Symposium International CITADEL 2023

Centre d'intégration et d'analyse en données médicales

Montréal, Québec Canada

November 7, 2023

Aligning biomedical terminologies

From lexical models to supervised learning

Olivier Bodenreider, MD, PhD

Senior Scientist

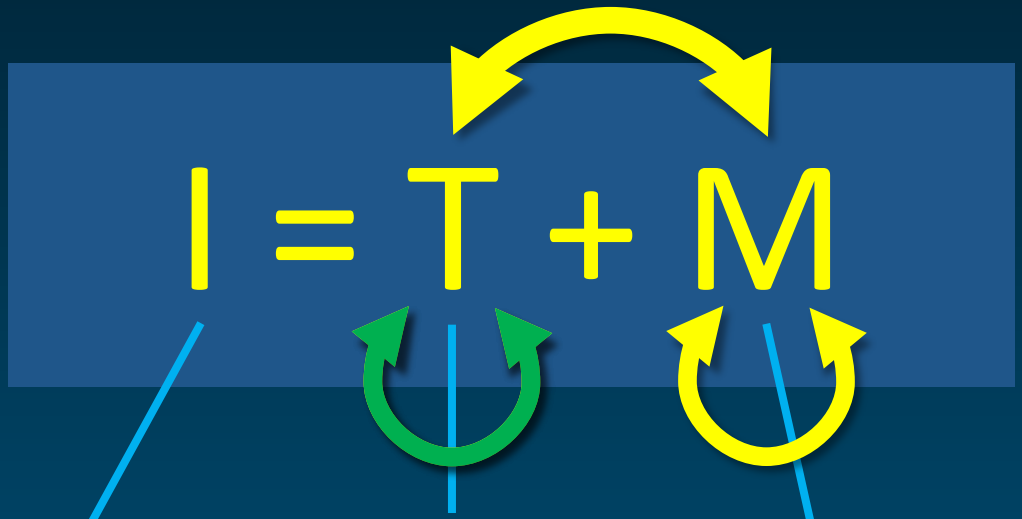
National Library of Medicine

Lister Hill National Center for Biomedical Communications

Disclaimer

The views and opinions expressed do not necessarily state or reflect those of the U.S. Government, and they may not be used for advertising or product endorsement purposes.

Fundamental theorem of clinical data interoprability



Interoperability

Terminologies

Vocabularies
Code sets
Terminology Standards

Information models

Content structure standards
Data models

Outline

- Introduction to the UMLS Metathesaurus
- Lexical model of synonymy
- Supervised machine learning for synonymy prediction

Introduction to the UMLS Metathesaurus

What does UMLS stand for?

- Unified
- Medical
- Language
- System

http://www.nlm.nih.gov/research/umls/

Motivation

- Started in 1986
- National Library of Medicine

«[…] the UMLS project is an effort to overcome two significant barriers to effective retrieval of machine-readable information.

- The first is the variety of ways the same concepts are expressed in different machine-readable sources and by different people.
- The second is the distribution of useful information among many disparate databases and systems.»

UMLS Metathesaurus (2023AA)

- 166 families of source vocabularies
  - Not counting translations
- 27 languages
- Broad coverage of biomedicine
  - 11.8M names (normalized)
  - ~3.3M concepts
  - >10M relations
- Common presentation

