Atrial fibrillations (from MSH)
	L0004327 Auricular fibrillation Auricular fibrillations (synonyms)	**S0016899** Auricular fibrillation (preferred)	**A0027930** Auricular fibrillation (from PSY)
		S0016900 Auricular fibrillations (plural variant)	**A0027932** Auricular fibrillations (from MSH)

Le métathesaurus UMLS

En réalité, bien plus d'informations :

https://uts.nlm.nih.gov/metathesaurus.html#C0004238;0;1;CUI;2016AA;EXACT_MATCH;CUI;*;

Le métathesaurus UMLS



http://genome.tugraz.at/MedicalInformatics/WinterSemester2011Holzinger/3_LV444_152_STRUCTURED_DATA_HOLZINGER_2011.pdf

Le metathesaurus UMLS

- 1.5 million concepts
- 120+ vocabulaires
- 5 millions de termes
- 7 million d’atomes
- 17 langues représentées

Construction et structure d'UMLS

- Un exemple tiré du tutoriel UMLS de Bodenreider, Willis et Hole, 2004

- https://mor.nlm.nih.gov/pubs/pres/20071204-KAIST-tutorial.pdf

Le metathesaurus NCI

- NCI : National Cancer Institute
- Leur metathesaurus NCIm est basé sur UMLS, plus des vocabulaires liés aux activités du NCI
- Contient 850,000 concepts, représentant environ 1.5 million de termes

Plan du cours

- Qu’est-ce qu’une base de connaissances ?
- Les principales bases de connaissances médicales (BCM)
- Annotation à l’aide de BCM
- Indexation à l’aide de BCM

L’annotation - définition

- Annoter signifie ajouter des méta-informations à un document :
  - Des entités (en référence à une BCM)
  - Des relations (implicites ou explicites)
    - Entre entités (*HK1 involved in glycolytic process*)
- Entre une entité et le document :
  - Les entités MeSH pour l’indexation de références sur MEDLINE : *PMID:3207429* est indexé par *Glucose/metabolism* et *Hexokinase/genetics**
  - Le code ICD10 dans un rapport médical

