ICF Conference
Mayo Clinic - June 21, 2005

Mapping New Vocabularies to the UMLS

Experience with ICF

Olivier Bodenreider

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

Outline

- Terminology integration
  - *The Unified Medical Language System*
- Methods
  - Normalizing terms
  - Categorizing terms
  - Recording relations
  - Editing and auditing
- Experience with ICF

Terminology integration

The Unified Medical Language System

Motivation

- Started in 1986
- National Library of Medicine
- “Long-term R&D project”
- Complementary to IAIMS

(Integrated Academic Information Management Systems)

«[…] 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.»

Source Vocabularies (2005AA)

- 134 source vocabularies
  - 132 contributing concept names
- ~80 families of vocabularies
  - multiple translations (e.g., MeSH, ICPC, ICD-10)
  - variants (American-English equivalents, Australian extension/adaptation)
  - subsequent editions usually considered distinct families (ICD: 9-10; DSM: IIIR-IV)
- Broad coverage of biomedicine
- Common presentation

Biomedical terminologies

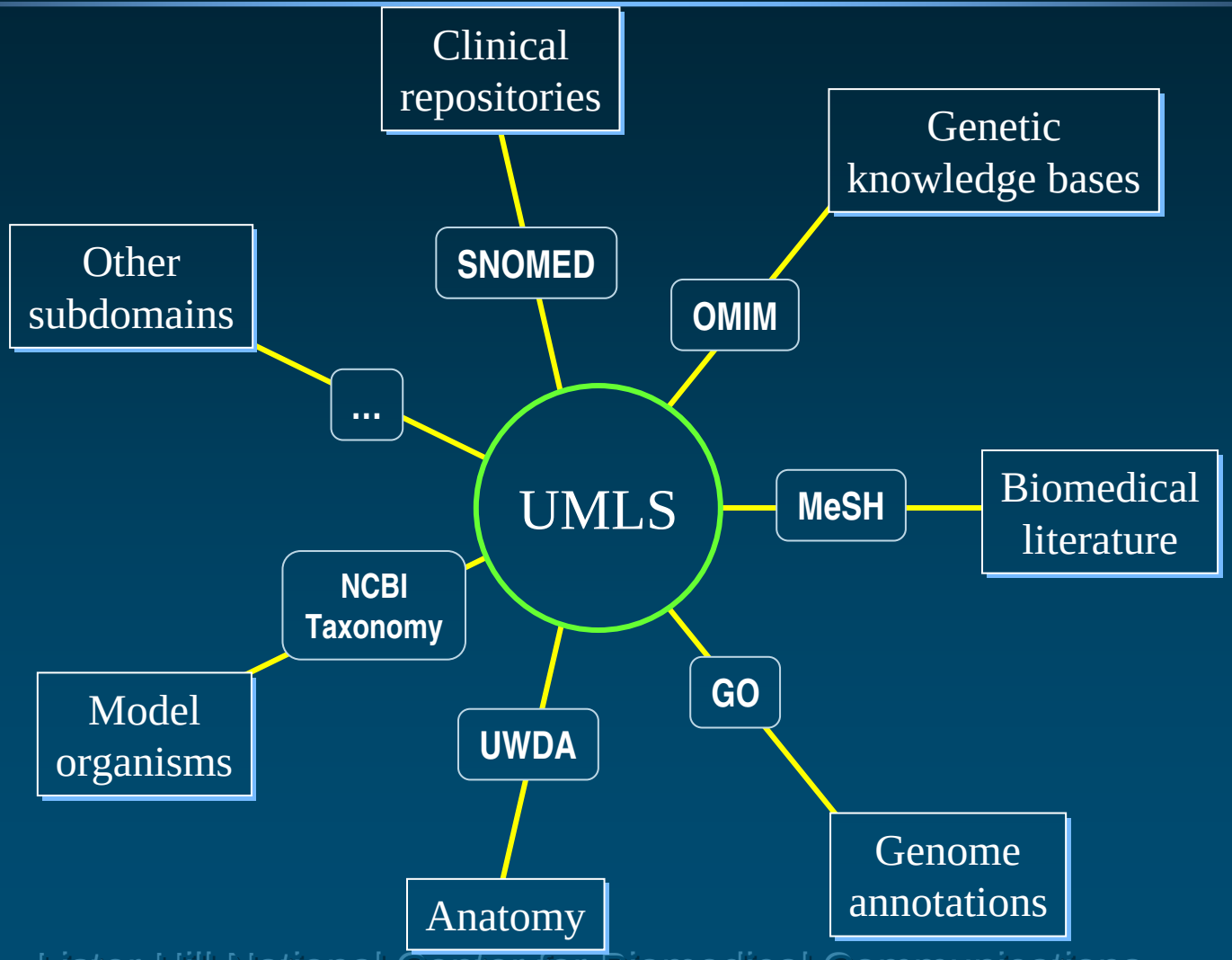
- General vocabularies
  - anatomy (UWDA, Neuronames)
  - drugs (RxNorm, First DataBank, Micromedex)
  - medical devices (UMD, SPN)
- Several perspectives
  - clinical terms (SNOMED CT)
  - information sciences (MeSH, CRISP)
  - administrative terminologies (ICD-9-CM, CPT-4)
  - data exchange terminologies (HL7, LOINC)

Biomedical terminologies (cont'd)

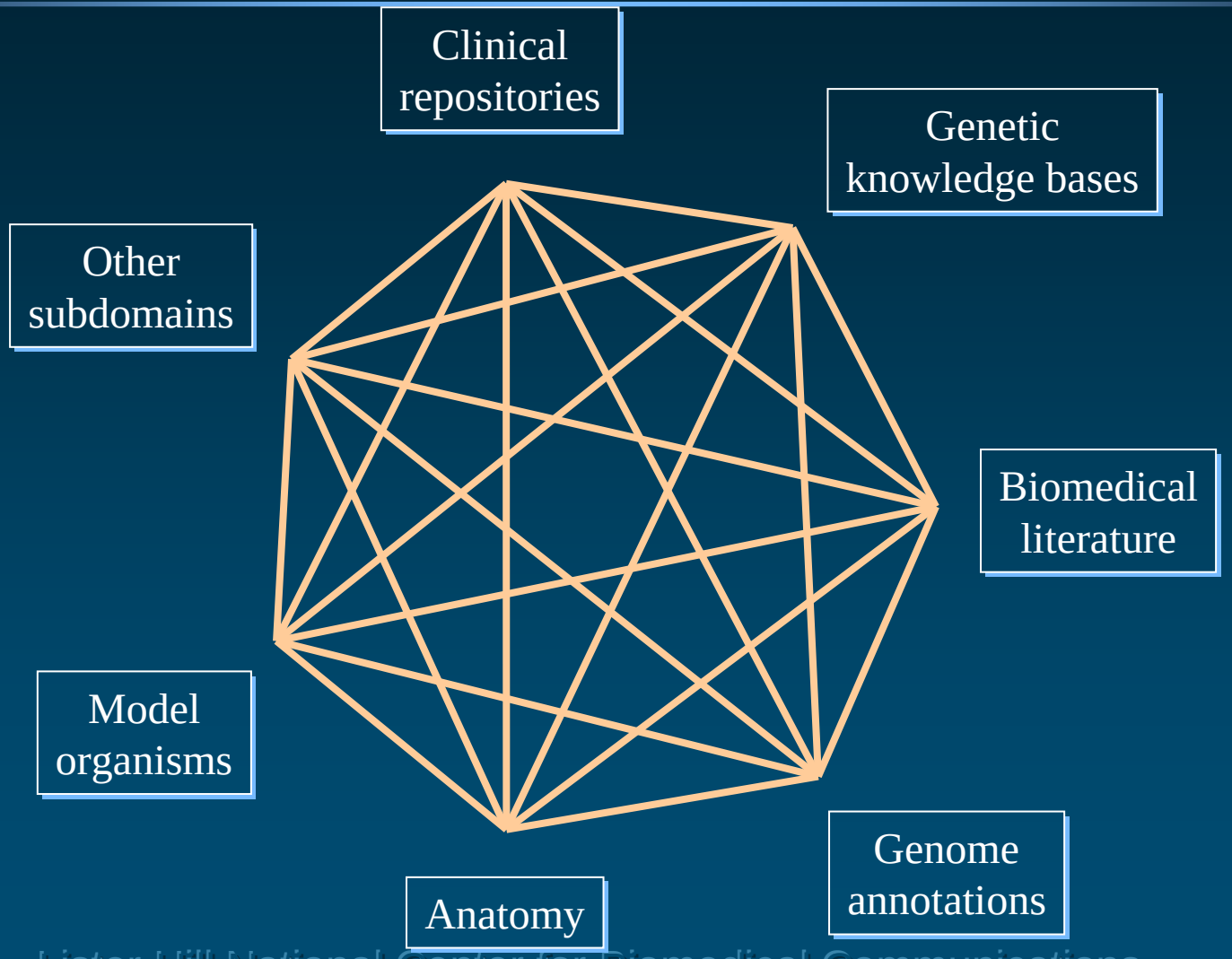
- Specialized vocabularies
  - nursing (NIC, NOC, NANDA, Omaha, PCDS)
  - dentistry (CDT)
  - psychiatry (DSM, APA)
  - adverse reactions (COSTART, WHO ART)
  - primary care (ICPC)
  - genomics (GO, OMIM, HUGO)
- Terminology of knowledge bases (AI/Rheum, DXplain, QMR)

The UMLS serves as a vehicle for the regulatory standards (HIPAA, CHI)

Integrating subdomains



Integrating subdomains



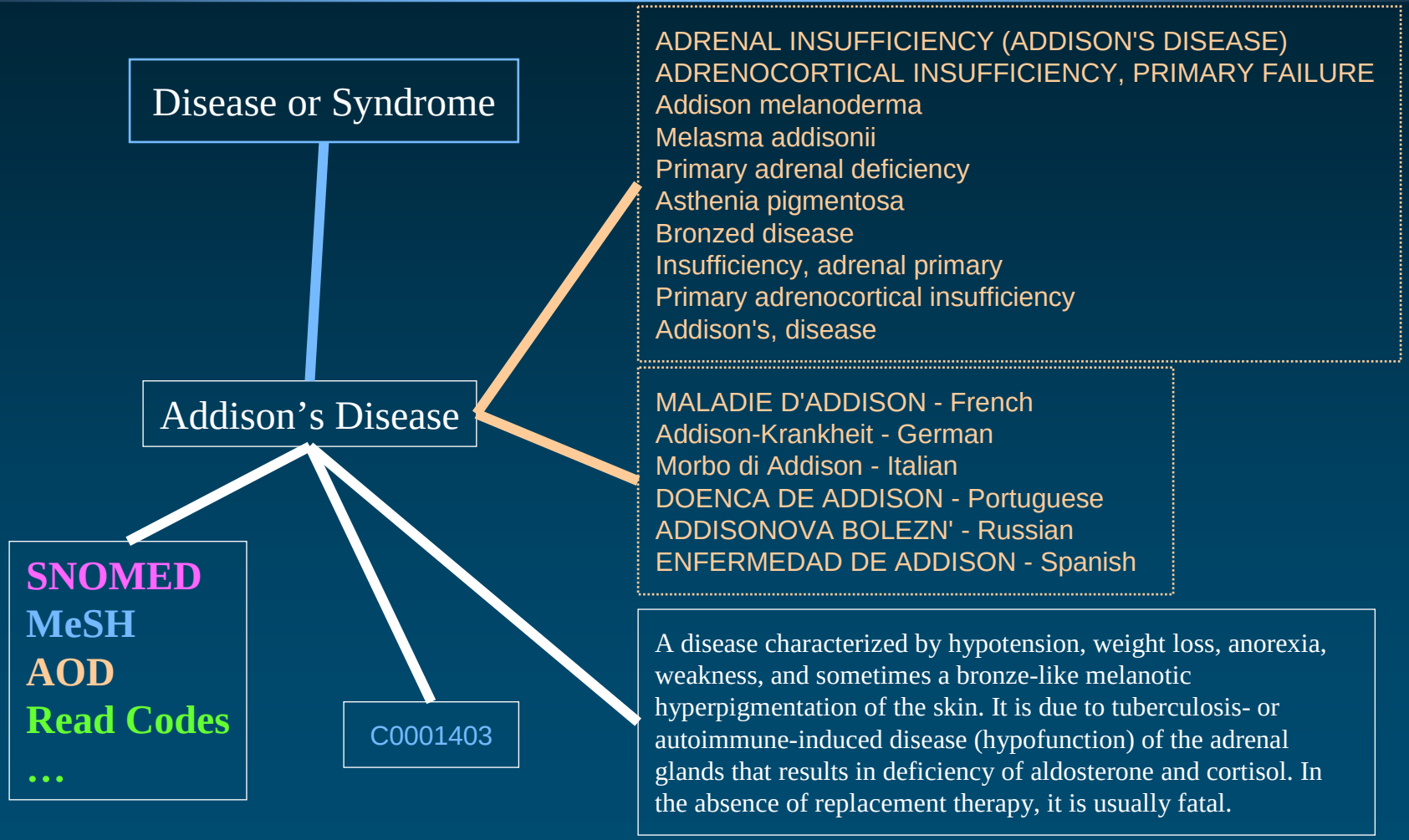
UMLS: 3 components

- Metathesaurus
  - Concepts
  - Inter-concept relationships
- Semantic Network
  - Semantic types
  - Semantic network relationships
- Lexical resources
  - SPECIALIST Lexicon
  - Lexical tools

