MAYO CLINIC

caCDE-QA: A Quality Assurance Platform for Cancer Study Common Data Elements

Guoqian Jiang, MD, PhD
Mayo Clinic

NCI ITCR PI Annual F2F Meeting
June 13-14, 2016

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.

Challenges = Heterogeneities

Composite Variable vs. Individual Components

Study A

Var:
Stage at initial diagnosis

Stage: I, II, III, IV

Study B

Var:
Stage at initial diagnosis

T stage: Tis, T1, T2, T3, T4

N stage: N0, N1, N2

M stage: M0, M1

MAYO CLINIC

Source from Drs. Qian Shi and Dan Sargent – ARCAD Project

Challenges = Heterogeneities

Coding List

Study A

Yes = 1
No = 0

Event = 1
Censoring = 0

Study B

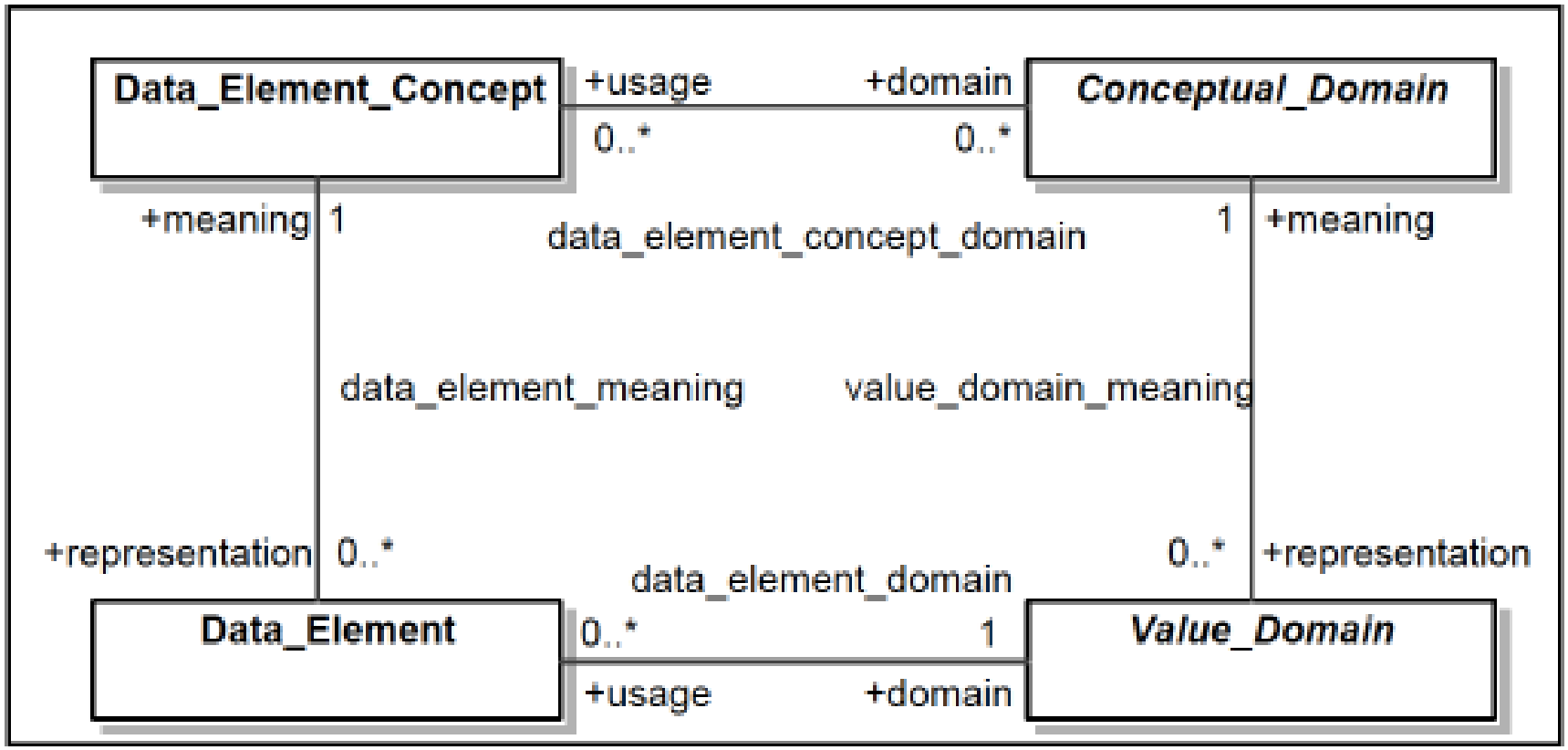
Yes = 1
No = 2

Event = 0
Censoring = 1

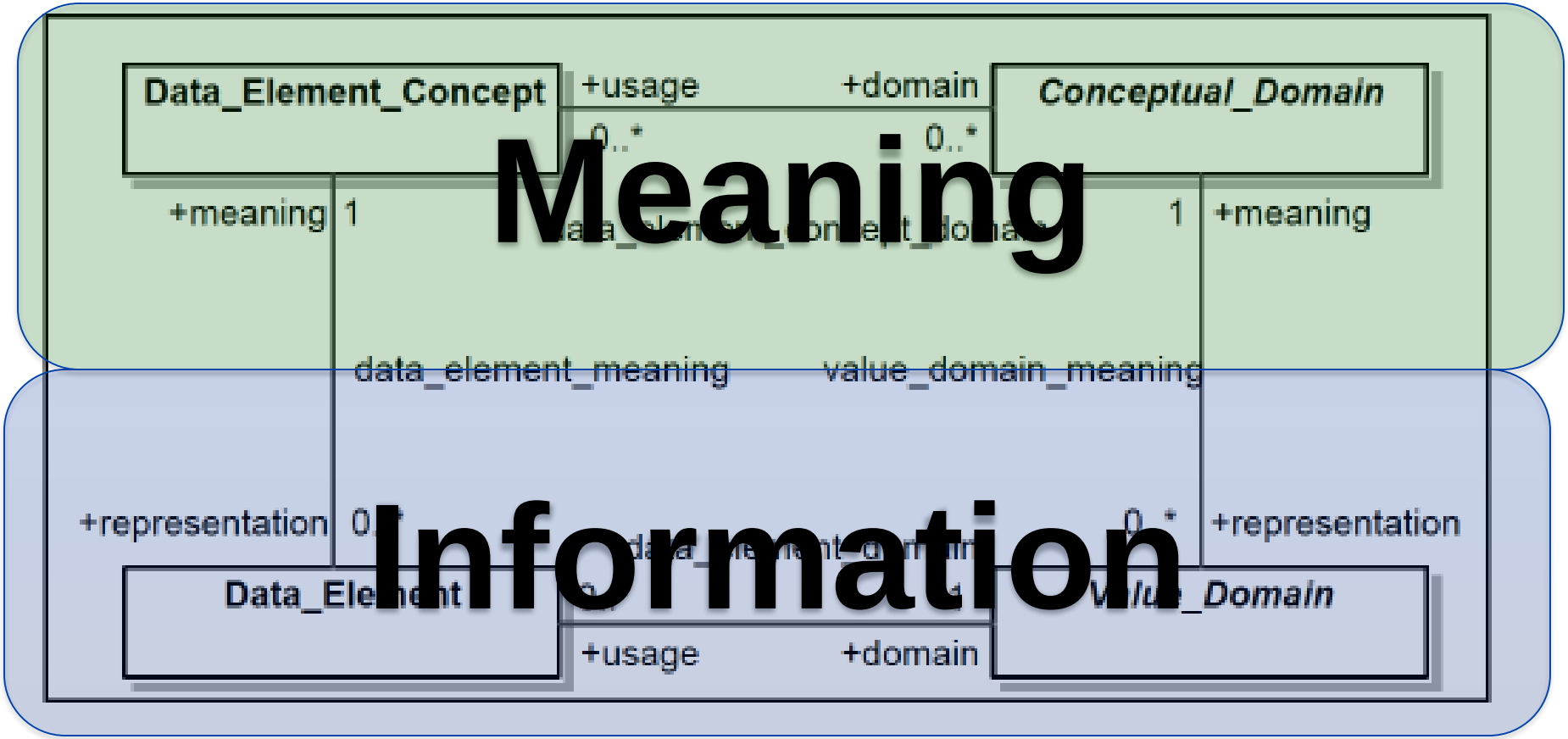
MAYO CLINIC

Source from Drs. Qian Shi and Dan Sargent – ARCAD Project

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



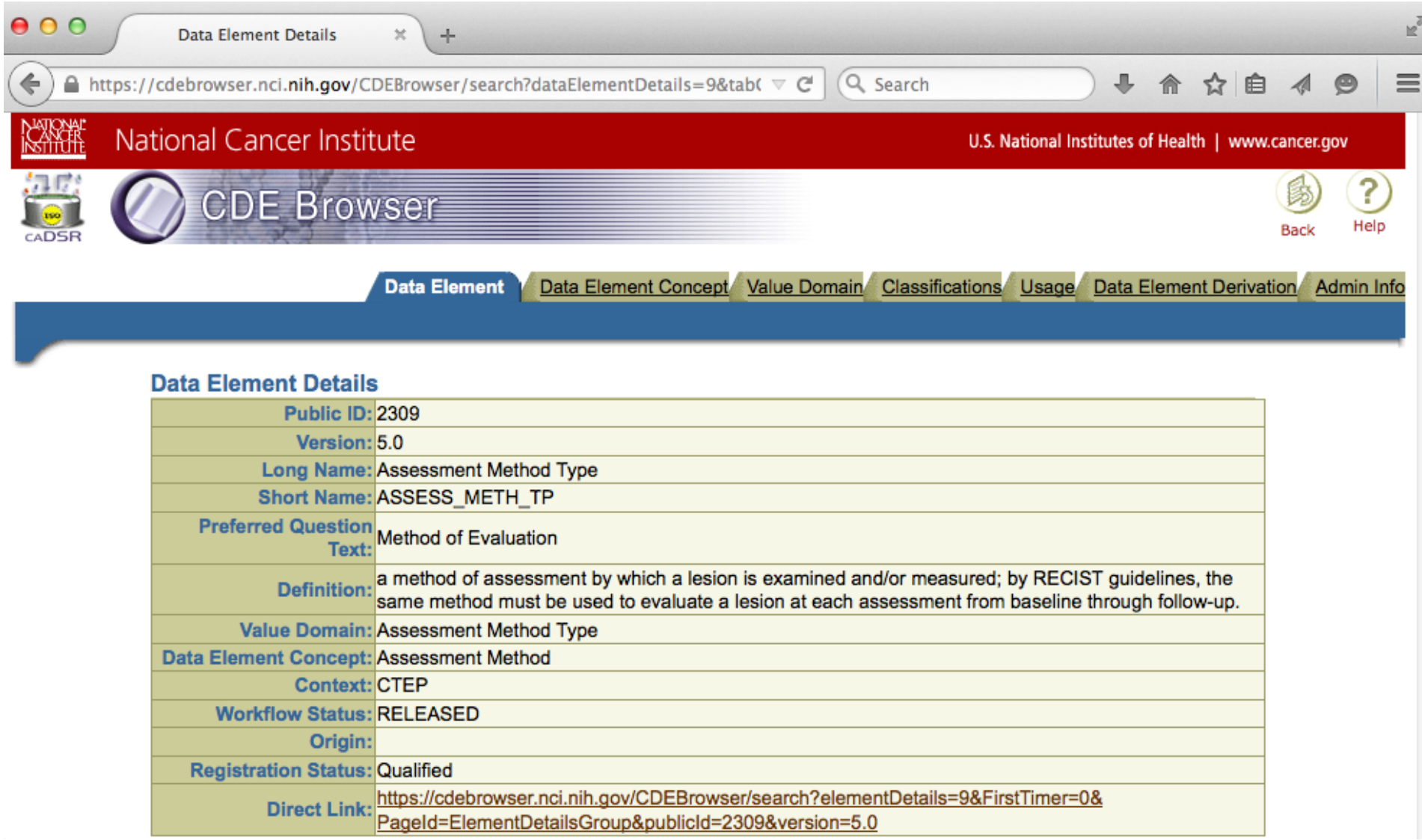
Information and Meaning



ISO/IEC 11179-based Metadata Repository

