
# caCDE-QA: A quality assurance platform for cancer study common data elements

Guoqian Jiang, MD, PhD
Mayo Clinic

NCI ITCR Monthly PIs Meeting
March 6, 2015

# Introduction

- Semantic interoperability among terminologies, data elements, and information model is fundamental and critical for sharing information from the scientific bench to the clinical bedside and back among systems.
- Domain-specific Common Data Elements (CDEs) are emerging as an effective approach to standards-based clinical research data storage and retrieval and have been broadly adopted.

# Introduction

- National Cancer Institute (NCI) created the Cancer Data Standards Repository (caDSR) based on the ISO/IEC 11179 standard for metadata repositories.
- In the ISO/IEC 11179, a *data element* is defined as a unit of data for which the definition, identification, representation and permissible values are specified by means of a set of attributes.
- The binding of controlled terminology provides the basis for semantic scaling of the CDEs.

# Challenges

- The potential of terminology-based scaling has not yet been fully explored.
- In particular, there is a very limited toolbox at present for quality assurance (QA) of meta-data registered in such a repository like the caDSR.
- CDE content errors can have a significant negative impact on downstream CDE uses.

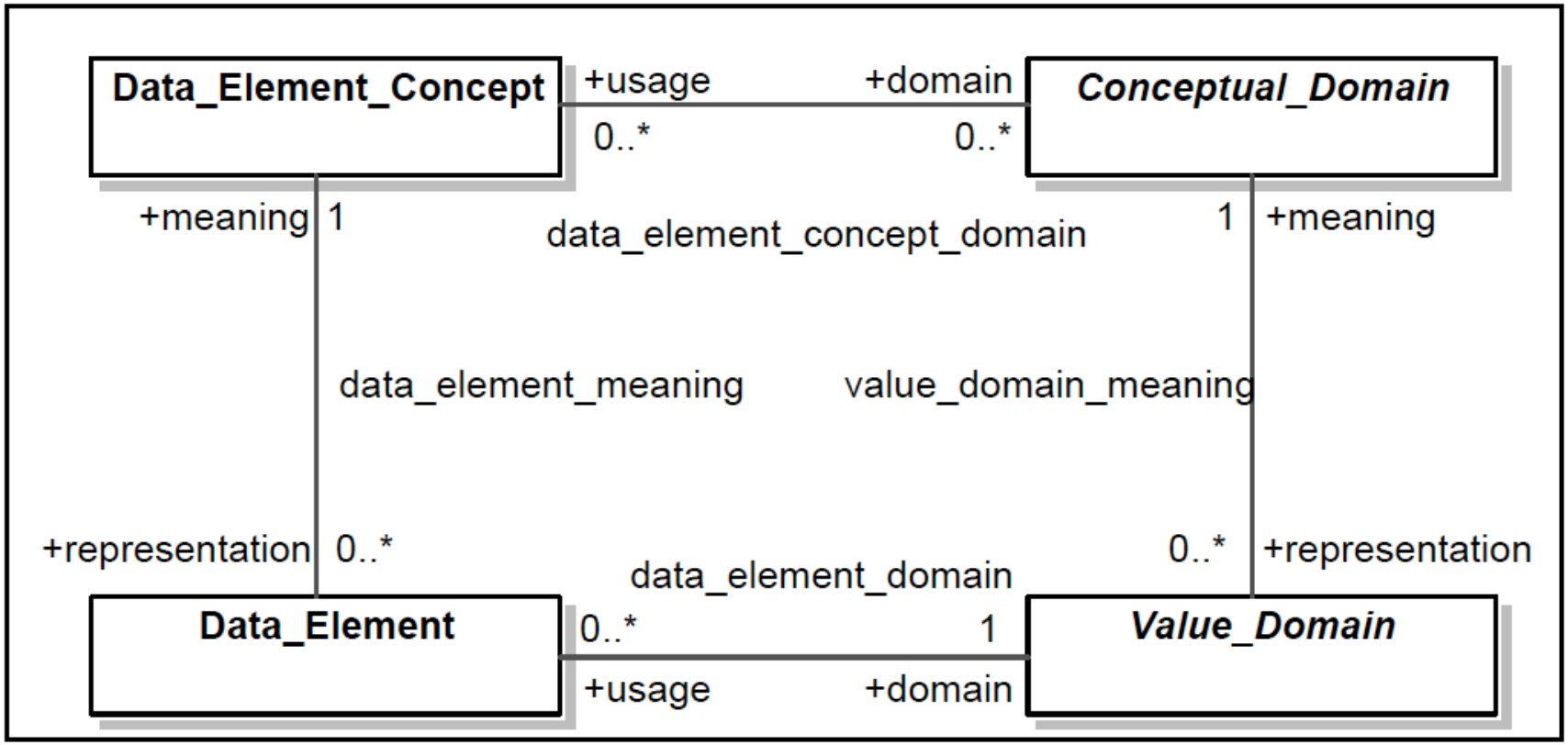

 

# Specific Aims

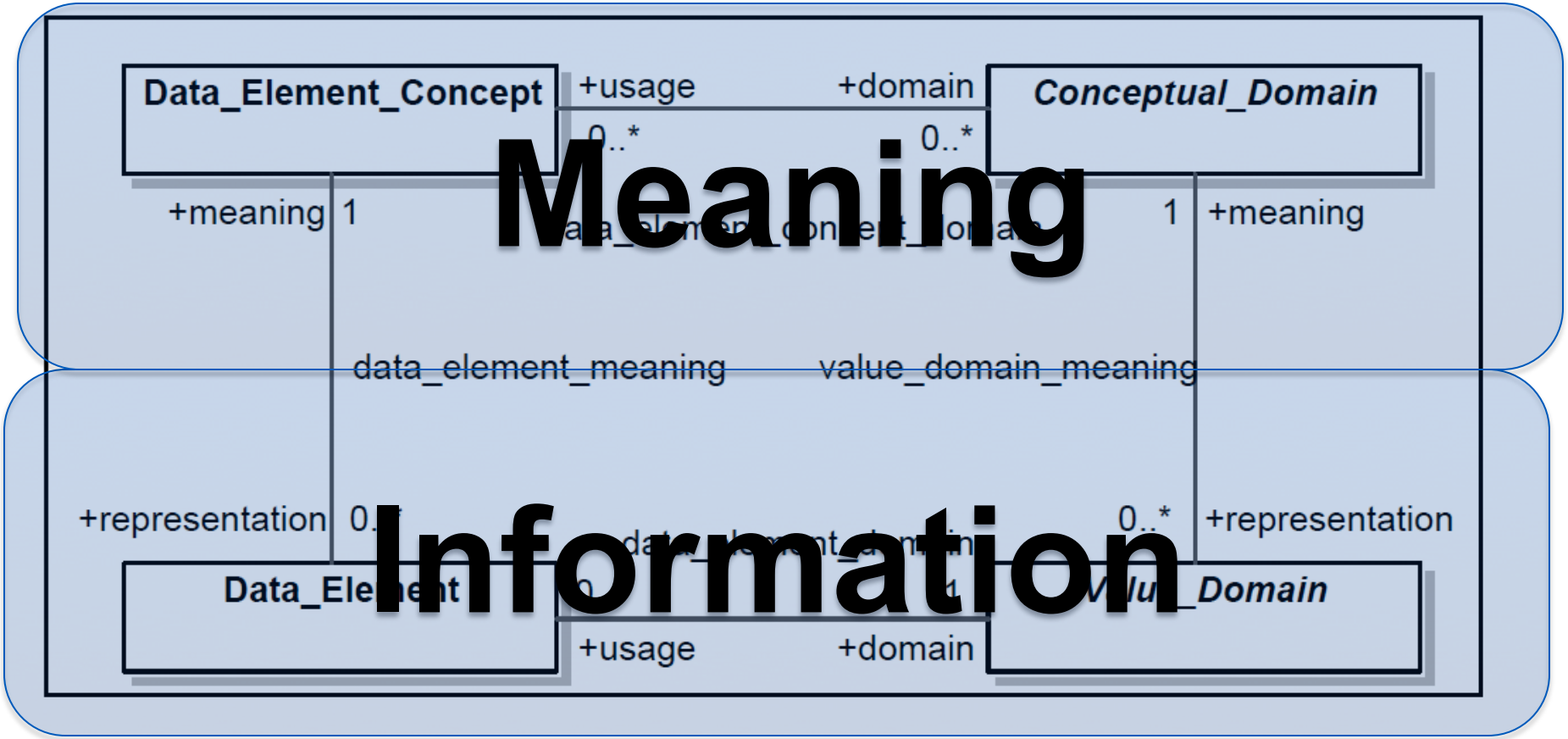
- Aim 1: To develop a suite of QA tools for validation and harmonization of cancer study CDEs;
  - UMLS Semantic Network-based approaches
  - Semantic Web-based approaches
- Aim 2: To apply the QA tools to audit experimental cancer study CDEs represented in a semantic web framework;
  - NCI caDSR
  - Preferred sets of CDEs from TCGA data dictionary
- Aim 3: To deploy and evaluate a QA web-portal for collaborative CDE review and harmonization.
  - Specification for Standard CDE Services

