Unified Medical Language System

Primo elaborato di Web Semantico

Veronika Folin - veronika.folin@studio.unibo.it

Settembre 2023

Contents

List of Figures

1 Introduzione

Oggigiorno, i ricercatori in campo biomedico hanno a disposizione un numero enorme di risorse: la letteratura è vasta ed in continua crescita. Ogni anno vengono pubblicati centinaia di migliaia di articoli scientifici in riviste biomediche di tutto il mondo. Esistono numerose basi di dati che raccolgono e indicizzano queste risorse, tra cui PubMed (gestito dalla National Library of Medicine degli Stati Uniti), Scopus, Web of Science e Google Scholar. Inoltre, la letteratura biomedica proviene da una varietà di fonti come università, istituti di ricerca, ospedali, aziende farmaceutiche e organizzazioni governative. Inoltre, la ricerca in questo ambito ha uno spettro internazionale e questo rende la letteratura ancora più diversificata. È fondamentale per i professionisti, i ricercatori e gli studiosi che lavorano nel campo medico e biomedico tenersi aggiornati su questa fonte di conoscenza in rapida evoluzione.

Per fare un esempio, il 14 settembre 2023, PubMed restituiva in totale 36,200,622 risultati e, nell'ultimo decennio, ogni anno sono stati aggiunti in media quasi 1 milione di nuovi record. Per verificare, si può inserire il seguente codice nel motore di ricerca di PubMed: `1800:2024[dp]`

Unified Medical Language System (UMLS) [1] è un progetto avviato nel 1986 dal medico Donald AB Lindberg e sviluppato dalla **US National Library of Medicine**. UMLS consiste in un sistema di ontologie, software e tools con lo scopo di mappare la terminologia biomedica in un formato unificato. Oltre a migliorare l'interoperabilità tra sistemi informatici biomedici, questo sistema può essere utilizzato per migliorare l'accesso alla letteratura scientifica, mediante lo sviluppo di applicazioni in grado di recuperare informazioni distribuite in molteplici basi di dati separate e di comprendere il linguaggio biomedico.

1.1 Obiettivi

UMLS ha diversi scopi e obiettivi, volti principalmente a facilitare l'integrazione e il recupero di informazioni biomediche e sanitarie da varie fonti. In particolare, viene applicato per:

- **Integrare informazioni biomediche**, armonizzando un'ampia gamma di terminologie, vocabolari e sistemi di codifica biomedici e clinici. Funge da ponte che collega e allinea terminologie disparate al fine di creare una rappresentazione unificata della conoscenza. Ad esempio, si potrebbe voler collegare, in riferimento ad una singola persona, i termini e codici utilizzati dal suo medico, dalla sua farmacia e dalla sua compagnia assicurativa.

- **Migliorare l'interoperabilità dei dati** tra diversi sistemi informativi sanitari. Mappando e collegando terminologie diverse, consente lo scambio di dati e la comunicazione tra varie applicazioni e database sanitari. Oltre ad essere tecnicamente interoperabili, i dati sono semanticamente interoperabili: possono essere scambiati e compresi con coerenza di significato tra diversi sistemi e organizzazioni. Ad esempio, si potrebbe coordinare l'assistenza ai pazienti tra diversi reparti all'interno di uno stesso ospedale.

- **Supportare applicazioni di Information Retrieval**: fornisce un vocabolario standardizzato e strutturato che aiuta gli utenti a recuperare in modo efficiente le informazioni rilevanti. Di conseguenza, consente ricerche più accurate e complete su diverse fonti e piattaforme.

- **Agevolare l'elaborazione del linguaggio naturale (NLP)**: UMLS offre una terminologia standardizzata e una rappresentazione semantica che può aiutare nella comprensione e nell'elaborazione di testo clinico non strutturato.

- **Supporto alle decisioni**: UMLS supporta lo sviluppo di sistemi di supporto alle decisioni cliniche, fornendo un linguaggio comune per codificare e rappresentare la conoscenza clinica ed agevolando la ricerca di informazioni rilevanti.

- **Ricerca e analisi dei dati**: consente l'analisi di grandi set di dati, l'esplorazione delle relazioni esistenti tra i concetti e lo sviluppo di soluzioni sanitarie innovative.

- **Istruzione e formazione**: funge da risorsa educativa per operatori sanitari, studenti e ricercatori. Aiuta nella comprensione dei concetti e della terminologia medica, promuovendo una comunicazione migliore all'interno della comunità sanitaria.

- **Sviluppo della terminologia**: fornisce una piattaforma per la creazione e il mantenimento di vocabolari, ontologie e sistemi di codifica controllati.

2 Descrizione del sistema

2.1 Componenti

Nelle successive sottosezioni vengono descritte le principali componenti del sistema UMLS.

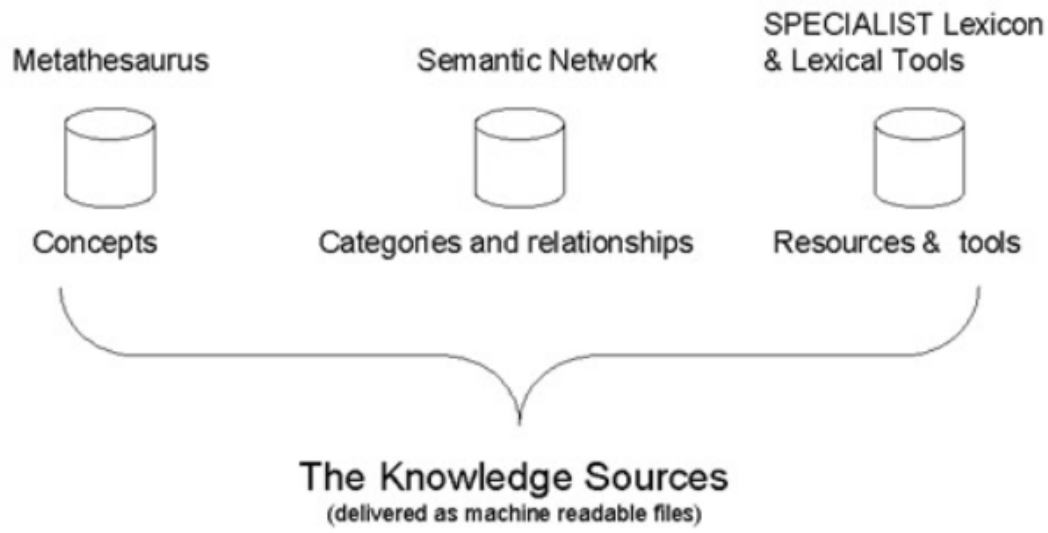

Figure 1: Overview della struttura di UMLS.

2.1.1 Metathesaurus

É il componente più grande di UMLS. É un ampio thesaurus biomedico, **organizzato per concetto o significato**, i cui dati derivano da quasi 200 vocabolari diversi (e.g., SNOMED CT, RxNorm, LOINC, MeSH, CPT, ICD-10-CM, MedDRA, Human Phenotype Ontology, etc; qui è disponibile la lista completa). Le risorse presenti appartengono a diverse categorie, come rappresentato nella Figura 2. Il principale obiettivo è quello di collegare concetti medici simili, provenienti da fonti diverse.

Ogni concetto ha **attributi** specifici che ne definiscono il significato ed è collegato ai concetti corrispondenti nei vari vocabolari di origine. Mediante Metathesaurus sono rappresentate numerose **relazioni** tra i concetti: ad esempio, quelle **gerarchiche** come "*is a*" per le sottoclassi e "*fa parte di*" per le sotto-unità, e quelle **associative** come "*è causato da*" o "*in letteratura spesso ricorre vicino a*". Tutti i concetti sono assegnati ad almeno un Tipo Semantico della Semantic Network.

