

BioCreative V Workshop  
Corvallis, OR  
August 1, 2016

## Annotations for biomedical research and healthcare – Bridging the gap



*Olivier Bodenreider*

Lister Hill National Center  
for Biomedical Communications  
Bethesda, Maryland - USA



U.S. National Library of Medicine









# 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.



# Outline

- ◆ Annotations
- ◆ Ontology integration systems
- ◆ Challenges in automated annotation
  - Equivalent mapping
    - Lexical variation
  - Partial mapping
    - Lexical approaches
    - Logical approaches



# Annotations

# Annotations What

- ◆ Metadata added to a document
  - Entities
    - In reference to some target ontology
  - Relations (implicit or explicit)
    - Among entities
      - GO annotations
        - » HK1 involved in glycolytic process
    - Between the entity and the document
      - MeSH indexing of MEDLINE citations
        - » PMID:3207429 indexed with
          - » Glucose/metabolism
          - » Hexokinase/genetics\*
      - ICD10-CM codes in a patient record



# Annotations How

- ◆ Metadata added to a document
  - Assigned automatically
    - Automated NER and normalization
    - Automatic indexing
  - Assigned by humans
    - Manual annotation, curation
    - Human indexing (e.g., most of MEDLINE indexing)
    - Billing codes added to patient records
  - Derived from other annotations
    - Through mapping