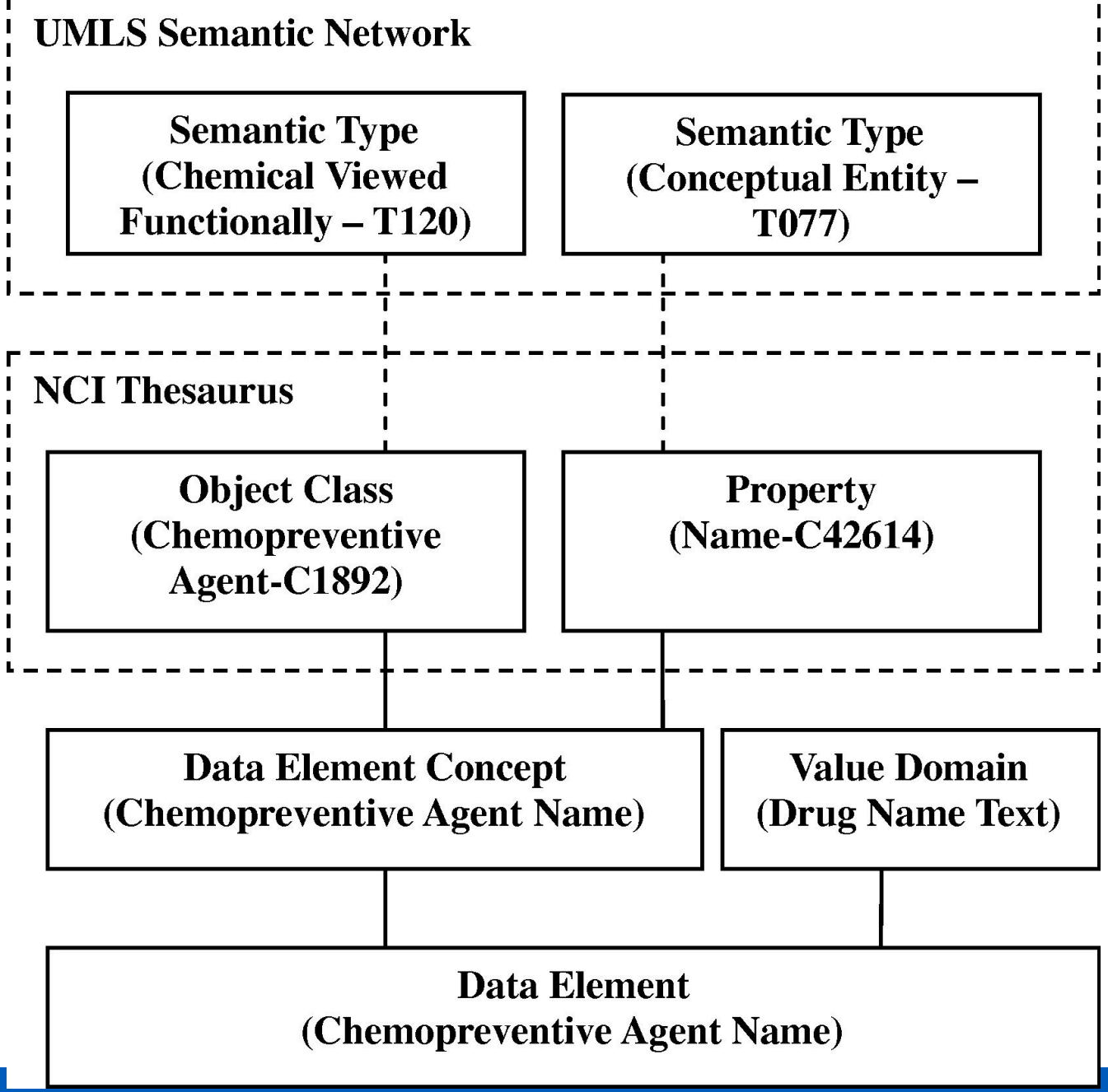

 