UMLS Metathesaurus Example

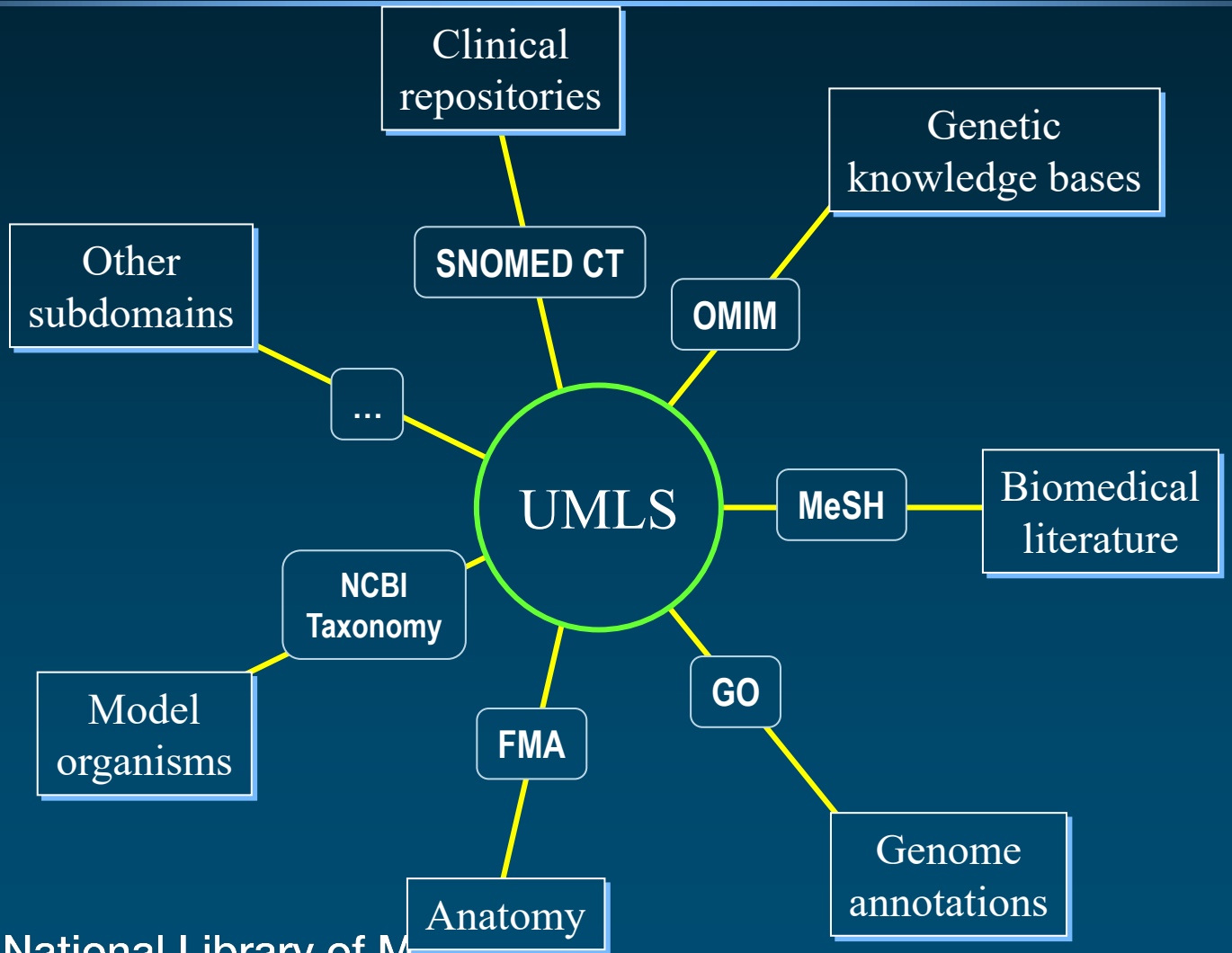
- Synonymous terms clustered into a concept
- Preferred term
- Unique identifier (CUI)

Addison Disease	MeSH	D000224
Primary hypoadrenalism	MedDRA	10036696
Primary adrenocortical insufficiency	ICD-10	E27.1
Addison's disease (disorder)	SNOMED CT	363732003