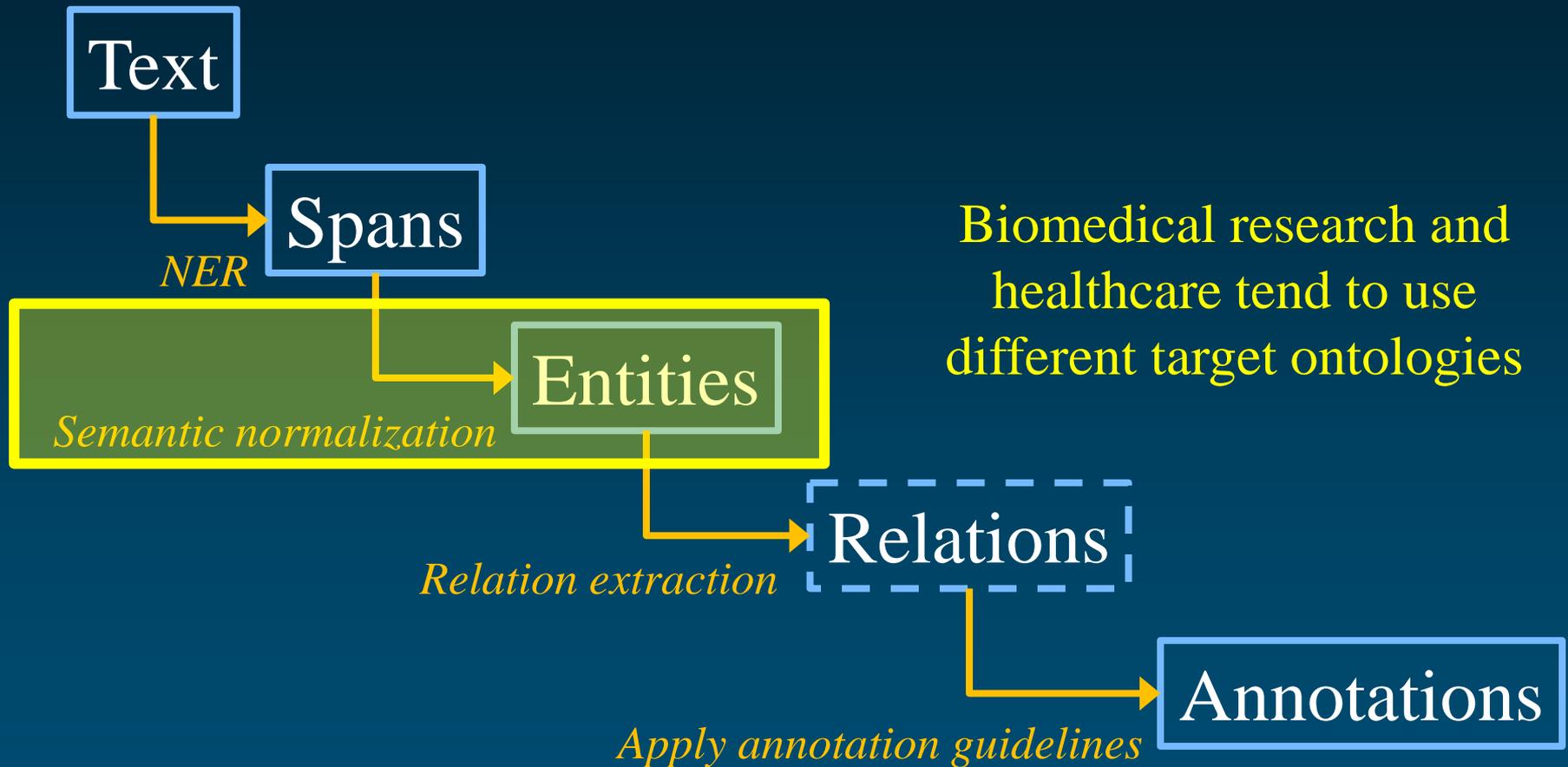


# Annotations Why

- ◆ Metadata added to a document
  - For a given purpose and according to specific guidelines
    - GO annotations
      - Extracting actionable, interoperable knowledge
        - » Curation rules
    - MeSH indexing
      - Supporting retrieval
        - » Indexing rules (checktags, “rule of 3”, etc.)
    - ICD10-CM coding
      - Supporting billing
        - » Billing rules (use most specific codes)