Quando due diversi vocabolari utilizzano lo stesso nome per concetti diversi, il Metathesaurus rappresenta entrambi i significati e indica quale significato è presente in quale fonte. Quando lo stesso concetto appare in diversi contesti gerarchici in diversi vocabolari di origine, il Metathesaurus include tutte le gerarchie. Quando relazioni con-

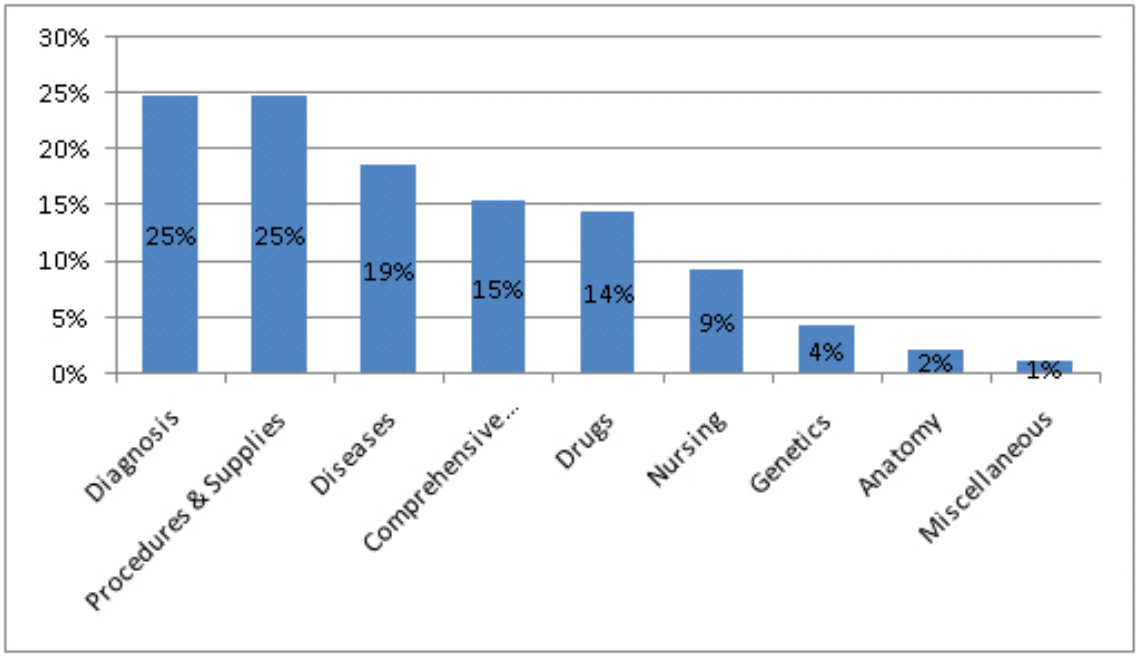

Figure 2: Percentuale delle diverse categorie di vocabolario rappresentate nel Metathesaurus.

trastanti tra due concetti compaiono in vocabolari diversi, entrambe le visioni sono incluse. Dunque, il Metathesaurus **non rappresenta una singola vista consistente** dell'ontologia biomedica: preserva le numerose visioni del mondo presenti nei vocabolari di partenza, in quanto possono essere utili per task diversi.

Per la costruzione del Metathesaurus, il formato nativo di ogni vocabolario viene studiato attentamente e poi convertito in un **formato comune**. Inoltre, la struttura del Metathesaurus facilita la traduzione dei suoi vocabolari in lingue diverse dall'inglese: difatti, l'attuale versione del sistema contiene numerose traduzioni.

Poiché è una risorsa multi-purpose che include concetti e termini provenienti da vocabolari di origine diversa, sviluppati per scopi molto diversi, il Metathesaurus deve essere **personalizzato** per un utilizzo efficace in applicazioni specifiche, attraverso l'inclusione di sottoinsiemi selezionati. Le fonti essenziali per alcuni scopi, ad esempio LOINC per lo scambio standardardizzato di dati di laboratorio, possono risultare svantaggiose per altri, come NLP. MetamorphoSys è la procedura di installazione guidata di UMLS e il tool di personalizzazione di Metathesaurus, incluso in ogni versione. Altri modi per accedere a Metathesaurus sono il download diretto dei dati, il Metathesaurus Browser o APIs.

Il Metathesaurus consiste generalmente in 40 file che possono seguire due formati relazionali diversi: **Rich Release Format (RRF)** (preferibile) e **Original Release**

Format (ORF).

Gli **index files** vengono prodotti per aiutare gli sviluppatori a creare applicazioni che cercano parole o gruppi di parole specifici e forniscono tre indici per ogni stringa.

1. *Word Index*: per ogni lingua, ogni parola trovata in ciascuna stringa univoca viene collegata a tutti i relativi identificatori di stringa, termine e concetto.
2. *Normalized Word Index*: collega ogni singola parola (inglese) normalizzata a tutti i relativi identificatori di stringa, termine e concetto.
3. *Normalized String Index*: collega la forma normalizzata di ogni stringa del Metathesaurus a tutti i relativi identificatori di stringa, termine e concetto.

I **metadata files** contengono informazioni su ciascuna versione del Metathesaurus, come:

- le caratteristiche della versione attuale;
- le modifiche tra la versione attuale e quella precedente;
- lo storico degli identificatori di concetto (CUI).

I **data files** contengono le informazioni ottenute dai vocabolari, come mostrato in Figura 3.

Metadata File Name	Contents
MRCONSO.RRF	Names, Synonyms, Terms, Term Types, Codes
MRREL.RRF	Relationships
MRHIER.RRF	Hierarchies
MRSAT.RRF	Attributes
MRDEF.RRF	Definitions
MRMAP.RRF	Mappings
MRSMAP.RRF	Simplified Mappings
MRSTY.RRF	Semantic Types

Figure 3: Contenuto dei data files in Metathesaurus.

Quando un concetto viene aggiunto al Metathesaurus riceve un identificatore univoco, per quattro livelli di specificità (vedi Figura 4):

- **Concept Unique Identifiers (CUI)** [C]: un concetto è un significato che può avere molti nomi diversi. L'obiettivo è comprendere il significato previsto da ciascun vocabolario e collegare tutti i nomi che significano la stessa cosa (sinonimi);
- **Lexical (term) Unique Identifiers (LUI)** [L]: collega stringhe che sono varianti lessicali, rilevate utilizzando il programma Lexical Variant Generator (LVG), uno dei Lexical Tools di UMLS;

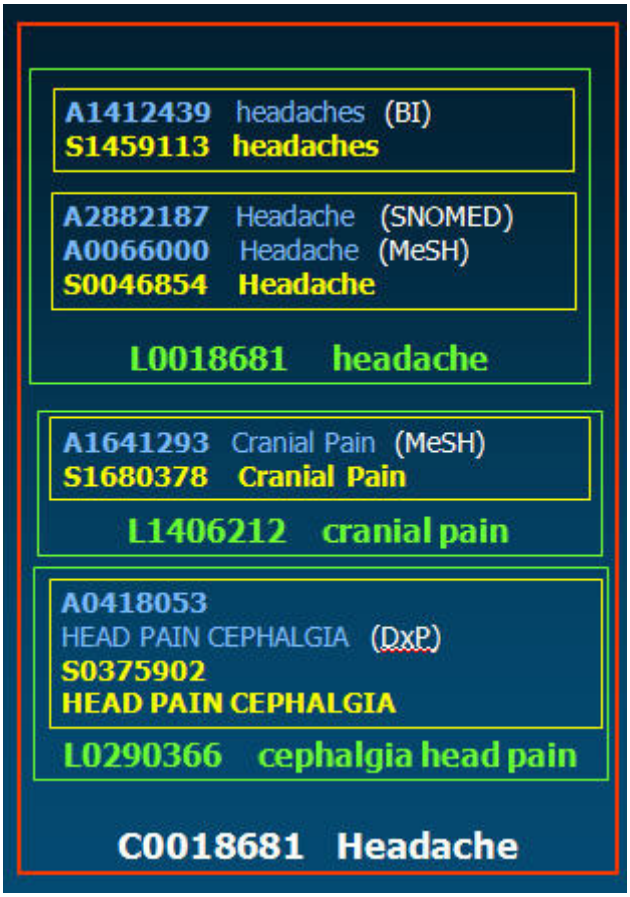

Figure 4: Occorrenze del concetto "Headache" all'interno del Metathesaurus.

- **String Unique Identifiers (SUI)** [S]: ogni stringa univoca, in ciascuna lingua, ha un identificatore univoco e permanente. Qualsiasi variazione nel set di caratteri, nelle lettere maiuscole e minuscole o nella punteggiatura costituisce una stringa separata;
- **Atom Unique Identifiers (AUI)** [A]: gli "atomi" da cui è costruito il Metathesaurus sono i record di ciascuno dei vocabolari di origine. A ogni occorrenza viene assegnato un identificatore di atomo univoco. Se esattamente la stessa stringa appare più volte nello stesso vocabolario, ad esempio come nome alternativo per concetti diversi, per ogni occorrenza viene assegnato un AUI diverso. L'abbreviazione della fonte che ha contribuito a ciascuna stringa è annotata tra parentesi dopo la stringa.

In alcuni casi, lo stesso nome può applicarsi a concetti diversi (vedi esempio in Figura 5). Questa eventualità viene gestita mediante appositi file che contengono tutti i **termini ambigui** noti.