- National Cancer Institute (NCI) created the Cancer Data Standards Repository (caDSR) based on the ISO/IEC 11179 metadata standard.
- 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.

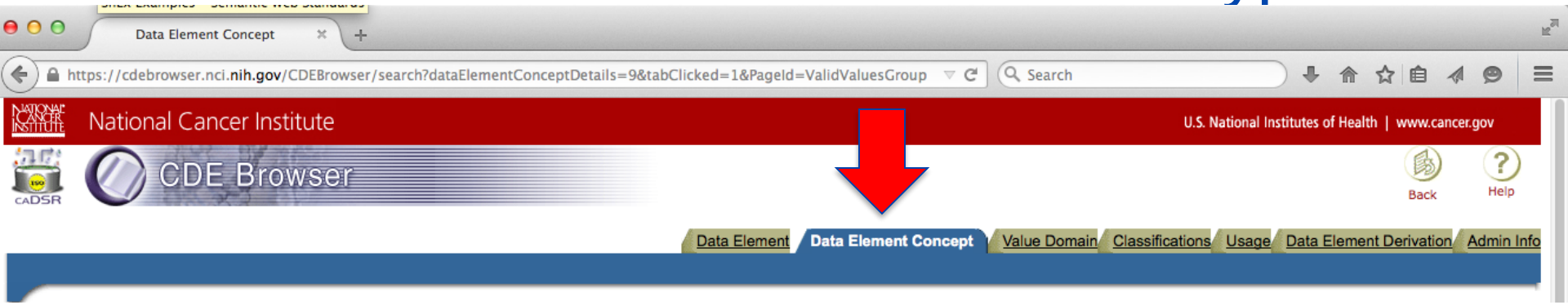
caDSR CDE: Assessment Method Type



Data Element Details

Public ID:	2309
Version:	5.0
Long Name:	Assessment Method Type
Short Name:	ASSESS_METH_TP
Preferred Question Text:	Method of Evaluation
Definition:	a method of assessment by which a lesion is examined and/or measured; by RECIST guidelines, the same method must be used to evaluate a lesion at each assessment from baseline through follow-up.
Value Domain:	Assessment Method Type
Data Element Concept:	Assessment Method
Context:	CTEP
Workflow Status:	RELEASED
Origin:	
Registration Status:	Qualified
Direct Link:	https://cdebrowser.nci.nih.gov/CDEBrowser/search?elementDetails=9&FirstTimer=0&PageId=ElementDetailsGroup&publicId=2309&version=5.0

caDSR CDE: Assessment Method Type



Object Class More Details

Public ID:	2176954
Version:	1.0
Long Name:	Assessment
Short Name:	Assessment
Context:	CTEP
Qualifier:	