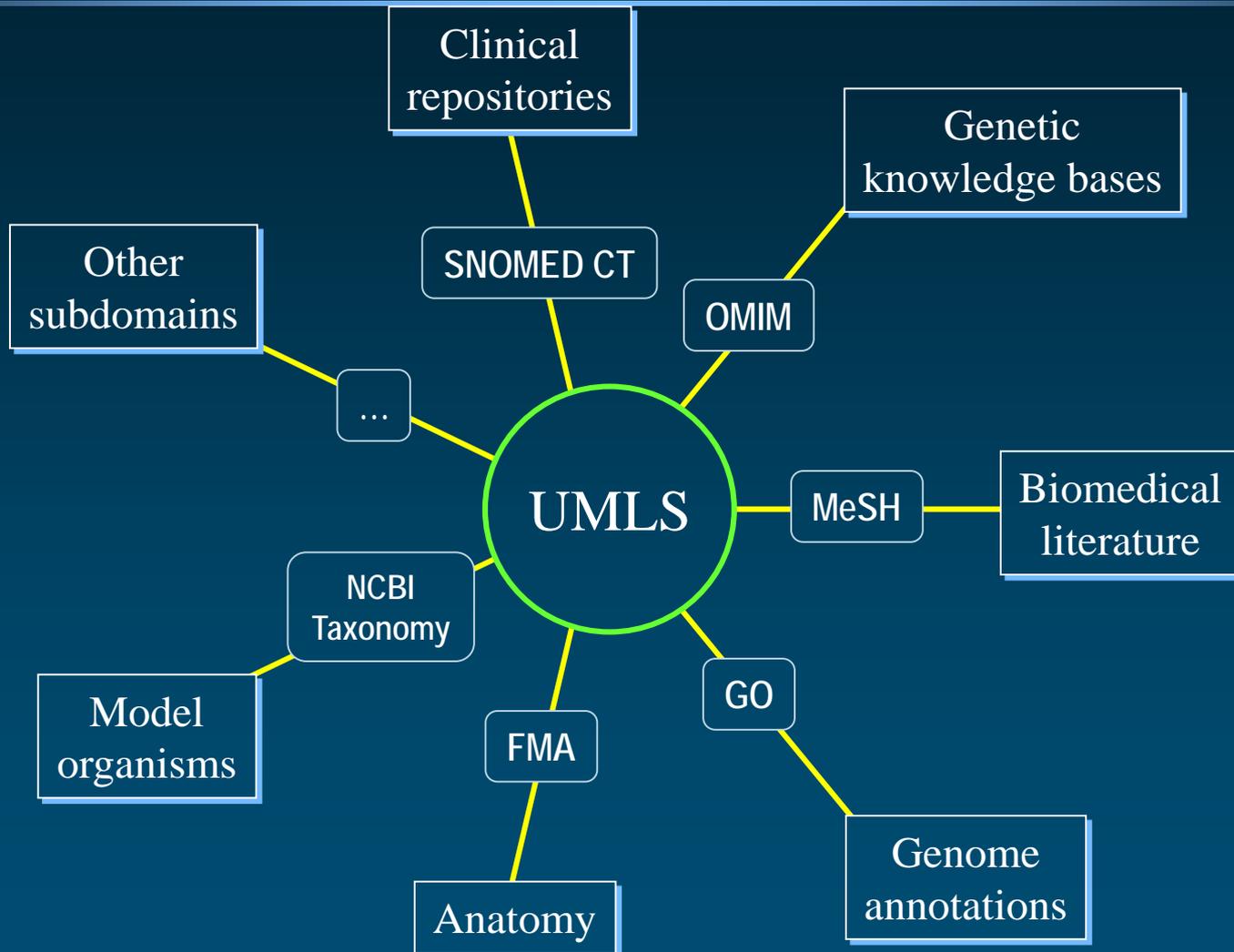
# Text annotation pipeline



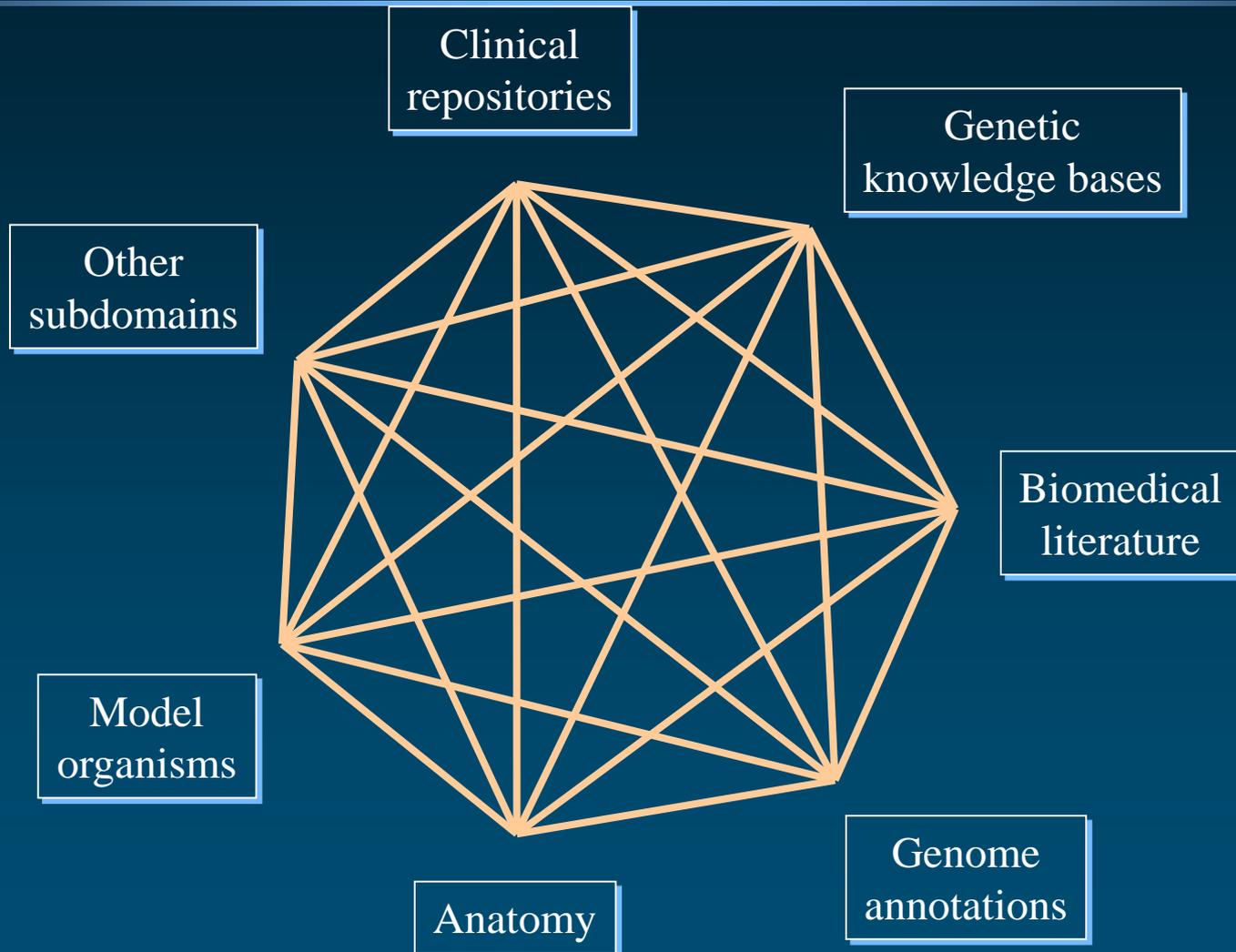
# Ontology integration systems

*Unified Medical Language System (UMLS)*

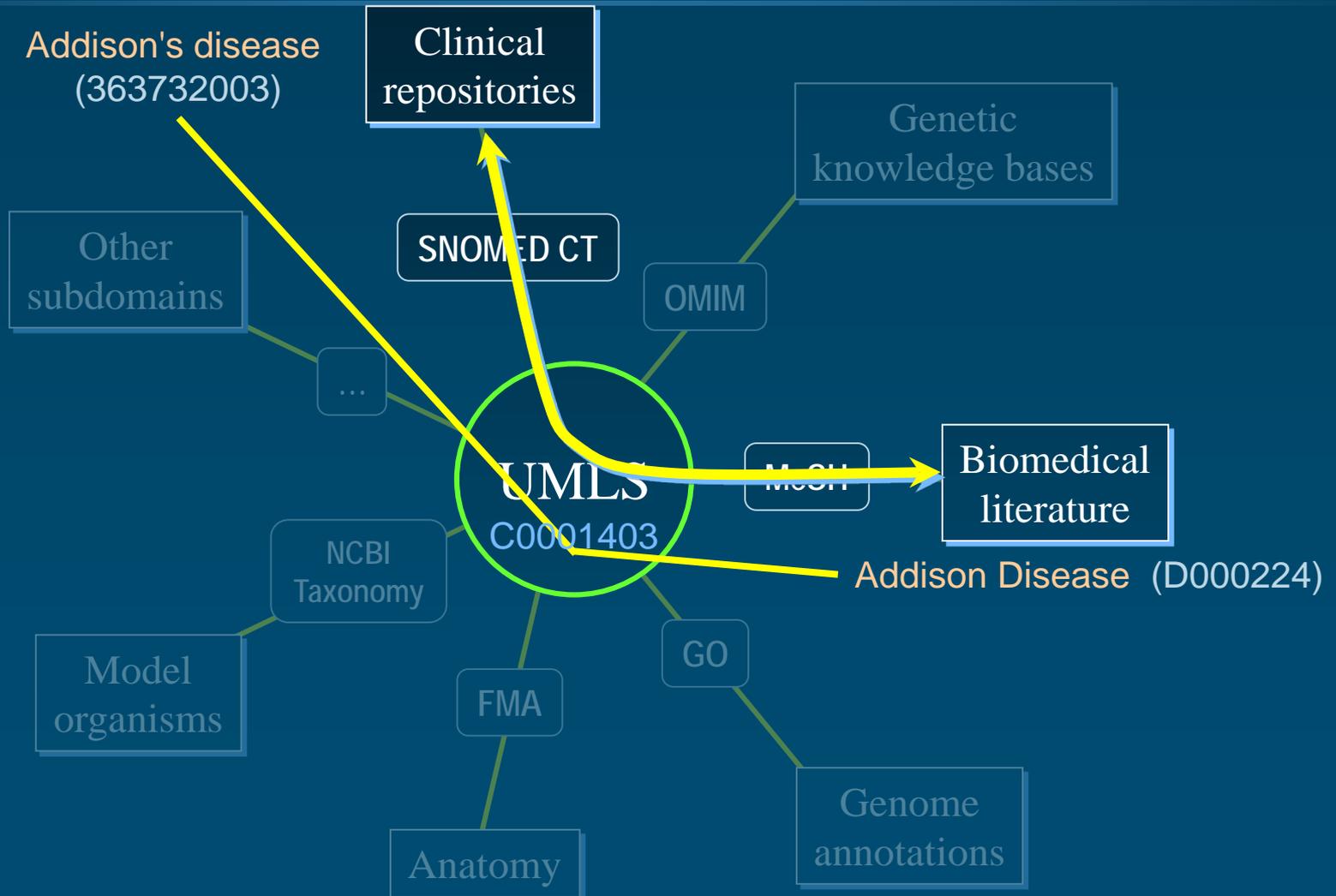
# Integrating subdomains



# Integrating subdomains



# Trans-namespace integration



# Source Vocabularies

(2016AA)

- ◆ 150 families of source vocabularies
  - Not counting translations
- ◆ Broad coverage of biomedicine
  - 9.9M names (normalized)
  - ~3.2M concepts
  - > 10M relations
- ◆ Mappings are curated by the Metathesaurus editors



# Source Vocabularies

- ◆ Major healthcare terminologies
  - SNOMED CT, LOINC, RxNorm
- ◆ Selected for extending coverage
