June 24, 2008

Using Semantic Web Technologies for Terminology Integration

Olivier Bodenreider

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

Motivation Translational research

- "Bench to Bedside"
- Integration of clinical and research activities and results
- Supported by research programs
  - NIH Roadmap
  - Clinical and Translational Science Awards (CTSA)
- Requires the effective integration and exchange and of information between
  - Basic research
  - Clinical research

Translational research NIH Roadmap

OPASI Office of Portfolio Analysis and Strategic Initiatives
National Institutes of Health

Search GO

OPASI Home | NIH Roadmap | Council | Governance | Divisions | News & Events | About OPASI

Back to: OPASI Home > Division of Strategic Coordination (DSC) > NIH Roadmap > Re-engineering the Clinical Research Enterprise

NIH Roadmap FOR MEDICAL RESEARCH

Re-engineering the Clinical Research Enterprise

- Overview
- Implementation Group Members
- Funding Opportunities
- Funded Research
- Meetings
- Mid-course Reviews

- CTSAweb.org EXIT Disclaimer

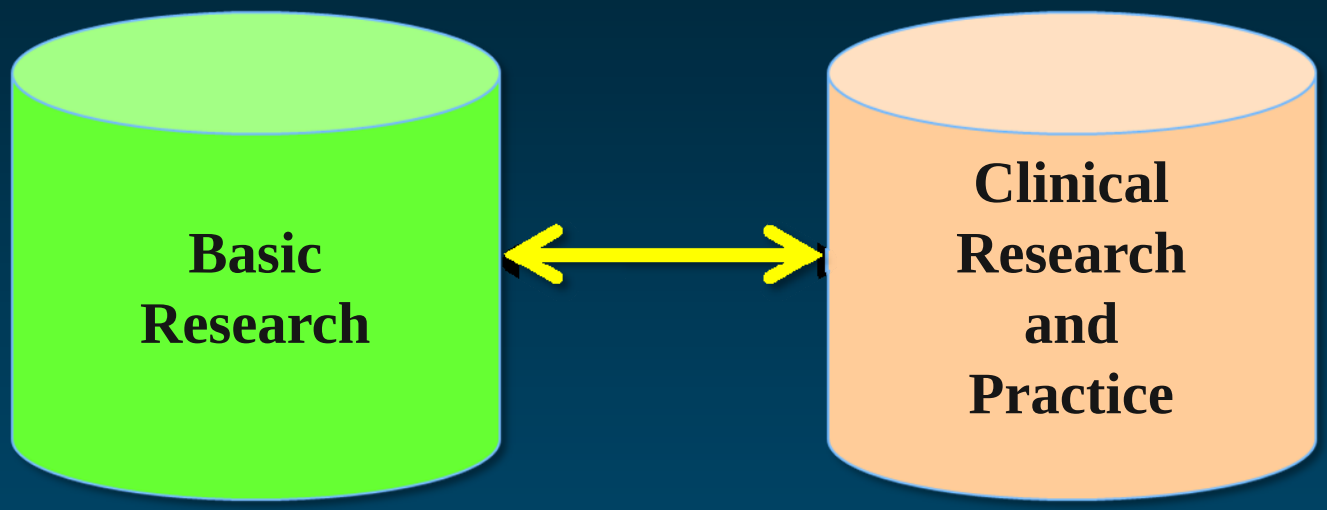
TRANSLATIONAL RESEARCH

OVERVIEW

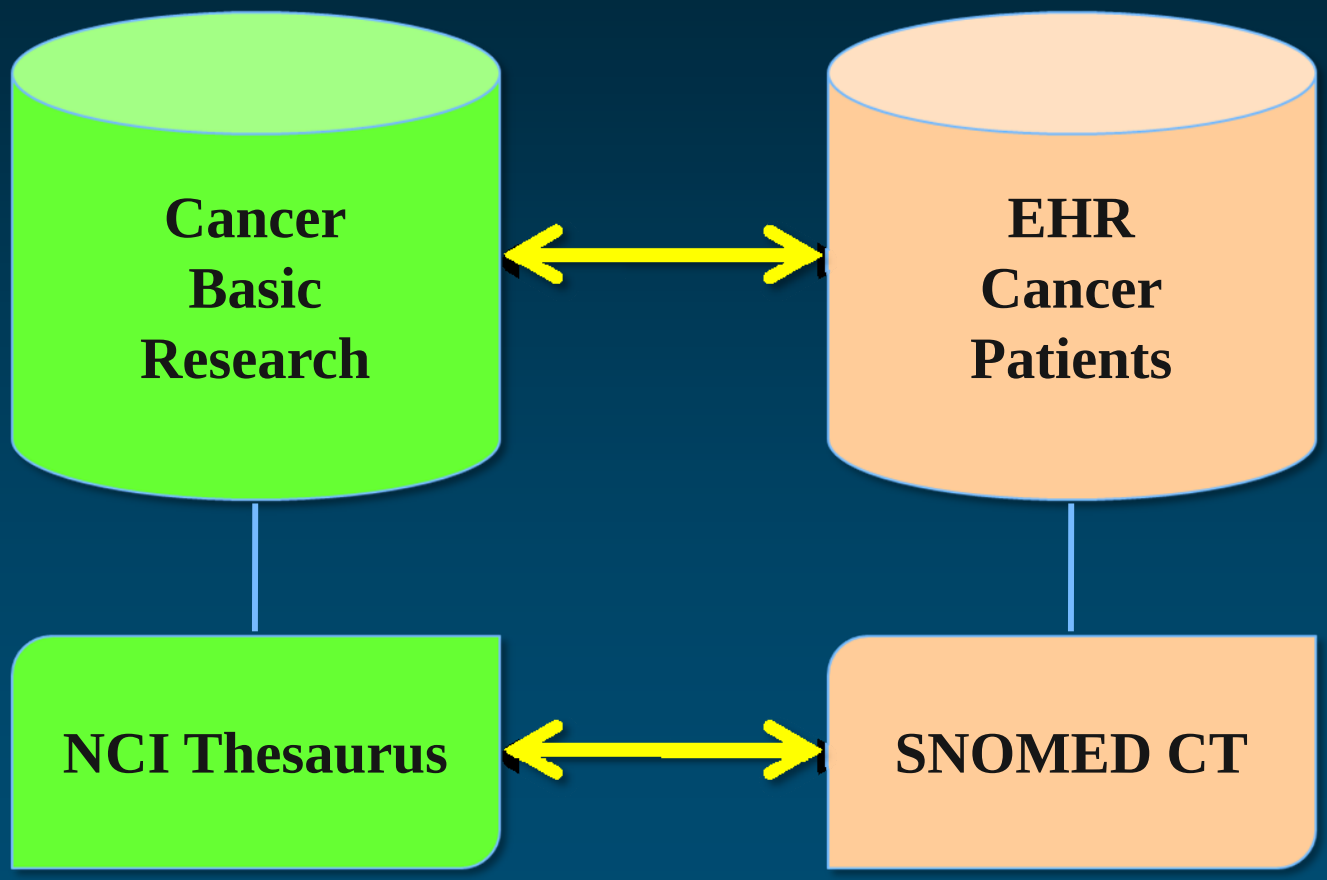
To improve human health, scientific discoveries must be translated into practical applications. Such discoveries typically begin at "the bench" with basic research — in which scientists study disease at a molecular or cellular level — then progress to the clinical level, or the patient's "bedside."

Scientists are increasingly aware that this bench-to-bedside approach to translational research is really a two-way street. Basic scientists provide clinicians with new tools for use in patients and for assessment of their impact, and clinical researchers make novel observations about the nature and progression of disease that often stimulate basic investigations.