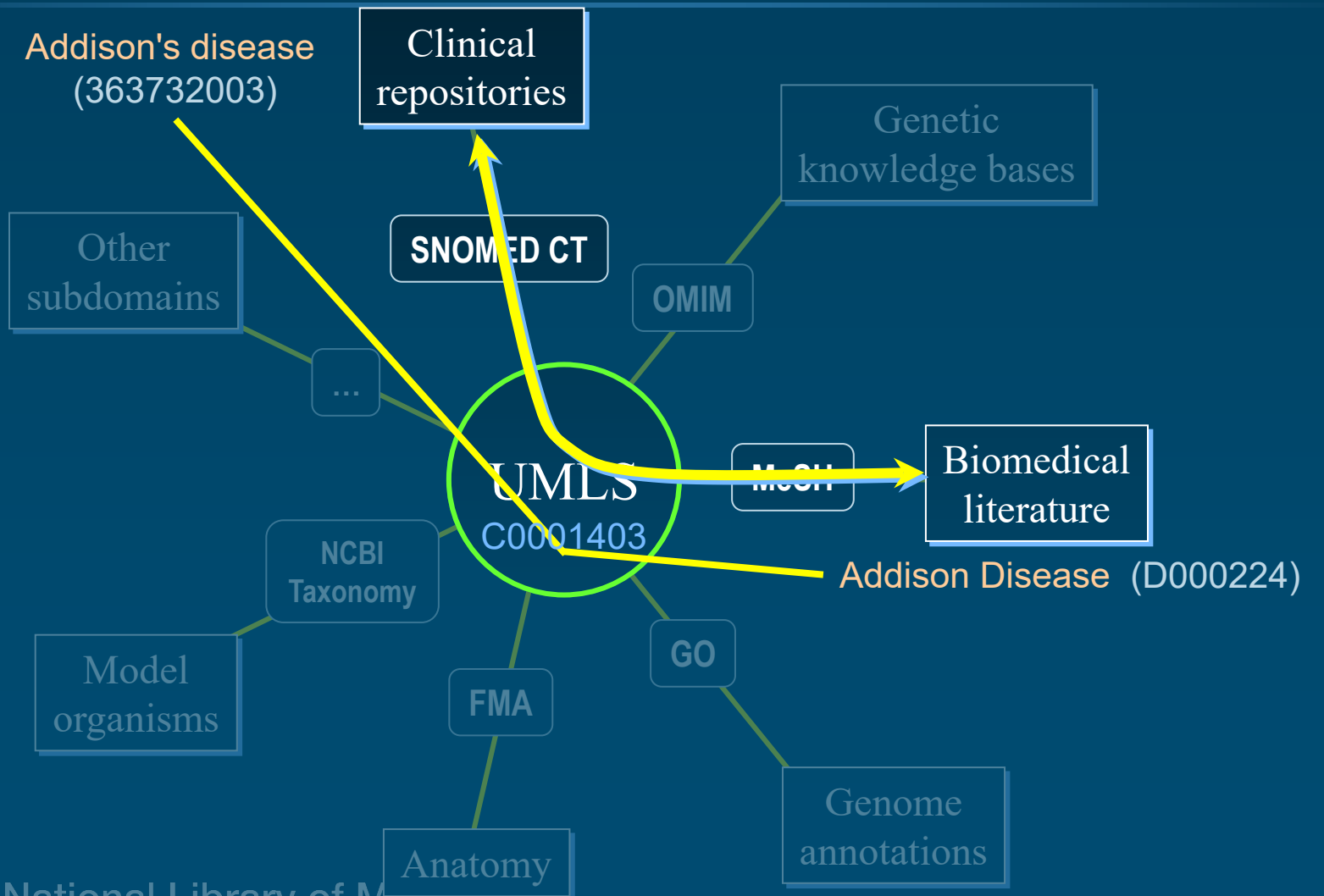
C0001403

Addison's disease

Integrating subdomains

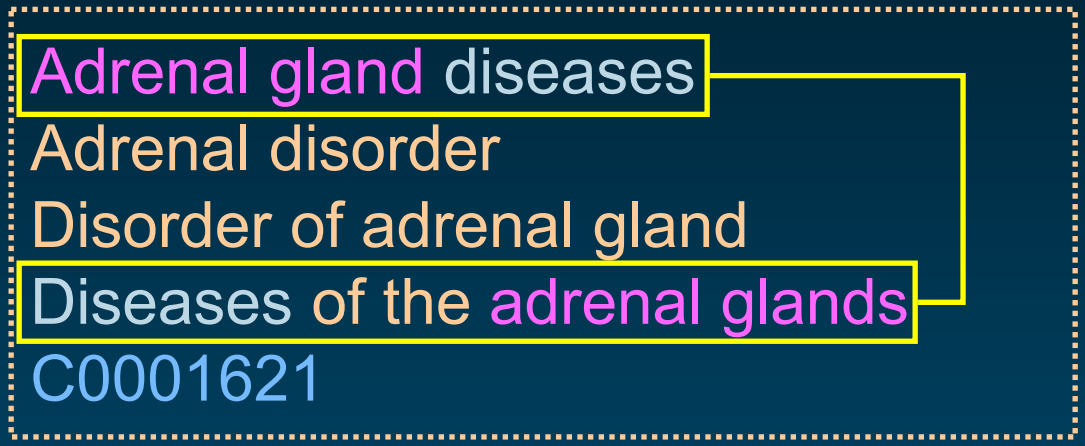


Trans-namespace integration



Lexical model of synonymy

From lexical features to synonymy



SPECIALIST Lexicon

- Content
  - English lexicon
  - Many words from the biomedical domain
- Over 500,000 lexical items
- Word properties
  - morphology
  - orthography
  - syntax
- Used by the lexical tools

Morphology

- Inflection