# High-level data description meta-model in ISO 11179 specification



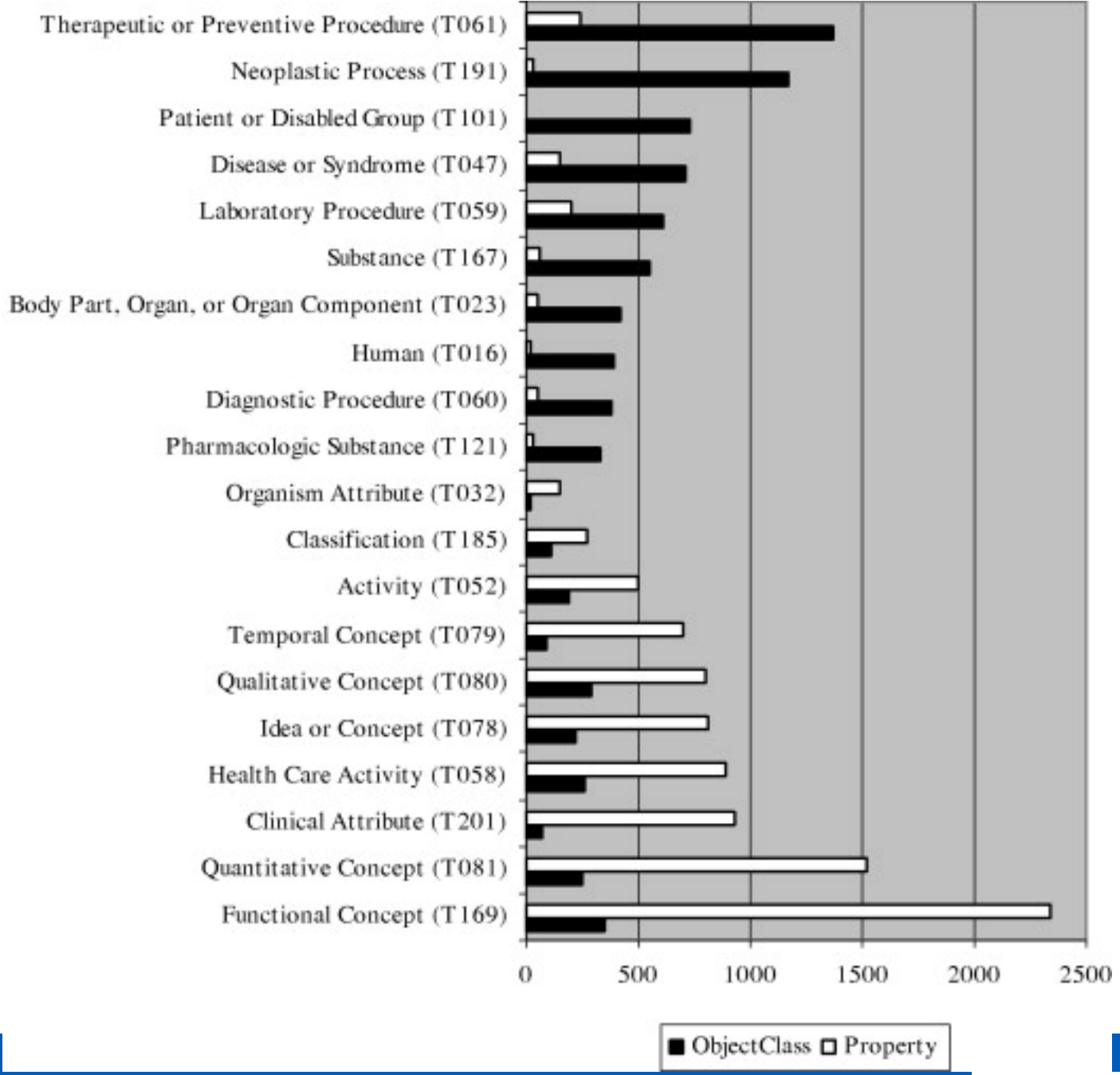
# Information and Meaning



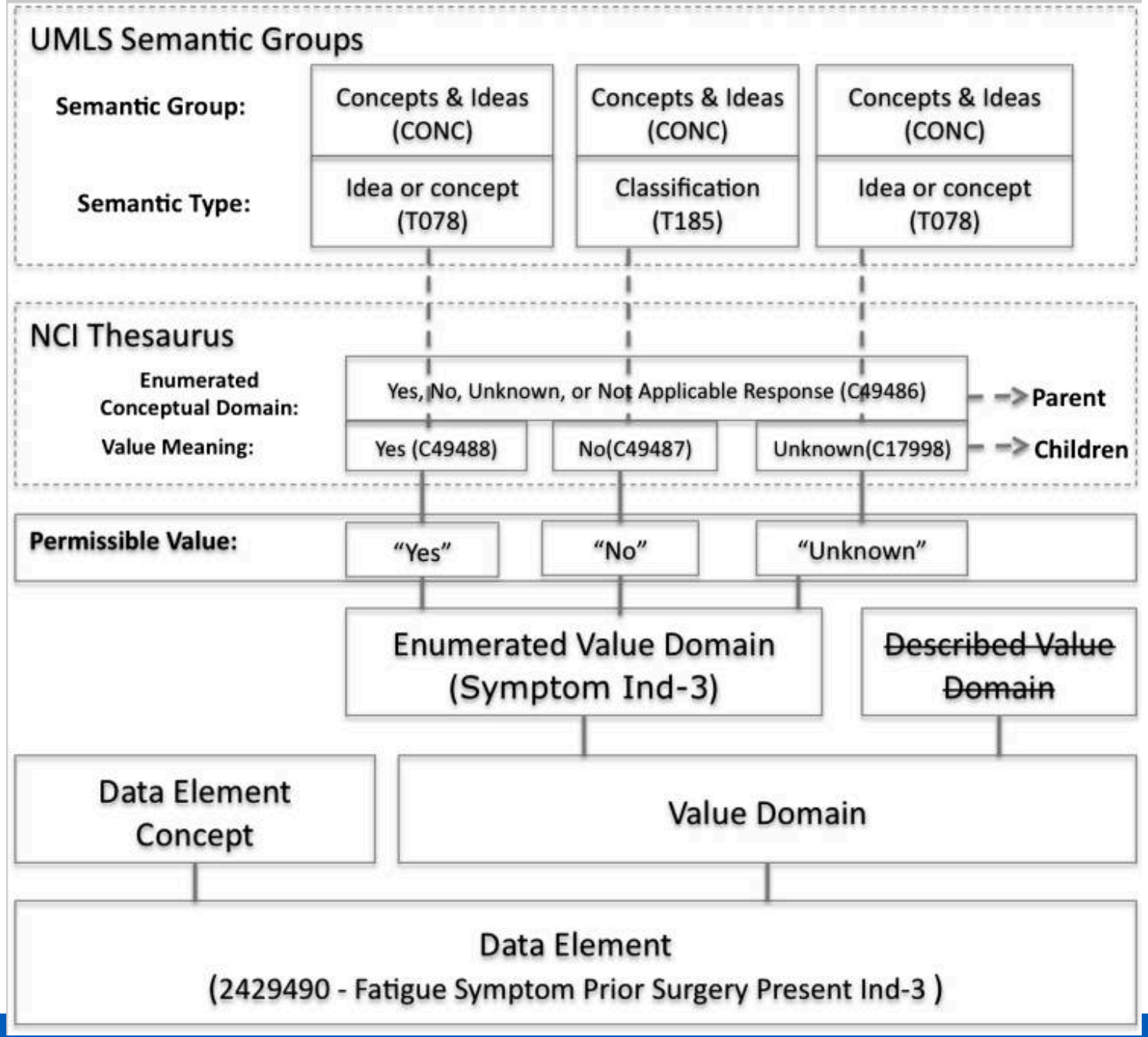



Jiang G, et al. JBI, 2011

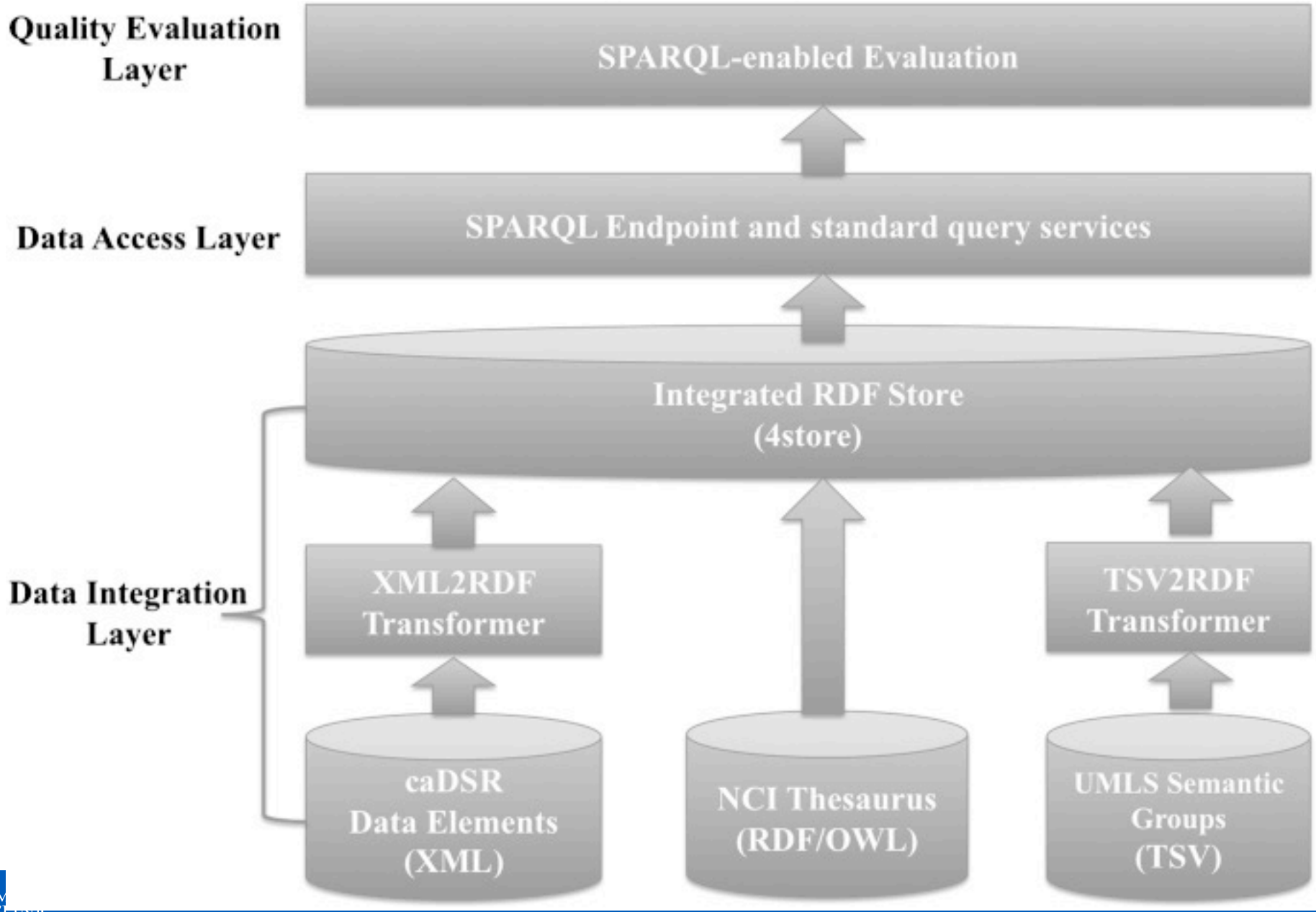

 




Jiang G, et al. JBI, 2011




Jiang G, et al. JAMIA, 2012




Jiang G, et al. JAMIA, 2012

**Table 3**

Permissible values of the first example from the data elements identified with inconsistent codes (highlighted in bold and italic)

| Valid value | Value meaning | NCIt code (meaning concept) | NCIt concept label | Semantic type | Type code | Semantic group | Group code |
|---|---|---|---|---|---|---|---|
| Palpatation | Palpation | C16950 | Palpation | Diagnostic Procedure | T060 | Procedures | PROC |
| CT | Computed Tomography | C17204 | Computed Tomography | Diagnostic Procedure | T060 | Procedures | PROC |
| Chest x-ray | Chest Radiography | C38103 | Chest Radiography | Diagnostic Procedure | T060 | Procedures | PROC |
| Spiral CT | Spiral CT | C20645 | Spiral CT | Diagnostic Procedure | T060 | Procedures | PROC |
| ***Plain x-ray*** | ***X-Ray*** | **C17262** | ***X-Ray*** | ***Natural Phenomenon or Process*** | ***T070*** | ***Phenomena*** | ***PHEN*** |
| MRI | Magnetic Resonance Imaging | C16809 | Magnetic Resonance Imaging | Diagnostic Procedure | T060 | Procedures | PROC |

Data element: Malignant Neoplasm Measurable Disease Evaluation Method Clinical Trial Eligibility Criteria Type (3179024). Value domain: Malignant Neoplasm Measurable Disease Evaluation Method Type (3179022).

NCIt, National Cancer Institute Thesaurus.