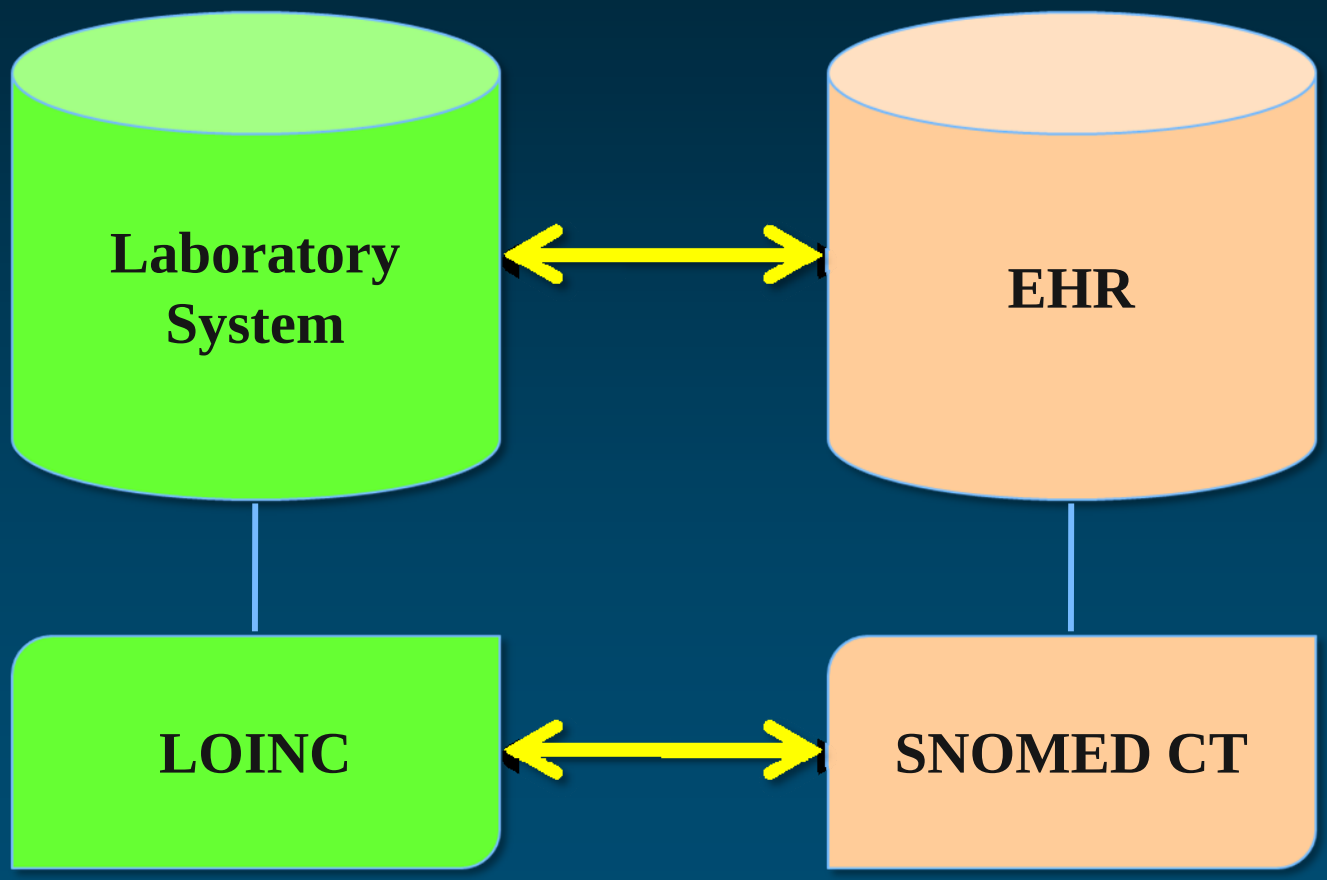
Motivation Translational research



Terminology and translational research

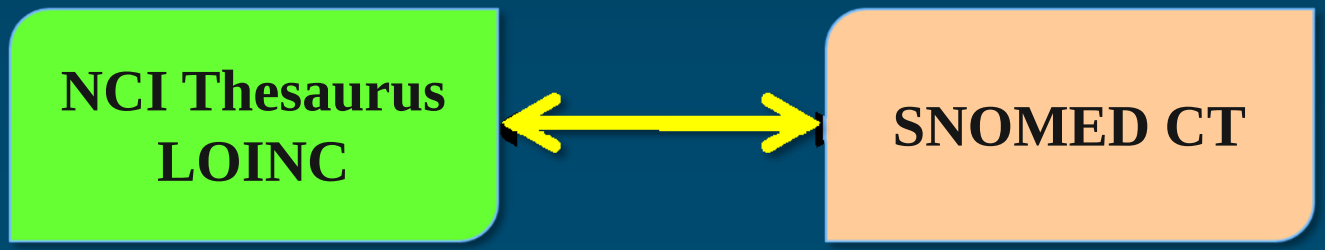


Terminology and data integration



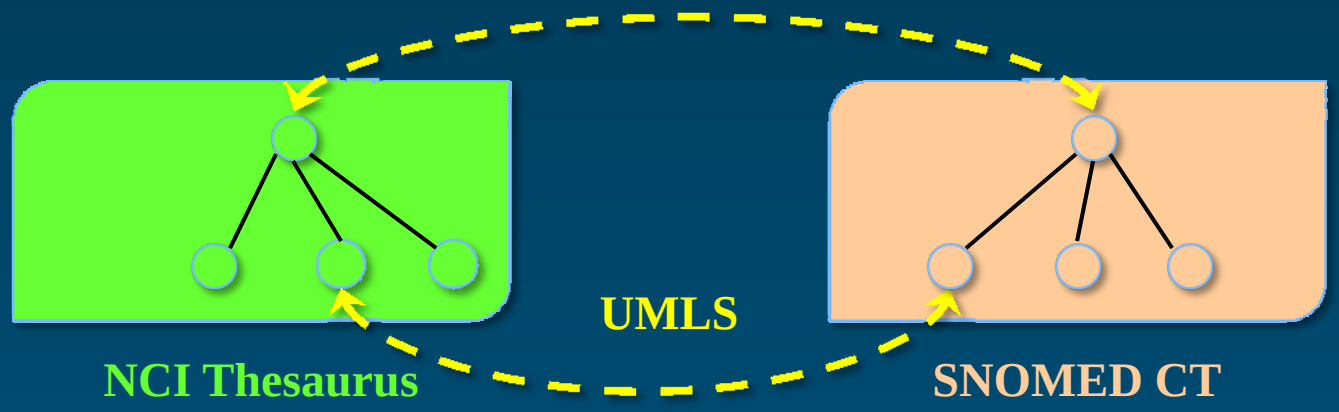
Objectives

- To evaluate interoperability between data repositories at the terminological level
  - Compare formal definitions
    - NCI Thesaurus (NCIt) / SNOMED CT [Bodenreider, KR-MED 2008]
    - [NCI Thesaurus (NCIt) / SNOMED CT / ICD-O] [Burgun, APIII 2007]
    - LOINC / SNOMED CT [Bodenreider, AMIA 2008]
  - Evaluate Semantic Web technologies in this context



