



International Conference May 31 - June 2, 2008, Phoenix, Arizona, USA

KR-MED 2008



June 1, 2008

# Comparing SNOMED CT and the NCI Thesaurus through Semantic Web Technologies



*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



## 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

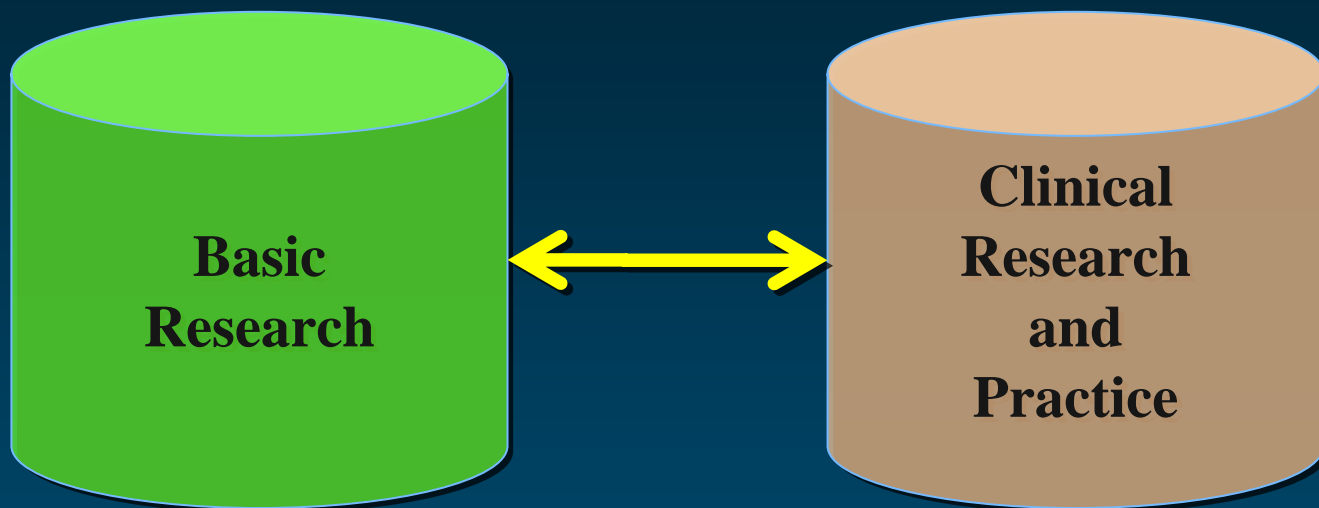
##### OVERVIEW

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."