Concepts (CUIs)	Terms (LUIs)	Strings (SUIs)	Atoms (AUIs) * RRF only
C0009264 **Cold Temperature**	**L0215040** cold temperature	**S7669511** Cold Temperature	**A15594156** Cold Temperature (from MTH)
	L0009264 cold	**S0026353** Cold	**A0040709** Cold (from LCH)
			A4711382 Cold (from SNOMEDCT)
C0009443 **Common Cold**	**L0009443** cold common	**S0026747** Common Cold	**A0041261** Common Cold (from MSH)
	L0009264 cold	**S0026353** Cold	**A0040708** Cold (from COSTAR)
			A2880095 Cold (from SNOMEDCT)
C0024117 **Chronic Obstructive Airway Disease**	**L0498186** airway chronic disease obstructive	**S0837575** Chronic Obstructive Airway Disease	**A0896021** Chronic Obstructive Airway Disease (from MSH)
	L0008703 chronic disease lung obstructive	**S0837576** Chronic Obstructive Lung Disease	**A0896023** Chronic Obstructive Lung Disease (from MSH)
	L0009264 cold	**S0474508** COLD	**A10765219** COLD (from NCI)
			A0539536 COLD (from SNMI)

Figure 5: Ambiguità della stringa "cold" in Metathesaurus.

Il Metathesaurus include molte relazioni tra concetti diversi (oltre alle relazioni tra sinonimi gestite tramite gli identificatori univoci) dello stesso vocabolario (intra-source vocabulary relationships) e di vocabolari diversi (inter-source vocabulary relationships). Le **relazioni intra-source** si verificano nelle disposizioni gerarchiche o nei contesti espliciti o impliciti di un vocabolario di origine, nelle strutture cross-reference, nelle regole per l'applicazione di qualificatori o nelle connessioni tra i diversi nomi per lo stesso concetto (ad esempio, abbreviazioni e forme complete). Le **relazioni inter-source** possono essere generate durante la costruzione del Metathesaurus per collegare specifici concetti "orfani" ad un contesto più ricco di un altro vocabolario, possono essere segnalate dagli utenti oppure vengono generate delle mappature tra due diversi vocabolari tramite progetti ad-hoc, generalmente sotto la supervisione del National Library of Medicine (NLM), l'ente responsabile del Metathesaurus.

Tutte le relazioni nel Metathesaurus portano un'etichetta (**Relationship Labels**, REL), che ne descrive la natura di base (ad esempio, CHD sta per "has child relationship in a Metathesaurus source vocabulary"). Circa un quarto delle relazioni portano anche un'etichetta aggiuntiva (RELA) che spiega più esattamente la natura della relazione (ad esempio, "cause_of" o "diagnosed_by"). Ogni relazione presente nel Metathesaurus ha un identificatore di relazione univoco (Relationship Identifiers, RUI) e permette, principalmente, di rilevare dei cambiamenti nelle relazioni tra le versioni del Metathesaurus: la comparsa o la scomparsa di un identificatore indica un cambiamento nelle relazioni presenti nel Metathesaurus.

I gruppi di relazioni (**Relationship Groups**, RG) sono associazioni di relazioni dichiarate o implicite che possono essere utilizzate per aggiungere significato o chiarezza quando sono presenti più relazioni.

Il Metathesaurus può includere anche attributi di:

- *concetto*, che si applicano a tutti i nomi di un concetto. Ad esempio, "Pathologic Function" e "Finding" sono attributi del concetto "Atrial Fibrillation".
- *atomo*, che provengono da un particolare vocabolario. Ad esempio, la definizione "Disorder of cardiac rhythm characterized by rapid, irregular atrial impulses and ineffective atrial contractions" è un attributo dell'atomo "Atrial Fibrillation" in MeSH.
- *relazione*, che descrivono caratteristiche speciali di una relazione in un determinato vocabolario.

A ciascuna occorrenza di ciascun attributo all'interno del Metathesaurus viene assegnato un identificatore univoco (ATUI), che permette di identificare eventuali cambiamenti nel contenuto.

Esistono una serie di diagrammi ER-like chiamati **query diagrams**, ognuno dei quali considera un caso d'uso particolare e mostra come i file possono essere uniti ed interrogati per trovare tutti i dati rilevanti (vedi Figura 6).

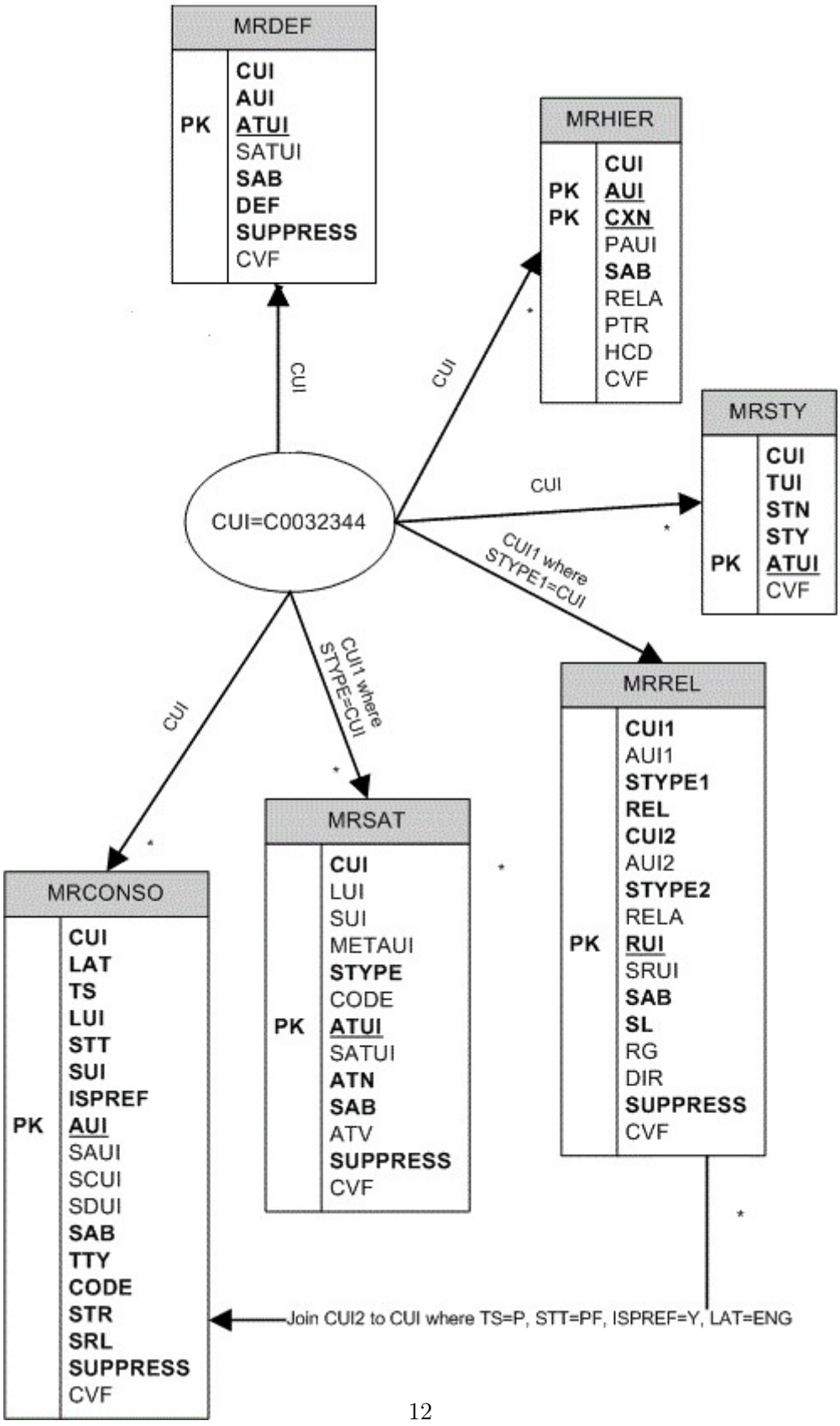

Figure 6: Query Diagram per trovare tutte le informazioni associate a un particolare concetto UMLS.

2.1.2 Semantic Network

Consiste in un insieme di **Tipi Semantici** (nodi della rete) che forniscono una categorizzazione di tutti i concetti rappresentati nel Metathesaurus (e.g., Disease, Syndrome, Clinical Drug) e **Relazioni Semantiche** (collegamenti tra i nodi) che esistono tra i tipi (e.g., Clinical Drug treats Disease or Syndrome) (vedi Figura 7). L'ultima versione risalente al 2023 riporta 127 tipi semantici e 54 relazioni semantiche.