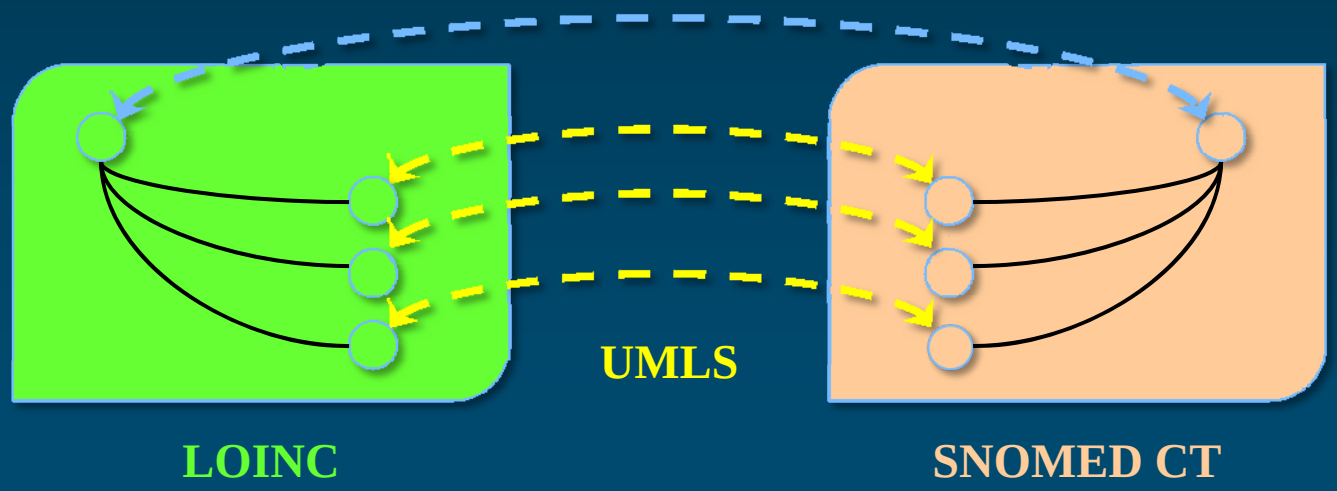
Hypotheses (1)

- Equivalent concepts should have similar formal definitions
  - Concept equivalence is provided by the UMLS



Hypotheses (2)

- Concepts having similar formal definitions should be equivalent
  - Manual mapping between relationships
  - Concept equivalence is provided by the UMLS



Semantic Web technologies

- Collection of languages, formalisms and tools developed to support the Semantic Web
  - XML
  - Resource Description Framework (RDF)
  - Web Ontology Language (OWL)

- RDF
  - Relations expressed as triples, forming a graph
  - NCIt, LOINC and SNOMED CT can be easily converted to RDF

NCI Thesaurus

NCI thesaurus Characteristics (1)

- Current version: 08.04d (~monthly releases)
- Type: Controlled terminology / ontology
- Domain: Cancer
- Developer: NCI Center for Bioinformatics
- Funding: NCI
- Availability
  - Publicly available: Yes
  - Repositories: UMLS / OBO / NCBO BioPortal
- URL: http://nciterms.nci.nih.gov/

NCI thesaurus Characteristics (2)

- Number of
  - Concepts: 58,868 (2007_05E)
  - Terms: 2.68 per concept
- Major organizing principles:
  - Subsumption hierarchy
  - Rich set of associative relationships
  - Small proportion of defined concepts (many primitives)
  - Links to many external resources
- Formalism: OWL Lite

NCI thesaurus Top level

NCI_Thesaurus Taxonomy

- Abnormal Cell
- Activity
- Anatomic Structure, System, or Substance
- Biochemical Pathway
- Biological Process
- Chemotherapy Regimen or Agent Combination
- Conceptual Entity
- Diagnostic, Therapeutic, and Research Equipment
- Diagnostic or Prognostic Factor
- Disease, Disorder or Finding
- Drug, Food, Chemical or Biomedical Material
- Experimental Organism Anatomical Concept
- Experimental Organism Diagnosis
- Gene
- Gene Product
- Molecular Abnormality
- NCI Administrative Concept
- Organism
- Property or Attribute
- Retired Concept

NCI thesaurus Example

Concept Details

URI: http://nciterms.nci.nih.gov:80/NCIBrowser/ConceptReport.jsp?dictionary=NCI_Thesaurus&code=C2919
Version: June 2007 (07.06d)

Prostate Adenocarcinoma

Identifiers:

name	Prostate_Adenocarcinoma
code	C2919

Relationships to other concepts:

Disease_Has_Finding	Invasive Lesion
Disease_Has_Abnormal_Cell	Adenocarcinoma Cell
Disease_Has_Normal_Tissue_Origin	Prostatic Epithelium
Disease_May_Have_Finding	Serum Prostate Specific Antigen Increased
Disease_Has_Finding	Carcinomatous Component Present
Disease_Excludes_Abnormal_Cell	Neoplastic Smooth Muscle Cell
Disease_Excludes_Abnormal_Cell	Malignant Squamous Cell
Disease_Has_Primary_Anatomic_Site	Prostate Gland
Disease_Has_Associated_Anatomic_Site	Male Reproductive System
Disease_Excludes_Abnormal_Cell	Malignant Stromal Cell
Disease_Has_Associated_Anatomic_Site	Prostate Gland
Disease_Has_Normal_Cell_Origin	Epithelial Cell

Superconcepts:

- Adenocarcinoma
- Common Carcinoma
- Invasive Prostate Carcinoma

Subconcepts:

- Acinar Prostate Adenocarcinoma
- Metastatic Prostatic Adenocarcinoma
- Moderately Differentiated Prostate Adenocarcinoma
- Poorly Differentiated Prostate Adenocarcinoma
- Prostate Adenocarcinoma with Focal Neuroendocrine Differentiation
- Prostate Ductal Adenocarcinoma
- Stage III Prostate Adenocarcinoma
- Stage II Prostate Adenocarcinoma
- Stage I Prostate Adenocarcinoma
- Well Differentiated Prostate Adenocarcinoma

Information about this concept:

- DEFINITION
- Synonym with source data
- Synonym with source data
- Synonym with source data
- Preferred_Name
- Semantic_Type
- Synonym
- Synonym
- Synonym
- Unified Medical Language System Concept Identifier

Logical Observation Identifiers, Names and Codes (LOINC)

LOINC®

Logical Observation Identifiers Names and Codes

LOINC Characteristics (1)

- Current version: 2.22 (Dec. 2007)
- Type: Controlled terminology*
- Domain: Laboratory and clinical observations
- Developer: Regenstrief Institute
- Funding: NLM
- Availability
  - Publicly available: Yes
  - Repositories: UMLS
- URL: www.regenstrief.org/loinc/loinc.htm

LOINC Characteristics (2)

- Number of
  - Concepts: 50k active codes (2.18)
  - Terms: n/a*
- Major organizing principles:
  - No hierarchical structure among the main codes
  - 6 axes
    - Component (analyte [+ challenge] [+ adjustments])
    - Property
    - Timing
    - System
    - Scale
    - [Method]
- Formalism: “DL-like”

LOINC Example

- *Sodium:SCnc:-Pt:Ser/Plas:Qn*
  [the molar concentration of sodium is measured in the plasma (or serum), with quantitative result]

Axis	Value
Component	Sodium
Property	SCnc – Substance Concen-tration (per volume)
Timing	Pt – Point in time (Random)
System	Ser/Plas – Serum or Plasma
Scale	Qn – Quantitative
Method	--

SNOMED Clinical Terms

SNOMED CT Characteristics (1)

- Current version: January 31, 2008 (2 annual releases)
- Type: Reference terminology / ontology
- Domain: Clinical medicine
- Developer: IHTSDO
- Funding: IHTSDO
- Availability
  - Publicly available: Yes* (in member countries)
  - Repositories: UMLS
- URL: http://www.ihtsdo.org/

SNOMED CT Characteristics (2)

- Number of
  - Concepts: 311,313 active concepts (Jan. 31, 2008)
  - Terms: 794,061 active “descriptions”
- Major organizing principles:
  - Utility for clinical medicine (e.g., assertional + definitional knowledge)
  - Model of meaning (incomplete)
  - Rich set of associative relationships
  - Small proportion of defined concepts (many primitives)
- Formalism: Description logics (KRSS)

SNOMED CT Top level

Hierarchy | Subtype hierarchy

- 138875005 SNOMED CT Concept
  - 362981000 qualifier value
  - 106237007 linkage concept
  - 370115009 special concept
  - 48176007 social context
  - 419891008 record artifact
  - 363787002 observable entity
  - 308916002 environment or geographical location
  - 123038009 specimen
  - 254291000 staging and scales
  - 123037004 body structure
  - 272379006 event
  - 78621006 physical force
  - 404684003 clinical finding
  - 260787004 physical object
  - 410607006 organism
  - 71388002 procedure
  - 373873005 pharmaceutical / biologic product
  - 243796009 situation with explicit context
  - 105590001 substance

SNOMED CT Example

Hierarchy | Subtype hierarchy

- 27010001 partial excision of large intestine
- 8613002 operation on appendix
  - 80146002 appendectomy
    - 82730006 incidental appendectomy
    - 49438003 appendectomy with drainage
    - 174036004 emergency appendectomy
    - 174045003 interval appendectomy
    - 6025007 laparoscopic appendectomy
    - 235313004 non-emergency appendectomy
    - 235314005 inversion appendectomy
    - 1299000 excision of appendiceal stump

Definition: Fully defined by ...