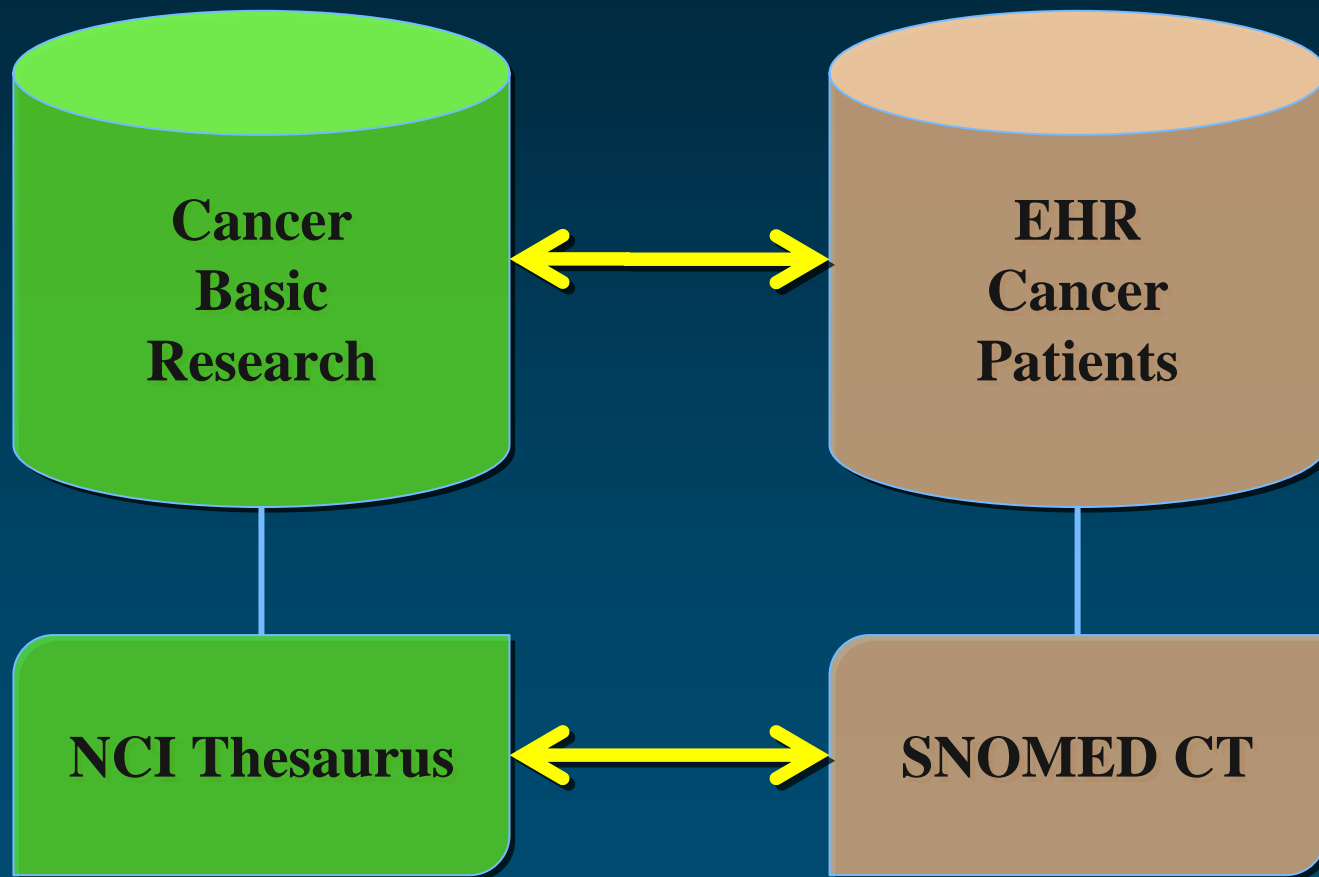
Scientists are increasingly aware that this bench-to-bedside approach to translational research is really a two-way street. Basic scientists provide clinicians with new tools for use in patients and for assessment of their impact, and clinical researchers make novel observations about the nature and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