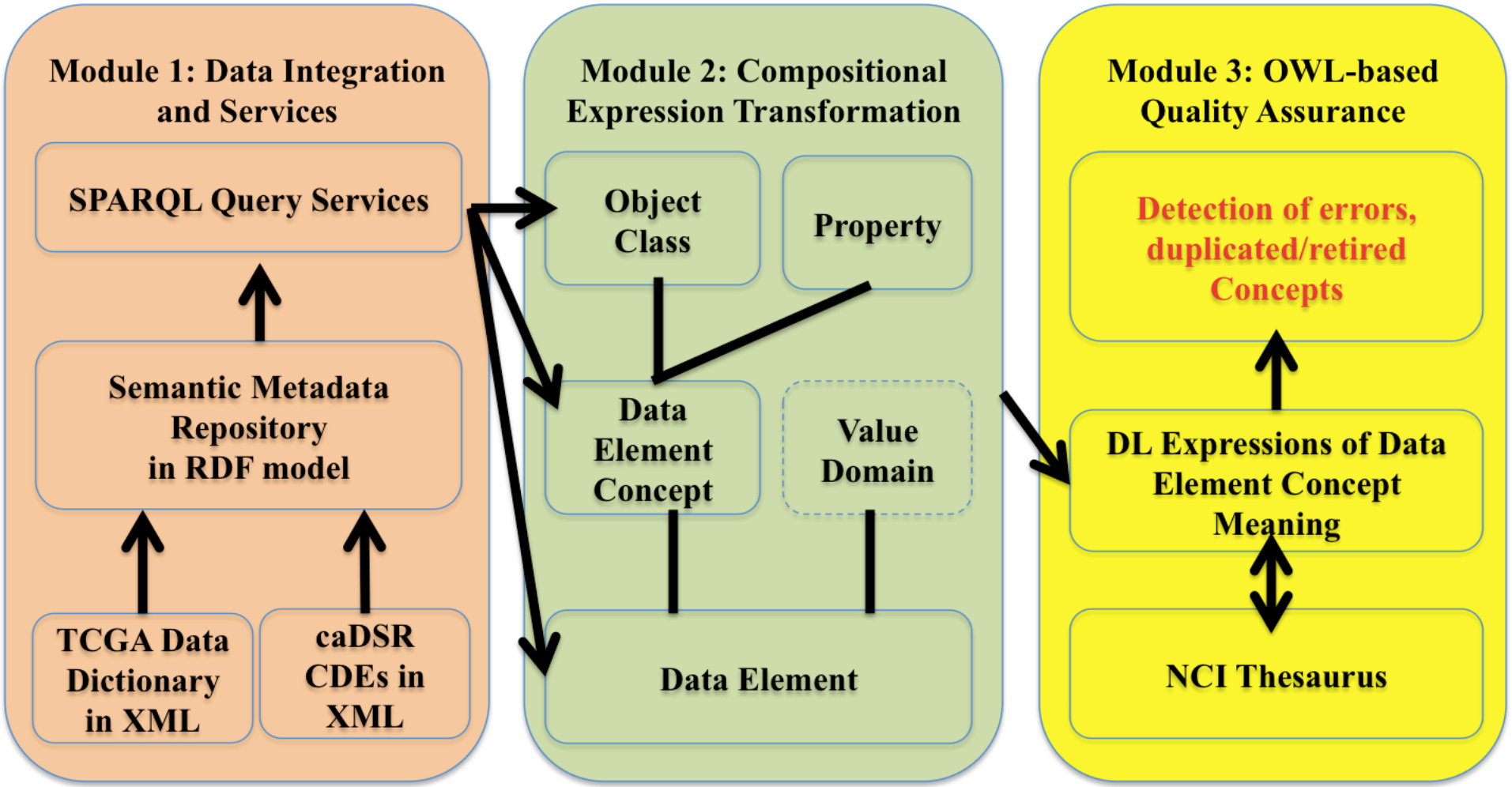
**C17262|X-Ray -> C38101|Radiography**

A radiographic procedure using the emission of X-Rays to form an image of the structure penetrated by the radiation

MAYO CLINIC

# Progress to date

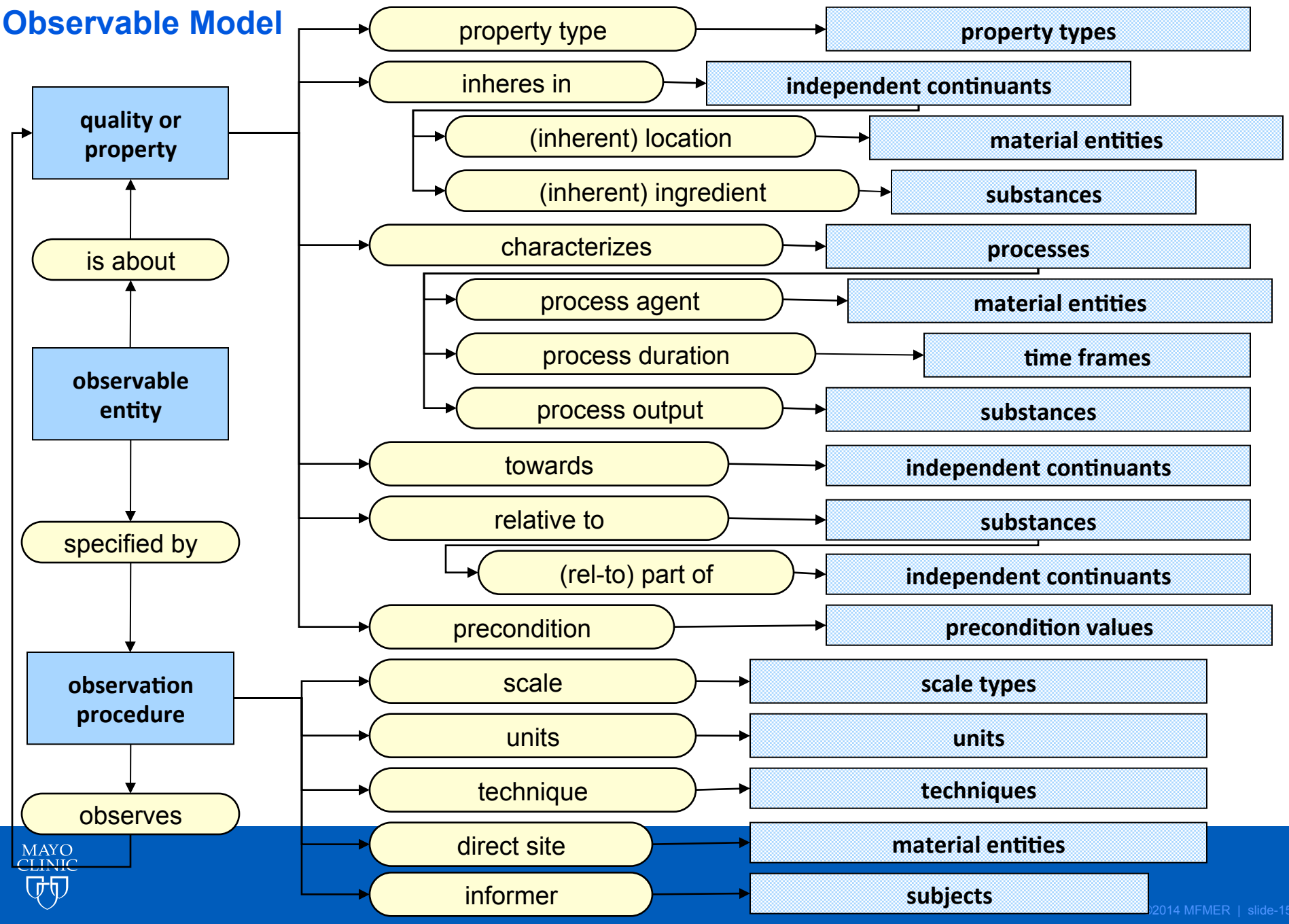
# QA of cancer study CDEs using a post-coordination approach