Meaning of the CDE
Defined in NCI Thesaurus Concept Codes

Object Class Concepts

Concept Name	Concept Code	Public ID	Definition Source	EVS Source	Primary
Physical Examination	C20989	2202921	NCI	NCI_CONCEPT_CODE	Yes

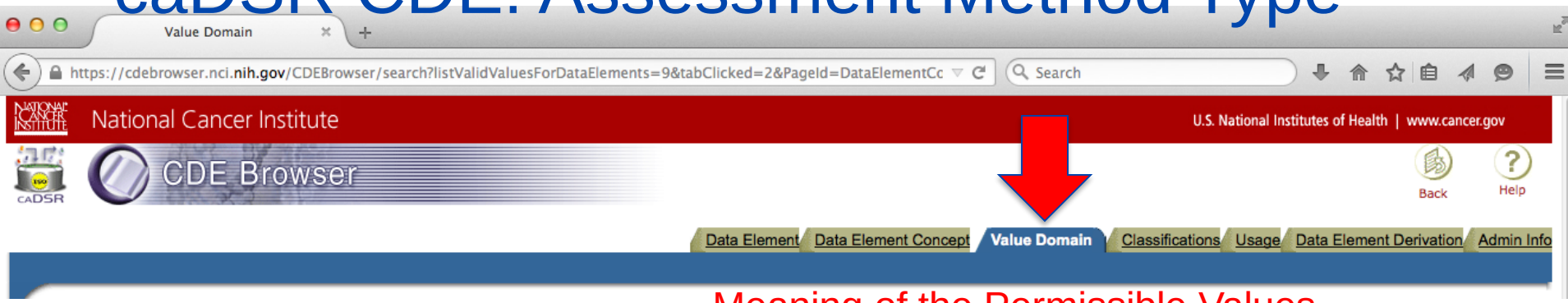
Property

Public ID:	2177492
Version:	1.0
Long Name:	Method
Short Name:	Method
Context:	CTEP
Qualifier:	

Property Concepts

Concept Name	Concept Code	Public ID	Definition Source	EVS Source	Primary
Technique	C16847	2203854	NCI	NCI_CONCEPT_CODE	Yes

caDSR CDE: Assessment Method Type



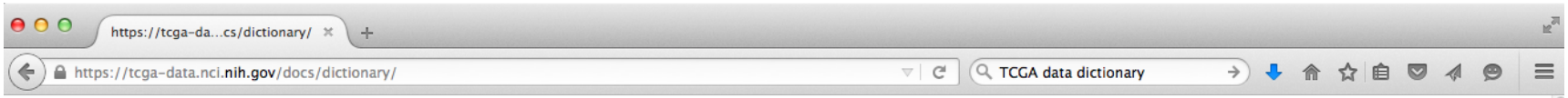
Meaning of the Permissible Values Defined in NCI Thesaurus Concept Codes

Permissible Values

1 - 20 of 164 Next 20

PV	PV Meaning	PV Meaning Concept Codes	PV Meaning Description	PV Begin Date	PV End Date	VM Public ID	VM Version
2-D Echocardiogram	2-D Echocardiogram		2-D Echocardiogram	2009-03-09		2847201	1.0
Abdominal CT	Abdominal CT-scan		Abdominal CT-scan	2007-08-06		2667062	1.0
Abdominal MRI	Abdominal MRI		Magnetic Resonance Imaging of the abdomen.	2007-08-17		2673036	1.0
Antigen	Antigen	C268	Any substance that appears foreign or potentially adverse to the body and elicits immune response.	2014-04-03		3167779	1.0
Aspergillus galactomannan assay	ELISA	C16553	A highly sensitive technique for detecting and measuring antigens or antibodies in a solution; the solution is run over a surface to which immobilized antibodies specific to the substance have been attatched, and if the substance is present, it will bind to the antibody layer, and its presence is verified and visualized with an application of antibodies that have been tagged in some way. (BioTech Life Science Dictionary)	2011-01-06		3173905	1.0
Autopsy	Autopsy	C25153	Autopsy; an examination and dissection of a dead body to determine cause of death or the changes produced by disease.	2002-02-11		2567287	1.0
Axillary dissection	AXILLARY DISSECTION		AXILLARY DISSECTION	2002-02-11		2558265	1.0
Bilateral mammogram	Bilateral mammogram		Bilateral mammogram	2002-11-06		2563569	1.0
Bimanual examination	BIMANUAL EXAMINATION		BIMANUAL EXAMINATION	2002-02-11		2558295	1.0
Biopsy	Biopsy	C15189	Removal and pathologic examination of specimens in the form of small pieces of tissue from the living body.	2002-02-11		2574093	1.0
Biopsy other	Biopsy	C15189	Removal and pathologic examination of specimens in the form of small pieces of tissue from the living body.	2002-12-19	2012-07-10	2574093	1.0
Bone Marrow Aspirate	Aspirate, Bone Marrow	C15644	(as-per-AY-shun) The removal of a small sample of bone marrow (usually from the hip) through a needle for examination under a microscope.	2009-08-20		2573943	1.0
Bone Marrow Biopsy	BONE MARROW BIOPSY	C15193	BONE MARROW BIOPSY	2009-08-20		2562182	1.0

Data Dictionary Using caDSR CDEs



The Cancer Genome Atlas

Data Dictionary

as of January 2016

Click CDE ID link to visit the entry's NCI CDE Browser page.
Listed in Alphabetical Sequence by Data Element name.
Tumor-specific entries are annotated with their associated tumor types.