  - noun — nucleus, nuclei
  - verb — cauterize, cauterizes, cauterized, cauterizing
  - adjective — red, redder, reddest
- Derivation
  - verb ⟺ noun — cauterize -- cauterization
  - adjective ⟺ noun — red -- redness

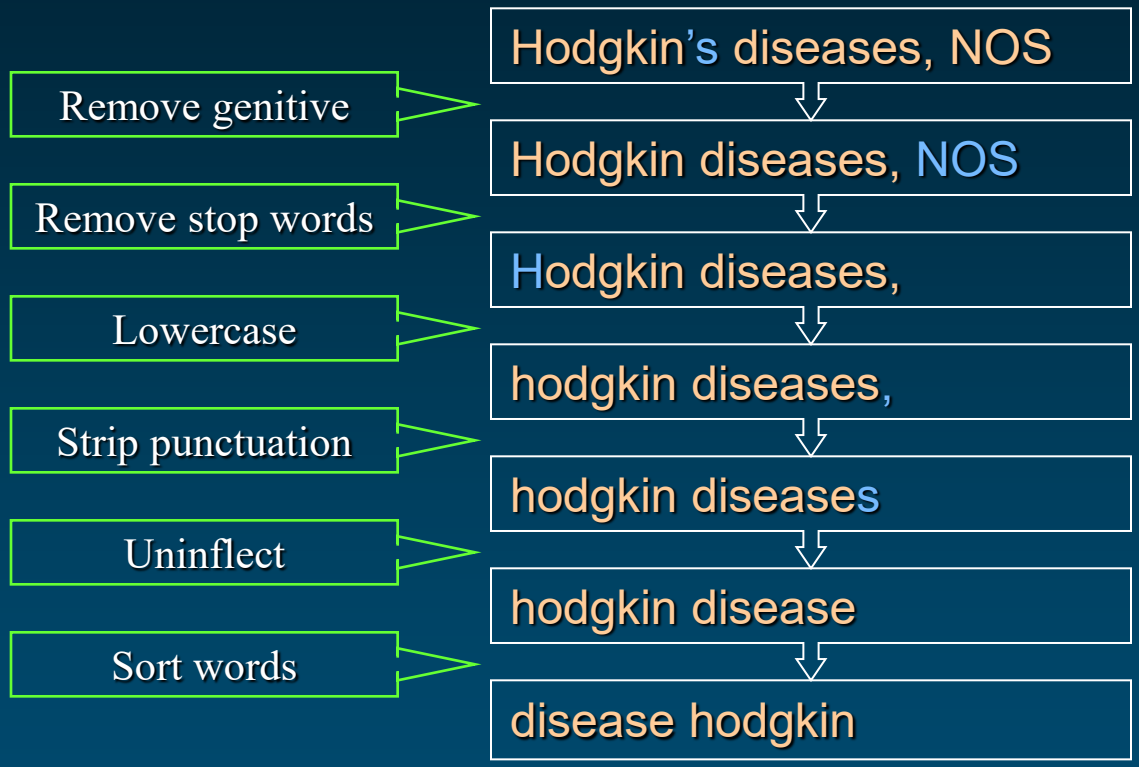
Orthography

- Spelling variants
  - oe/e — oesophagus - esophagus
  - ae/e — anaemia - anemia
  - ise/ize — cauterise - cauterize
  - genitive mark — Addison's disease
Addison disease
Addisons disease