Addison's Disease in medical vocabularies

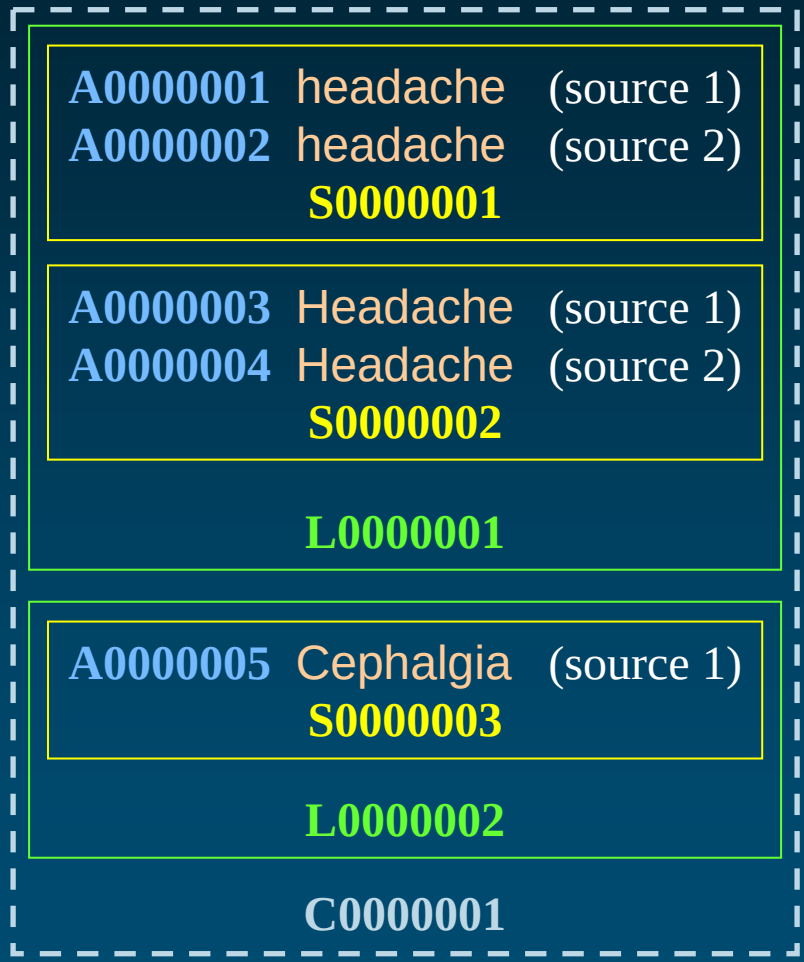
- Synonyms: different terms
  - Addisonian syndrome — eponym
  - Bronzed disease — symptoms
  - Addison melanoderma — symptoms
  - Asthenia pigmentosa — symptoms
  - Primary adrenal deficiency — clinical variants
  - Primary adrenal insufficiency — clinical variants
  - Primary adrenocortical insufficiency — clinical variants
  - Chronic adrenocortical insufficiency — clinical variants
- Contexts: different hierarchies

Addison's Disease: Concept

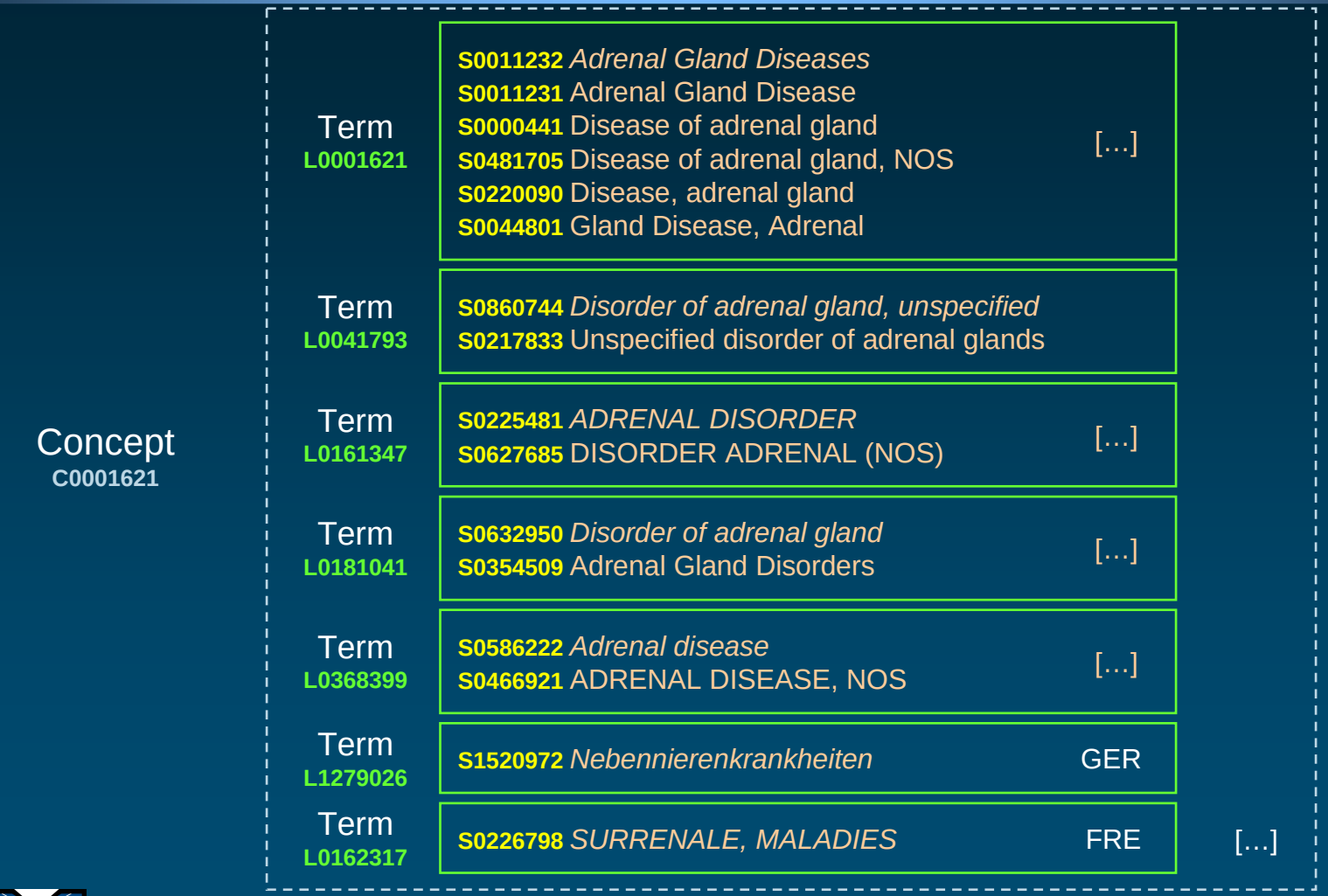


Metathesaurus Concepts (2005AA)

- Concept (~ 1.2M) CUI
  - Set of synonymous concept names
- Term (~ 4.2 M) LUI
  - Set of normalized names
- String (~ 4.7M) SUI
  - Distinct concept name
- Atom (~ 5.5M) AUI
  - Concept name in a given source

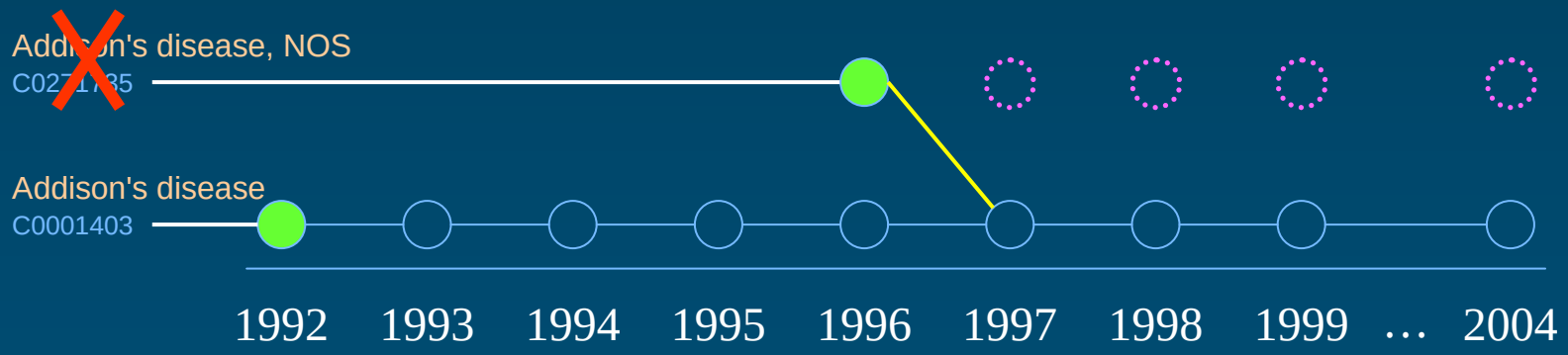


Cluster of synonymous terms



Metathesaurus Evolution over time

- Concepts never die (in principle)
  - CUIs are permanent identifiers
- What happens when they do die (in reality)?
  - Concepts can merge or split
  - Resulting in new concepts and deletions