# Motivation Translational research

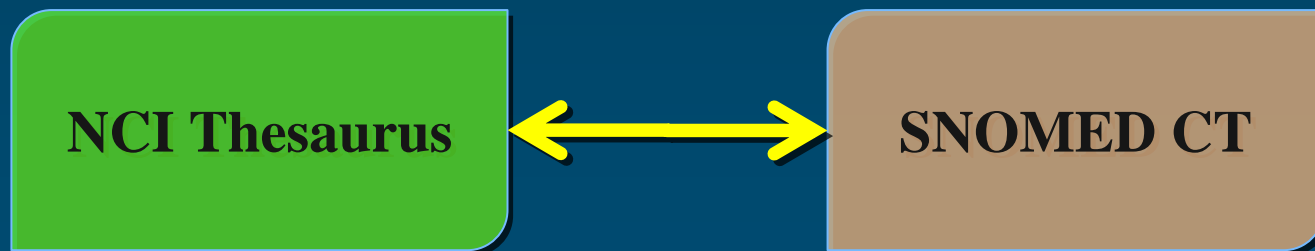


# Terminology and translational research



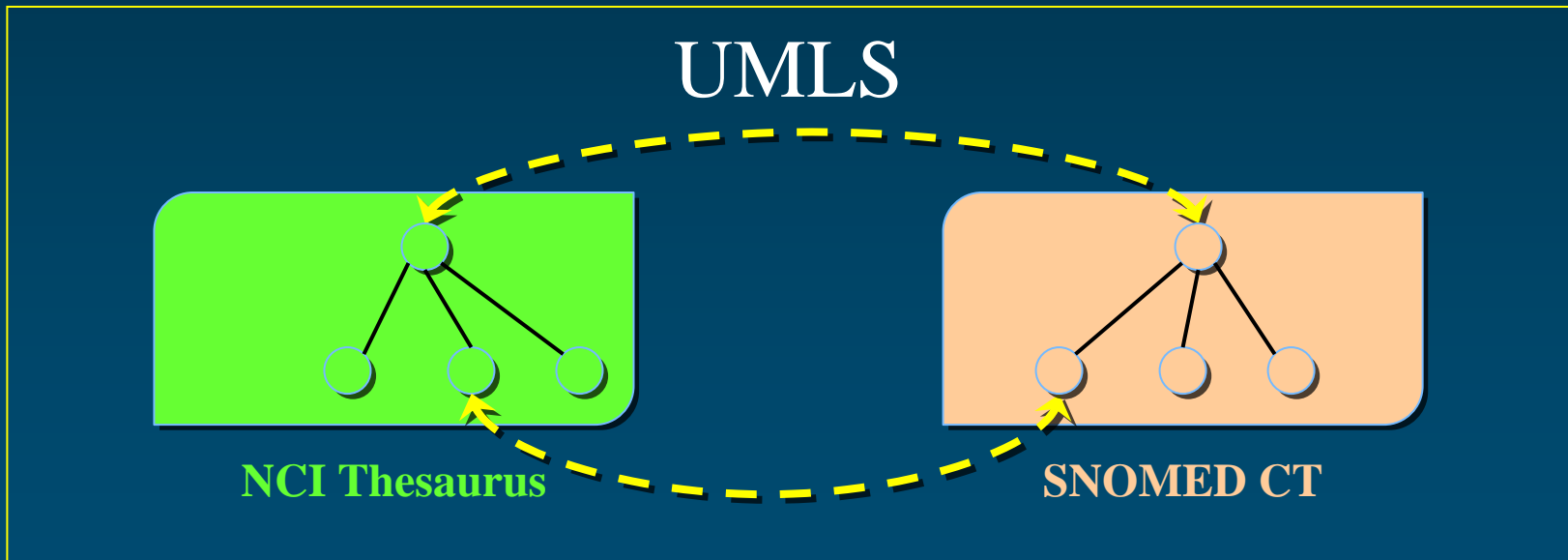
# Objectives

- ◆ To evaluate interoperability between data repositories for cancer research (basic/clinical) at the terminological level
  - To compare the formal definitions of NCI Thesaurus (NCIt) and SNOMED CT concepts
  - Evaluate Semantic Web technologies in this context



# Hypotheses

- ◆ Equivalent concepts should have similar formal definitions
  - 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
  - Both NCIIt and SNOMED CT can be easily converted to RDF



# Methods Outline

- ◆ Acquiring RDF triples
- ◆ Adding entailed statements through inference rules
- ◆ Querying the triple store (RDF graph)
- ◆ Comparing shared relata of NCI and SNOMED CT concepts