Lexical tools

- To manage lexical variation in biomedical terminologies
- Major tools
  - Normalization
  - Indexes
  - Lexical Variant Generation program (lvg)
- Based on the SPECIALIST Lexicon
- Used by noun phrase extractors, search engines

Normalization



Normalization: Example

Hodgkin Disease
HODGKINS DISEASE
Hodgkin's Disease
Disease, Hodgkin's
Hodgkin's, disease
HODGKIN'S DISEASE
Hodgkin's disease
Hodgkins Disease
Hodgkin's disease NOS
Hodgkin's disease, NOS
Disease, Hodgkins
Diseases, Hodgkins
Hodgkins Diseases
Hodgkins disease
hodgkin's disease
Disease, Hodgkin

Normalization Applications

- Model for lexical resemblance
- Help find lexical variants for a term
  - Terms that normalize the same usually share the same LUI
- Help find candidates to synonymy among terms
- Help map input terms to UMLS concepts

Metathesaurus building process

- All terms from source vocabularies are processed
  - Terms that have the same normalized for are candidates for synonymy
    - Unless they bear different semantics
  - Synonymy indicated by source vocabularies tends to be preserved
- *All candidates (from normalization or sources) are reviewed manually*
  - Labor-intensive and error-prone
- Synonyms are assigned the same CUI

Example

String	Source	SCUI	AUI	LUI
Headache	MSH	M0009824	A0066000	L0018681
Headaches	MSH	M0009824	A0066008	L0018681
Cranial Pains	MSH	M0009824	A1641924	L1406212
Cephalodynia	MSH	M0009824	A26628141	L0380797
Cephalodynia	SNOMEDCT_US	25064002	A2957278	L0380797
Headache (finding)	SNOMEDCT_US	25064002	A3487586	L3063036

Supervised machine learning for synonymy prediction

Intuition

- Large collection of synonymy assertions in Metathesaurus can be used for supervised learning
  - Positive examples: terms from the same concept
  - Negative examples: terms from different concepts
- Possible features
  - Lexical (words in a term)
  - Semantic (semantics of the source)
  - Relations to other terms

Synonymy function

Addison Disease
Primary hypoadrenalism
Primary adrenocortical insufficiency
Addison's disease (disorder)
[…]

C0001403

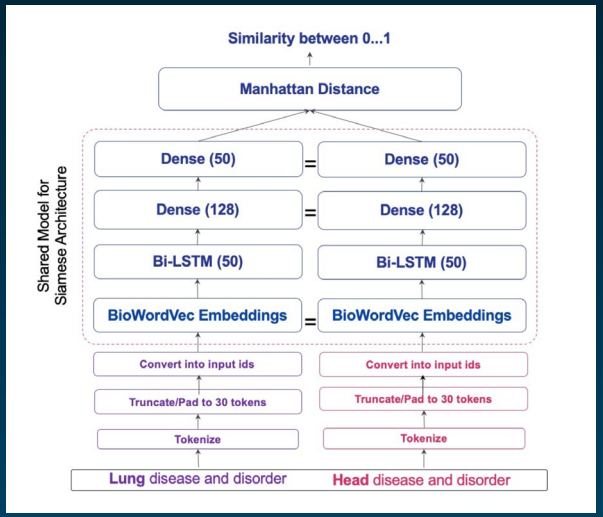
Hodgkin Disease
Granuloma, Malignant
Hodgkin lymphoma
Malignant lymphoma, Hodgkin's
[…]