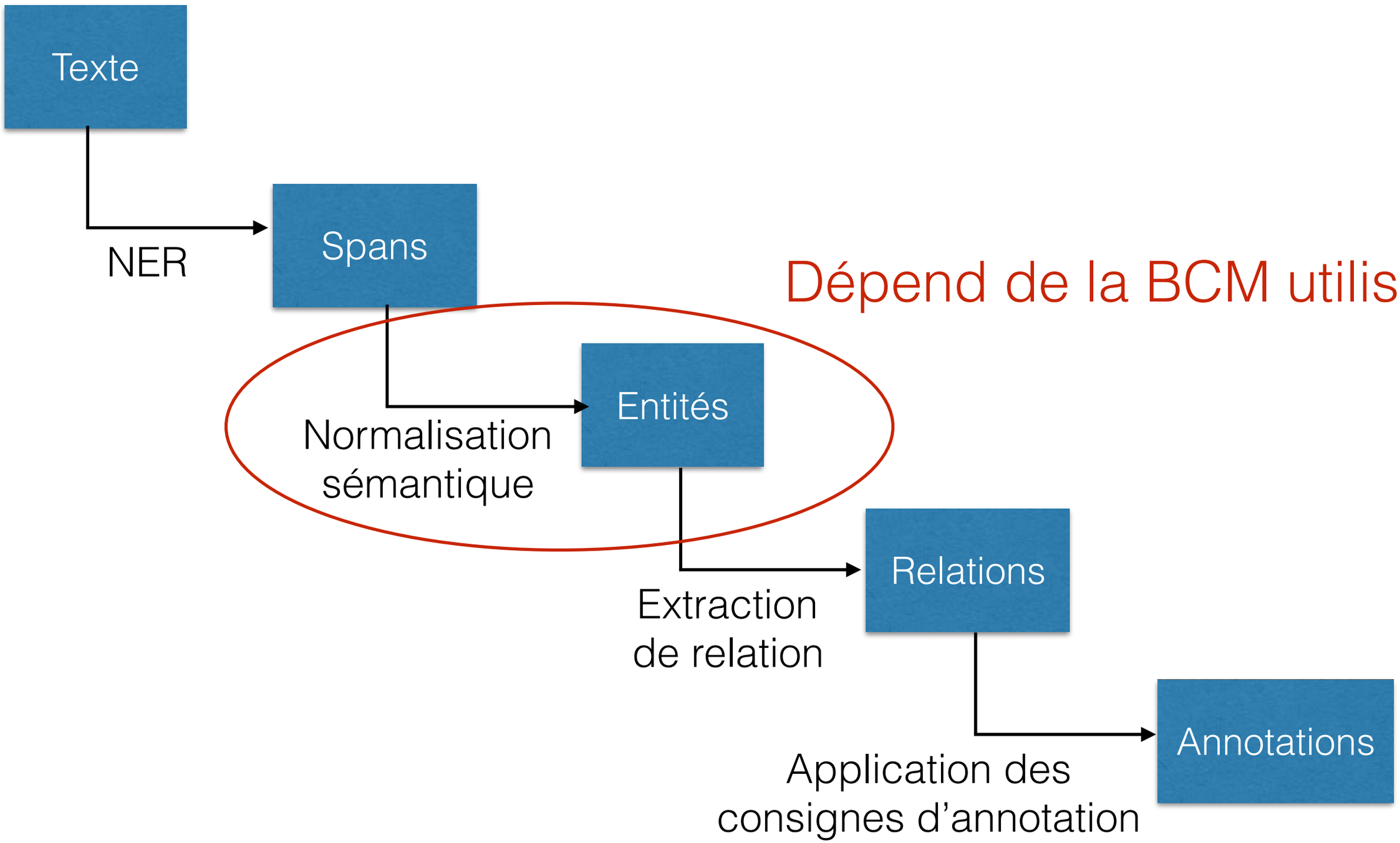
L’annotation - définition

- Les métadonnées assignées à un document peuvent l’être :
  - Automatiquement
    - Reconnaissance automatique d’entités nommées et normalisation
    - Indexation automatique
  - Manuellement
    - Curation des données, étiquetage manuel
    - Indexation manuelle (majorité sur MEDLINE)
    - Codes ajoutés aux dossiers patients lors de la facturation
  - Dérivées d’autres annotations
    - Grâce aux liens