  - HPO
    - 50% of HPO phenotypes were not represented in UMLS
- ◆ Selected for extending interoperability
  - ATC
  - DrugBank (upcoming)

# Challenges in automated annotation

*Lexical variation and equivalence mapping*

# Reference vs. interface terminologies

## ◆ Reference terminologies

- Focus on definitions and concept properties
- Usually contain a minimal number of terms
- No attempt to systematically represent lexical variants
- May not be sufficient for NER

## ◆ Interface terminologies

- Focus on the terms as they are used in practice
- Must represent synonyms, shortcuts, colloquialisms

*By integrating multiple sources, ontology integration systems generally better represent lexical variation*



# Addison's disease

Term	SNCT	I10	MeSH	HPO	MDR	OMIM	NCI	UMLS
Addison['s] disease	x	x	x	x	x	x	x	x
Primary adrenal deficiency	x							x
Primary adrenal insufficiency			x	x	x			x
Primary adrenocortical insufficiency	x	x	x					x
Primary adrenocortical failure				x		x		x
Chronic primary adrenal insufficiency							x	x

# Annotation/mapping strategy

- ◆ Resolve the mentions to the terminology integration system, not the target terminology
- ◆ Equivalent mappings to the target can be derived indirectly from synonyms in other terminologies
- ◆ Complement existing variation in terminologies with systematic variants
  - E.g., Roman/Arabic numerals (type II/type 2)