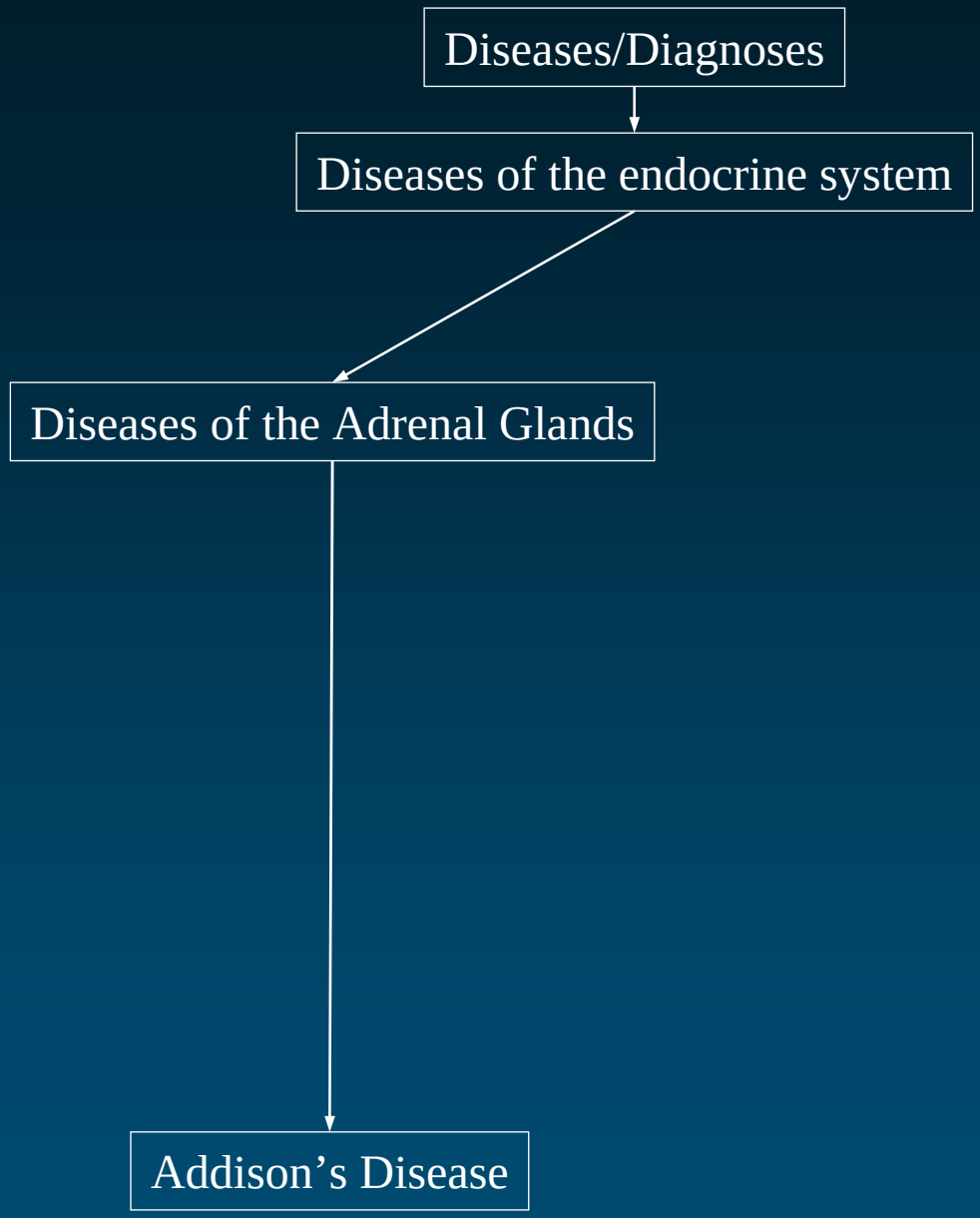


Metathesaurus Relationships

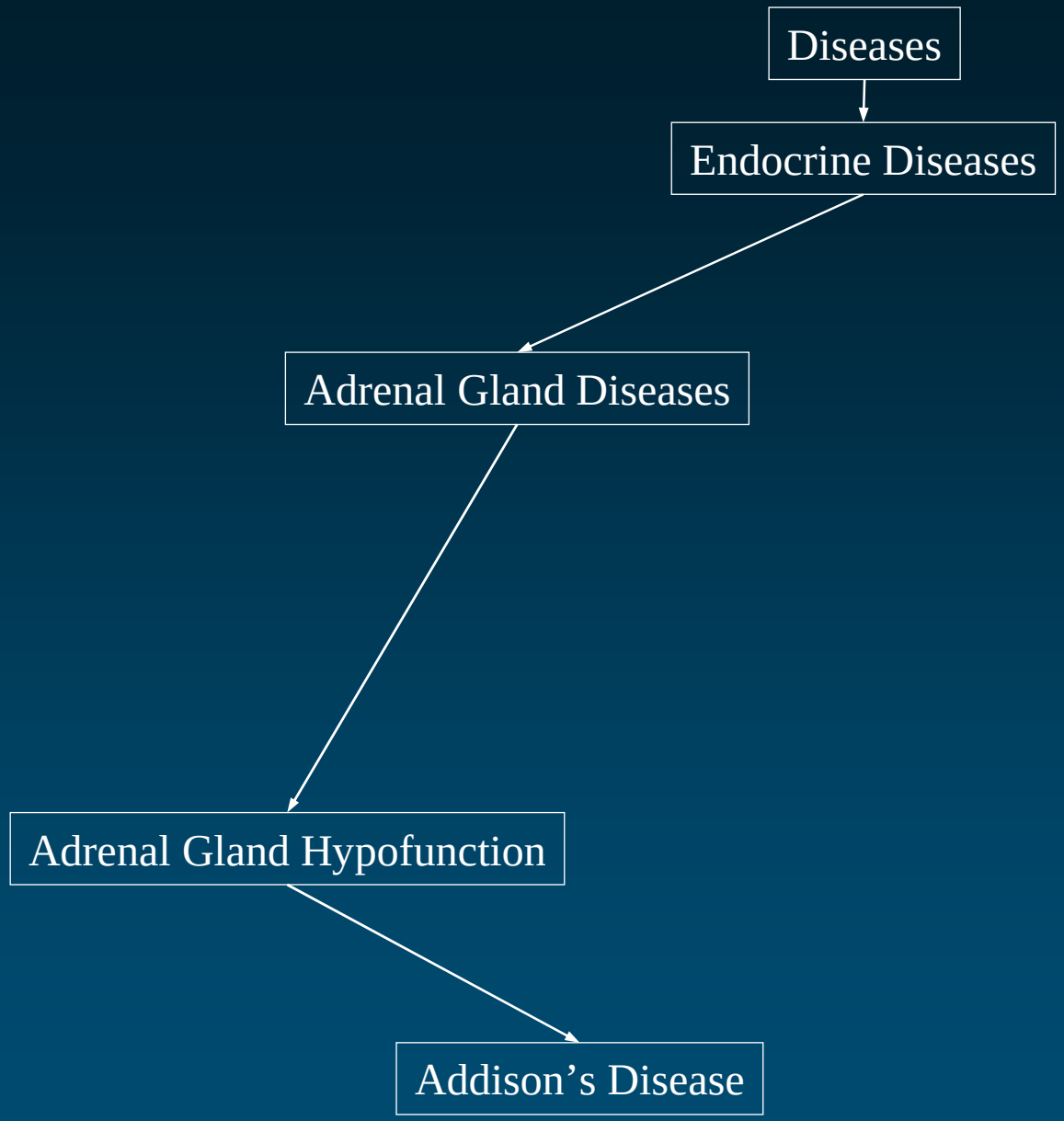
- Symbolic relations: ~9 M pairs of concepts
- Statistical relations : ~7 M pairs of concepts (co-occurring concepts)
- Mapping relations: 100,000 pairs of concepts

- Categorization: Relationships between concepts and semantic types from the Semantic Network

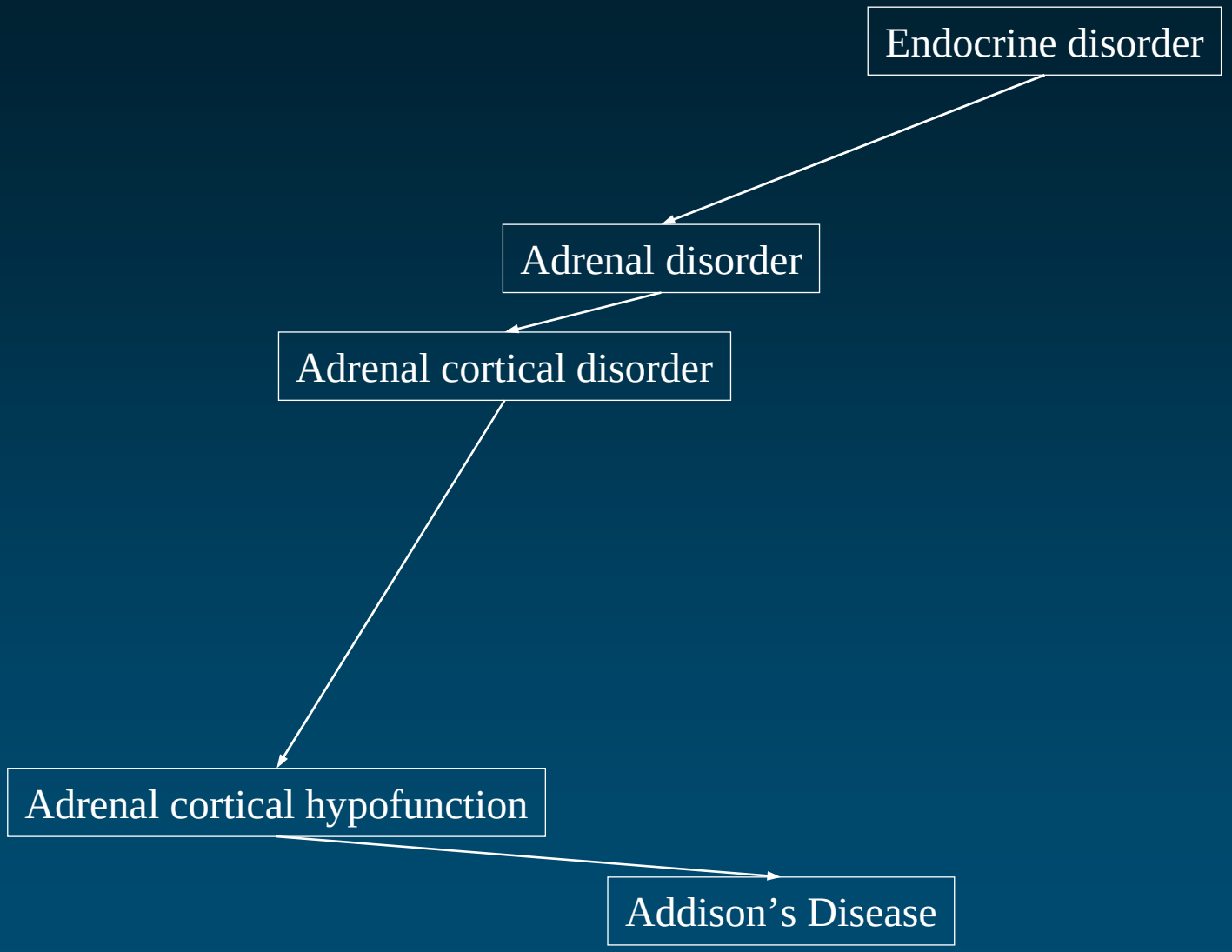
SNOMED International



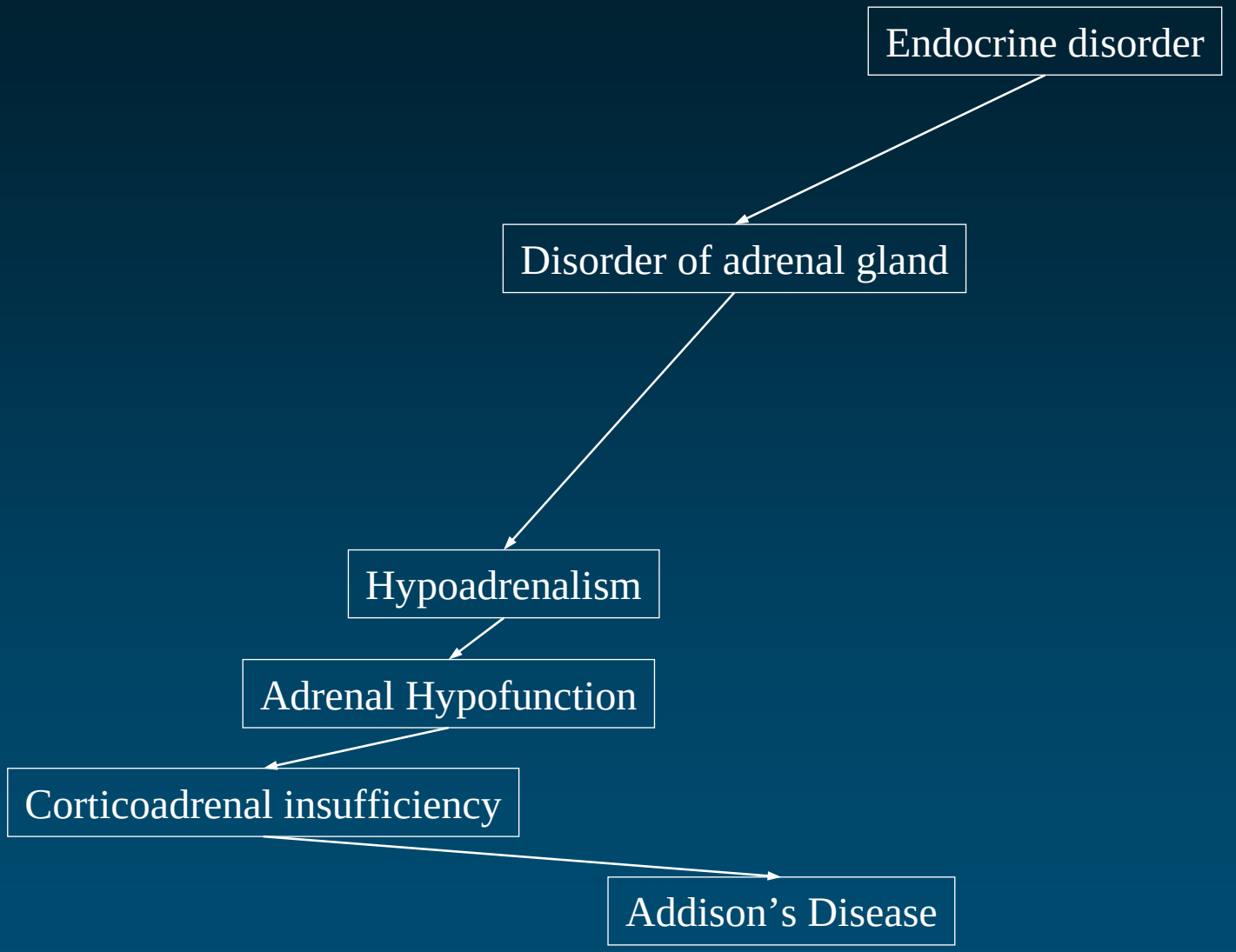
MeSH


Diseases
Endocrine Diseases
Adrenal Gland Diseases
Adrenal Gland Hypofunction
Addison’s Disease