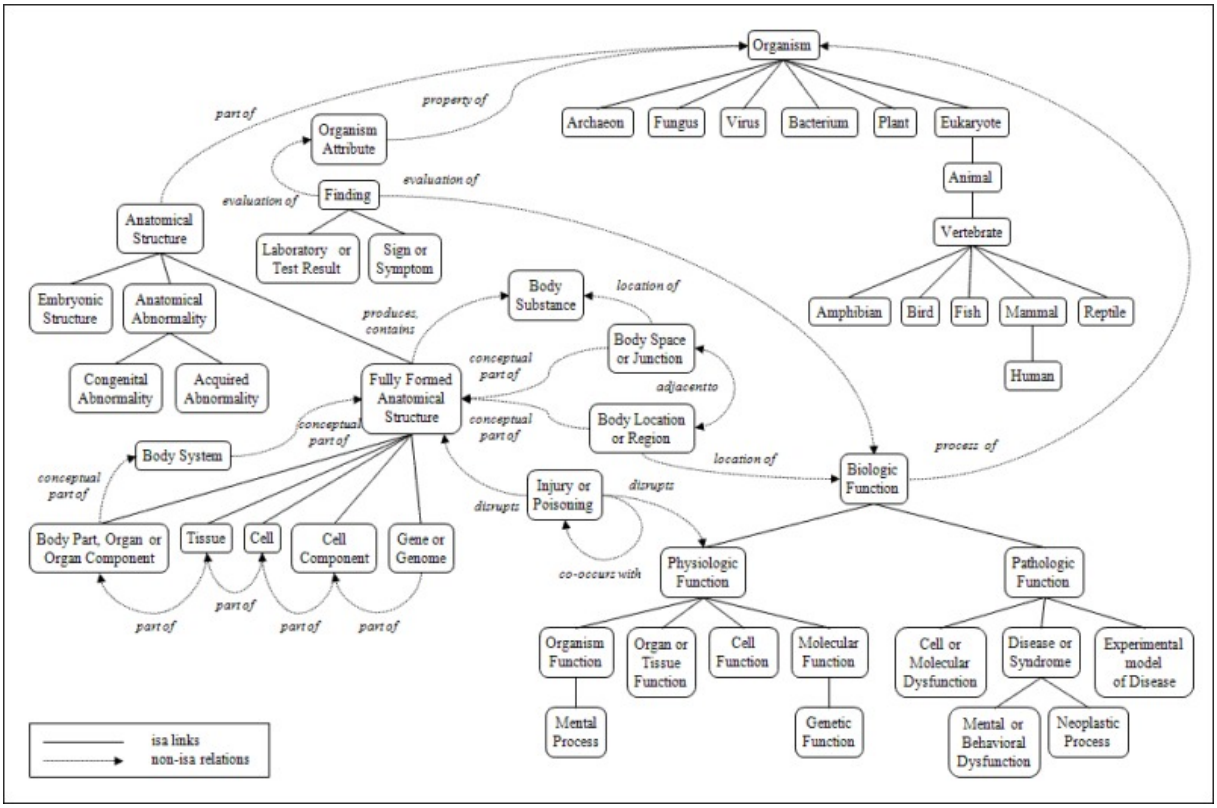

Figure 7: Porzione della Semantic Network: tipi e relazioni semantiche.

Esempi di tipi semantici sono organismi, strutture anatomiche, funzioni biologiche, sostanze chimiche, eventi, oggetti fisici, concetti o idee. Questi sono organizzati in una gerarchia che prevede due categorie principali (vedi Figura 10): **Entità** (e.g., "Amphibian", "Gene", "Carbohydrate") ed **Evento** ("Social Behavior", "Laboratory Procedure", "Mental Process").

Le relazioni **gerarchiche *is-a*** stabiliscono la gerarchia dei tipi all'interno della rete e vengono utilizzate per decidere il tipo semantico più specifico disponibile per l'assegnazione a un concetto del Metathesaurus (vedi Figura 8). Le **relazioni associative** possono appartenere alle categorie "*fisicamente correlati a*", "*spazialmente correlati a*", "*temporalmente correlati a*", "*funzionalmente correlati a*" e "*concettualmente correlati a*".

Anche le relazioni sono organizzate in maniera gerarchica (vedi Figura 9), vengono stabilite tra tipi semantici di alto livello nella rete e sono generalmente ereditate tramite il collegamento *is-a* da tutti i figli di quei tipi. Se le relazioni non si applicano a tutte le istanze di concetti che sono state assegnati a quei tipi semantici, si parla di **relazioni deboli**. In alcuni casi si verificherà un conflitto tra il posizionamento dei tipi nella rete e il collegamento da ereditare: in tal caso, si dice che l'ereditarietà del collegamento è bloccata. Ad esempio, per ereditarietà il tipo "Mental Process" sarebbe "process of" "Plant": poiché le piante non sono esseri senzienti, questo collegamento è bloccato.

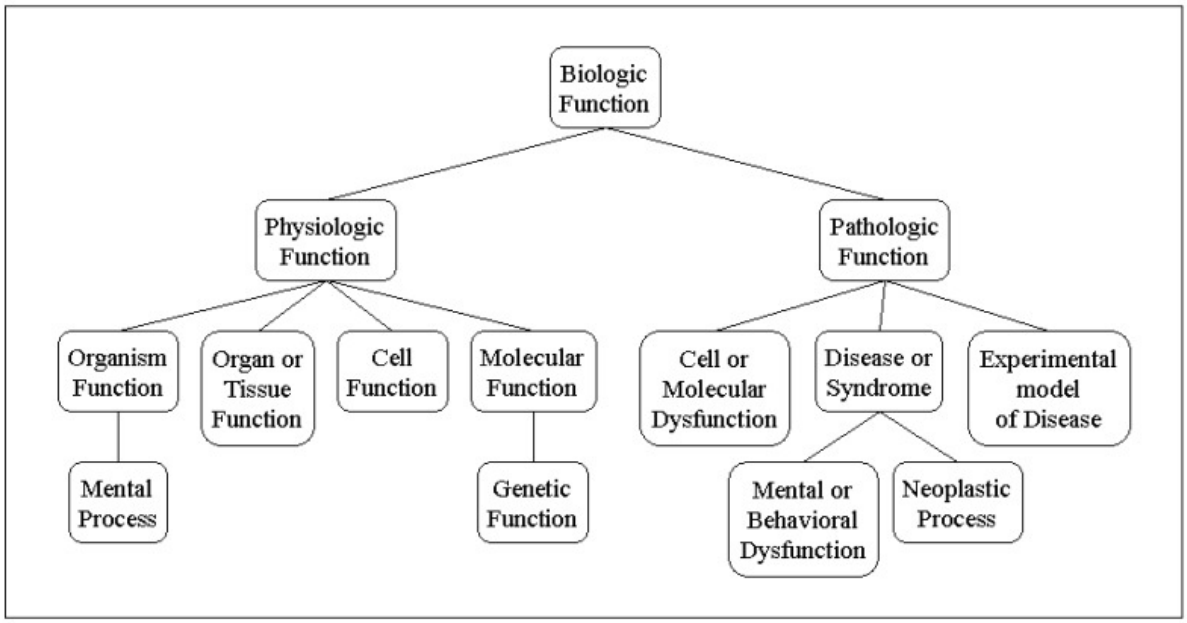

Figure 8: Gerarchia del tipo semantico "Biologic Function".

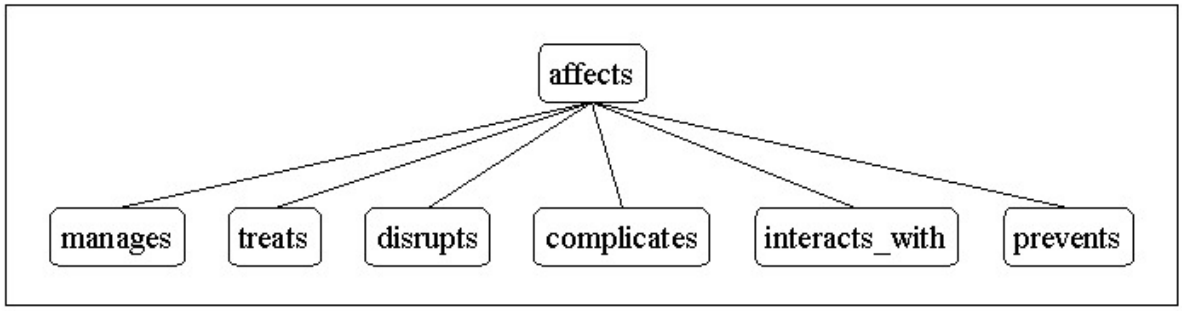

Figure 9: Gerarchia della relazione semantica "affects".