- is a
  - D partial excision of large intestine
  - D operation on appendix
- *Group*
  - method
    - D excision - action
  - procedure site - Direct
    - D appendix structure
- *Qualifiers*
  - access
    - surgical access values
  - priority
    - priorities

appendectomy - Definition

Concept Status: **Current**

- *Descriptions*
  - F appendectomy (procedure)
  - P appendectomy
  - S excision of appendix
  - U appendicectomy

Codes

- *Original SnomedId : P1-57450*
- *Read Code (Ctv3Id) : X20Wz*

Unified Medical Language System (UMLS)

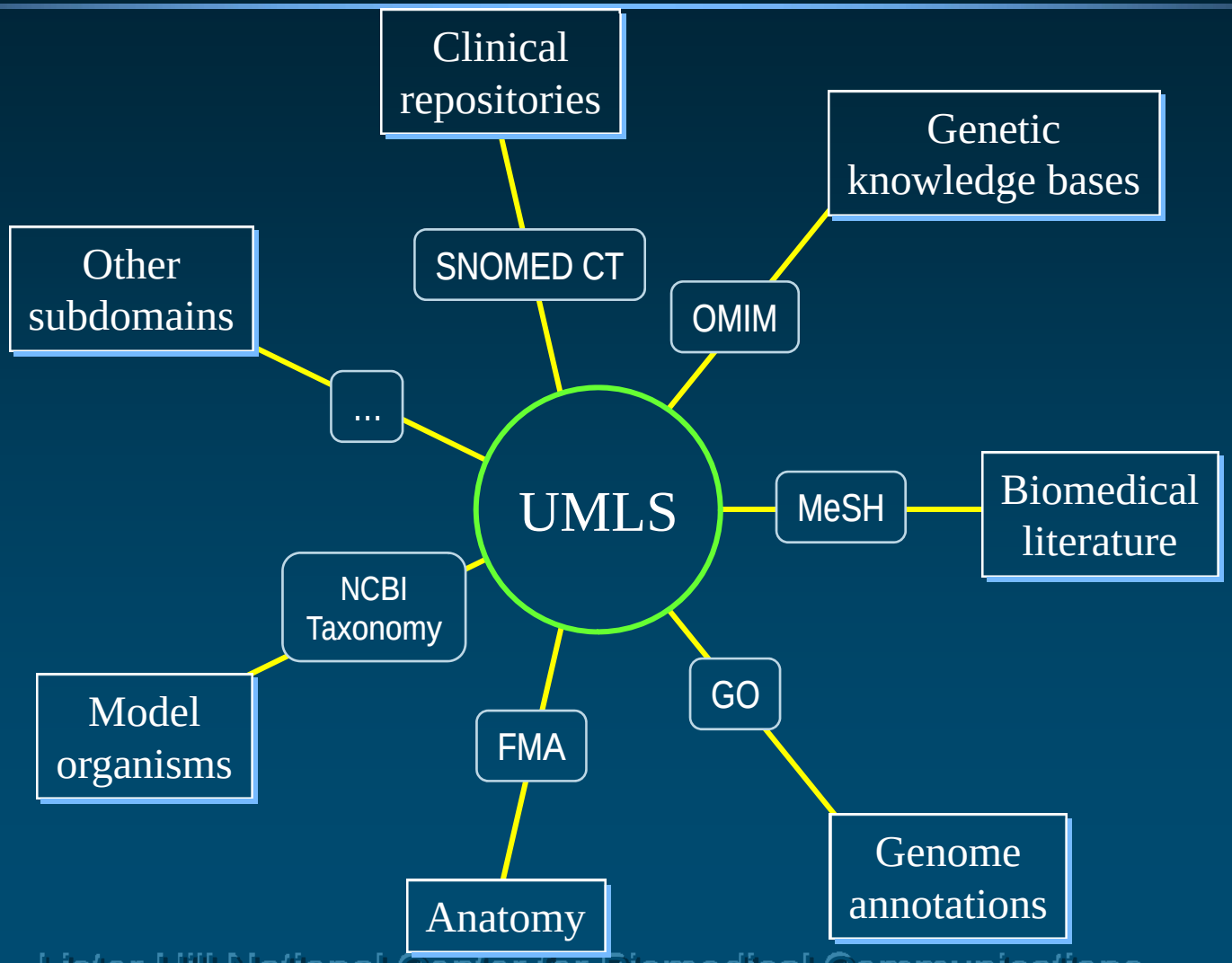
UMLS Characteristics (1)

- Current version: 2008AA (2-3 annual releases)
- Type: Terminology integration system
- Domain: Biomedicine
- Developer: NLM
- Funding: NLM (intramural)
- Availability
  - Publicly available: Yes* (cost-free license required)
  - Repositories: UMLS
- URL: http://umlsks.nlm.nih.gov/

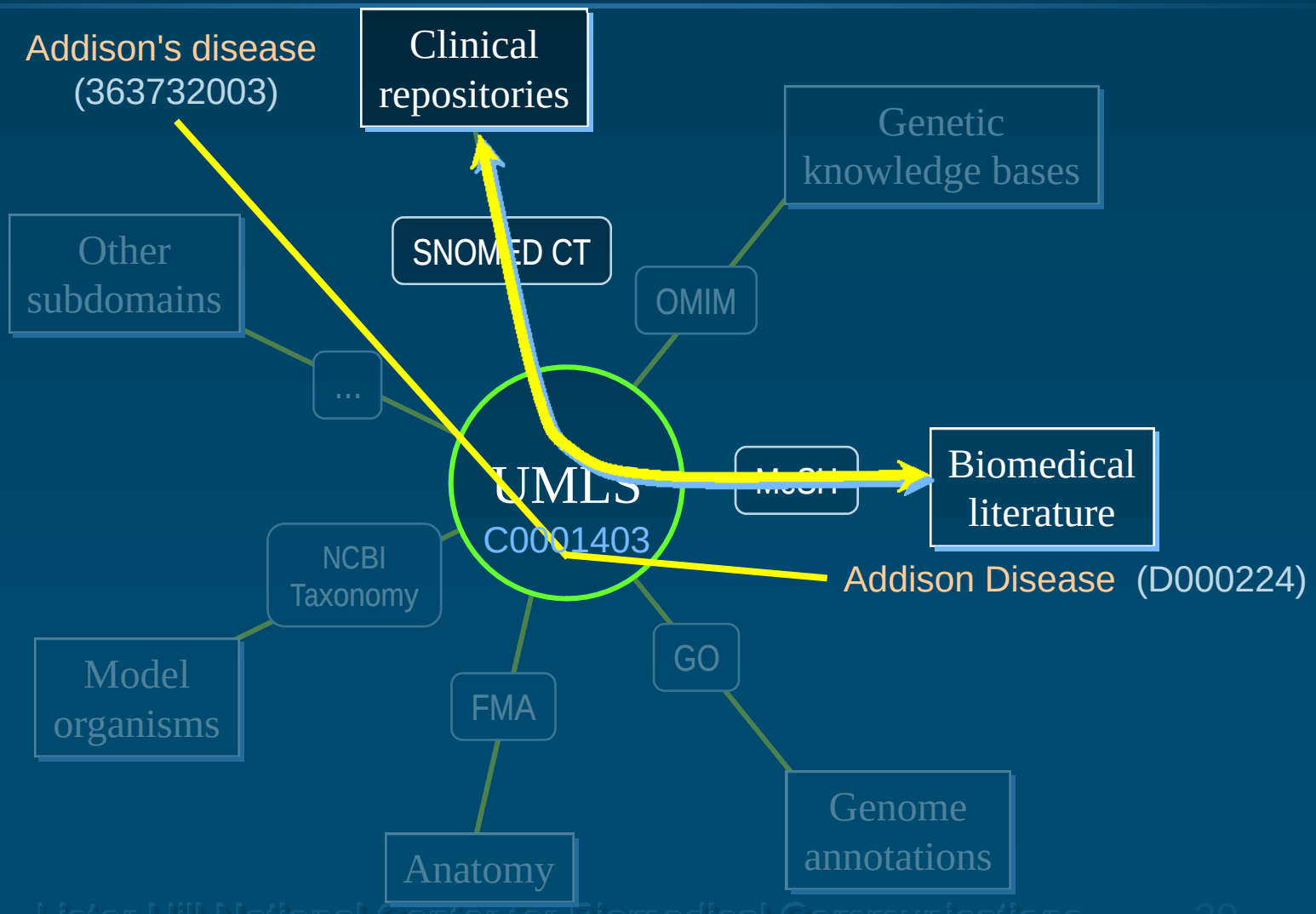
UMLS Characteristics (2)