CDE Public Id	CRF Question Text	Data Element	Definition	Valid Values	Tumor Types
3225946v1	1p/19q Status Per Report	1p And 19q Chromosome Status Type	Text that describes the deletion abnormality for the proximal (short) arm of chromosome 1 and the distal (long) arm of chromosome 19.	N/A I 1p/19q co-del I 1p/19q intact (non-del) I Only 19q del I Only 1p del	
2625737v1	Number of days of ATRA before registration	ATRA Agent Prior Clinical Trial Registration Administered Day Count	the number of days all-trans retinoic acid, a naturally-occurring acid of retinol, was administered prior to registration or enrollement in a controlled study performed in human subjects and intended to discover, evaluate, and/or verify safety, effectiveness, clinical and pharmacological effects, and adverse reactions.		
3121640v1	Was the patient exposed to ATRA prior to procurement?	ATRA Agent Prior Clinical Trial Registration Administered Indicator	Text indicator to signify whether all-trans retinoic acid, a naturally-occurring acid of retinol, was administered prior to registration or enrollement in a controlled study performed in human subjects and intended to discover, evaluate, and/or verify safety, effectiveness, clinical and pharmacological effects, and adverse reactions.	No I Unknown I Yes	
3225706v1	Day of Ablation	Ablation Performed Day Number	Numeric value that represents the day the ablation was performed.	22 I 21 I 20 I 19 I 18 I 17 I 16 I 15 I 14 I 13 I 1 I 10 I 11 I 12 I 2 I 3 I 4 I 5 I 6 I 7 I 8 I 9 I 31 I 30 I 29 I 28 I 27 I 26 I 25 I 24 I 23	
3225709v1	Were Ablation Techniques Utilized?	Ablation Performed Indicator	Text indicator to signify whether or not ablation techniques were used.	Unknown I Yes I No	
3225710v1	Number of Lesions Treated with Ablation	Ablation Performed Lesion Count	Numeric value that represents the number of lesions that were treated with ablation therapy.		
3225707v1	Month of Ablation	Ablation Performed Month Number	Numeric value that represents the month the ablation was performed.	12 I 11 I 10 I 9 I 8 I 7 I 6 I 5 I 4 I 3 I 2 I 1	
3225702v1	Ablation Type	Ablation Performed Type	Text term to decribe the type of ablation therpy.	Other I Radiofrequency Ablation I Microwave I Ethanol Injection I Radiosurgical Ablation	

Challenges

- The potential of the binding of a data element to controlled terminology 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 caDSR.
- CDE content errors can have a significant negative impact on downstream CDE uses.

Specific Aims of caCDE-QA

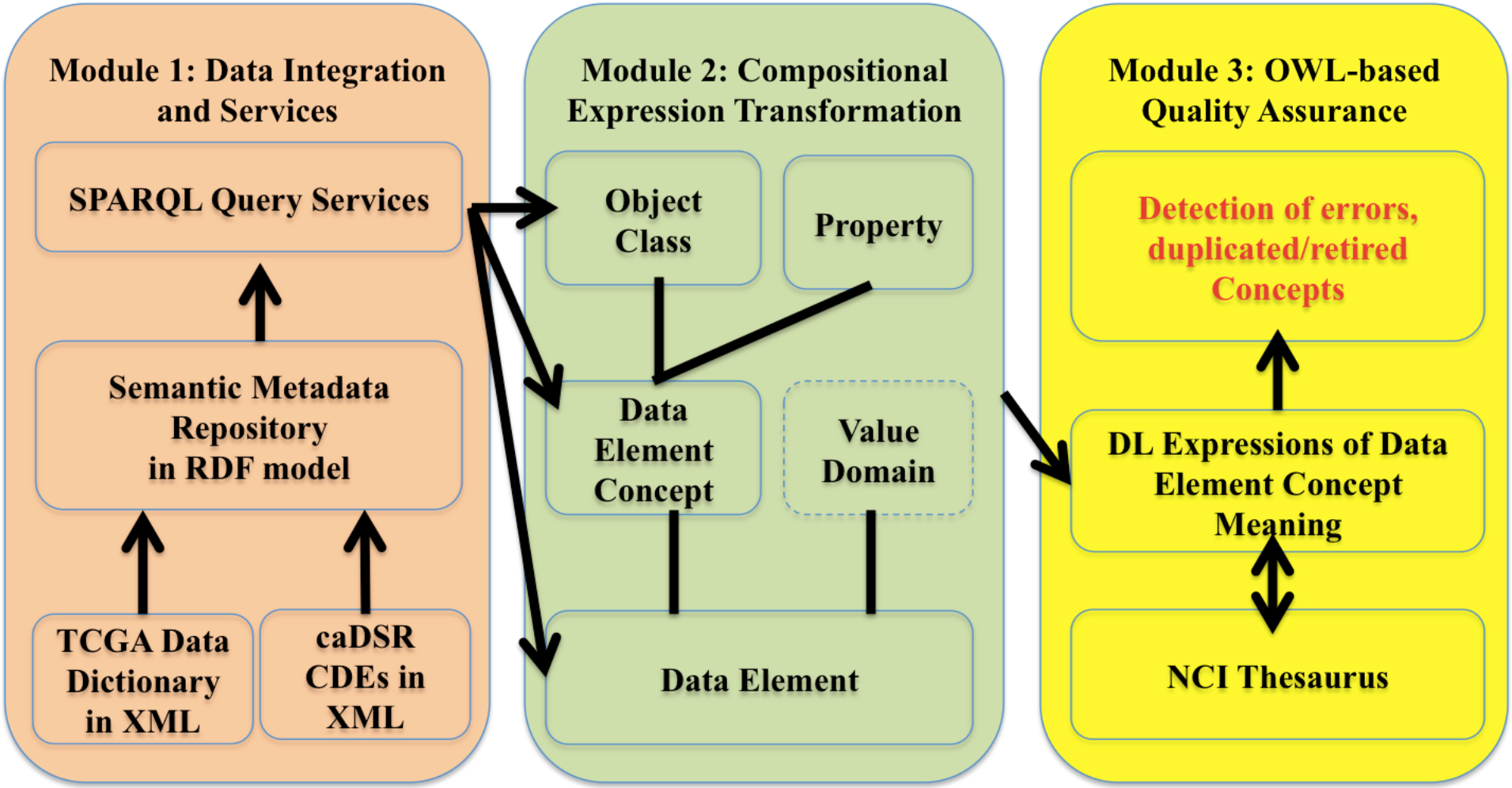
- Aim 1: To develop a suite of QA tools for validation and harmonization of cancer study CDEs;
  - UMLS Semantic Network-based approaches
  - Semantic Web-based approaches
- Aim 2: To apply the QA tools to audit experimental cancer study CDEs represented in a semantic web framework;
  - NCI caDSR
  - Preferred sets of CDEs from TCGA data dictionary
- Aim 3: To deploy and evaluate a QA web-portal for collaborative CDE review and harmonization.
  - CTS2-like standard CDE services