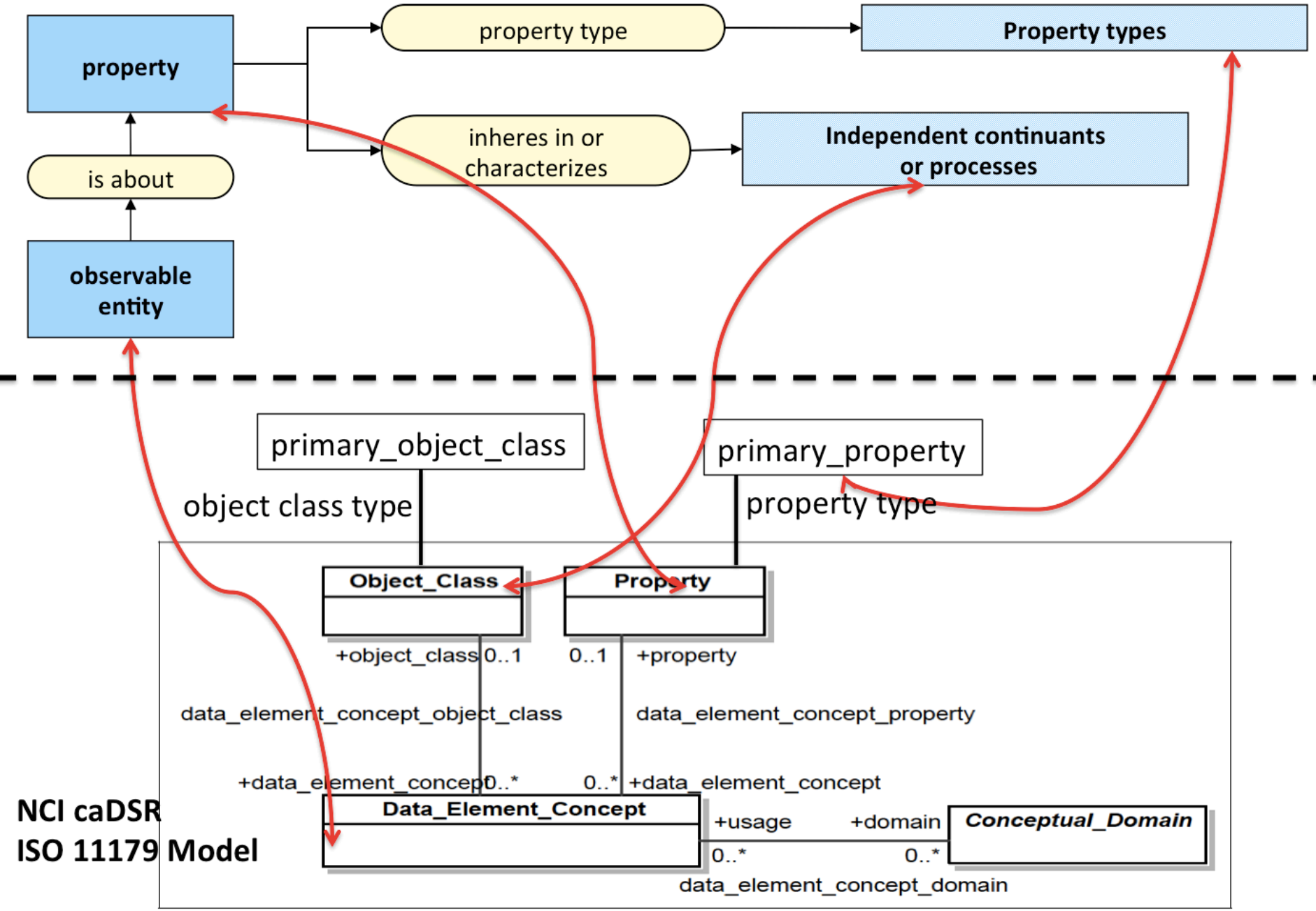


# SNOMED CT Observable Model



# Modified SNOMED CT Observable Model



# An example of compositional expression

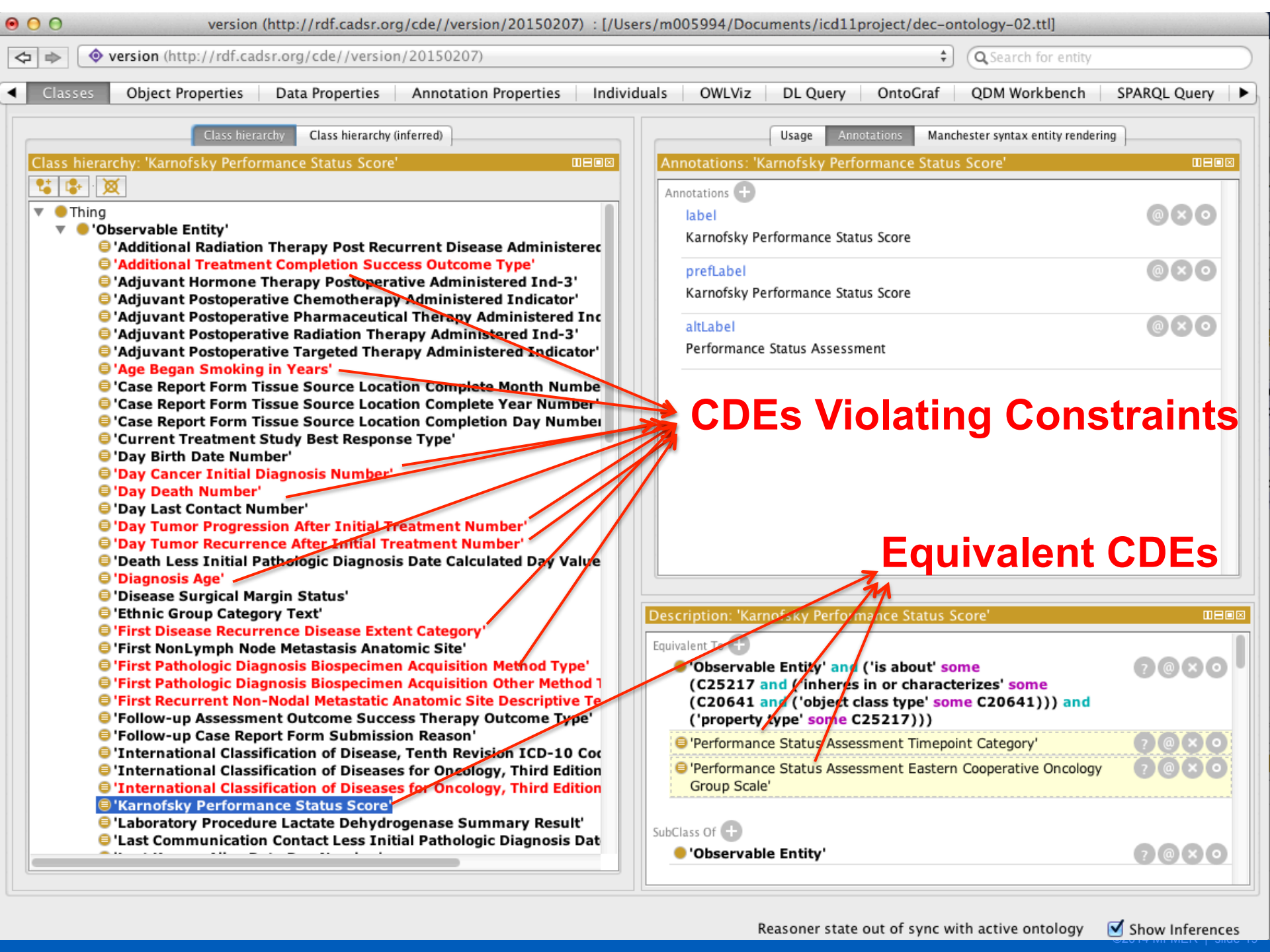
| Original Data Recorded in caDSR | Transformed Compositional Expression |
|---|---|
| Public Id: 3378323<br>CDE Name: Clinical Trial Drug Classification Name<br>Property Code: C25161<br>Property Name: Classification<br>Primary Property: C25161<br>Object Class code: C71104:C1708<br>Object Class Name: Clinical Trial Agent<br>Primary Object Class: C1708 | Class: 'Clinical Trial Drug Classification Name'<br>Annotations:<br>label "Clinical Trial Drug Classification Name"<br>EquivalentTo:<br>'Observable Entity'<br>and ('is about' some (Classification<br>and ('inheres in or characterizes' some ('Clinical Trial Agent'<br>and ('object class type' some Agent)))<br>and ('property type' some Classification)))<br>SubClassOf:<br>'Observable Entity' |