- Number of
  - Concepts: 1.5M (2008AA)
  - Terms: ~6M
- Major organizing principles (Metathesaurus):
  - Concept orientation
  - Source transparency
  - Multi-lingual through translation
- Formalism: Proprietary format (RRF)

UMLS Integrating subdomains



Trans-namespace integration



Methods

Methods Outline

- Acquiring RDF triples
- Mapping relationships between sources (LOINC/SNOMED CT only)
- Adding entailed statements through inference rules (NCIt/SNOMED CT only)
- Querying the triple store (RDF graph)
- Comparing shared relata of NCIt/LOINC and SNOMED CT concepts

Methods Acquiring RDF triples

- SNOMED CT
  - CONCEPTID as the identifier
  - Relationships: from the RELATIONSHIP table
- NCIt
  - CODE as the identifier
  - Relationships: from the OWL file
- LOINC
  - LOINC ID as the identifier
  - Relationships: from the 6-axis structure
- UMLS
  - Link between SCUI and CUI
  - Relationship: hasCUI

Methods Acquiring RDF triples

S:Disorder of pancreas —*S:finding site*→ S:Pancreatic structure

S:Disorder of endocrine pancreas —*S:is_a*→ S:Disorder of pancreas

N:Pancreatic disorder —*N:disease assoc. anat. site*→ N:Pancreas

N:Endocrine pancreas disorder —*N:subClassOf*→ N:Pancreatic disorder

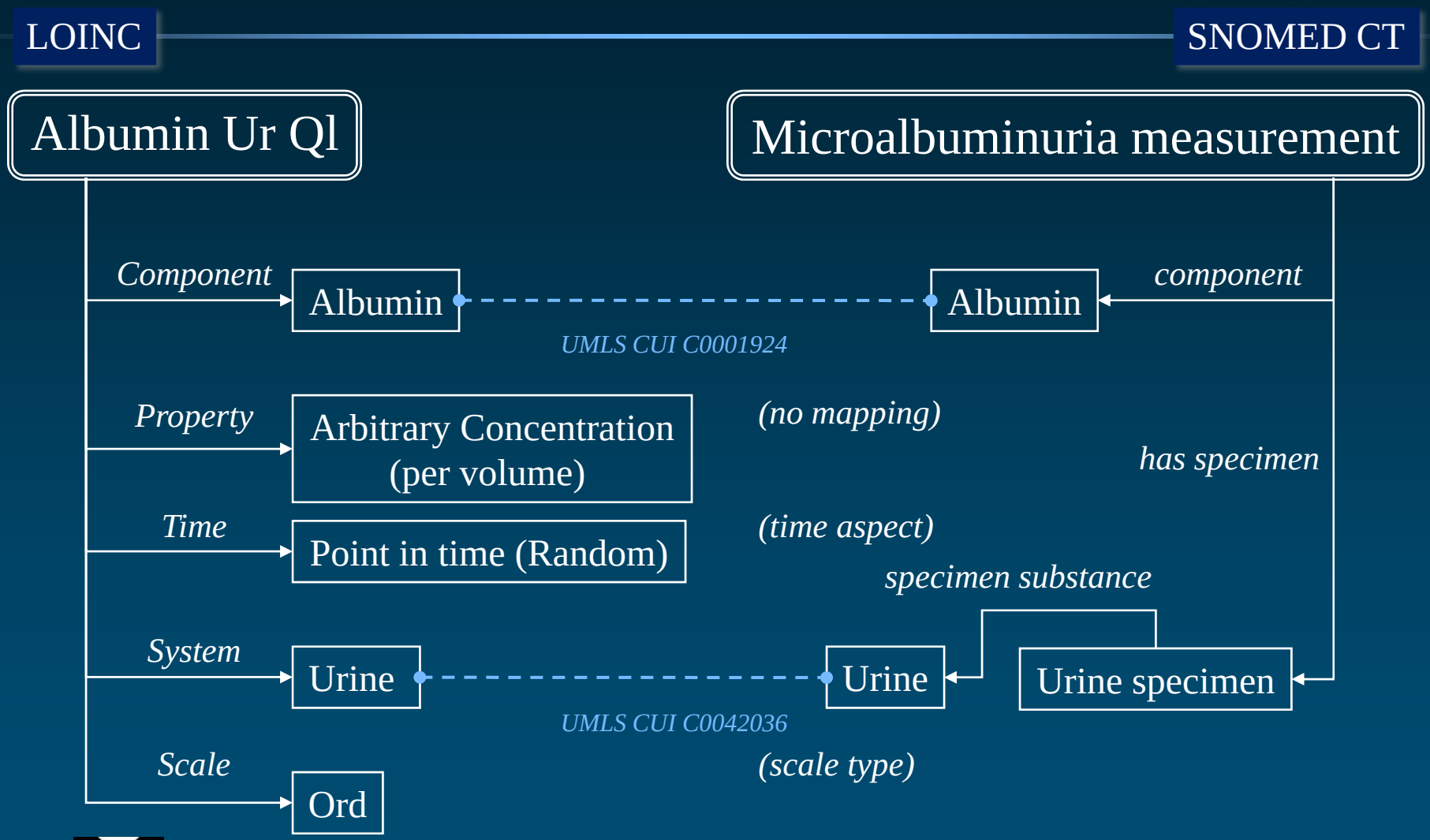
S:Disorder of pancreas —*U:hasCUI*→ ***C0030286***

N:Pancreatic disorder —*U:hasCUI*→ ***C0030286***

Methods Mapping relationships

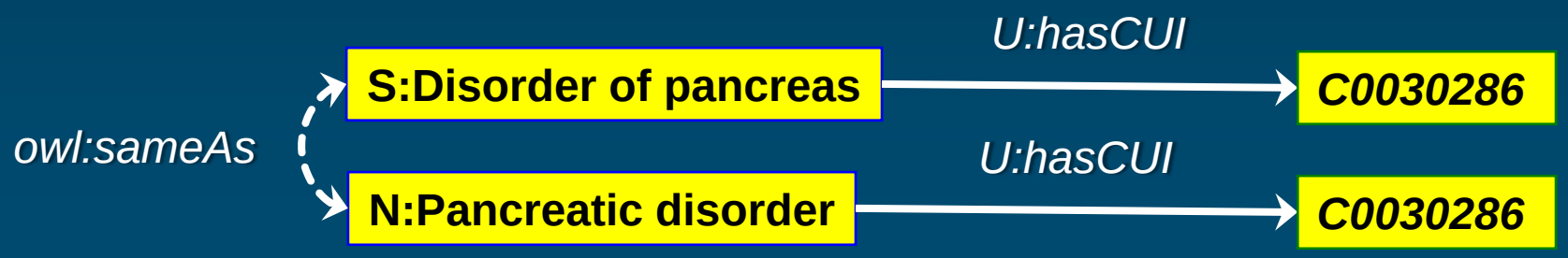
- NCIt/SNOMED CT
  - No constraints
  - Multiple correspondences in most cases
  - Shared relata allowed through any relationship
- LOINC/SNOMED CT
  - Each of the 6 LOINC relationships mapped to some relationship in SNOMED CT
    - Method axis is optional and was ignored
    - Direct mappings
    - Mappings through other types