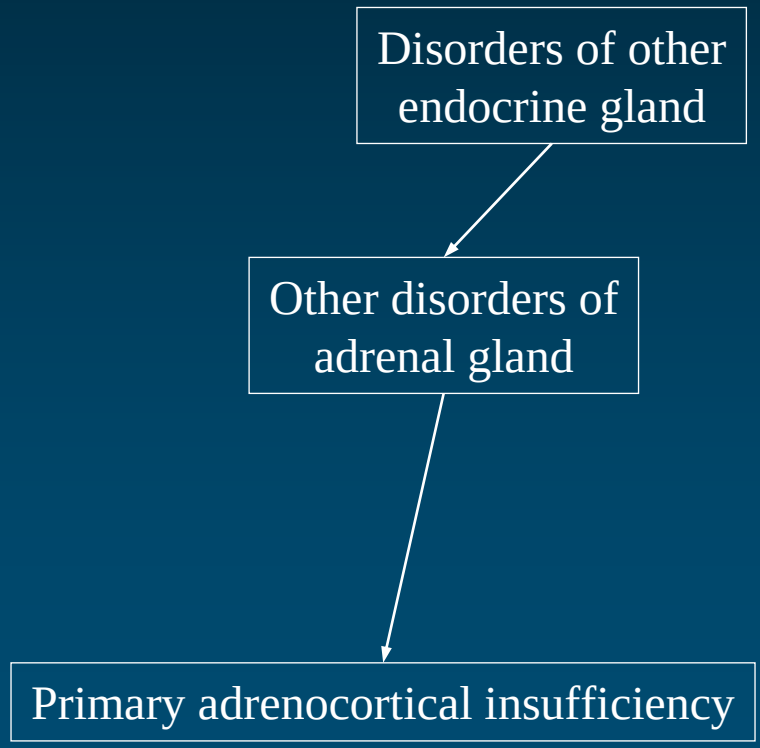
AOD


Endocrine disorder
Adrenal disorder
Adrenal cortical disorder
Adrenal cortical hypofunction
Addison’s Disease

Read Codes

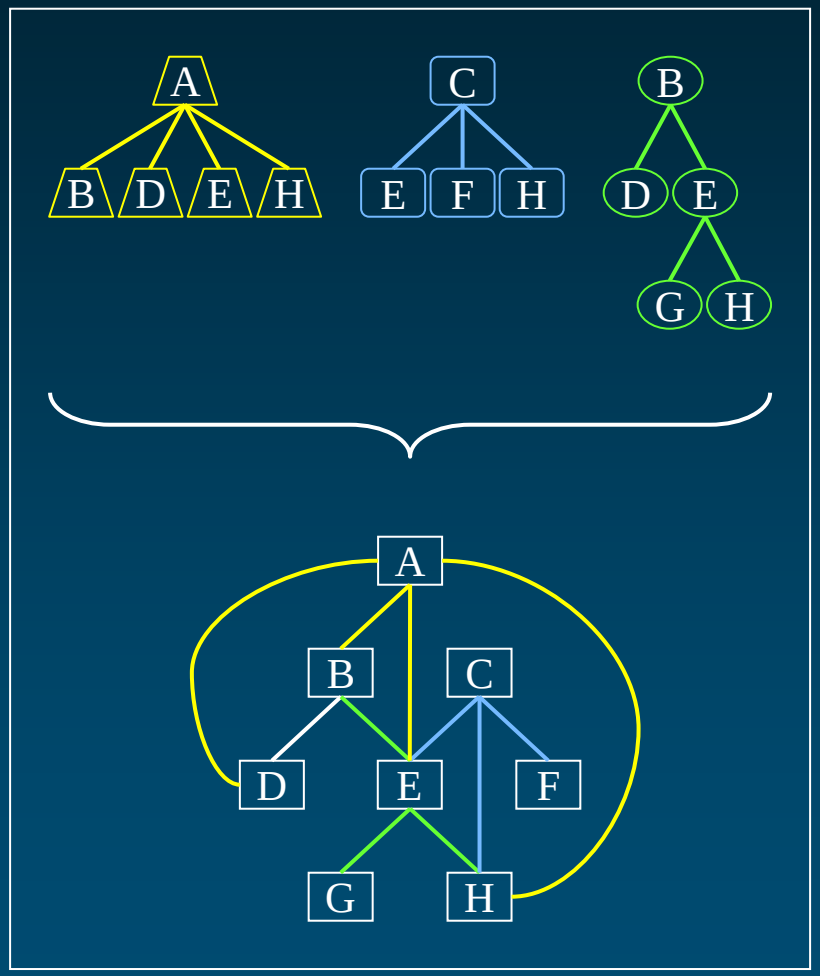

Endocrine disorder
Disorder of adrenal gland
Hypoadrenalism
Adrenal Hypofunction
Corticoadrenal insufficiency
Addison’s Disease

ICD-10


Disorders of other endocrine gland
Other disorders of adrenal gland
Primary adrenocortical insufficiency

Organize concepts

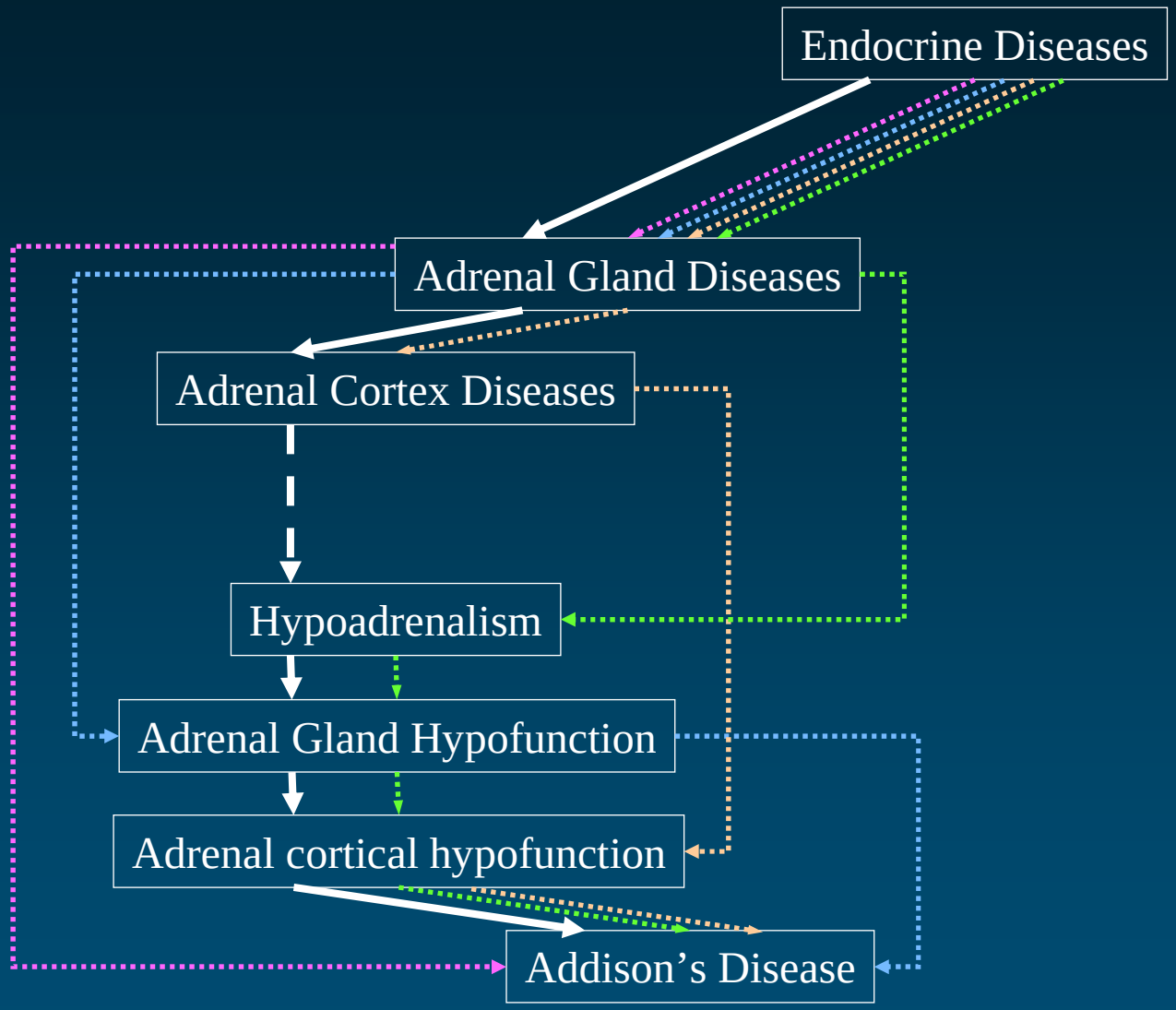
- Inter-concept relationships: hierarchies from the source vocabularies
- Redundancy: multiple paths
- One graph instead of multiple trees (multiple inheritance)