Annotation - définition

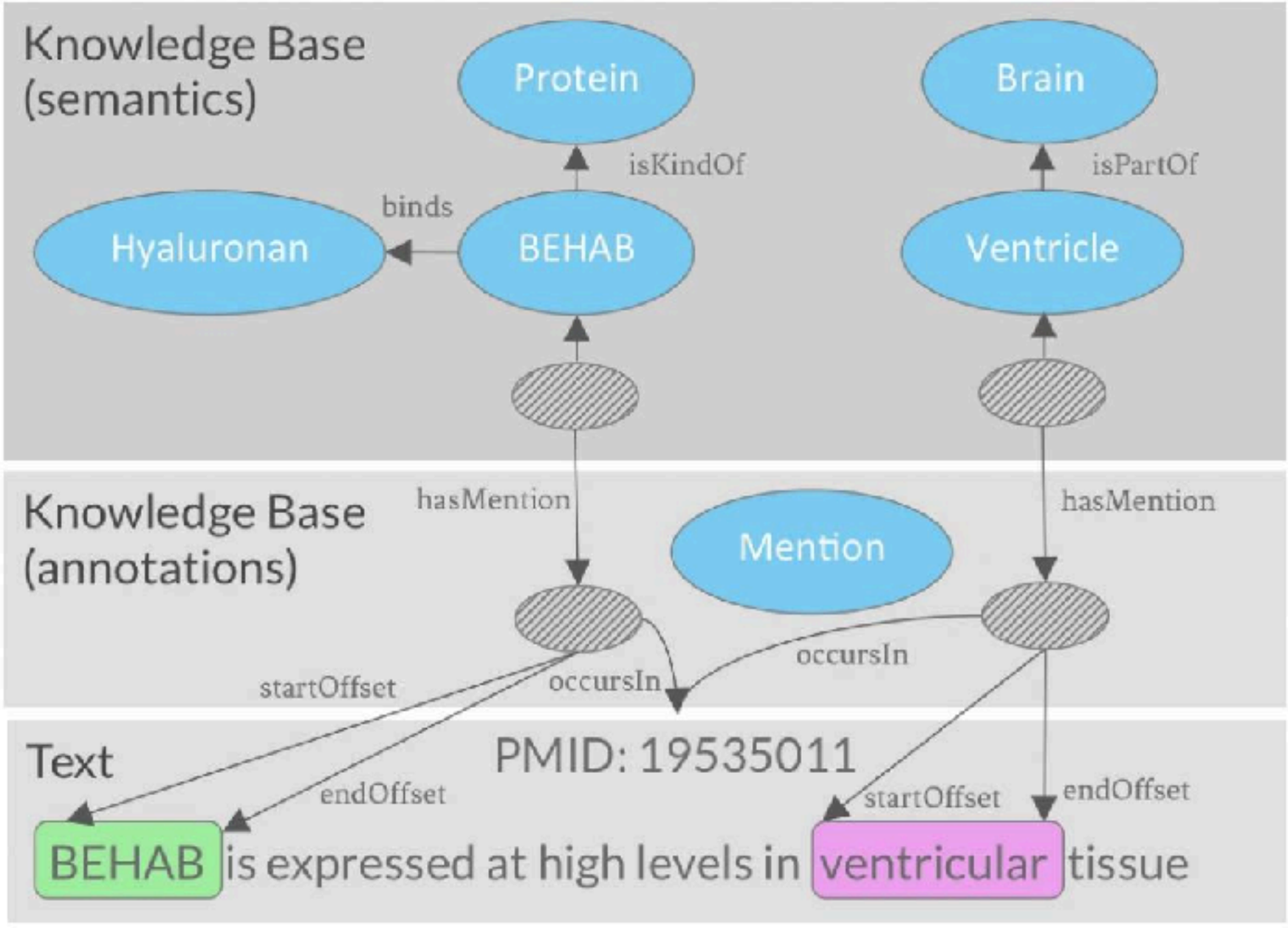
- Pourquoi ajouter des métadonnées à un document ?
  - Dans un but précis et en suivant des consignes précises
  - Annotation dans un document :
    - Extraire des connaissances interopérables, réutilisables
- Indexation MeSH
  - Support à la recherche d'information
- Codage ICD
  - Support à la création et gestion de dossiers patients