LOINC/SNOMED CT relationship mapping

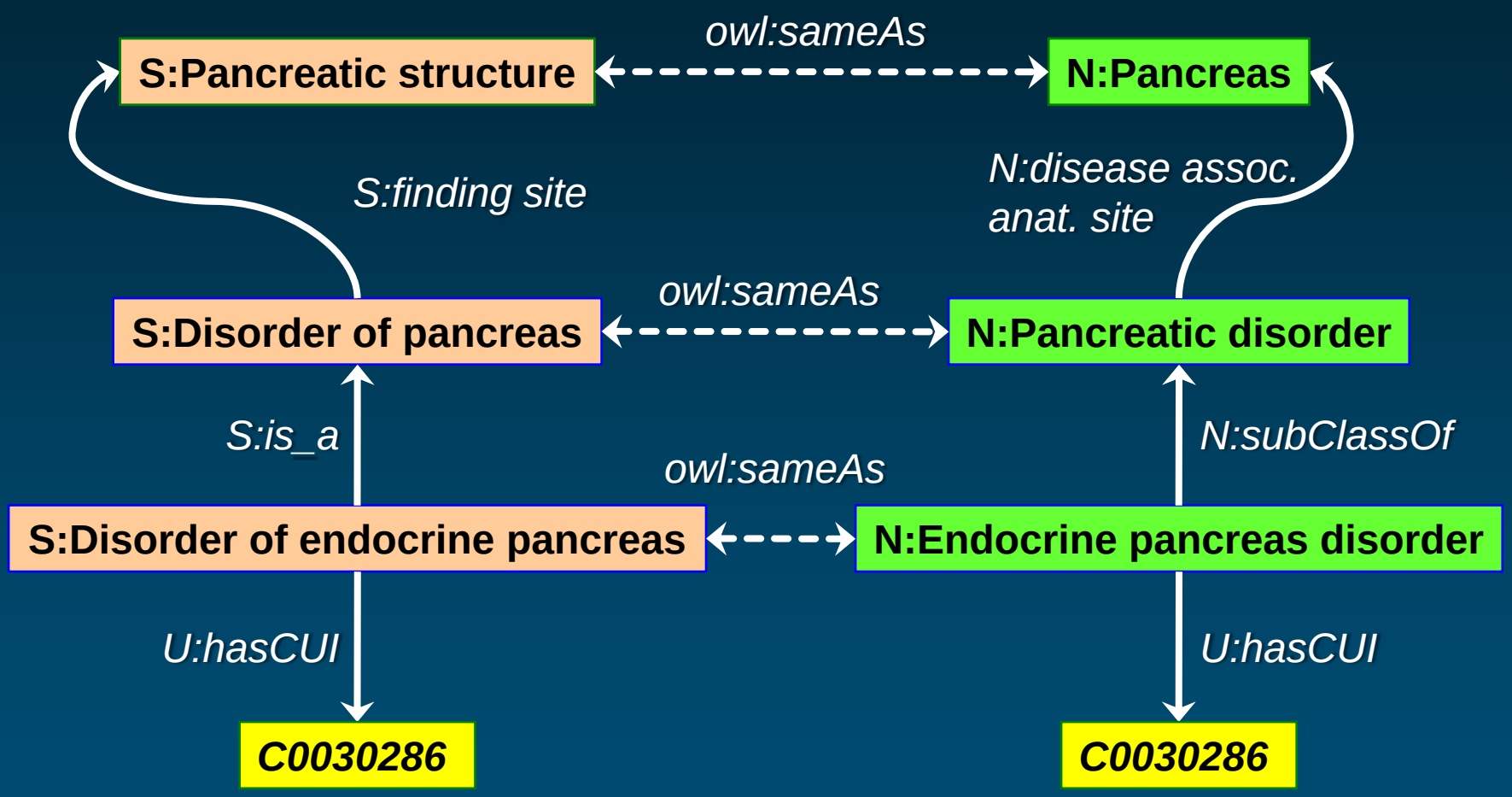


Methods Inference rules

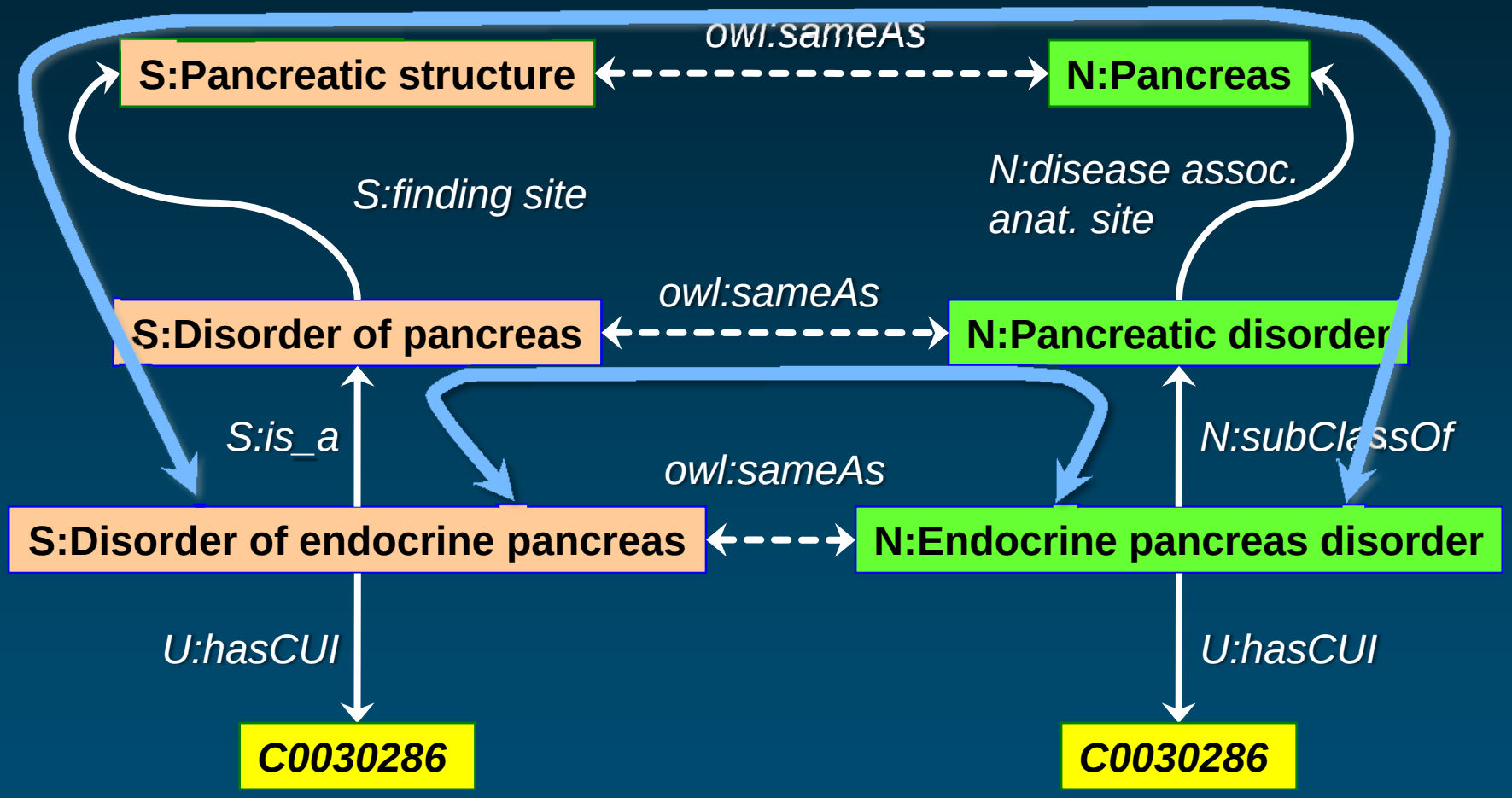
- Triple store: mulgara™ http://www.mulgara.org/
- Could not apply the rules for RDFS semantics on a large, heavily hierarchical graph
- Created one rule for the equivalence of concepts through UMLS
  - 2 concepts are equivalent if they have the same UMLS CUI