C0019829

$$syn(\text{“Addison Disease”}, \text{“Primary hypoadrenalism”}) = 1$$
$$syn(\text{“Addison Disease”}, \text{“Hodgkin Disease”}) = 0$$

Early experiments Pairwise similarity

- Types of embeddings
  - Word vectors for representing terms using BioWordVec (2021)
  - Knowledge Graph Embeddings for representing the context (2022)
- Siamese LSTM network



- Results: Best model
  - F1=0.765 (baseline: lexical similarity + source synonymy)
  - F1=0.906 (words)
  - F1=0.935 (context)

Recent experiments Vocabulary insertion

- Initial approach does not translate well to vocabulary insertion (inserting new terms into the Metathesaurus)
- Rethinking the approach as an entity linking problem
  - Given a new term, find the concept with which it should be associated
  - *Or indicate if there is no such concept*

Recent experiments Vocabulary insertion

	Accuracy
Rule Based Approximation (RBA)	70.1
LexLM	63.2
PubMedBERT	68.4
SapBERT	77.4
RBA + LexLM	80.4
RBA + PubMedBERT	83.7
RBA + SapBERT	90.7
Re-Ranker (PubMedBERT)	85.5
+ RBA Signal	93.2

Lexical similarity + source synonymy

Existing models

Existing models enriched with Lexical similarity + source synonymy

New models (re-ranking)

Discussion

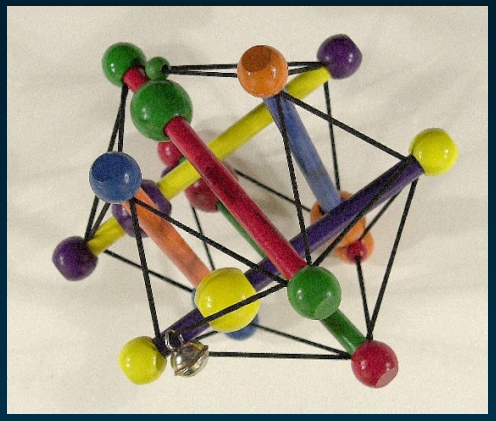
- Performance conserved
  - Across versions (UMLS insertion sets)
  - Across categories (UMLS semantic groups)
- Importance of extending entity linking with "null injection"
- The deep learning models improve when augmented with basic information (lexical similarity and source synonymy)

Overall summary

- The UMLS Metathesaurus is a biomedical terminology integration system
- Metathesaurus construction has relied on a lexical model for synonymy and human review
- Supervised machine learning approaches to predicting synonymy have shown promising results

References

- UMLS overview
  - Bodenreider O. (2004). The Unified Medical Language System (UMLS): Integrating biomedical terminology. *Nucleic Acids Research*; D267-D270. PMID: 14681409.
- Lexical approach
  - McCray AT, Srinivasan S, Browne AC. Lexical methods for managing variation in biomedical terminologies. *Proc Annu Symp Comput Appl Med Care*. 1994:235-9. PMID: 7949926.
- Supervised learning approach
  - Nguyen V, Yip HY and Bodenreider O. Biomedical vocabulary alignment at scale in the UMLS Metathesaurus. *Proceedings of the Web Conference 2021 (WWW'21)*; 2672-2683. PMID: 34514472.
  - Nguyen V, Yip HY, Bajaj G, Wijesiriwardene T, Javangula V, Parthasarathy S, Sheth A, Bodenreider O. Context-Enriched Learning Models for Aligning Biomedical Vocabularies at Scale in the UMLS Metathesaurus. *Proc Int World Wide Web Conf. 2022 (WWW'22)*. PMID: 36108322.
  - Jiménez Gutiérrez B, Mao Y, Nguyen V, Fung KW, Su Y, Bodenreider O. Solving the Right Problem is Key for Translational NLP: A Case Study in UMLS Vocabulary Insertion. *Findings of EMNLP 2023* (in press).



Medical Ontology Research

Contact: olivier@nlm.nih.gov
Web: mor.nlm.nih.gov

Olivier Bodenreider

National Library of Medicine
Lister Hill National Center for Biomedical Communications