Annotation - le processus



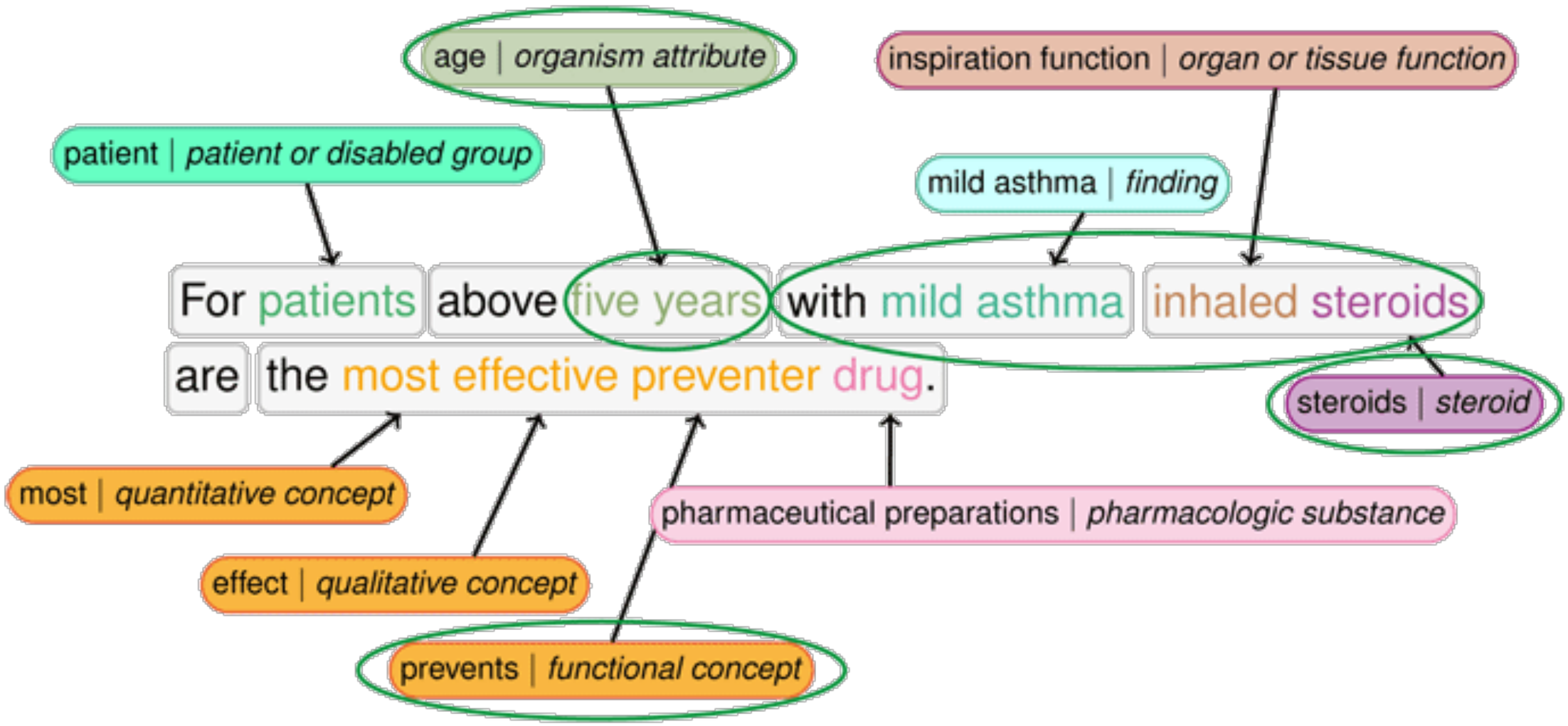
Dépend de la BCM utilisée

Annotation - le processus



http://kconnect.eu/sites/default/files/styles/gallery/public/images/kconnect_annotation2_june_2016.jpg?itok=m7jBnZL6

Annotation - le processus



http://ieg.ifs.tuwien.ac.at/~gschwand/mapface/project_page/img/corrections.gif

Metamap : un outil d'annotation

- Outil permettant d'annoter un texte avec les concepts UMLS

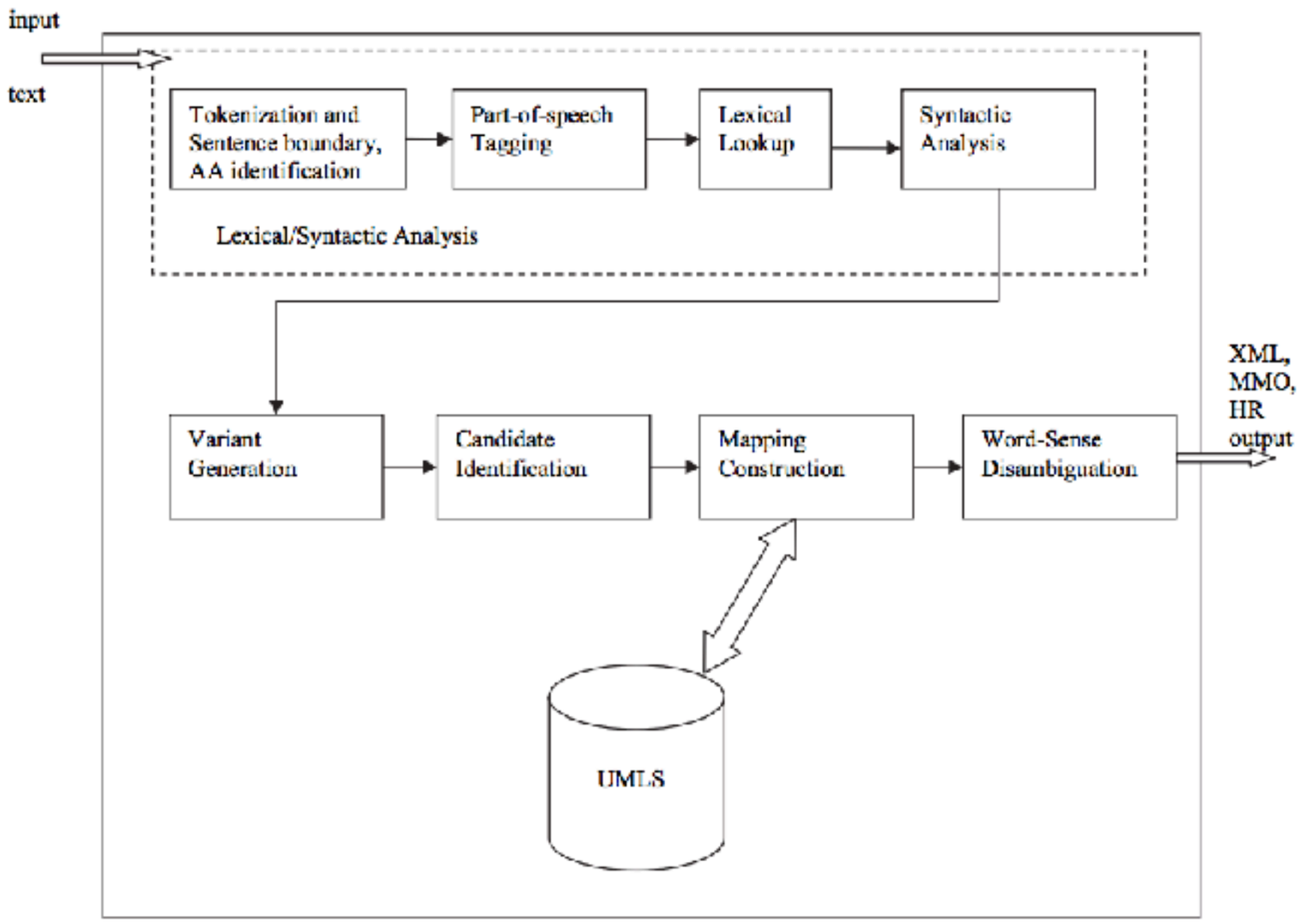
- Initialement conçu pour améliorer la recherche d'information sur MEDLINE (aller au delà des concepts annotés manuellement)