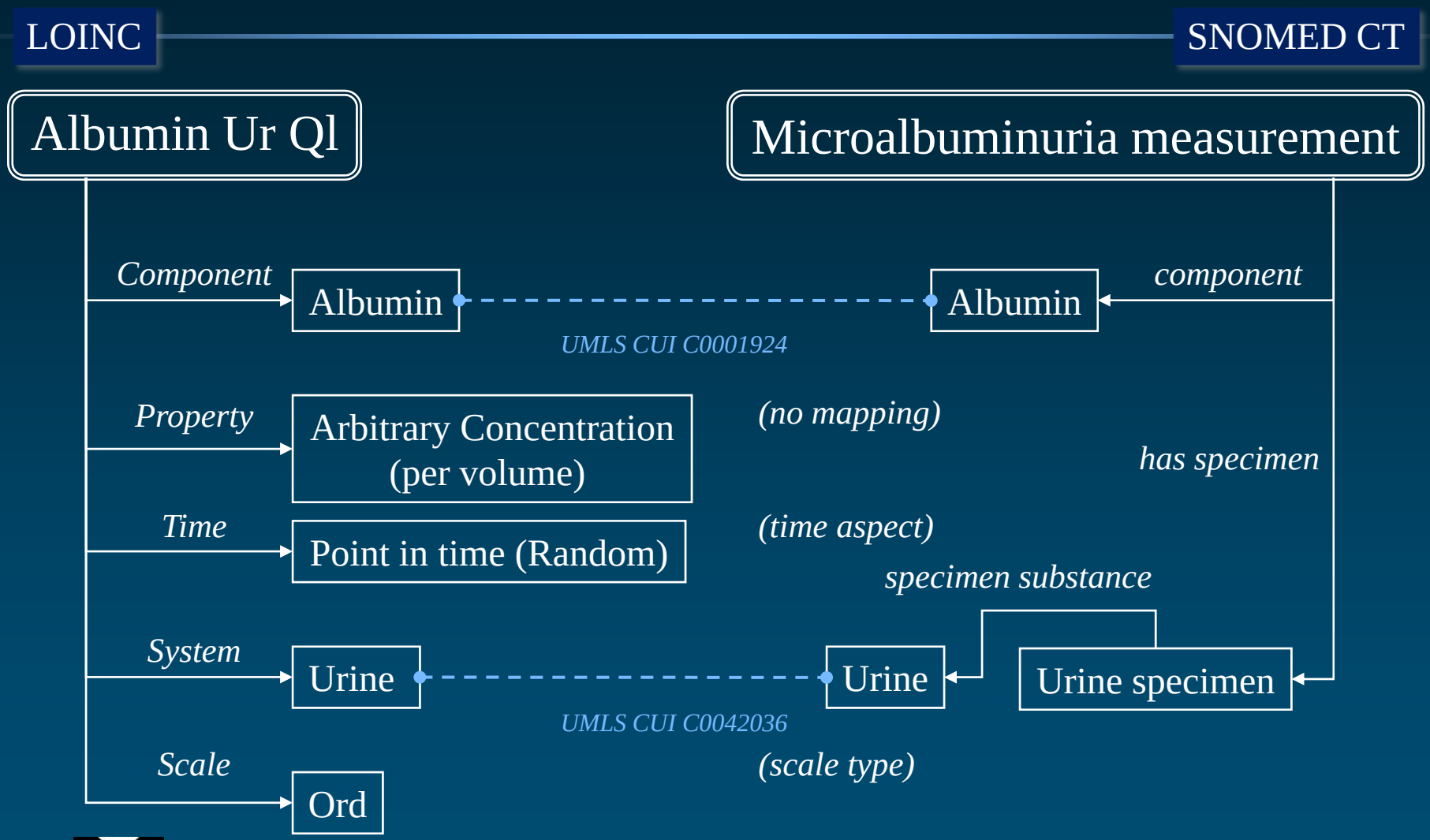
Methods RDF graph



Methods Querying the RDF graph



LOINC/SNOMED CT relationship mapping



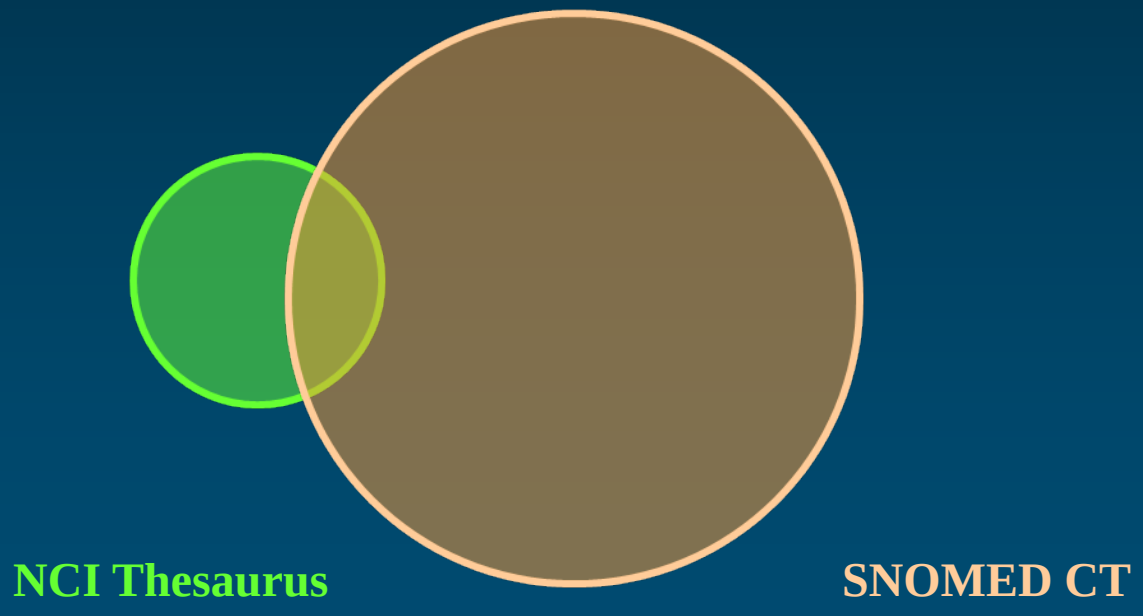
Methods Querying the RDF graph

- Comparing the shared relata of concepts
  - For various kinds of relationships (isa, part of, associative relationships)
  - Directly or recursively
- ITQL queries for each pair of equivalent concepts and for each type of relationship queried
  - Functionally equivalent to SPARQL
  - “walk” function compensates for the absence of transitive closure rule