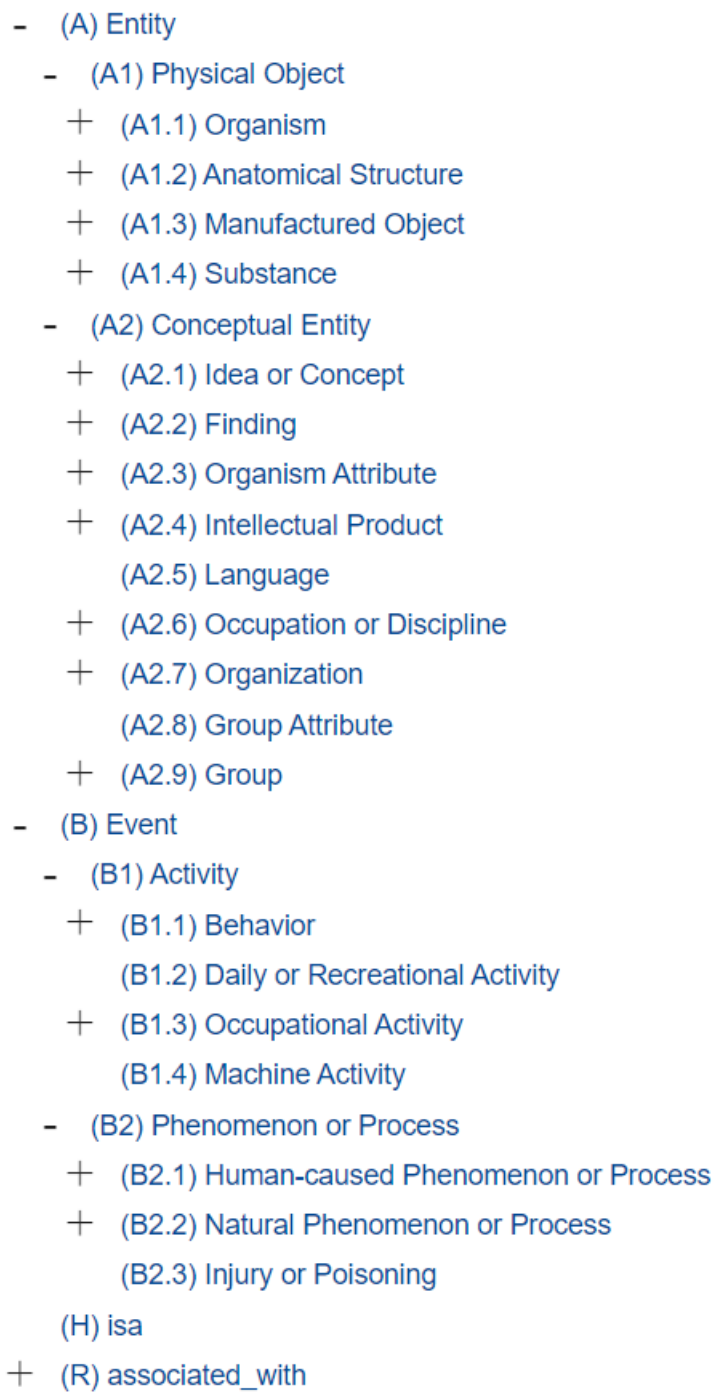

Figure 10: Classificazione dei tipi semantici in Semantic Network Browser.

La Semantic Network riduce la complessità del Metathesaurus raggruppando i concetti secondo i tipi semantici a loro assegnati (vedi esempio in Figura 11). I **gruppi semantici** forniscono una partizione del Metathesaurus per il 99,5% dei concetti.

Le informazioni della rete sono disponibili in due formati: tabella relazionale o unit record. Per accedervi è possibile scaricare localmente i dati, oppure è disponibile il Semantic Network Browser.

```
ANAT|Anatomy|T017|Anatomical Structure
ANAT|Anatomy|T029|Body Location or Region
ANAT|Anatomy|T023|Body Part, Organ, or Organ Component
ANAT|Anatomy|T030|Body Space or Junction
ANAT|Anatomy|T031|Body Substance
ANAT|Anatomy|T022|Body System
ANAT|Anatomy|T025|Cell
ANAT|Anatomy|T026|Cell Component
ANAT|Anatomy|T018|Embryonic Structure
ANAT|Anatomy|T021|Fully Formed Anatomical Structure
ANAT|Anatomy|T024|Tissue
```

Figure 11: Semantic Group "Anatomy".

2.1.3 SPECIALIST Lexicon e Lexical Tools

SPECIALIST Lexicon. É un ampia raccolta di lessico e sintassi inglese che comprende molti termini biomedici. Tra le sue fonti vi sono MEDLINE (Medical Literature Analysis and Retrieval System Online), Metathesaurus, Dorland's Illustrated Medical Dictionary e il vocabolario inglese. É stato sviluppato per fornire le informazioni lessicali necessarie per il sistema di elaborazione del linguaggio naturale SPECIALIST.

Il lessico è costituito da un insieme di voci lessicali, ognuna delle quali rappresenta una parola. La voce copre una o più forme ortografiche e ne descrive le proprietà morfologiche, ortografiche e sintattiche ma non quelle semantiche. Ogni voce comprende:

- forma base;
- parte del discorso;
- identificatore univoco;
- tutte le varianti ortografiche disponibili.

SPECIALIST Lexicon è disponibile in due formati:

- **unit record**, che comprende degli elementi, definiti *slot*, a cui possono essere attribuiti dei valori, definiti *fillers* (vedi Figura 12). Lo slot *spelling_ variants* indica le varianti ortografiche, *entry* registra l'identificatore univoco (EUI) del record, *cat* specifica la parte del discorso, *variant* la morfologia flessiva del termine e *position* la posizione nel discorso.

```
{ base=anaesthetic
      spelling_variant=anesthetic
      entry=E0008769
          cat=noun
          variants=reg
}
{ base=anaesthetic
      spelling_variant=anesthetic
      entry=E0008770
          cat=adj
          variants=inv
          position=attrib(3)
}
```

Figure 12: Termine "anesthetic" in formato unit record.

- **relational table** dove ogni voce lessicale è rappresentata in diverse tabelle, ciascuna memorizzata in un file (vedi Figura 13). Questo formato contiene una grande quantità di dati ridondanti in quanto non è normalizzata. Oltre a ciò, sono in fase di sviluppo i cosiddetti Lexical Databases, ossia database che contengono informazioni lessicali ritenute utili per l'elaborazione del linguaggio naturale (e.g., sinonimi).

SPECIALIST Lexical Tools. Sono un insieme di programmi JAVA progettati per aiutare gli utenti a gestire le **variazioni lessicali** in testi biomedici, facilitando lo sviluppo di applicazioni NLP o la risoluzione di task come quello di *indexing*. Tra l'altro, i termini composti da più parole nel Metathesaurus e in altri vocabolari possono avere varianti nell'ordine delle parole: i tools consentono all'utente di astrarre da questo tipo di variazione.

Il pacchetto è composto da tre applicazioni principali: un normalizzatore (vedi Figura 14), un generatore di word index e un generatore di varianti lessicali. Le funzioni includono il recupero di varianti flesse, forme non flesse, varianti ortografiche, sinonimi, contrari, termini normalizzati, conversi da UTF-8 ad ASCII, lettere minuscole, abbreviazioni e acronimi, etc...

É possibile utilizzare questi tool online o in locale, previa installazione.

Table Name	Descriptions
./	
LEXICON	Lexical records
LRABR	Abbreviations and Acronyms
LRAGR	Agreement and Inflection
LRCMP	Complementation
LRFIL	File Description
LRFLD	Field Description
LRMOD	Modifiers
LRNEG	Negatives
LRNOM	Nominalizations
LRPRN	Pronouns
LRPRP	Properties
LRSPL	Spelling Variants
LRTRM	Trade Marks
LRTYP	Inflection Type
LRWD	Word Index
./LEX_DB/	
AM.DB	Antonyms
DM.DB	Derivations
NC.DB	Neoclassical Compounds
SM.DB	Synonymys
./MISC/	
inflection.table	Inflectional variants
inflVars.data	Inflectional variants
negCueWords	Negation detection cue words
prevariants	Prevariants
./NUMBERS/	
NRNUM	Numbers
NRVAR	Variants of Numbers

Figure 13: Tabelle relazioni presenti in Lexicon.

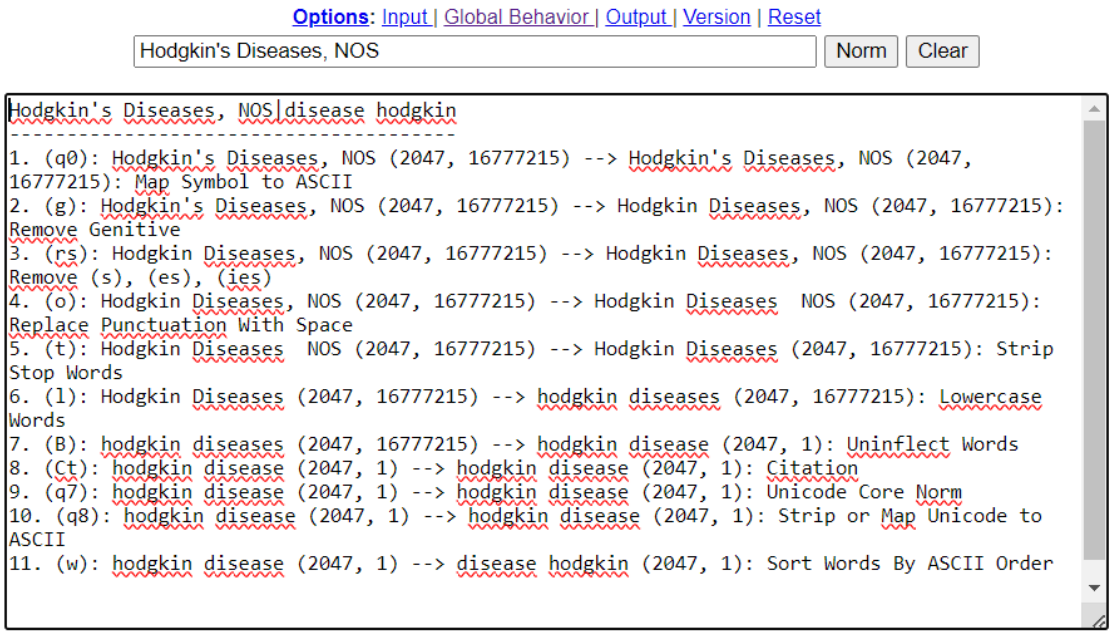

Figure 14: Esempio di utilizzo della funzione Norm di Lexical Tools.