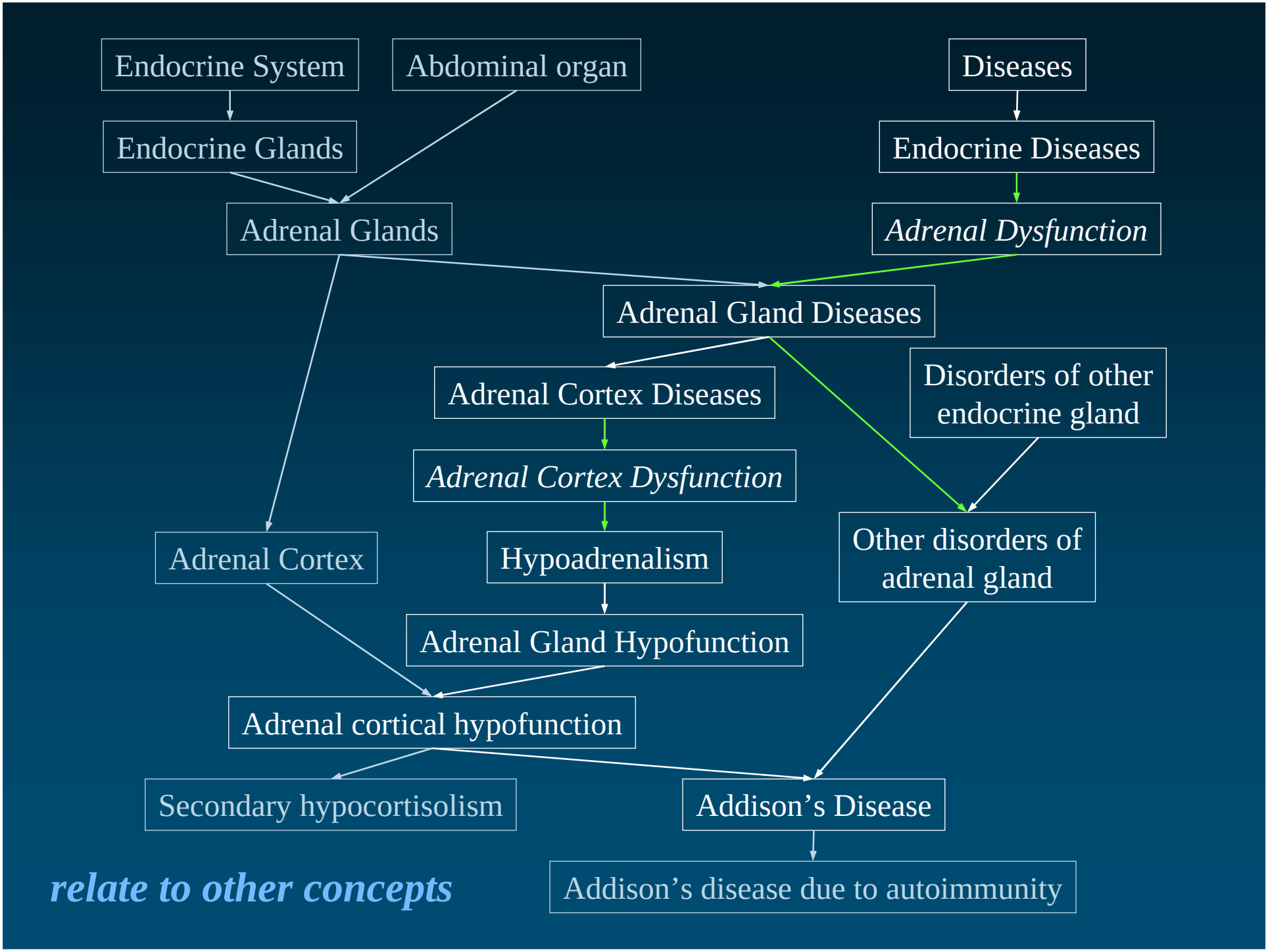


organize concepts

SNOMED
MeSH
AOD
Read Codes

UMLS




Endocrine System
Abdominal organ
Diseases
Endocrine Glands
Endocrine Diseases
Adrenal Glands
Adrenal Dysfunction
Adrenal Gland Diseases
Adrenal Cortex Diseases
Disorders of other endocrine gland
Adrenal Cortex Dysfunction
Adrenal Cortex
Hypoadrenalism
Other disorders of adrenal gland
Adrenal Gland Hypofunction
Adrenal cortical hypofunction
Secondary hypocortisolism
Addison's Disease
relate to other concepts
Addison's disease due to autoimmunity

Symbolic relations

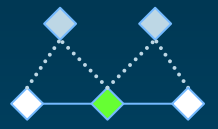
- Relation
  - Pair of “atom” identifiers
  - Type
  - Attribute (if any)
  - List of sources (for type and attribute)
- Semantics of the relationship: defined by its type [and attribute]

Source transparency: the information is recorded at the “atom” level

Symbolic relationships Type

- Hierarchical
  - Parent / Child — **PAR/CHD**
  - Broader / Narrower than — **RB/RN**

- Derived from hierarchies
  - Siblings (children of parents) — **SIB**

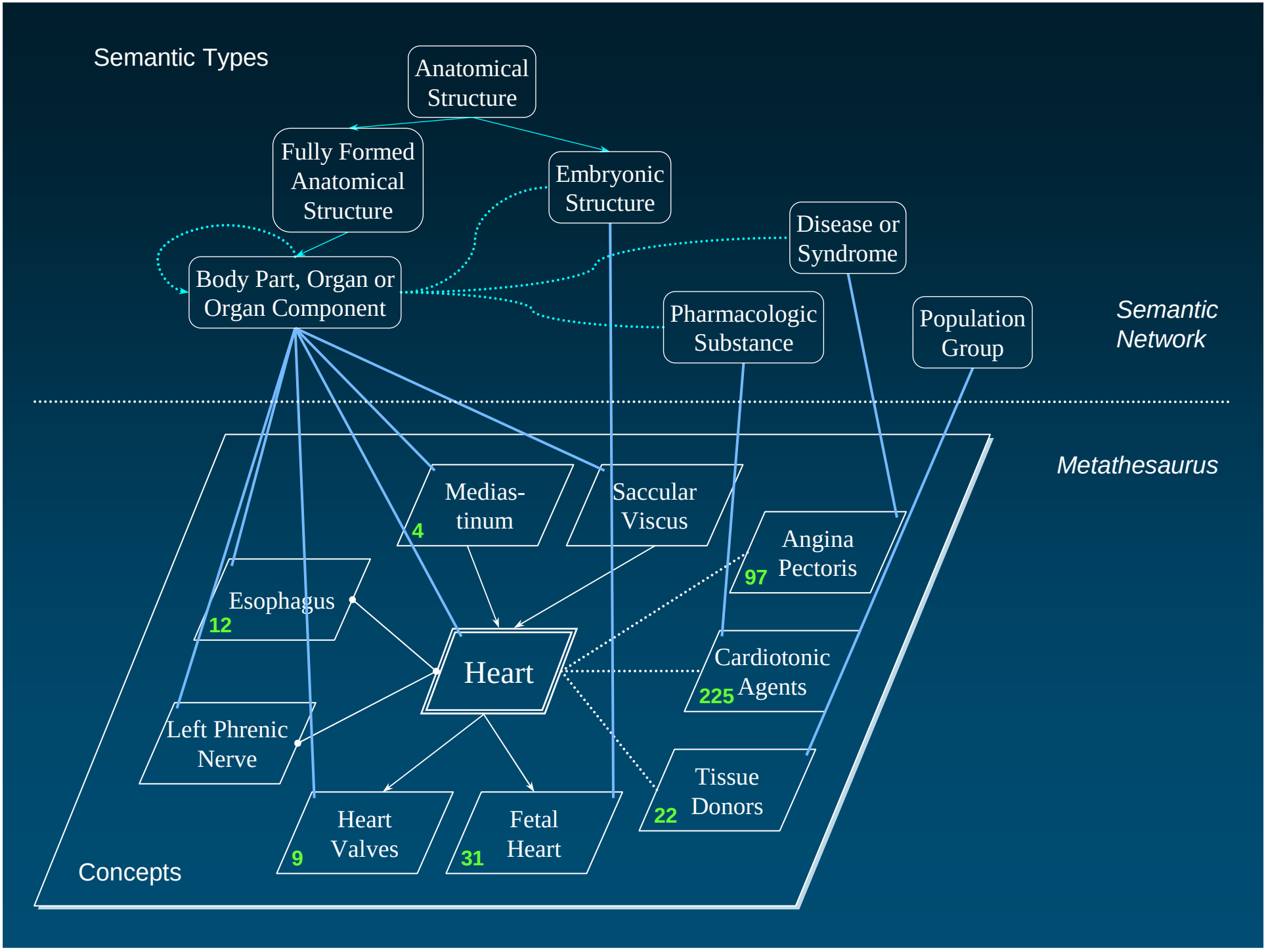


- Associative
  - Other — **RO**

- Various flavors of near-synonymy
  - Similar — **RL**
  - Source asserted synonymy — **SY**
  - Possible synonymy — **RQ**

Symbolic relationships Attribute

- Hierarchical
  - isa (is-a-kind-of)
  - part-of
- Associative
  - location-of
  - caused-by
  - treats
  - …
- Cross-references (mapping)


Semantic Types
Anatomical Structure
Fully Formed Anatomical Structure
Embryonic Structure
Disease or Syndrome
Body Part, Organ or Organ Component
Pharmacologic Substance
Population Group
Semantic Network
Metathesaurus
Medias-tinum
4
Saccular Viscus
Angina Pectoris
97
Esophagus
12
Heart
Cardiotonic Agents
225
Left Phrenic Nerve
Tissue Donors
22
Heart Valves
9
Fetal Heart
31
Concepts

Terminology integration methods

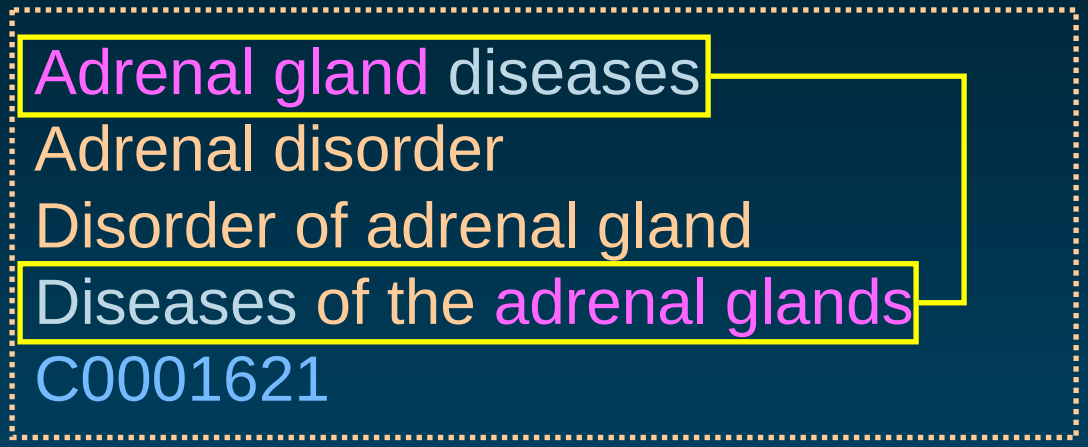
How do they do that?

- Integrating terms
  Lexical knowledge
- Categorizing concepts
  Semantic pre-processing
- Integrating relations
  Recording relations
- Editing and auditing
  UMLS editors

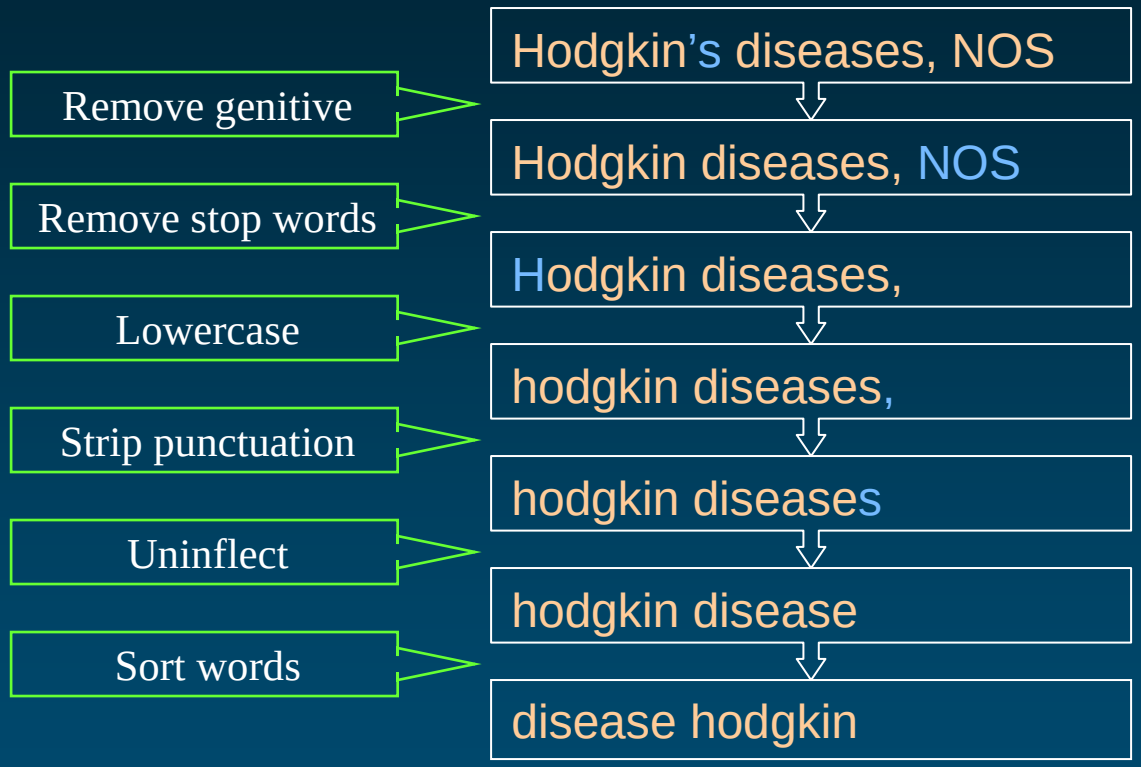
Terminology integration methods

Lexical knowledge

Lexical knowledge



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

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

Integrating terms Examples

- Exact match
  - Original term: Pain in back (b28013)
  - Concept mapped to: Back Pain (C0003862)