# Methods Acquiring RDF triples

## ◆ SNOMED CT

- CONCEPTID as the identifier
- Relationships: from the RELATIONSHIP table

## ◆ NCI<sub>t</sub>

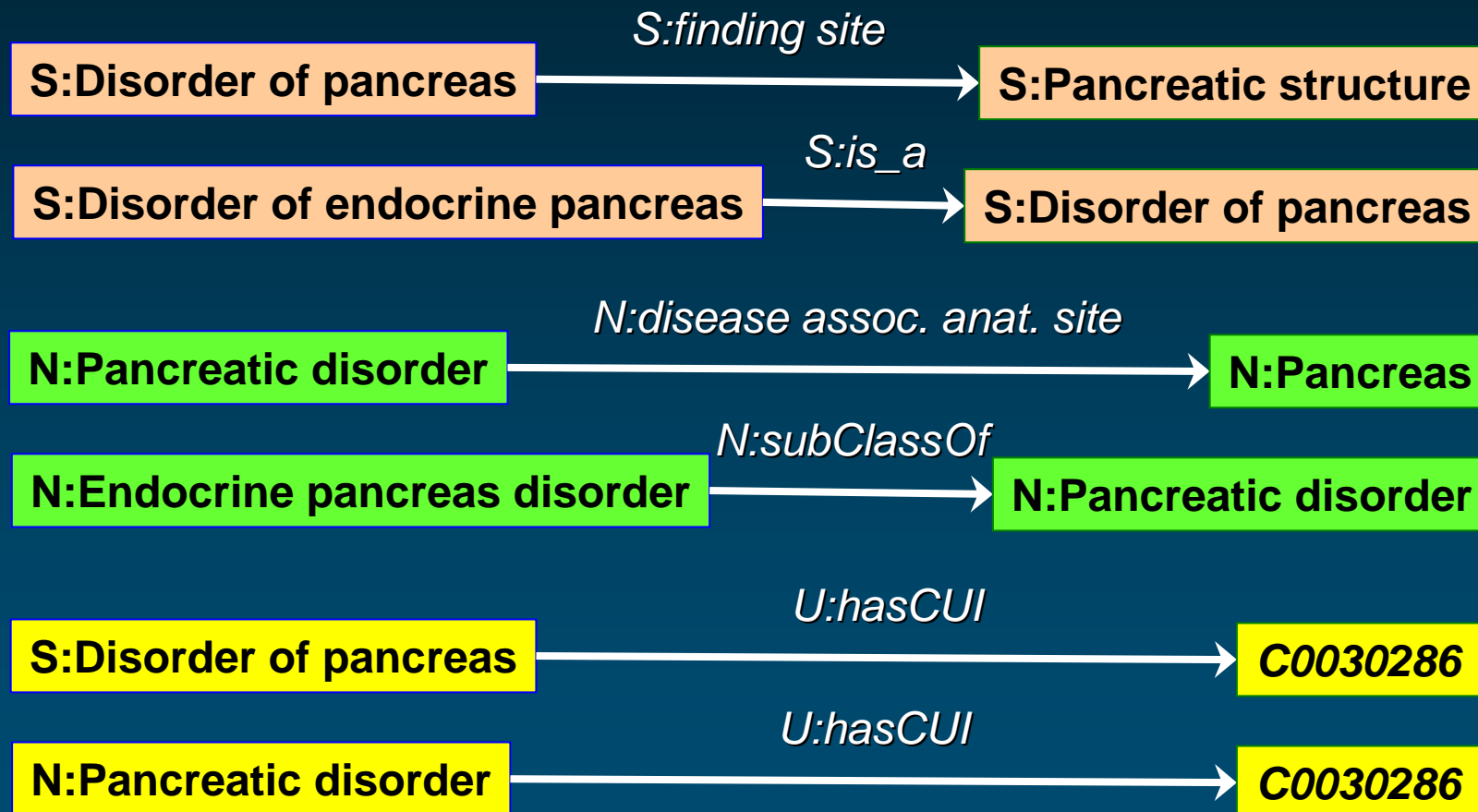
- CODE as the identifier
- Relationships: from the RELATIONSHIP table