# OWL-based constraints

- Disjointness constraint (owl:disjointWith)
  - The disjointness of a set of classes guarantees that an individual (or a subclass) that is a member (or a subclass) of one class cannot simultaneously be an instance (or a subclass) of a specified other class
- Domain/range constraint (rdfs:domain, rdfs:range)
  - restricts each object property linking between the instances of asserted classes


version (http://rdf.cadsr.org/cde//version/20150207) : [/Users/m005994/Documents/icd11project/dec-ontology-02.ttl]
version (http://rdf.cadsr.org/cde//version/20150207)
Search for entity
Classes
Object Properties
Data Properties
Annotation Properties
Individuals
OWLViz
DL Query
OntoGraf
QDM Workbench
SPARQL Query
Class hierarchy
Class hierarchy (inferred)
Class hierarchy: 'Karnofsky Performance Status Score'
Thing
'Observable Entity'
'Additional Radiation Therapy Post Recurrent Disease Administered
'Additional Treatment Completion Success Outcome Type'
'Adjuvant Hormone Therapy Postoperative Administered Ind-3'
'Adjuvant Postoperative Chemotherapy Administered Indicator'
'Adjuvant Postoperative Pharmaceutical Therapy Administered Inc
'Adjuvant Postoperative Radiation Therapy Administered Ind-3'
'Adjuvant Postoperative Targeted Therapy Administered Indicator'
'Age Began Smoking in Years'
'Case Report Form Tissue Source Location Complete Month Numbe
'Case Report Form Tissue Source Location Complete Year Number'
'Case Report Form Tissue Source Location Completion Day Numbe
'Current Treatment Study Best Response Type'
'Day Birth Date Number'
'Day Cancer Initial Diagnosis Number'
'Day Death Number'
'Day Last Contact Number'
'Day Tumor Progression After Initial Treatment Number'
'Day Tumor Recurrence After Initial Treatment Number'
'Death Less Initial Pathologic Diagnosis Date Calculated Day Value
'Diagnosis Age'
'Disease Surgical Margin Status'
'Ethnic Group Category Text'
'First Disease Recurrence Disease Extent Category'
'First NonLymph Node Metastasis Anatomic Site'
'First Pathologic Diagnosis Biospecimen Acquisition Method Type'
'First Pathologic Diagnosis Biospecimen Acquisition Other Method T
'First Recurrent Non-Nodal Metastatic Anatomic Site Descriptive Te
'Follow-up Assessment Outcome Success Therapy Outcome Type'
'Follow-up Case Report Form Submission Reason'
'International Classification of Disease, Tenth Revision ICD-10 Coc
'International Classification of Diseases for Oncology, Third Edition
'International Classification of Diseases for Oncology, Third Edition
'Karnofsky Performance Status Score'
'Laboratory Procedure Lactate Dehydrogenase Summary Result'
'Last Communication Contact Less Initial Pathologic Diagnosis Dat
Usage
Annotations
Manchester syntax entity rendering
Annotations: 'Karnofsky Performance Status Score'
Annotations
label
Karnofsky Performance Status Score
prefLabel
Karnofsky Performance Status Score
altLabel
Performance Status Assessment
CDEs Violating Constraints
Equivalent CDEs
Description: 'Karnofsky Performance Status Score'
Equivalent To
'Observable Entity' and ('is about' some (C25217 and ('inheres in or characterizes' some (C20641 and ('object class type' some C20641))) and ('property type' some C25217)))
'Performance Status Assessment Timepoint Category'
'Performance Status Assessment Eastern Cooperative Oncology Group Scale'
SubClass Of
'Observable Entity'
Reasoner state out of sync with active ontology
Show Inferences

# Future Plan for OWL-based QA tool

- 1) developing a systematic QA approach leveraging upper level ontologies;
- 2) refining the compositional expression model by aligning with SNOMED observable model;
- 3) making the reasoning-based explanation more user-friendly; and
- 4) incorporating the value domain in the scope.

# Specific Aims

- Aim 1: To develop a suite of QA tools for validation and harmonization of cancer study CDEs;
  - UMLS Semantic Network-based approaches
  - Semantic Web-based approaches
- Aim 2: To apply the QA tools to audit experimental cancer study CDEs represented in a semantic web framework;
  - NCI caDSR
  - Preferred sets of CDEs from TCGA data dictionary
- Aim 3: To deploy and evaluate a QA web-portal for collaborative CDE review and harmonization.
  - Specification for Standard CDE services

# (SWAT4LS 2014, Paper Session)

## Building A Semantic Web-based Metadata Repository for Facilitating Detailed Clinical Modeling in Cancer Genome Studies

Guoqian Jiang[1], Deepak K. Sharma[1], Harold R. Solbrig[1], Cui Tao[2], Chunhua Weng[3], Christopher G. Chute[1]

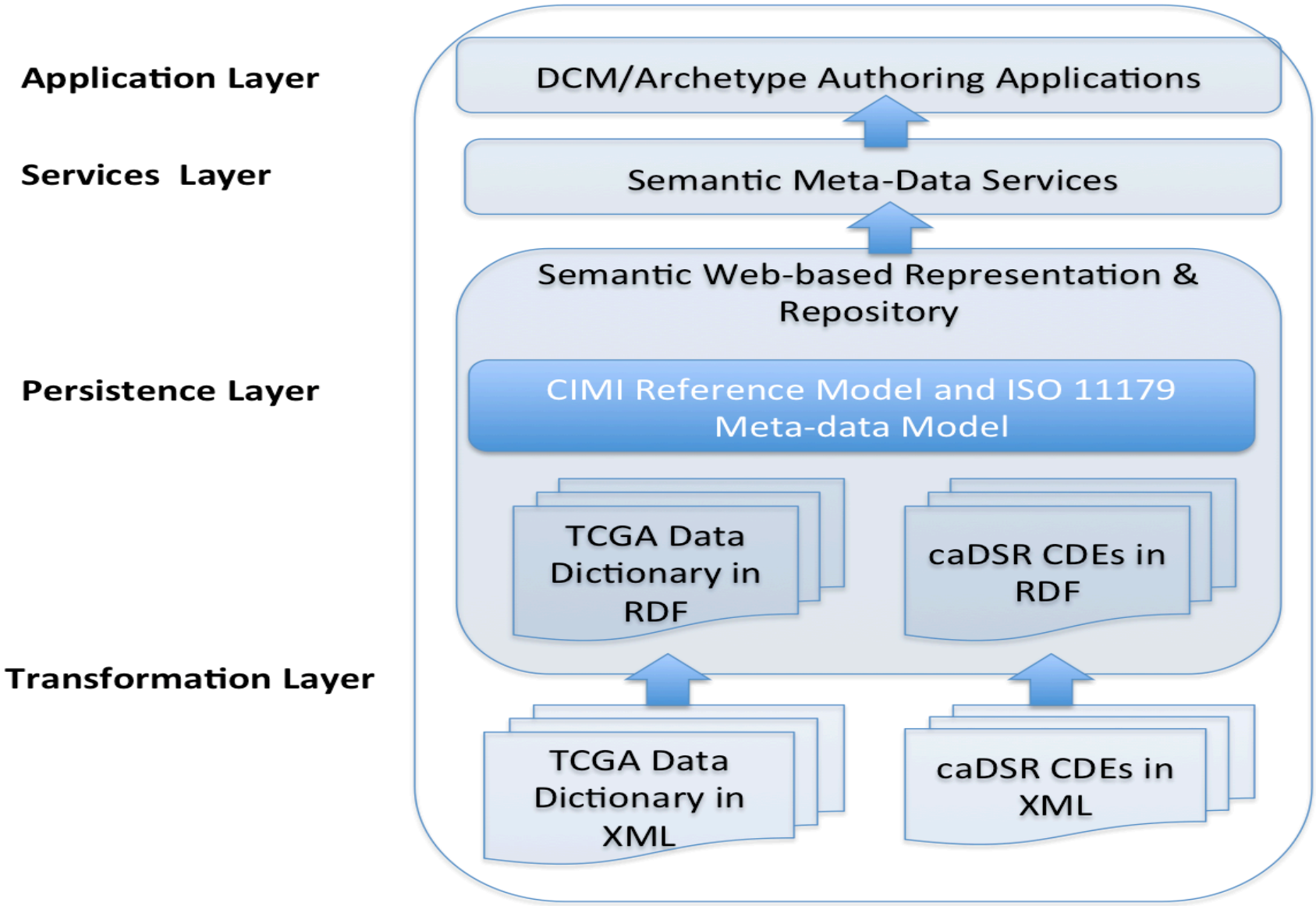
[1] Department of Health Sciences Research, Mayo Clinic College of Medicine, Rochester, MN
{jiang.guoqian, sharma.deepak2, solbrig.harold, chute}@mayo.edu
[2] University of Texas Health Science Center at Houston Houston, TX
cui.tao@uth.tmc.edu
[3] Columbia University, New York City, NY
cw2384@cumc.columbia.edu