# Challenges in automated annotation

*Partial mapping vs. equivalent mapping*

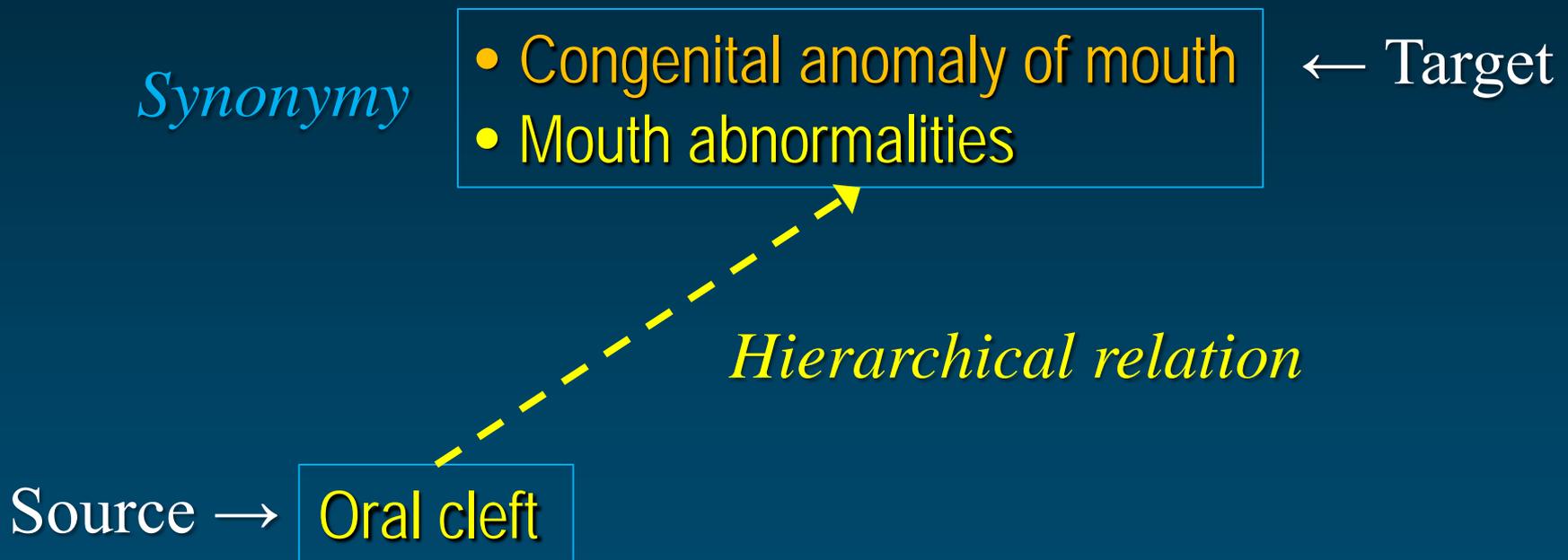
# Partial mappings

- ◆ Generally reflect a difference in granularity between
  - Source (more specific)
  - Target (more generic)
- ◆ Partial mappings may be sufficient depending on the use case
  - E.g., indexing (or abstraction, more generally)

# Partial lexical mappings

- ◆ **Bilateral renal atrophy** (HPO)
  - **Renal atrophy** (SNOMED CT)
- ◆ Approaches
  - Longest span from an HPO term found in SNOMED CT (agnostic of linguistic roles)
  - “Demodification” – specifically removing modifiers (linguistically-motivated)

# Partial logical mappings

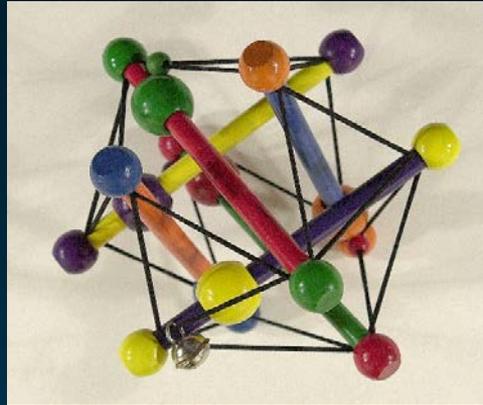


# Annotation/mapping strategy

- ◆ Consider partial mappings
- ◆ Logical partial mappings can be inferred by leveraging both
  - Synonymy relations
  - Hierarchical relations

# Summary

- ◆ Metadata are generated through annotation
  - But in reference to different target ontologies
  - Especially in healthcare and biomedical research
- ◆ Interoperability between datasets is key to knowledge discovery
- ◆ Mappings across annotations can be provided by ontology integration systems
  - Equivalent mappings whenever possible
  - Partial mappings otherwise



# Medical Ontology Research

Contact: [olivier@nlm.nih.gov](mailto:olivier@nlm.nih.gov)

Web: <http://mor.nlm.nih.gov>



*Olivier Bodenreider*

Lister Hill National Center  
for Biomedical Communications  
Bethesda, Maryland - USA