## ◆ UMLS

- Link between SCUI and CUI
- Relationship: hasCUI

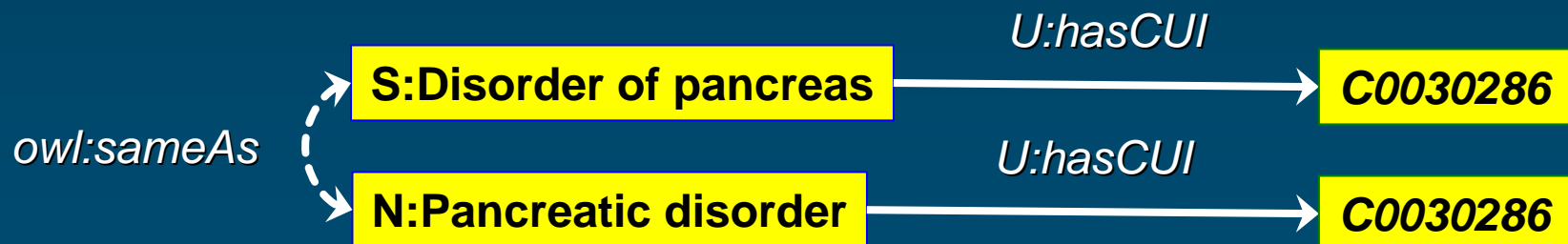


# Methods Acquiring RDF triples

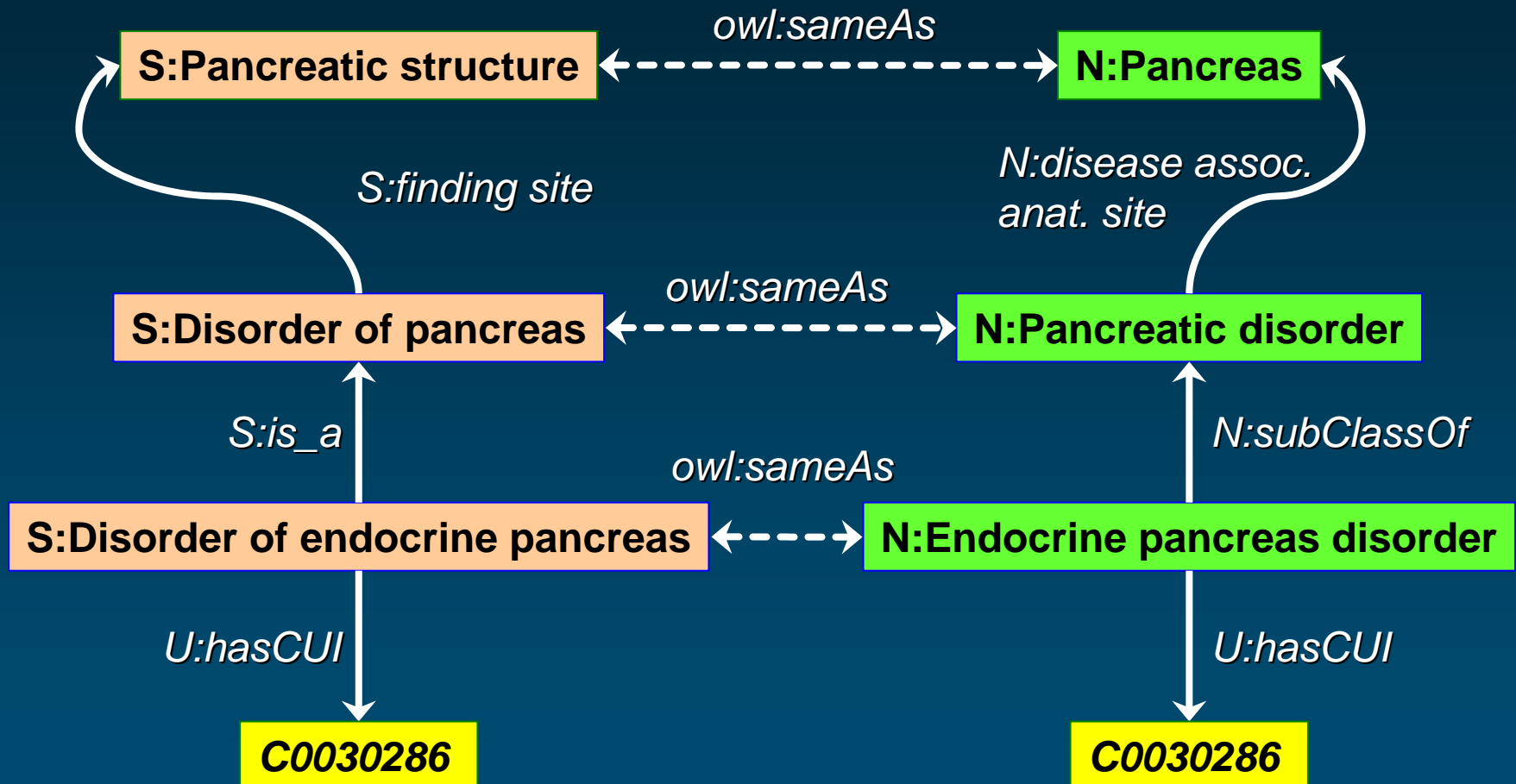


# Methods Inference rules

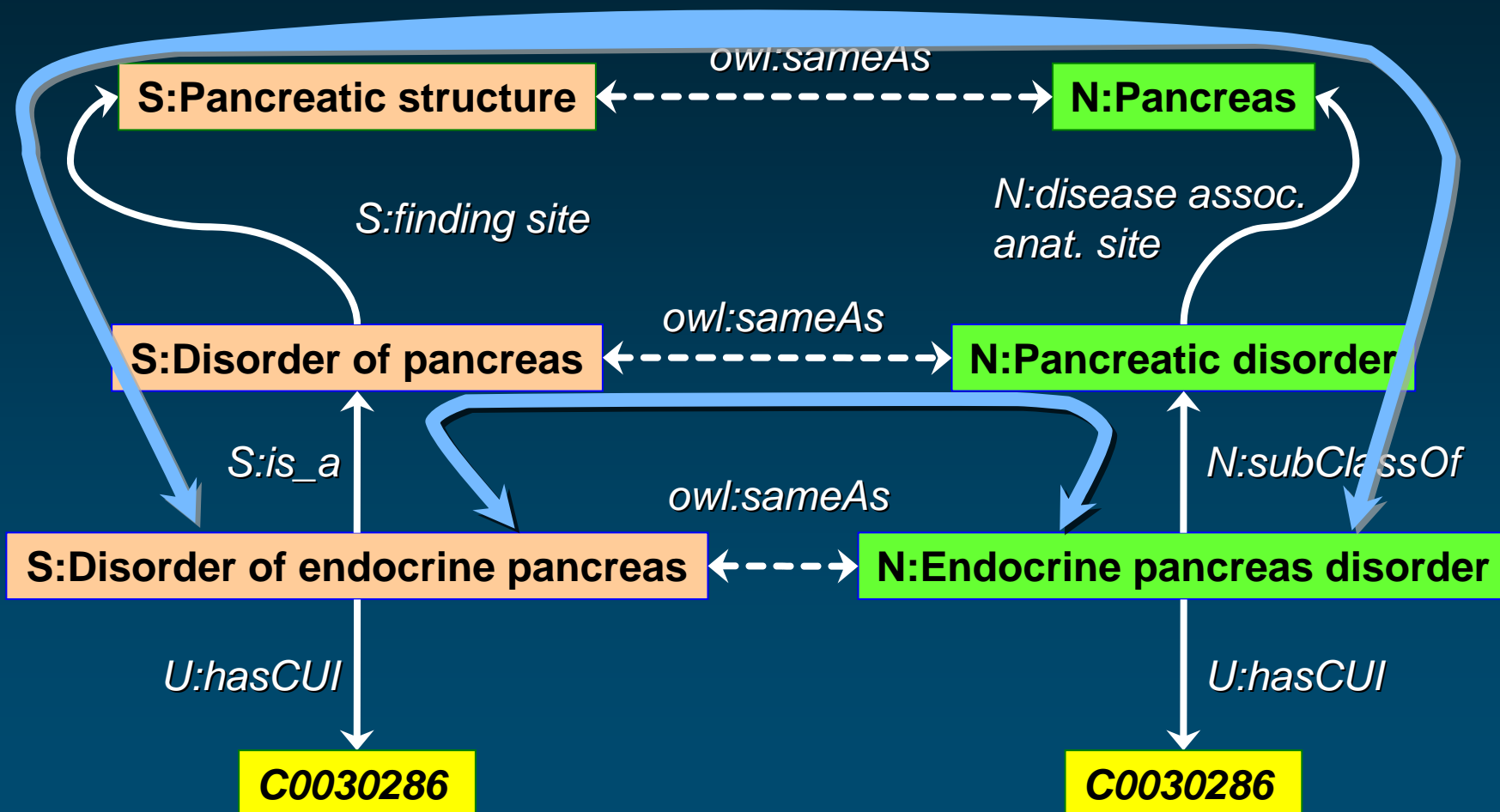
- ◆ Triple store: mulgara™ <http://www.mulgara.org/>
- ◆ Could not apply the rules for RDFS semantics on a large, heavily hierarchical graph
- ◆ Created one rule for the equivalence of concepts through UMLS
  - 2 concepts are equivalent if they have the same UMLS CUI