2.2 Tecnologie e linguaggi

UMLS è un sistema complesso che incorpora varie tecnologie e linguaggi per raggiungere i suoi obiettivi. Di seguito vengono elencate quelle chiave.

- **Database Management Systems (DBMS)**, vengono utilizzati per archiviare e gestire la vasta raccolta di terminologie, concetti e relazioni;
- **Data Transformation**, per convertire e mappare dati da diverse fonti in un formato comune all'interno del Metathesaurus UMLS;
- **Java**, alcuni componenti dell'UMLS (e.g., UMLS Knowledge Source Server [2]) e i SPECIALIST Lexical Tools sono sviluppati con questo linguaggio;
- **Python** è comunemente utilizzato per l'elaborazione dei dati, lo scripting e lo sviluppo di applicazioni che interagiscono con i dati UMLS;
- **RESTful Web Services**, ossia API che consentono di accedere ai servizi e a recuperare i dati;
- **Web-Based User Interfaces** che consentono agli utenti di esplorare e interrogare i propri dati, come i UMLS Terminology Services (UTS);
- **Web Ontology Language (OWL)**, per rappresentare e modellare relazioni complesse tra concetti.

3 Applicazioni

L'UMLS è fornito gratuitamente: gli utenti di SPECIALIST Lexicon & Lexical Tools e Semantic Network devono leggere i termini di utilizzo prima di utilizzare uno dei prodotti; gli utilizzatori di Metathesaurus sono tenuti ad accettare un contratto di licenza prima di scaricare i file.

Il sistema viene aggiornato periodicamente, ogni trimestre, ed è possibile accedervi mediante i **UMLS Terminology Services (UTS)**:

1. Metathesaurus Browser o Semantic Network Browser;
2. Installazione locale;
3. Web Services APIs.

3.1 Analisi degli utenti e degli utilizzi del sistema

La National Library of Medicine (NLM) raccoglie regolarmente dati riepilogativi sull'uso diretto delle risorse UMLS, che includono informazioni riguardo la registrazione degli utenti e i report inviati dagli utenti, oltre alle statistiche sui download e sulle chiamate tramite le interfacce delle applicazioni. NLM richiede annualmente un **report** da parte di chi possiede una licenza UMLS: la mancata presentazione del rapporto annulla automaticamente la licenza, impedendo così l'accesso alle nuove versioni.

Un articolo [3] risalente al 2020 riporta un **aumento significativo nell'utilizzo** dal 2013 al 2017, misurato in base al numeri di download, richieste API e risposte ai sondaggi (vedi Figura 15).

	2013	2014	2015	2016	2017	2018
Downloads	2020	2248	2249	2691	4898	4402
API Requests (millions)	15.0	53.0	32.7	29.1	68.4	66.1
Survey responses	3366	4096	4169	4500	5145	5043

Figure 15: Utilizzo di UMLS dal 2013 al 2018.

Nel 2018, gli utenti si sono identificati come ricercatori (42%), sviluppatori di software (28%), amministratori/manager (20%), operatori sanitari (7%), educatori (5%) , analisti (5%) e studenti (3%). In particolare, vengono segnalate più comunemente affiliazioni con istituzioni accademiche (32%), enti a scopo di lucro (23%), enti senza scopo di lucro (16%) e governo (7%).

In Figura 16, viene riassunto come hanno risposto gli utenti alla domanda "Per quali **scopi** hai utilizzato UMLS?".

For what purpose(s) did you use the UMLS?	Count (n= 5043)	%
Processing of texts to extract concepts, relationships or knowledge	2553	51%
Facilitate mapping between terminologies	2486	49%
Extract specific terminologies from the Metathesaurus (eg, MedDRA, MeSH, NDF-RT)	1442	29%
Develop an information retrieval system	949	19%
Creation and maintenance of local terminology	943	19%
Research terminologies and ontologies beyond any of the above categories	917	18%
Other	405	8%
Support of a terminology server or service	353	7%

Figure 16: Domanda: "Per quali scopi hai utilizzato UMLS?"

Inoltre, lo stesso studio ha analizzato un campione casuale della **letteratura** disponibile che utilizza UMLS o prodotti correlati come strumento metodologico (vedi Figure 17 e 18).

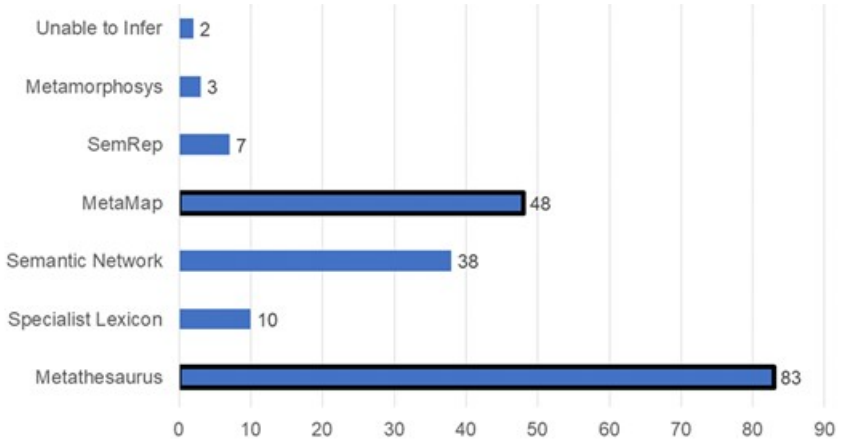

Figure 17: Utilizzo dei prodotti UMLS nella letteratura scientifica.

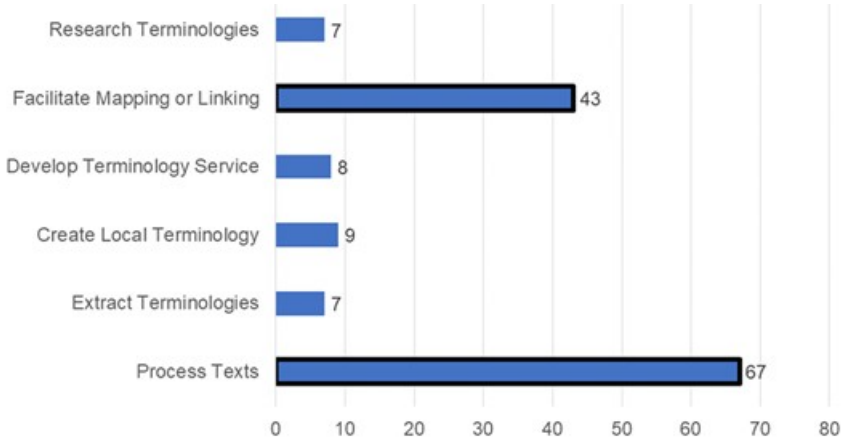

Figure 18: Casi d'uso di UMLS nella letteratura scientifica.

Il grafico in Figura 19 mostra la progressione del numero di licenze che è avvenuta tra gli anni 2010 e 2022 (fonte: National Institutes of Health (NIH)).