Semantic Web-based QA Approaches

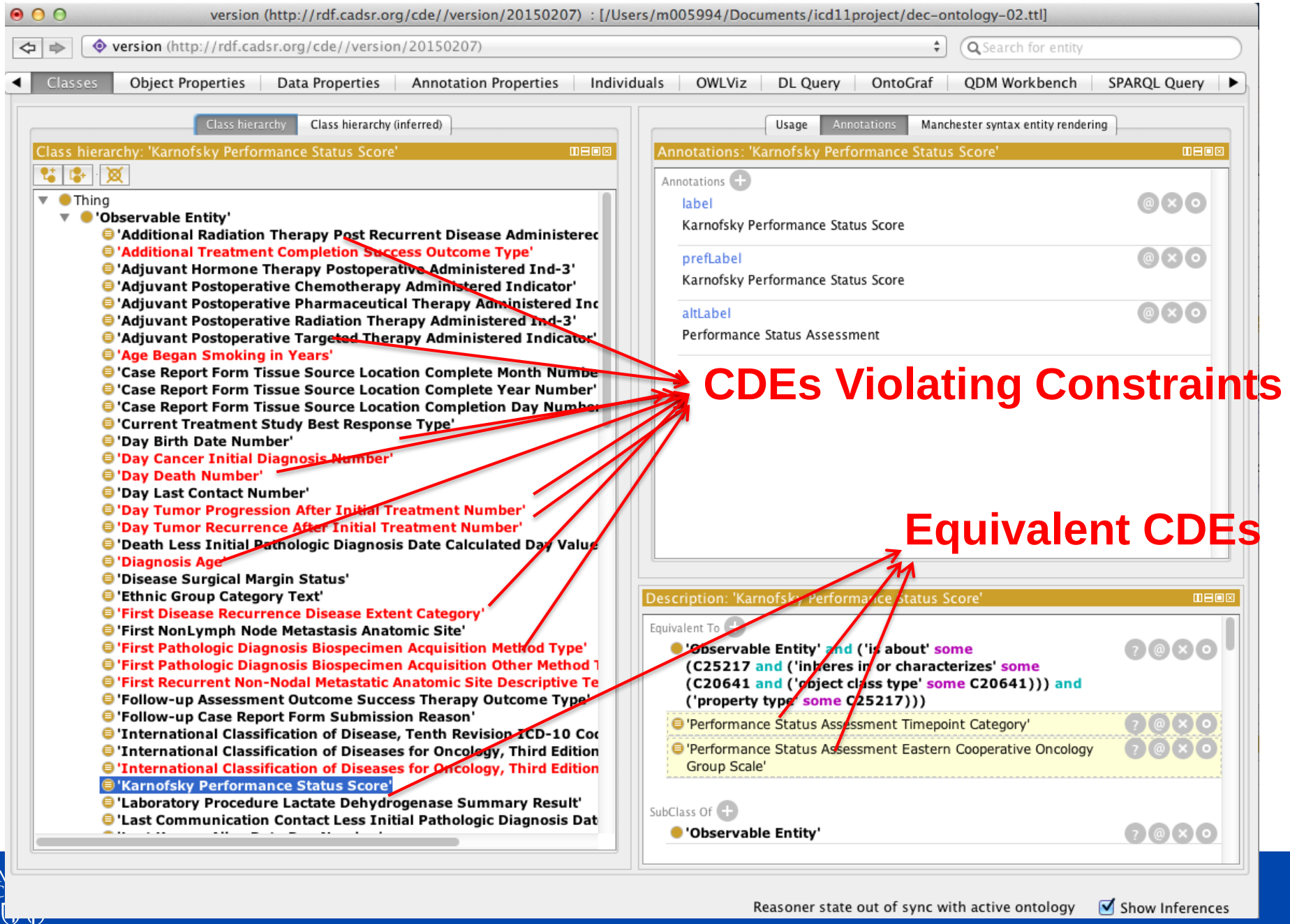
Semantic Web Technologies

- The W3C standards
- Resource Description Framework (RDF)
  - A model of directed, labeled graphs
  - Using a set of triples (subject, predicate, object) - statements
- SPARQL
  - A query language for RDF graphs
- Web Ontology Language (OWL)
  - A standard ontology language used for ontology modeling
  - Leveraging Description Logic for reasoning
- Shape Expressions (ShEx) Language
  - An emerging standard

OWL-based QA Tools



Jiang, et al. AMIA Annu Symp Proc. 2015


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

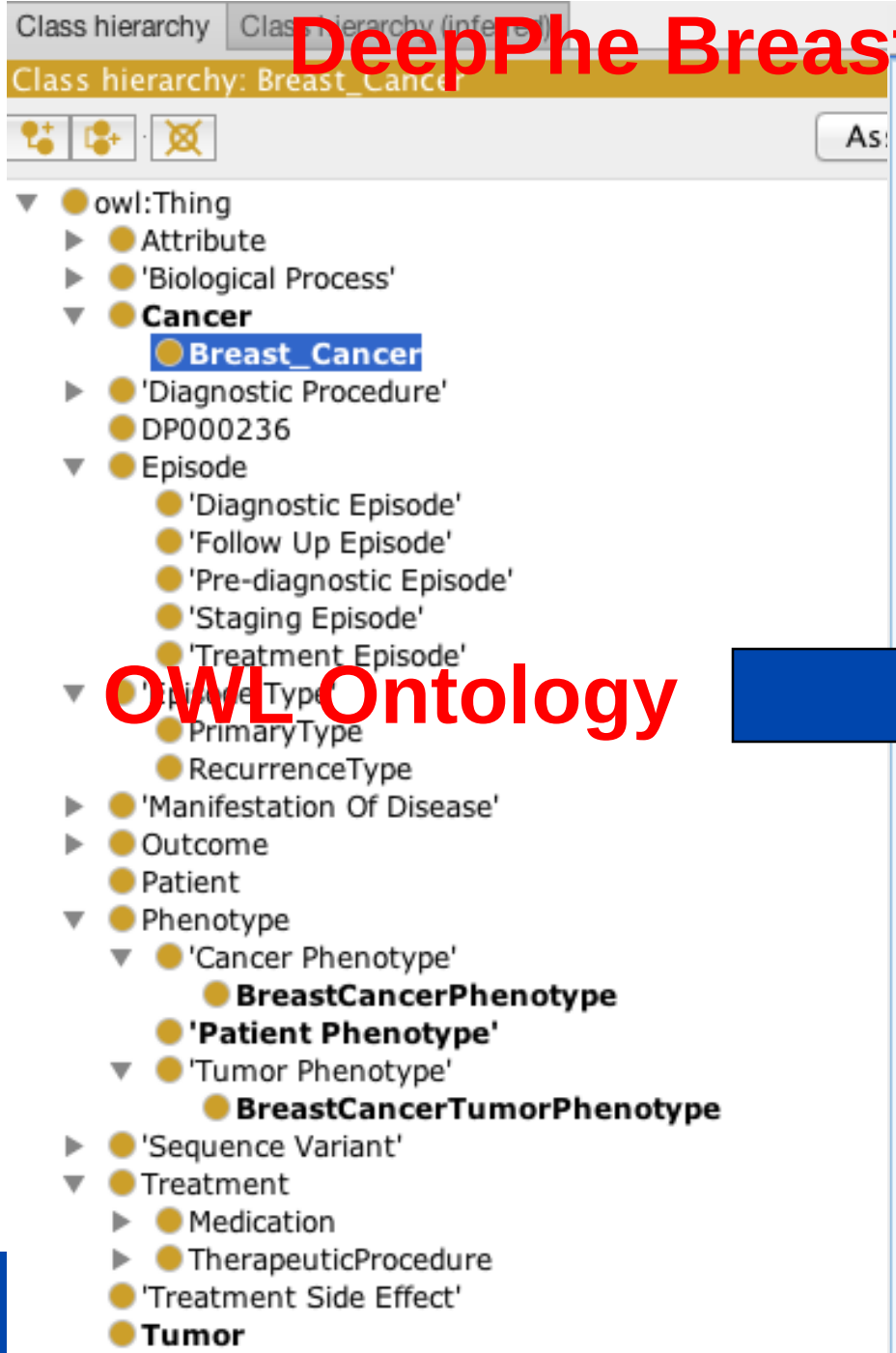
Shape Expressions (ShEx)-based QA Tools

MAYO CLINIC

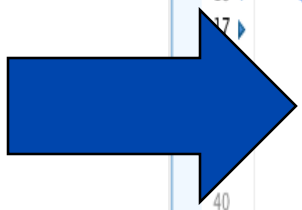
Video (5 min)