# Methods RDF graph



# Methods Querying the RDF graph

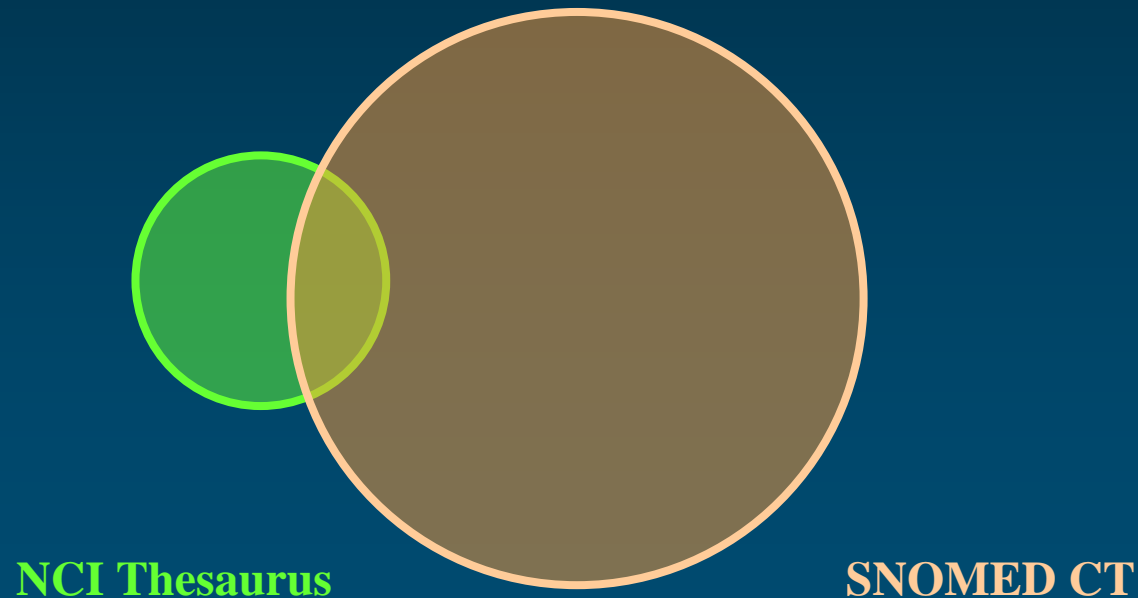


# Methods Querying the RDF graph

- ◆ 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

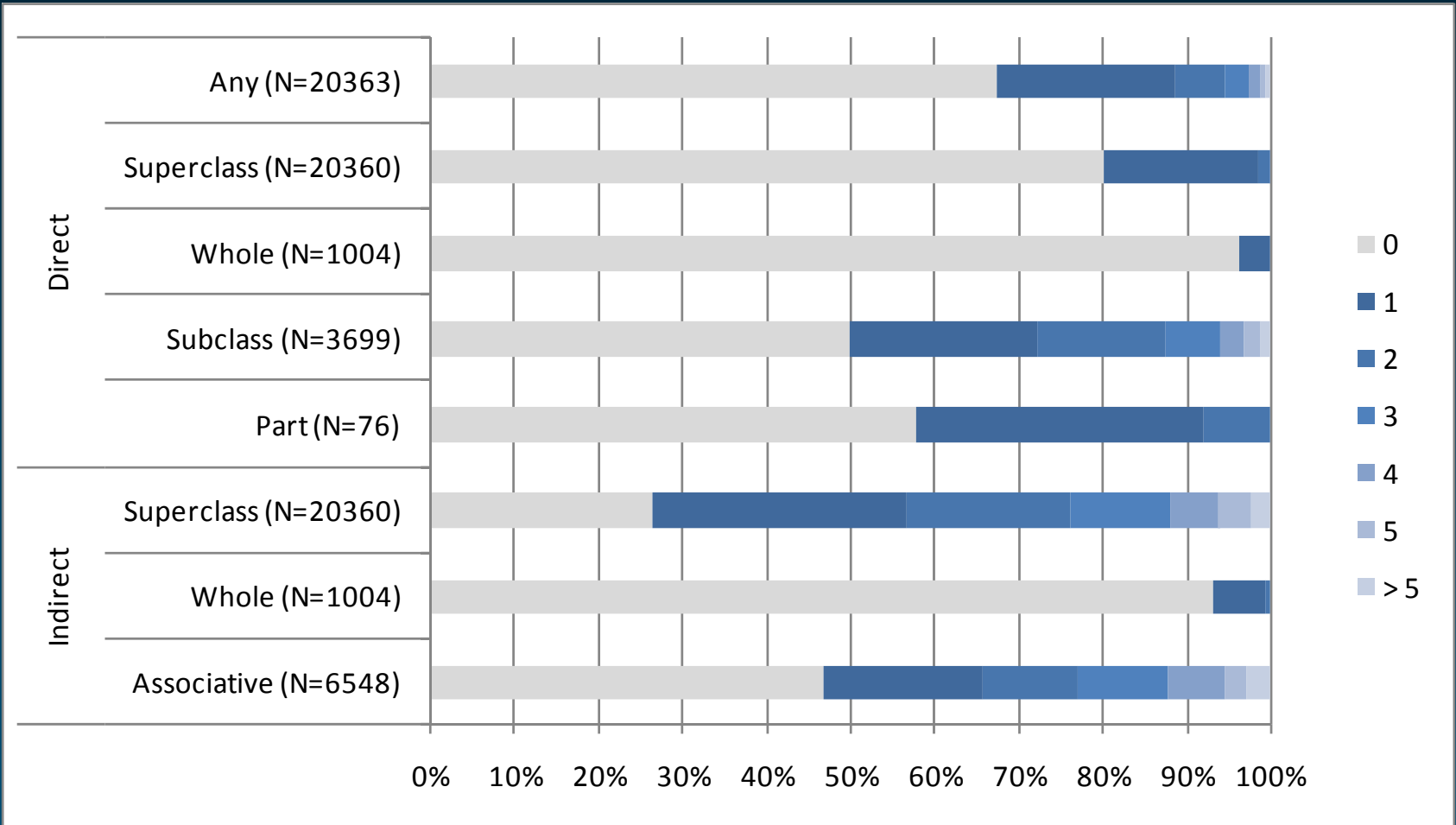
# Results Overlap NCIt/SNOMED CT

- ◆ 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)

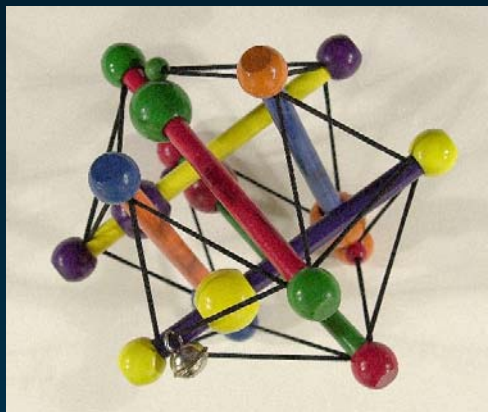


# Comparing formal definitions ☹️

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

# 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](mailto:olivier@nlm.nih.gov)

Web: [mor.nlm.nih.gov](http://mor.nlm.nih.gov)



*Olivier Bodenreider*

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