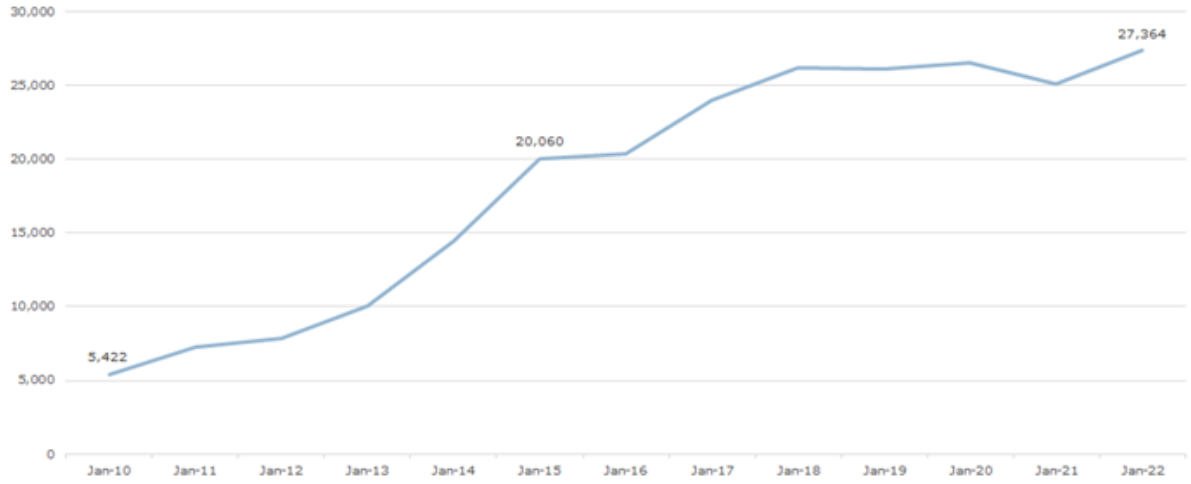

Figure 19: UMLS Licensee Count 2010-2022.

A gennaio 2023 sono stati registrati più di 30.000 utenti attivi in tutto il mondo, provenienti da 132 paesi.

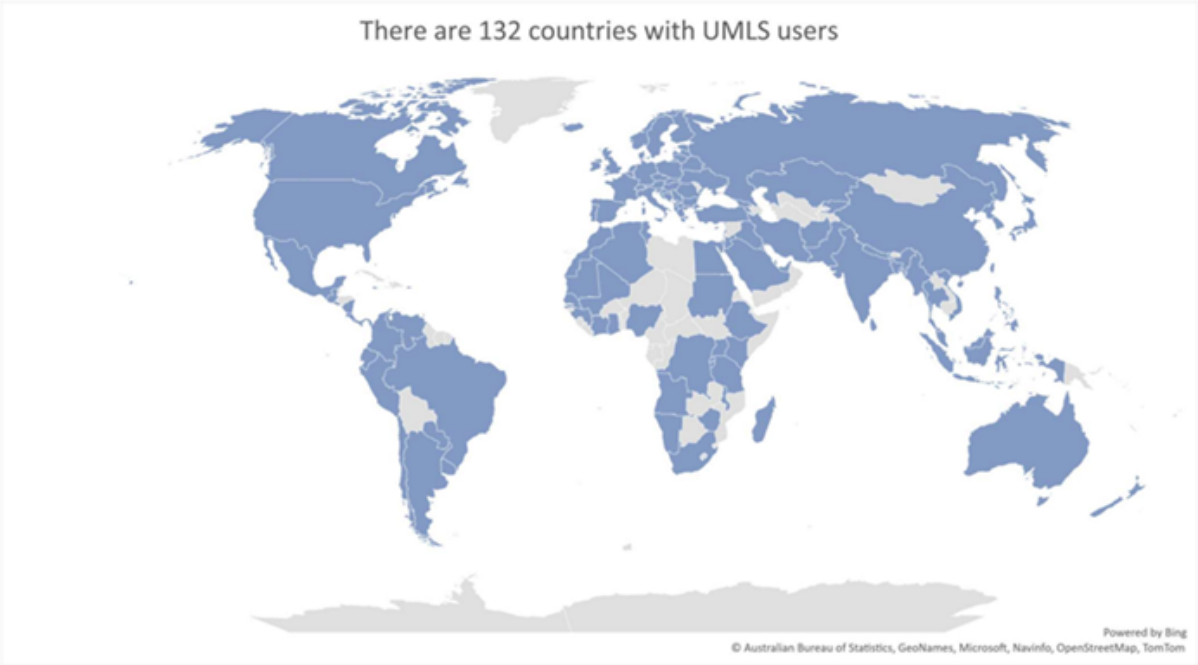

Figure 20: Paesi di origine degli utenti (a gennaio 2023).

Le migliaia di utenti di UMLS sono principalmente ricercatori informatici e sviluppatori, che utilizzano il sistema per **creare o migliorare risorse e applicazioni**. I prodotti sviluppati che si basano su UMLS sono utilizzati da milioni di persone in tutto il mondo, il cui impatto non è al momento quantificabile. La Figura 21 mostra la gamma di utenti che hanno utilizzato un tool basato su UMLS (fonte: Annual Report 2021).

Number of users who use your tool	Application count
1-10	421
11-100	204
101-1,000	168
1,001-10,000	126
More than 10,000	77

Figure 21: Popolarità dei tools basati su UMLS.

3.2 Progetti basati su UMLS

La comunità ha contribuito a estendere le funzionalità di UMLS sviluppando APIs, script automatici e tools di NLP, riassunti in questa pagina. Ad esempio, troviamo:

- **MetaMap**: è un tool di NLP, sviluppato dalla National Library of Medicine, che mappa il testo in concetti medici standardizzati. Aiuta nel recupero delle informazioni, nell'estrazione dei dati e nel supporto alle decisioni cliniche.

Altri esempi di progetti reali sono:

- **RxNorm** è un progetto che utilizza UMLS per standardizzare e normalizzare nomi e codici di farmaci. Aiuta gli operatori e i sistemi a scambiare informazioni in modo accurato, migliorando la sicurezza dei pazienti e l'interoperabilità dei sistemi.
- **ClinicalTrials.gov** è un database di studi clinici e utilizza UMLS per migliorare le funzionalità di ricerca.
- **PubMed** è repository di letteratura biomedica ampiamente utilizzato. Consente di ricercare articoli utilizzando la terminologia medica standardizzata di UMLS.
- **OpenInfobutton** è suite open-source di web services che consente l'integrazione di risorse online di evidenza scientifica con i sistemi di cartelle cliniche elettroniche (EHR). Fornisce un supporto al processo decisionale, chiarificando il contesto clinico del paziente e offrendo raccomandazioni basate sui dati della persona in cura. Dunque sfrutta la terminologia e la conoscenza intrinseca di UMLS.
- **i2b2 (Informatics for Integrating Biology and the Bedside)** è una piattaforma open-source di ricerca analitica e di data warehousing che consente la condivisione, l'integrazione, la standardizzazione e l'analisi di dati eterogenei in ambito sanitario e di ricerca.
- **Value Set Authority Center (VSAC)** è un repository e un tool di creazione per *value set*, ossia elenchi di codici e termini, tratti da vocabolari clinici standard. I concetti definiti all'interno permettono uno scambio di informazioni sanitarie efficace e interoperabile (e.g., nella cartelle cliniche elettroniche).

3.3 Progetti correlati

Esistono numerose ontologie mediche e sistemi terminologici nel campo della sanità e della biomedicina: ognuno di essi ha uno scopo specifico, la cui adozione aiuta a garantire coerenza e interoperabilità nello scambio e nella ricerca di informazioni. Di seguito vengono elencati i sistemi più conosciuti ed utilizzati:

- **SNOMED CT (Systematized Nomenclature of Medicine - Clinical Terms)** è un sistema di codifica e di terminologia clinica completo e riconosciuto a livello internazionale. Viene utilizzato per acquisire, archiviare e scambiare documentazione clinica e informazioni sui pazienti. Inoltre, è progettato per supportare sistemi EHR

e di supporto alle decisioni, rendendolo uno strumento essenziale per gli operatori sanitari, i ricercatori e lo sviluppo di sistemi IT sanitari.

- **LOINC (Logical Observation Identifiers Names and Codes)** è uno standard per identificare e scambiare osservazioni cliniche e di laboratorio. Viene comunemente utilizzato per codificare test di laboratorio, misurazioni cliniche e altre osservazioni sanitarie.

- **ICD (International Classification of Diseases)** è un sistema ampiamente utilizzato per classificare e fornire conoscenza rispetto la portata, le cause e le conseguenze delle malattie umane. É gestito dall'Organizzazione Mondiale della Sanità (OMS) e viene utilizzato per codificare diagnosi e cause di morte.

- **MedDRA (Medical Dictionary for Regulatory Activities)** è un dizionario terminologico medico standardizzato sviluppato dall'International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH). Ha il fine di facilitare lo scambio di informazioni nell'ambito della regolamentazione internazionale per prodotti medicali per uso umano e si applica a tutte le fasi dello sviluppo di questi. La terminologia riguarda anche gli effetti dei dispositivi sulla salute.

- **National Cancer Institute (NCI) Thesaurus** è un'ontologia utilizzata per la ricerca sul cancro.

- **MeSH (Medical Subject Headings)** è un progetto sviluppato dalla National Library of Medicine (NLM). Viene utilizzato principalmente per l'indicizzazione e la ricerca di letteratura biomedica in PubMed: facilita il recupero degli articoli in base al loro contenuto.