https://youtu.be/D93_p9QAomw

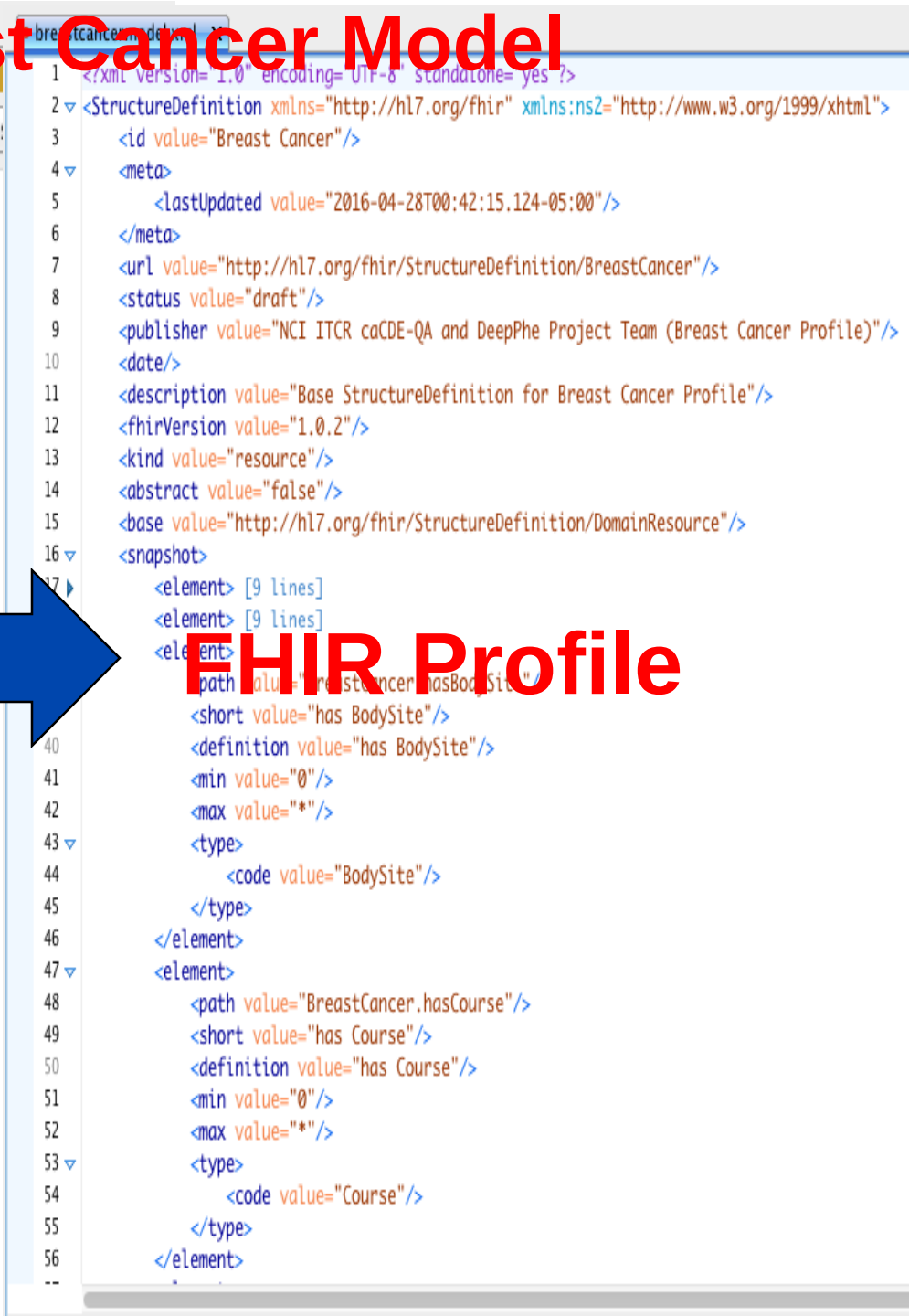
DeepPhe Breast Cancer Model



OWL Ontology



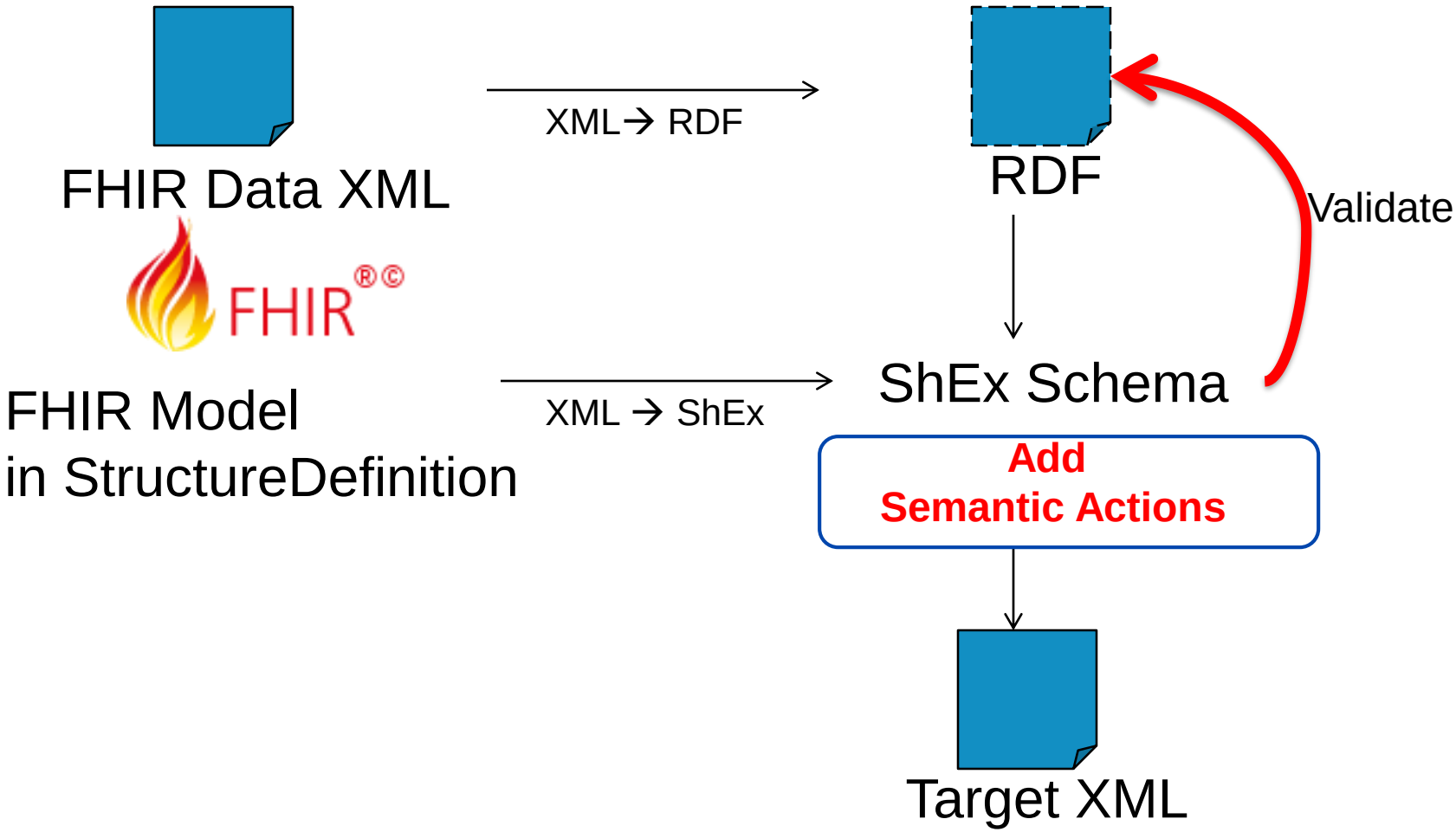
FHIR Profile



Shape Expressions (ShEx) Language