**Abstract.**

Detailed Clinical Models (DCMs) have been regarded as the basis for retaining computable meaning when data are exchanged between heterogeneous computer systems. To better support clinical cancer data capturing and reporting, there is an emerging need to develop informatics solutions for standards-based clinical models in cancer study domains. The objective of the study is to develop and evaluate a use case-driven approach that enables a Semantic Web-based cancer study metadata repository based on both ISO11179 metadata standard and Clinical Information Modeling Initiative (CIMI) Reference Model (RM). We used the common data elements (CDEs) defined in The Cancer Genome Atlas (TCGA) data dictionary, and extracted the metadata of the CDEs




Application Layer
Services Layer
Persistence Layer
Transformation Layer
DCM/Archetype Authoring Applications
Semantic Meta-Data Services
Semantic Web-based Representation & Repository
CIMI Reference Model and ISO 11179 Meta-data Model
TCGA Data Dictionary in RDF
caDSR CDEs in RDF
TCGA Data Dictionary in XML
caDSR CDEs in XML

# Study Plan for QA of cancer detailed clinical models

- Goal
  - To build QA tool for cancer detailed clinical models
- Leveraging existing CIMI modeling efforts
  - CIMI full reference model
  - CIMI archetype modeling language (AML)
  - CTS2-based terminology binding
- Looking into HL7 FHIR modeling efforts
  - FHIR resources and profiles
  - FHIR terminology binding (value sets)
  - FHIR Ontology (HL7 RDF workgroup)
- Collaboration with clinical cancer research community
  - DeepPhe project (PI: Drs. Guergana Savova and Rebecca Crowley)
  - Mayo cancer registries (e.g., breast cancer, ovarian cancer)

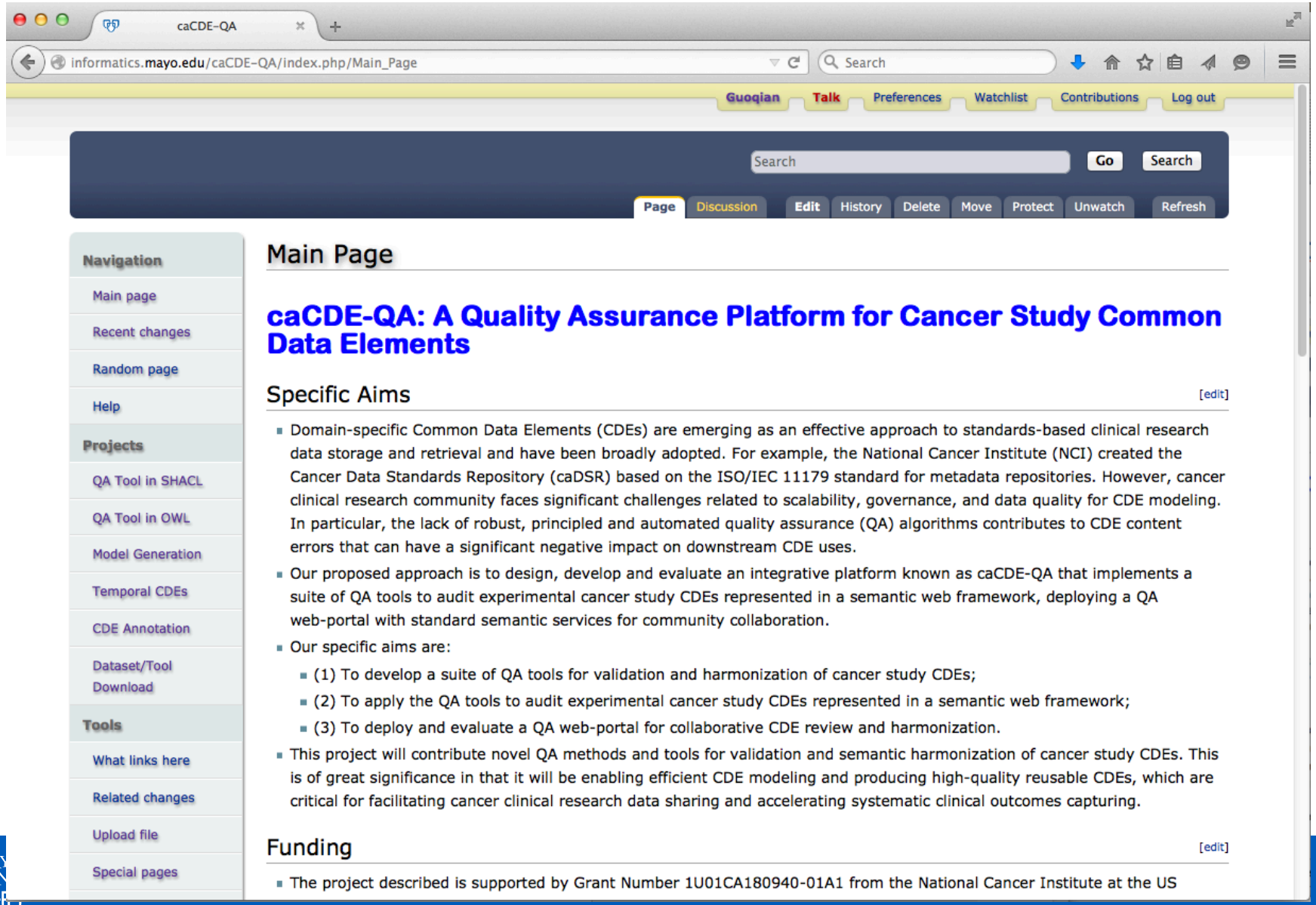

 

# Collaboration study plan

- CDE discovery tool
  - Collaboration with Dr. Chunhua Weng from Columbia University
  - ClinicalTrials.gov
- CDE temporal pattern identification
  - Collaboration with Dr. Cui Tao from UTHealth
  - CNTRO Temporal Ontology
- QA tool using Shape Constraint Language
  - Dr. Tim Berners-Lee - W3C/MIT
- CDISC / CIMI
- NCI caDSR team and TCGA project

# http://informatics.mayo.edu/caCDE-QA/index.php/Main_Page




# Questions & Discussion