Pain in back present in the Metathesaurus (from the Read Codes)

- Match after normalization
  - Original term: Pain in joints (b28016)
  - Normalized term: joint pain
  - Concept mapped to: Arthralgia (C0003862)

Joint pain is a synonym for Arthralgia

Integrating terms Examples

- No match found
  - Radiating pain in body part (b2801) *Too general*
  - Radiating pain in a dermatome (b2803) *Too specific*
  - Pain in stomach or abdomen (b28012) *Coordination*

e215	Population	(→ C0032659)
e2150	Demographic change	(→ C0681668)
e2151	Population density	(→ C0032665)
e2158	Population, **other specified**	
e2159	Population, **unspecified**	

Integrating terms Examples

- Multiple matches
  - Impulse control (b1304)
    - Impulse control (C0150632)
    - Impulse control training (C0262701)
    - Ability to control impulses (C0517616)
  - Frontal lobe (s11000)
    - frontal lobe (C0016733)
    - Entire frontal lobe (C1268977)

    } *SNOMED CT distinction*
  - Bites (b5101)
    - Biting (C0005658)
    - 2-(4-ethoxybenzyl)-1-diethylaminoethyl-5-isothiocyanatobenzimidazole (C0045724)

synonym for BIT alkylating agent

Terminology integration methods

Semantic pre-processing

Semantic pre-processing

- Metadata in the source vocabularies
- Tentative categorization
- Positive (or negative) evidence for tentative synonymy relations based on lexical features

Semantic pre-processing in practice

- Mapping between
  - Semantic types (UMLS Semantic Network)
  - Semantics of a given subset of a terminology
- Semantic Network
  - 135 semantic types (high-level categories)
  - 2 hierarchies for Entity and Event
  - Examples
    - Disease or Syndrome
    - Body Part, Organ, or Organ Component
    - Mental Process

UMLS Semantic Groups

- ACTI Activities & Behaviors
- ANAT Anatomy
- CHEM Chemicals & Drugs
- CONC Concepts & Ideas
- DEVI Devices
- DISO Disorders
- GENE Genes & Molecular Sequences
- GEOG Geographic Areas
- LIVB Living Beings
- OBJC Objects
- OCCU Occupations
- ORGA Organizations
- PHEN Phenomena
- PHYS Physiology
- PROC Procedures

- Acquired Abnormality
- Anatomical Abnormality
- Cell or Molecular Dysfunction
- Congenital Abnormality
- Disease or Syndrome
- Experimental Model of Disease
- Finding
- Injury or Poisoning
- Mental or Behavioral Dysfunction
- Neoplastic Process
- Pathologic Function
- Sign or Symptom

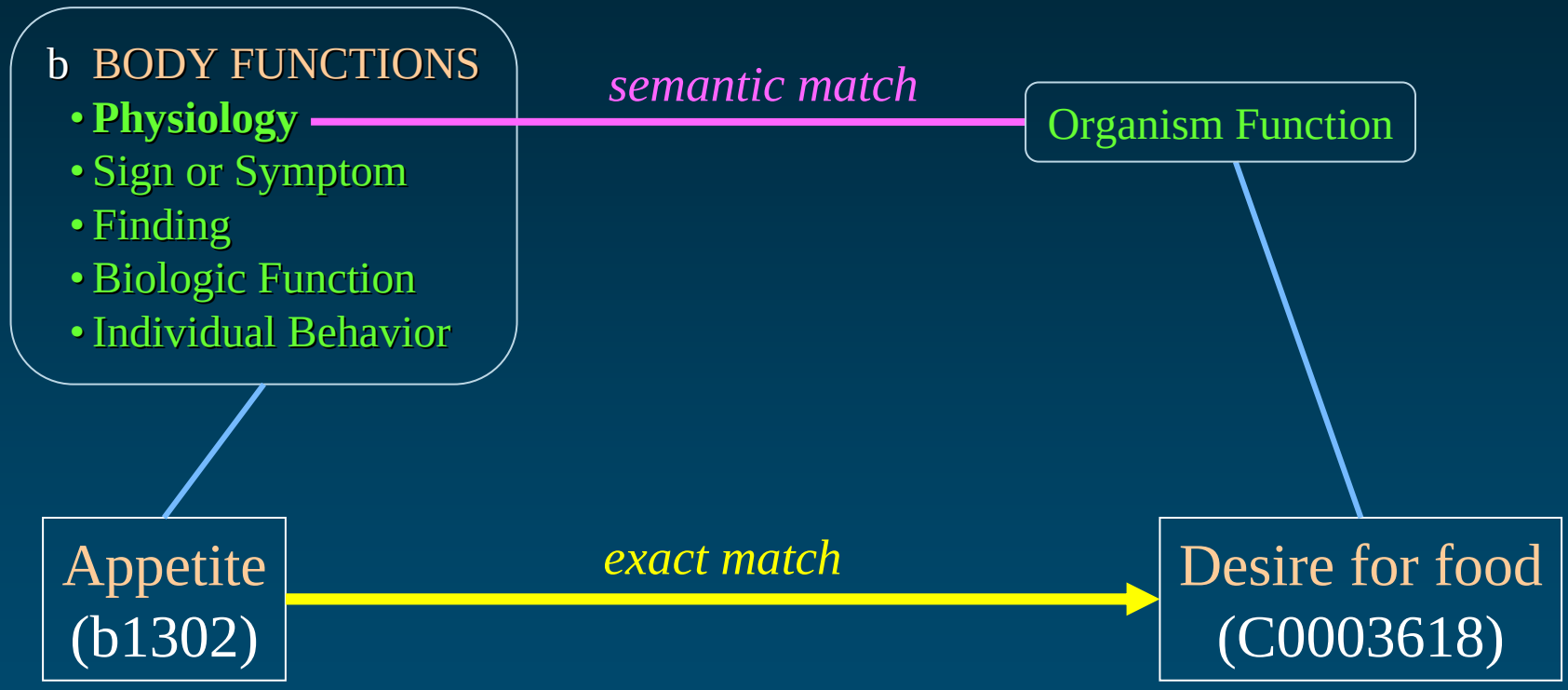
Semantic areas in ICF

- b BODY FUNCTIONS
  - **Physiology**
  - Sign or Symptom
  - Finding
  - Biologic Function
  - Individual Behavior
- s BODY STRUCTURES
  - **Anatomy**
- d ACTIVITIES AND PARTICIPATION
  - **Physiology**
  - **Activities & Behaviors**
  - Machine Activity
  - Sign or Symptom
  - Finding
  - Educational Activity

Semantic areas in ICF

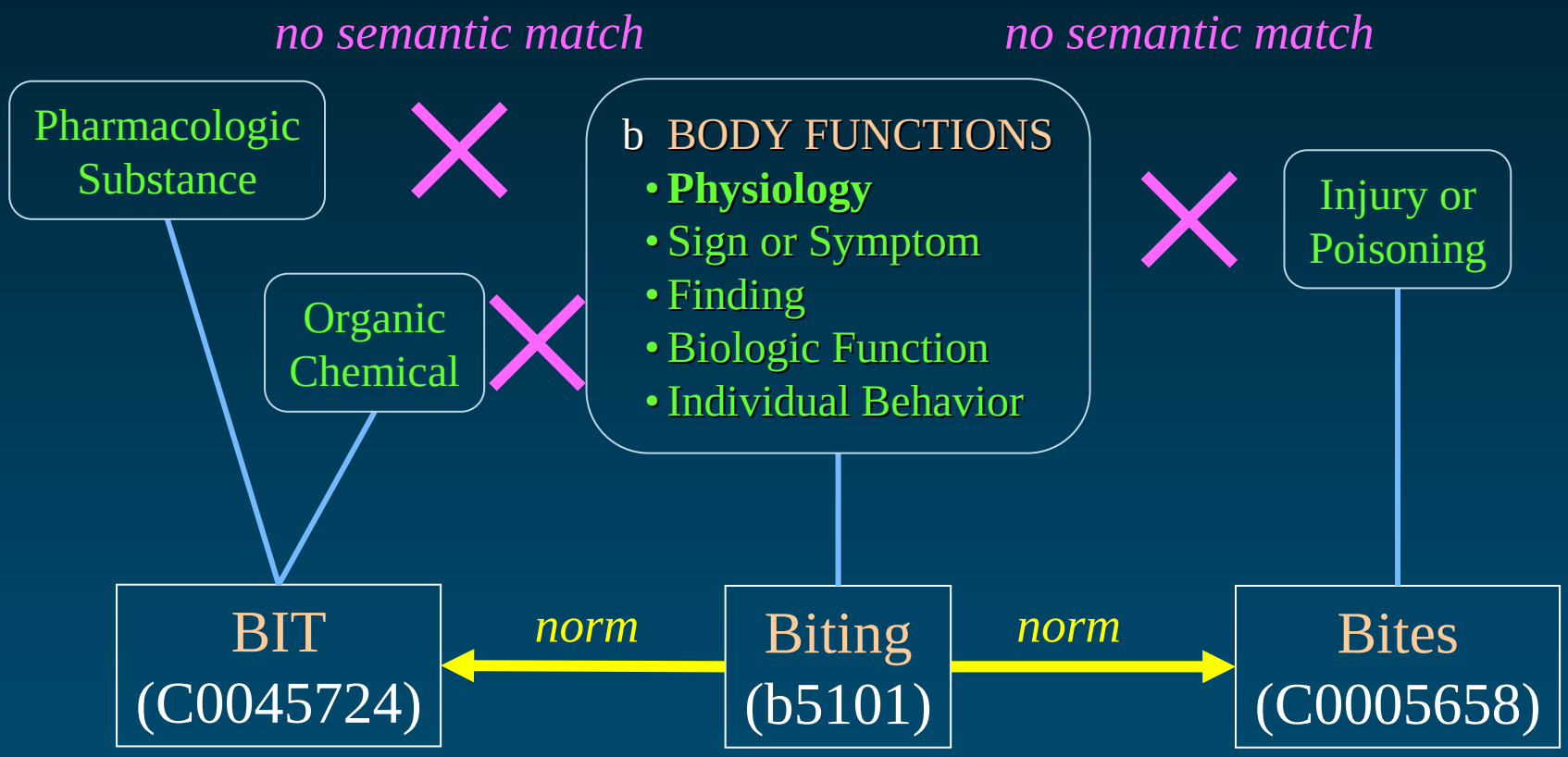
- e1 PRODUCTS AND TECHNOLOGY
  - ????
- e2 NATURAL ENVIRONMENT AND HUMAN-MADE CHANGES TO ENVIRONMENT
  - **Phenomena**
- e3 SUPPORT AND RELATIONSHIPS
  - Family Group
  - Population Group
  - Professional or Occupational Group
- e4 ATTITUDES
  - ????
- e5 SERVICES, SYSTEMS AND POLICIES
  - Governmental or Regulatory Activity
  - Regulation or Law