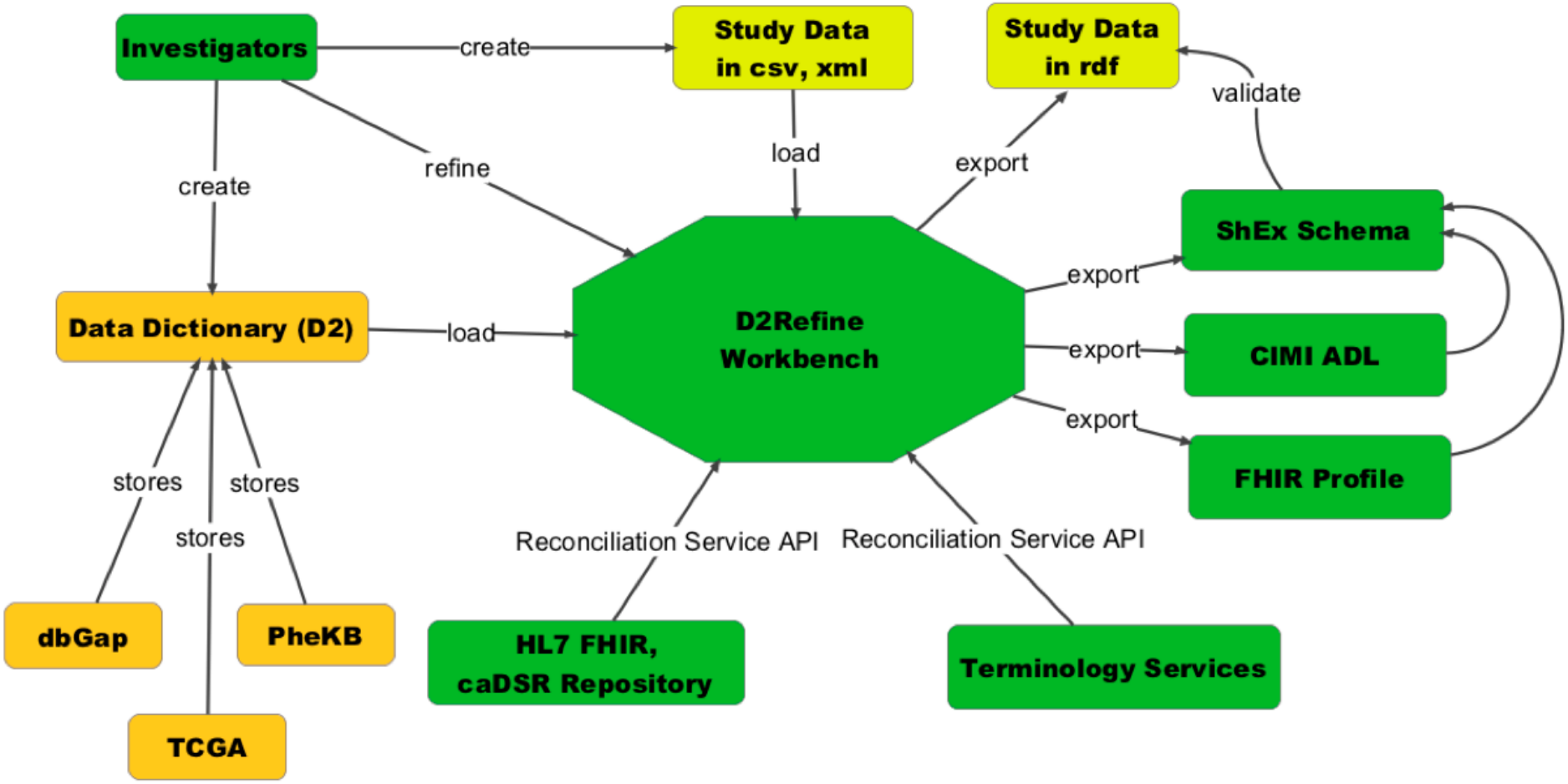
- Input to the W3C RDF Data Shapes working group
- Schema language for RDF
  - Formally describes RDF *structures*
  - RDF analog to XML schema
- ShEx – validate and traverse models