- Disponible en ligne, utilisable par les utilisateurs en possession d'une licence UMLS

- https://metamap.nlm.nih.gov/

Metamap : un outil d'annotation

Figure 1 MetaMap system diagram. HR, human readable; MMO, MetaMap machine output; UMLS, unified medical language system.



[Aronson & Lang, JAMIA 2010]

Metamap : un outil d'annotation

Input Text:

```
Hypertension and valvular heart disease are the most common alterable risk factors for AF.
```

Results:

```
Processing 00000000.tx.1: Hypertension and valvular heart disease are the most common alterable risk factors for AF.

Phrase: Hypertension
>>>>> Phrase
hypertension
<<<<< Phrase
>>>>> Mappings
Meta Mapping (1000):
  1000   HYPERTENSION (Hypertensive disease) [Disease or Syndrome]
Meta Mapping (1000):
  1000   Hypertension (Hypertension Adverse Event) [Finding]
<<<<< Mappings

Phrase: and
>>>>> Phrase
<<<<< Phrase

Phrase: valvular heart disease
>>>>> Phrase
valvular heart disease
<<<<< Phrase
>>>>> Mappings
Meta Mapping (1000):
  1000   Heart Disease, Valvular (Heart valve disease) [Disease or Syndrome]
Meta Mapping (1000):
  1000   Valvular heart disease (Valvular Heart Disease Adverse Event) [Finding]
<<<<< Mappings
```

Plan du cours

- Qu’est-ce qu’une base de connaissances ?
- Les principales bases de connaissances médicales (BCM)
- Annotation à l’aide de BCM
- Indexation à l’aide de BCM

Indexation - rappels

- Indexation manuelle : attribution humaine de concepts à un document (ex : MEDLINE)
  - Inconvénients : prix, temps, erreurs
- Indexation des mots (voir cours 1)
  - Inconvénients : ne traite pas la synonymie, polysémie, similarité sémantique, l'ambiguïté, la morphologie
  - (tout de même la plus utilisée)

Indexation sémantique

- Traite des textes annotés
- Indexe les concepts plutôt que les mots/racines
- Les CUI sont stockés dans l'index
- Permet de gérer les phénomènes de synonymie, polysémie, similarité sémantique
- (lors de la recherche les requêtes sont annotées)