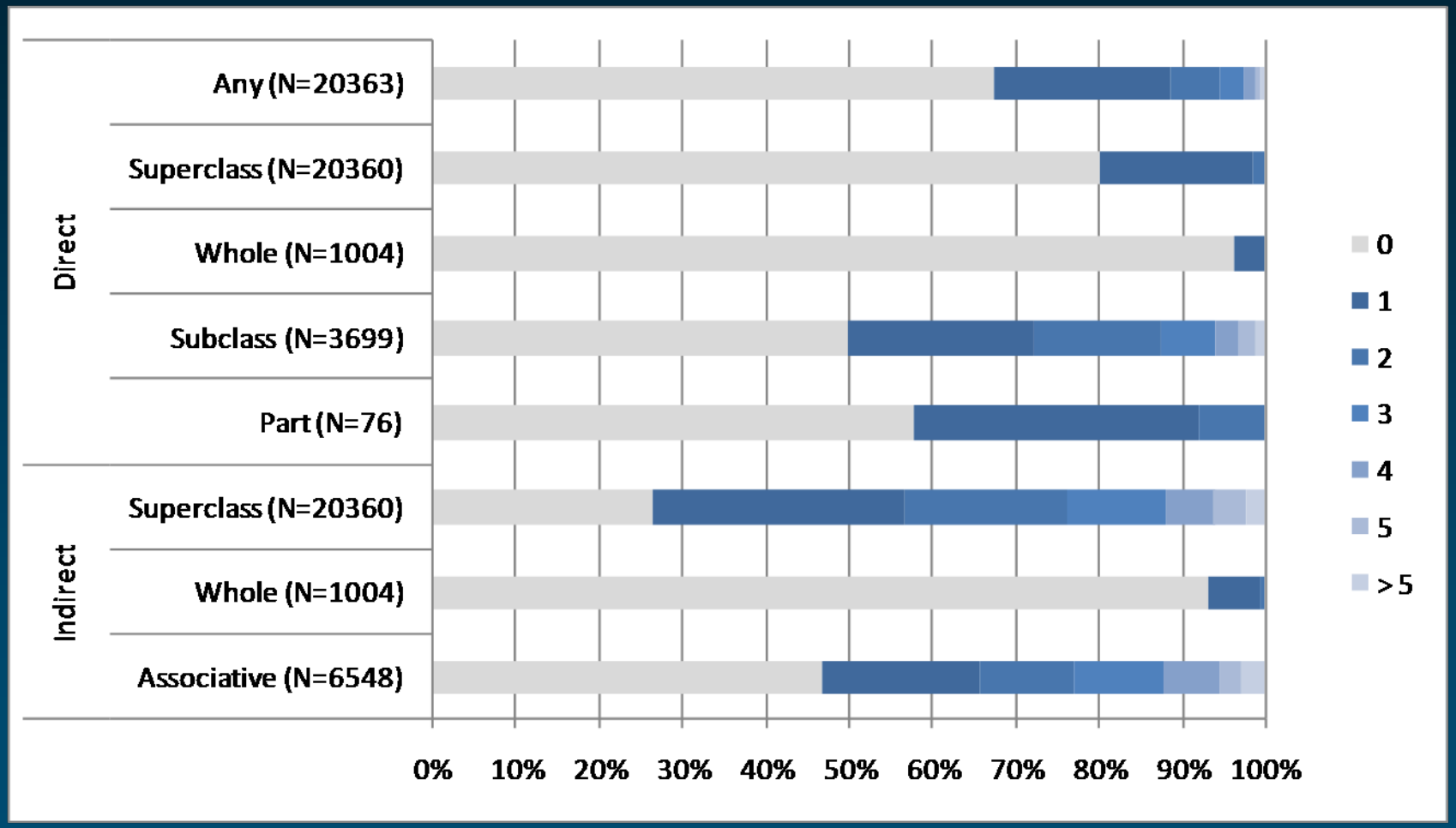
Results 1

Results Overlap NCIt/SNOMED CT

- 20,369 pairs of equivalent concepts
  - 23.9% of all (active) NCIT concepts
  - 6.3% of all (active) SNOMED CT concepts



Results Shared relata (quantitative)



Results Findings

- Relatively small proportion of relata in common between equivalent concepts from NCIt and SNOMED CT
- Large number of primitive concepts in NCIt and SNOMED CT (70-80%)
- Insufficient for effectively comparing definitions
  - Could not be used for validating the mapping provided by the UMLS

Results 2

LOINC®

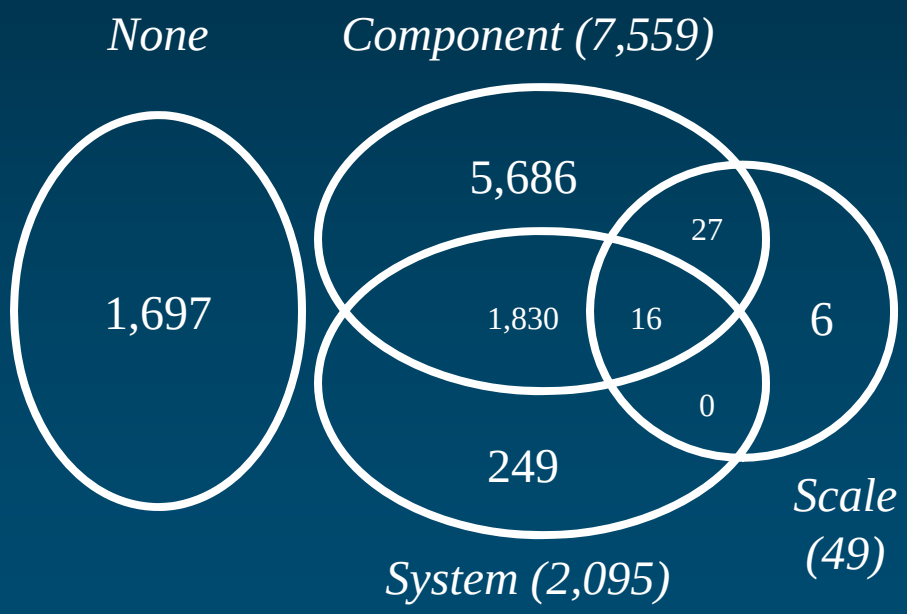
Logical Observation Identifiers Names and Codes

Results Mapping LOINC/SNOMED CT

- Laboratory procedures
  - 34,767 in LOINC (simple lab tests)
  - 9,511 in SNOMED CT
- LOINC "parts"
  - 44,314 total
  - 13,794 are used as value for the 6 main axes
  - 4,501 (33%) are mapped to SNOMED CT through the UMLS

Results Mapping LOINC/SNOMED CT

- Most LOINC concepts cannot be mapped to SNOMED CT
- Most SNOMED CT concepts are only partially modeled
- Most LOINC lab concept are described with more specific concepts than in SNOMED CT
- Recall vs. precision issues



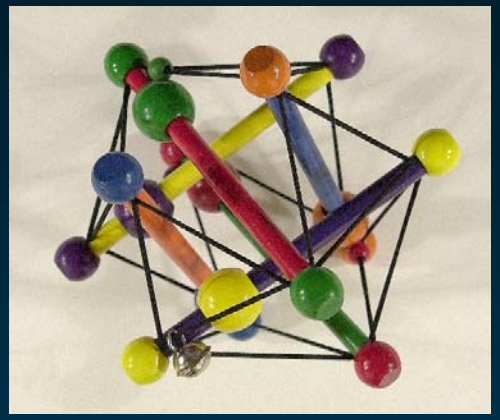
Conclusions

Terminology integration and formal definitions ☹

- Issues
  - Large number of primitive concepts in NCIt and SNOMED CT (70-80%)
  - Difference in granularity between LOINC and SNOMED CT
  - Differences in modeling
- Insufficient for effectively comparing definitions
  - Possible use as part of a semi-automatic approach

RDF approach to comparing concepts ☺

- RDF is suitable to the comparison of terminologies
  - Large terminologies
  - One is not available in OWL
- Mulgara supported the load
  - Except for applying RDFS rules
- RDF approach
  - No *ad hoc* programming necessary
  - Well adapted to recursive traversal of relationships (notoriously difficult with relational databases)



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