Semantic pre-processing Examples



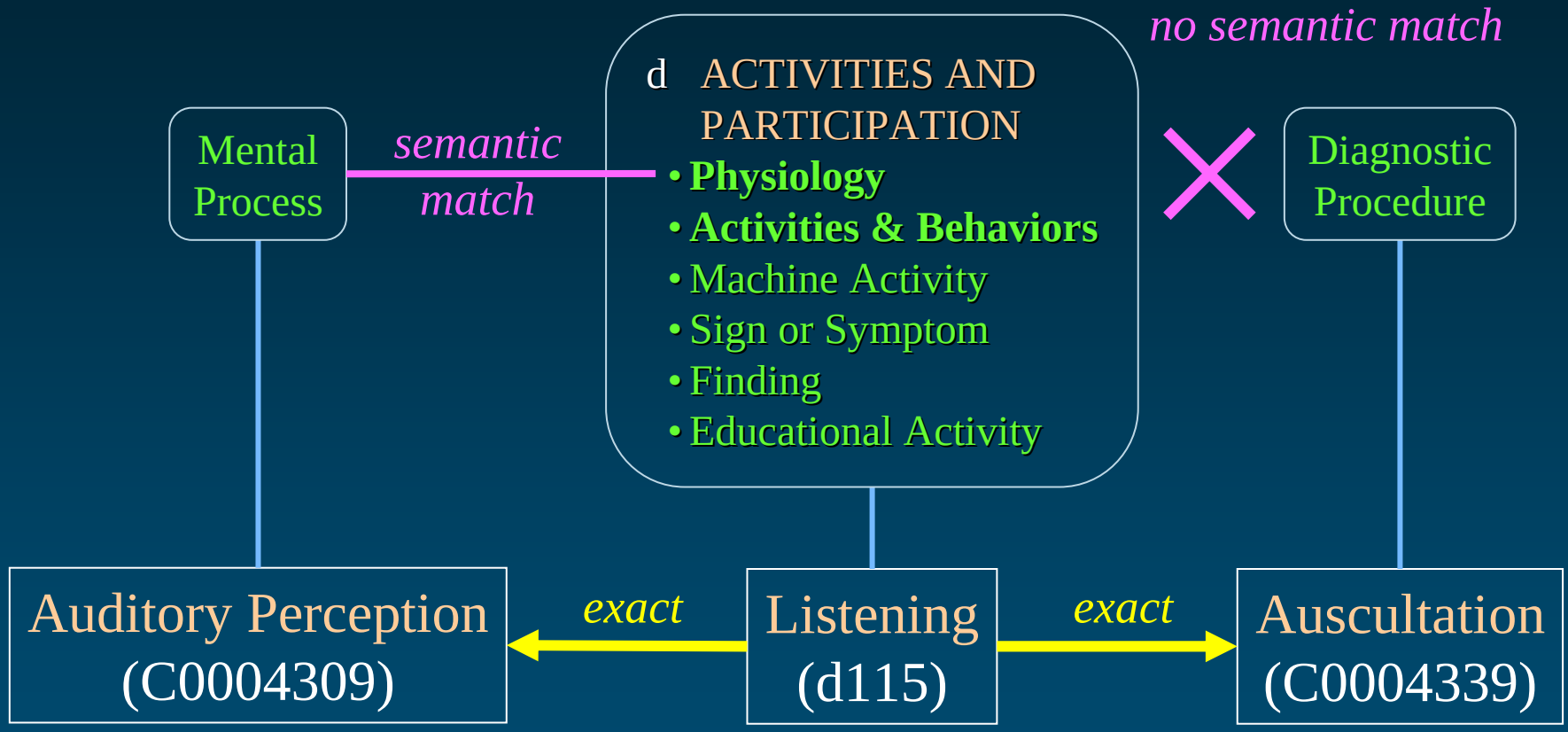
Validation

Semantic pre-processing Examples



No match

Semantic pre-processing Examples



Disambiguation

Terminology integration methods

Recording relations

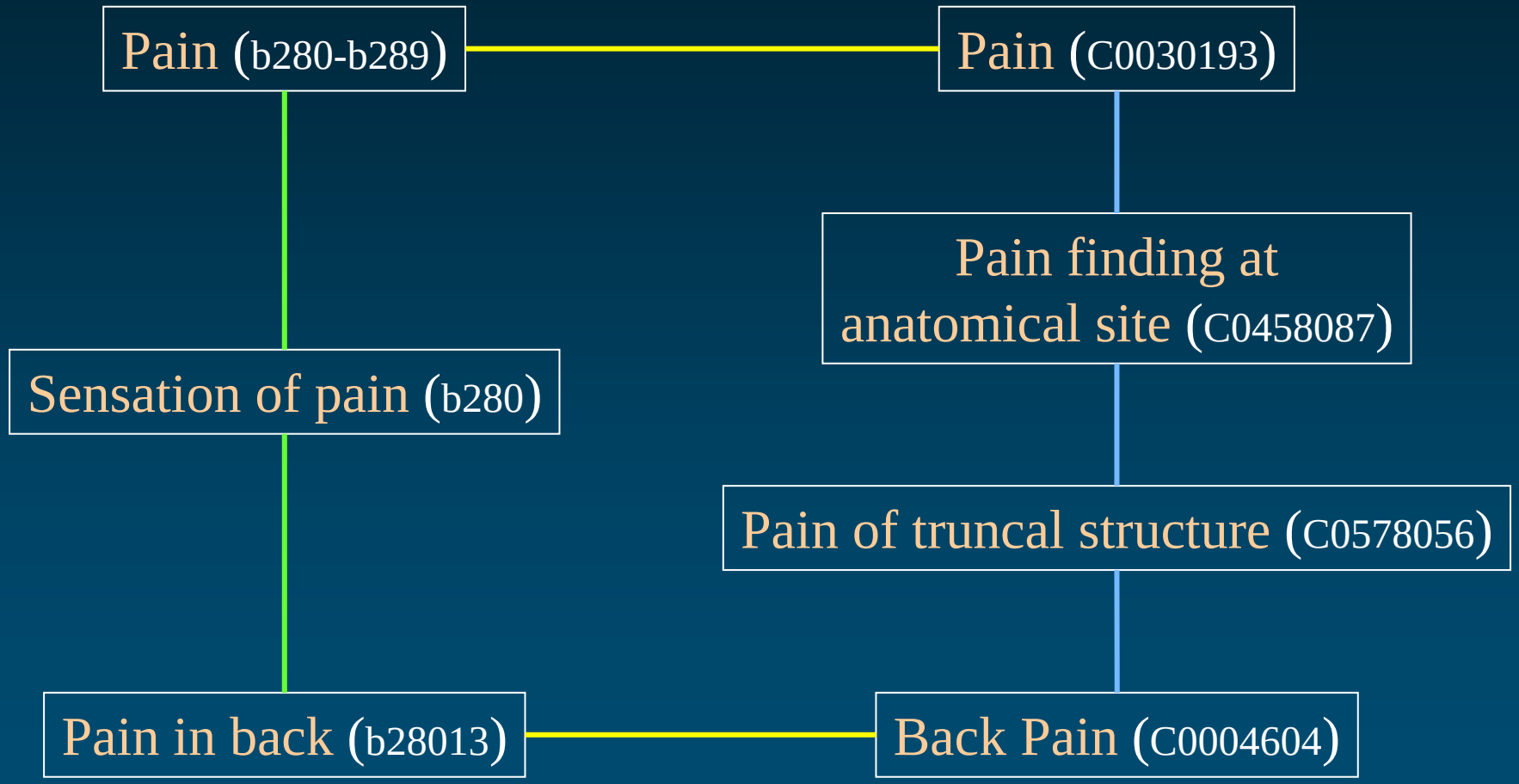
Recording relations

- Relations
  - Recorded at the term (atom) level
  - Aggregated at the concept level
- Once integrated into the UMLS, ICF relations participate to the Metathesaurus graph
- Possibly redundant with relations from other sources

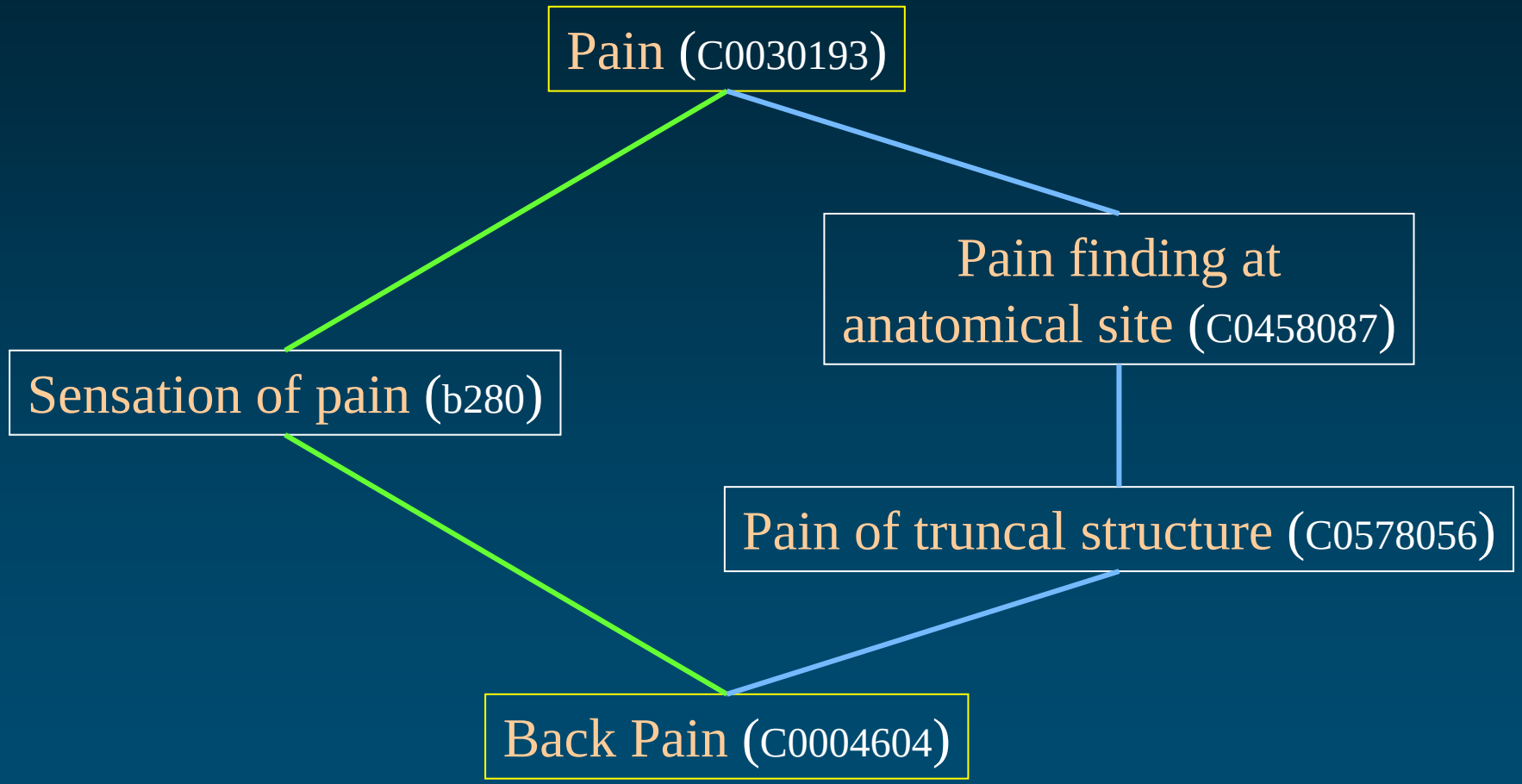
ICF relations in the Metathesaurus

- ICF hierarchical relations in the UMLS
  - REL: parent/child
  - RELA: none
  - SAB: ICF
- Other relations?