- **RADLEX** è un'ontologia specificatamente progettata per la radiologia e le immagini mediche. Standardizza la terminologia utilizzata nei referti e negli studi radiologici.

4 Valutazione del sistema

4.1 Criticità

UMLS è una risorsa importante nell'ambito della ricerca biomedica e del Natural Language Processing. Tuttavia, date le sue dimensioni e la sua complessità, presenta dei problemi e delle limitazioni.

UMLS integra centinaia di vocabolari, necessariamente non consistenti l'uno con l'altro, causando possibili **inconsistenze**. Gli errori riscontrati includono **ambiguità** e **ridondanza**, *cicli di relazione gerarchica* (un concetto è sia antenato che discendente di un altro), a*ntenati mancanti* (i tipi semantici dei concetti genitore e figlio non sono correlati) e **inversione semantica** (la relazione figlio/genitore dei tipi semantici non è coerente con i concetti) [4].

Inoltre, possiamo identificare le seguenti criticità:

- **Coverage & Completeness**: UMLS potrebbe non comprendere tutti i vocabolari e termini medici (ad esempio, quelli emergenti), determinando delle lacune e limitando l'utilità in determinati domini specializzati [5] [6].
- **Sensitivity & Specificity:** la mappatura dei termini potrebbe non sempre catturare le sottili differenze nel significato date, ad esempio, dal contesto, portando a problemi di sensibilità (falsi negativi) o specificità (falsi positivi) durante la ricerca di termini o concetti [7].
- **Updates & Maintenance**: la conoscenza biomedica è in continua evoluzione per cui UMLS potrebbe non sempre tenere il passo con le terminologie, i codici o i concetti medici più recenti. Gli aggiornamenti e la manutenzione possono essere dispendiosi e richiedere molto tempo.
- **Semantic Ambiguity**: UMLS mira ad armonizzare diverse terminologie mediche, ma non sempre riesce a risolvere l'ambiguità semantica o a catturare le diverse sfumature dei concetti medici, portando potenzialmente a interpretazioni errate.
- **Complexity & Learning Curve**: UMLS è un sistema complesso con una curva di apprendimento ripida. Gli utenti, soprattutto quelli inesperti nel campo dell'informatica biomedica, potrebbero trovare difficile navigare e utilizzare questa risorsa [8].
- **Licensing & Accessibility**: l'accesso ai dati di UMLS potrebbe essere soggetto a restrizioni e licenze, dunque a termini e condizioni che potrebbero rappresentare un ostacolo per ricercatori e sviluppatori che desiderano utilizzare questo sistema nei loro progetti.
- **Data Quality**: la qualità dei dati all'interno di UMLS può variare, poiché si basa su input provenienti da varie fonti. Imprecisioni, incoerenze o informazioni obsolete nei dati di origine possono propagarsi a UMLS.

- **Terminology Mapping**: mappare i concetti da una terminologia all'altra può essere un compito complesso, soggetto ad errori o imprecisioni che possono influenzare l'interoperabilità dei dati e i risultati della ricerca.
- **Concept Hierarchies**: UMLS utilizza una gerarchia complessa di concetti, che potrebbe non essere perfettamente in linea con il modo in cui gli operatori sanitari o i ricercatori concettualizzano la conoscenza medica. Ciò può portare a difficoltà nella navigazione e nell'interrogazione del sistema.
- **Limited Multilingual Support**: UMLS si concentra principalmente sulla lingua inglese e può rappresentare una limitazione per gli utenti regioni non anglofone [6].

Gli **audit manuali** possono richiedere molto tempo e denaro: i ricercatori hanno studiato molteplici metodi (manuali, automatizzati ed euristici) per affrontare il problema [9] [10] [11].

4.2 Considerazioni finali

UMLS è un sistema estremamente importante e potente nel campo della medicina e della sanità. Si tratta di una risorsa fondamentale per l'organizzazione e l'integrazione delle informazioni medico-sanitarie, consentendo interoperabilità fra i vari sistemi ed applicazioni, e favorendo la comunicazione fra i vari professionisti del settore. Allo stesso modo, promuove la standardizzazione, evitando ambiguità nei dati medici e facilitando l'interpretazione delle informazioni. UMLS rappresenta uno strumento a supporto della ricerca: gli studiosi lo utilizzano per identificare risorse e dati rilevanti per le loro ricerche. Questo sistema agevola la scoperta di nuove informazioni e l'avanzamento della conoscenza medica. É utilizzato anche nell'assistenza sanitaria per migliorare la documentazione clinica, la codifica delle diagnosi e dei procedimenti, e per aiutare i professionisti a prendere decisioni attraverso l'accesso a dati aggiornati e a risorse informative.

Nonostante i suoi benefici, UMLS affronta alcune sfide, tra cui la necessità di mantenere i dati aggiornati, la gestione delle ambiguità nei concetti medici e la comprensione delle sfumature nei diversi vocabolari. Oltre a ciò, con l'aumento della condivisione di dati sanitari, la privacy e la sicurezza delle informazioni diventano fondamentali. Tuttavia, il sistema è soggetto a costanti miglioramenti e aggiornamenti per affrontare queste criticità.

References

[1] Olivier Bodenreider. "The Unified Medical Language System (UMLS): Integrating Biomedical Terminology". In: *Nucleic acids research* 32 (Feb. 2004), pp. D267–70. DOI: 10.1093/nar/gkh061.

[2] Anantha Bangalore et al. "The UMLS Knowledge Source Server: An object model for delivering UMLS data". In: *AMIA Annual Symposium Proceedings*. Vol. 2003. American Medical Informatics Association. 2003, p. 51.

[3] Liz Amos et al. "UMLS users and uses: a current overview". In: *Journal of the American Medical Informatics Association* 27.10 (July 2020), pp. 1606–1611. ISSN: 1527-974X. DOI: 10.1093/jamia/ocaa084. eprint: https://academic.oup.com/jamia/article-pdf/27/10/1606/34152940/ocaa084.pdf. URL: https://doi.org/10.1093/jamia/ocaa084.

[4] James Geller et al. "Comparing Inconsistent Relationship Configurations Indicating UMLS Errors". In: *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium* 2009 (Nov. 2009), pp. 193–7.

[5] Carol Friedman. "The UMLS coverage of clinical radiology". In: *Proceedings / the ... Annual Symposium on Computer Application [sic] in Medical Care. Symposium on Computer Applications in Medical Care* (Feb. 1992), pp. 309–13.

[6] Olivier Bodenreider et al. "Evaluation of the Unified Medical Language System as a Medical Knowledge Source". In: *Journal of the American Medical Informatics Association : JAMIA* 5 (Jan. 1998), pp. 76–87. DOI: 10.1136/jamia.1998.0050076.

[7] Prakash Nadkarni, Roland Chen, and Cynthia Brandt. "UMLS Concept Indexing for Production Databases: A Feasibility Study". In: *Journal of the American Medical Informatics Association* 8.1 (Jan. 2001), pp. 80–91. ISSN: 1067-5027. DOI: 10.1136/jamia.2001.0080080. eprint: https://academic.oup.com/jamia/article-pdf/8/1/80/2152103/8-1-80.pdf. URL: https://doi.org/10.1136/jamia.2001.0080080.

[8] Srinivasan Fung Hole. "Who is using the UMLS and how - insights from the UMLS user annual reports". In: *AMIA Annual Symposium Proceedings* (2006), pp. 274–278. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1839427/.

[9] Xinxin Zhu et al. "A review of auditing methods applied to the content of controlled biomedical terminologies". In: *Journal of Biomedical Informatics* 42.3 (2009). Auditing of Terminologies, pp. 413–425. ISSN: 1532-0464. DOI: https://doi.org/10.1016/j.jbi.2009.03.003. URL: https://www.sciencedirect.com/science/article/pii/S1532046409000434.

[10] Ling Zheng et al. "A review of auditing techniques for the Unified Medical Language System". In: *Journal of the American Medical Informatics Association* 27.10 (Aug. 2020), pp. 1625–1638. ISSN: 1527-974X. DOI: 10.1093/jamia/ocaa108. eprint: https://academic.oup.com/jamia/article-pdf/27/10/1625/34153270/ocaa108.pdf. URL: https://doi.org/10.1093/jamia/ocaa108.

[11] Fengbo Zheng et al. “A transformation-based method for auditing the IS-A hierarchy of biomedical terminologies in the Unified Medical Language System”. In: *Journal of the American Medical Informatics Association* 27.10 (Oct. 2020), pp. 1568–1575. ISSN: 1527-974X. DOI: 10.1093/jamia/ocaa123. eprint: https://academic.oup.com/jamia/article-pdf/27/10/1568/34153215/ocaa123.pdf. URL: https://doi.org/10.1093/jamia/ocaa123.