ICF relations in the Metathesaurus



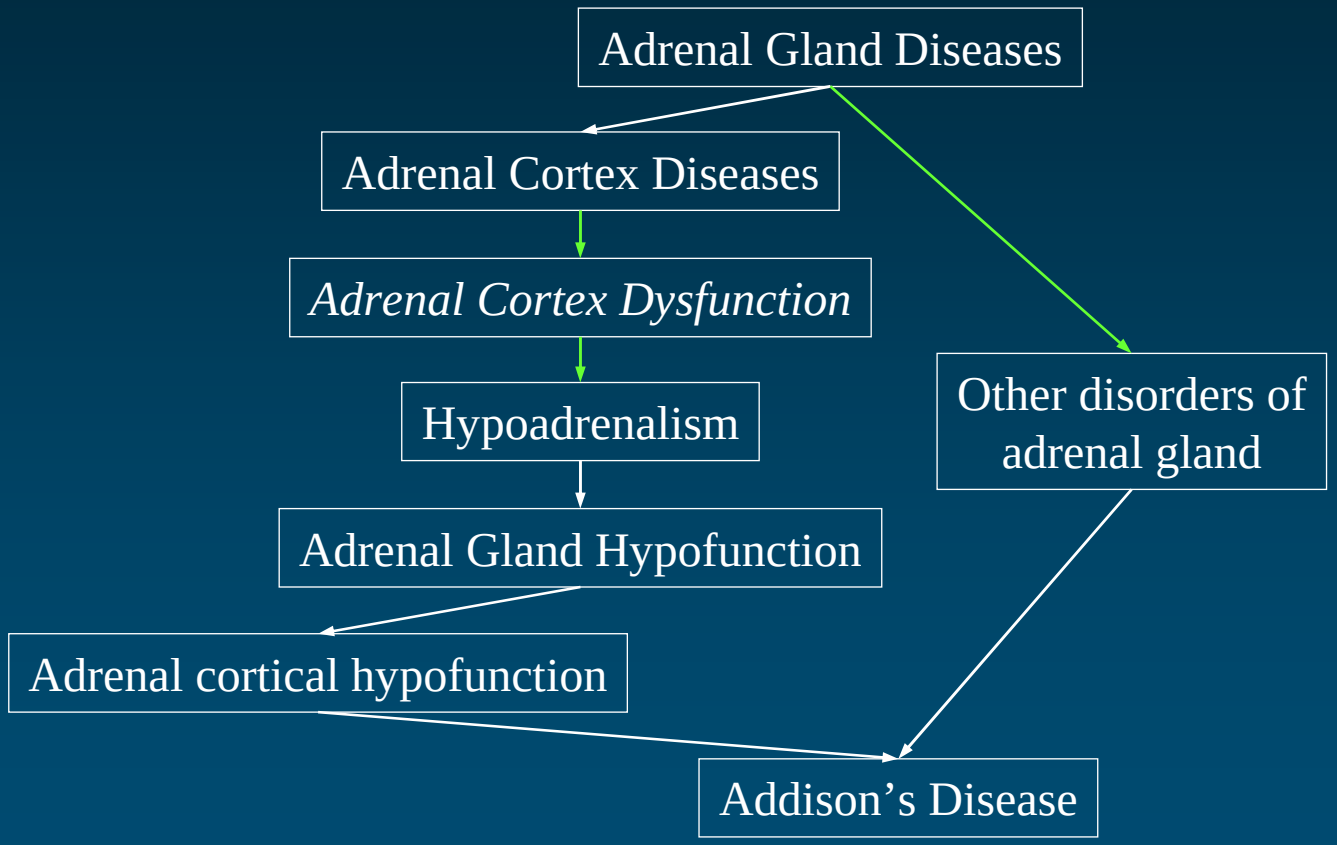
ICF relations in the Metathesaurus



Terminology integration methods

Editing and auditing

Additional knowledge: UMLS editors



Experience with ICF

Acknowledgments

- Marcy Harris
- Guergana Savova

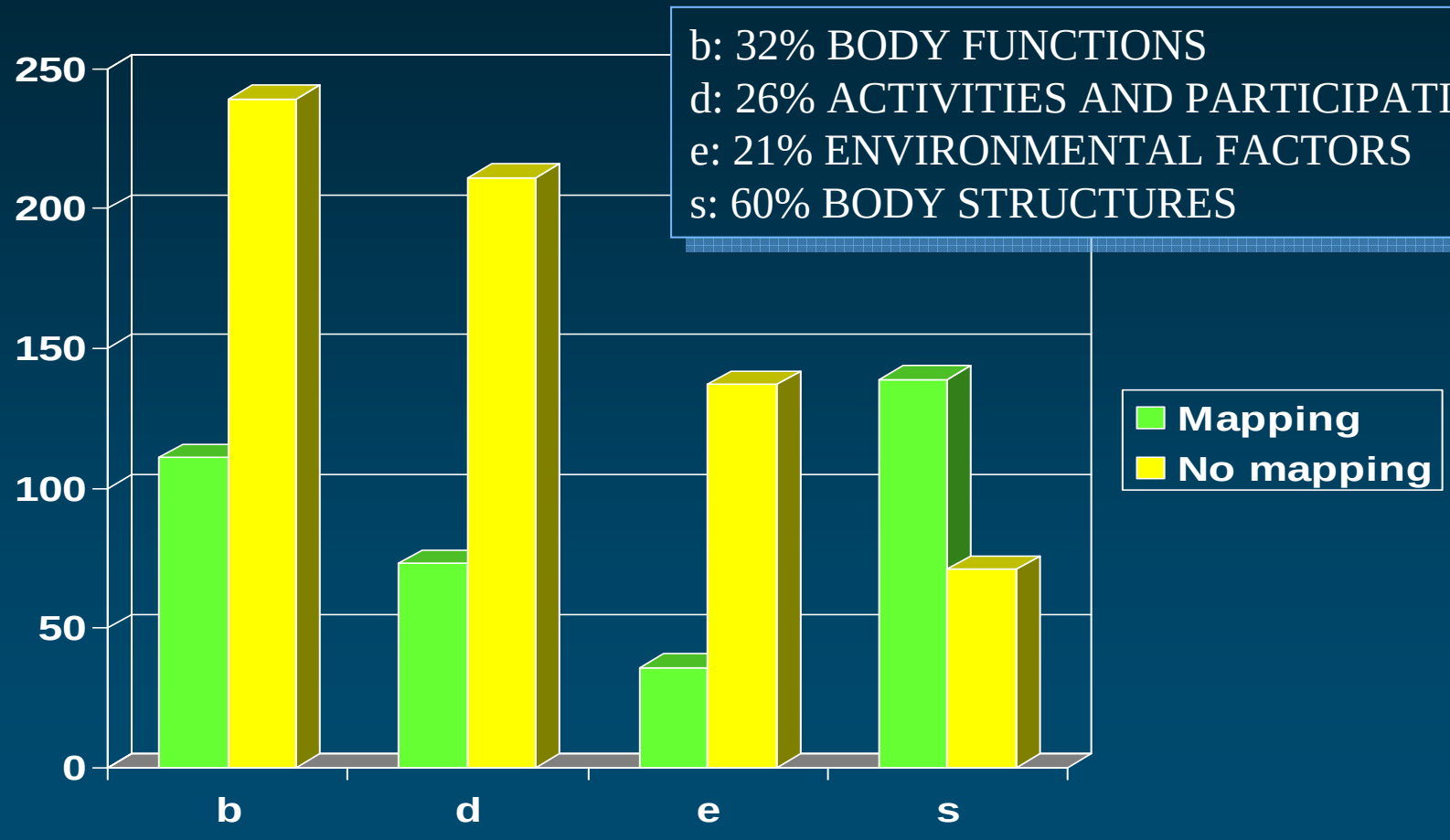
Materials

- ICF: 1495 terms
  - 478 terms filtered out
    - 218 terms with *other specified*
    - 217 terms with *unspecified*
    - 37 terms with *other specified* and *unspecified*
    - 2 terms with *specified* (alone)
    - 1 term with *other specified* (alone)
  - 1017 terms remaining
- UMLS: version 2004AA

Mapping to UMLS Metathesaurus

- Methods
  - Exact match first
  - Normalized match, if necessary
- Results
  - 359 ICF terms mapped (35%)
  - 658 terms without mapping

Mapping by category



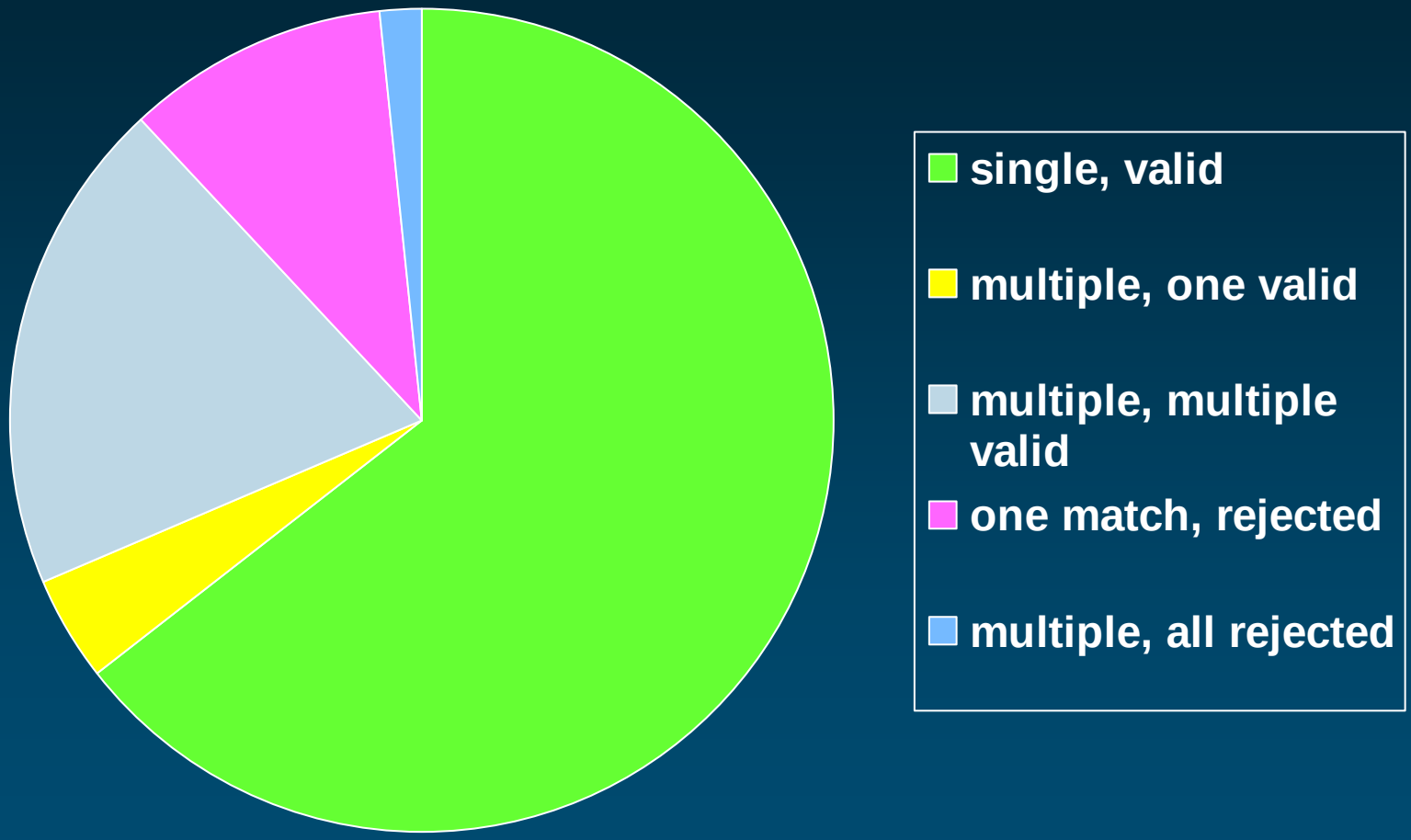
Issues with mapping

- Phenomena preventing the terms form being mapped:
  - coordination with *and* alone: 147
    - Education and training policies (e5852)
  - coordination with *or* alone: 7
    - Pain in stomach or abdomen (b28012)
  - coordination with both *and* and *or*: 2
    - Assistive products and technology for the practice of religion or spirituality (e1451)

Semantic validation

- Method
  - Correspondence between ICF chapters and UMLS semantic types/groups
- Issues
  - Correspondence difficult to establish for some subgroups in ENVIRONMENTAL FACTORS
    - PRODUCTS AND TECHNOLOGY
    - ATTITUDES

Semantic validation Results



Issues with semantic validation

- Multiple “valid” matches must be reviewed by experts and disambiguated
- Rejected mappings
  - Semantically invalid UMLS concepts

    or
  - Missing correspondence (ICF chapter/UMLS ST-SG)

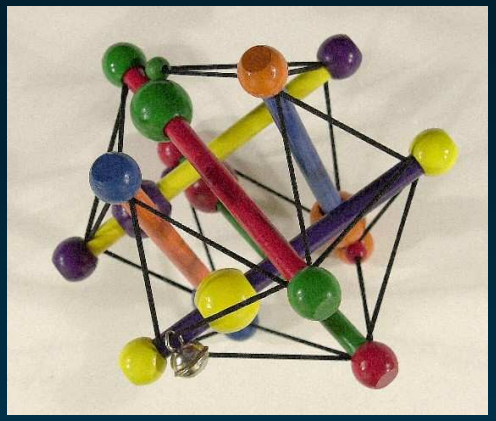
Conclusions

Conclusions (1)

- Integrating ICF into the UMLS
  - Should not be too difficult
    - Relatively small
    - Many concepts already present in UMLS
  - Challenges
    - Underspecified terms
    - Coordination
    - Specific perspective

Conclusions (2)

- Integrating ICF into the UMLS
  - Benefit for ICF
    - Links to other vocabularies
    - Facilitate downward extension
  - Benefit for UMLS
    - Adds specific perspective



Medical Ontology Research

Contact: olivier@nlm.nih.gov

Web: mor.nlm.nih.gov

Olivier Bodenreider

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