Approach

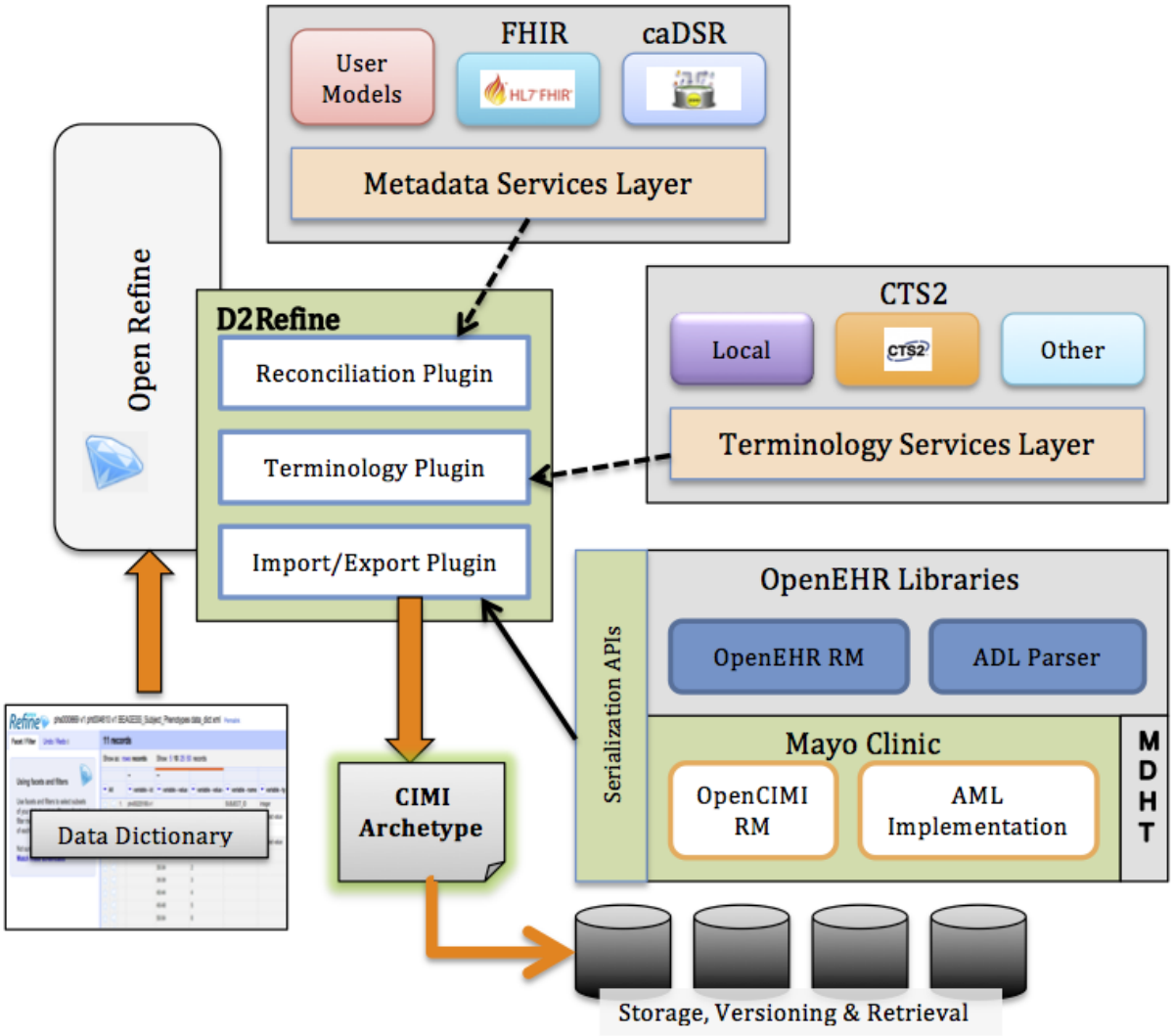

FHIR Data XML
XML→ RDF
RDF
Validate
FHIR Model
in StructureDefinition
XML → ShEx
ShEx Schema
Add
Semantic Actions
Target XML

D2Refine: A Metadata Harmonization and Validation Framework

D2Refine –
A Framework for Metadata Harmonization and Validation



D2Refine Plugins



NCI LexEVS CTS 2

https://wiki.nci.nih.gov/display/LexEVS/LexEVS+6.x+CTS2

NATIONAL CANCER INSTITUTE
NCI Wiki

New Account | Helpful Tips

Confluence | Spaces | Browse | Log in

Search

- LexEVS Documentation and Training
  - Deprecated LexEVS Documentation and Training
  - LexEVS 6.x Documentation
    - **LexEVS 6.x CTS2**
    - LexEVS 6.x Loader Use Guide
    - LexEVS 6.x Value Set and Pick List Definition Guide
    - LexEVS 6.x API Programmer's Guide
    - LexEVS 6.x Installation Guide
    - LexEVS 6.x Loader Guide
    - LexEVS 6.x Administration Guide
    - LexEVS 6.x Design and Architecture Guide
    - LexEVS 6.x OWL Export Guide
    - LexEVS 6.x Migration Tips for 5.x Users
  - LexEVS and EVS Javadocs Index
- LexEVS Installation and Downloads
- LexEVS Release Notes
- LexEVS Release Roadmap
- LexEVS Development Documents
- LexEVS Code Examples
- LexEVS Community Code Contribution
- LexGrid
- LexBIG
- Retrospectives
- How-to articles

EVS - LexEVS / ... / LexEVS 6.x Documentation

Tools

LexEVS 6.x CTS2

Created by Scott Bauer, last modified by Robert Wynne on Aug 24, 2015

LexEVS 6.x Documentation | LexEVS 6.x Loader Use Guide

LexEVS 6.x CTS2

LexEVS and its data model LexGrid comprise a full function terminology service meeting a wide range of requirements. The Common Terminology Service 2 is a comprehensive standard defined by a broad range of community interests.

- LexEVS 6.0 featured a comprehensive Java implementation of the CTS2 Service Functional Model in its preliminary form as a Draft Standard for Trial Use.
- The final specification of CTS2 was designed as a REST-specific API and LexEVS 6.1 has implemented the specification core functionality.
- The LexEVS 6.2 implementation features a new Value Set module giving it an implementation profile consistent with other CTS2 compliant services. Furthermore it implements the latest CTS2 1.1 specification providing up to date standardization complete with the canonical JSON output.
- The LexEVS 6.3 implementation is a minor update and provides minimal new features.

Contents of this Page

CTS2 Rest Client Quick Start

NCI EVS CTS 2

LexEVS CTS2 6.3 API Documentation (lexevscts2.nci.nih.gov/lexevscts2)

Toggle All Resources | Toggle All Methods

Service Metadata

List Methods | Expand Methods

- GET — Get CTS2 Service Metadata — /service

Entities

List Methods | Expand Methods

- GET — Search Entities — /entities
- HEAD — Count Entities — /entities
- GET — Read an Entity by URI — /entitybyuri
- GET — Read an Entity by ID — /entity/{entityid}

Code System Versions

List Methods | Expand Methods

- GET — Search Code System Versions — /codesystemversions
- GET — Read a Code System Version by Version ID — /codesystem/{codesystem}/version/{version}
- GET — Search Entities of a Code System Version — /codesystem/{codesystem}/version/{version}/entities
- GET — Read an Entity of a Code System Version by ID — /codesystem/{codesystem}/version/{version}/entity/{entityid}

Associations

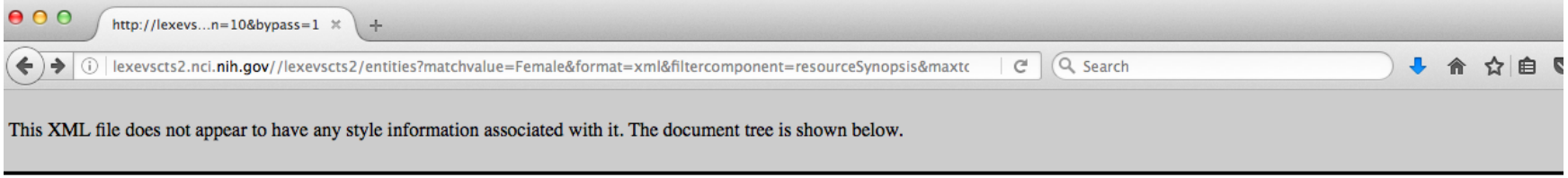
List Methods | Expand Methods

- GET — Get 'children' associations — /codesystem/{codesystem}/version/{version}/entity/{entityid}/children
- GET — Get 'subjectOf' associations — /codesystem/{codesystem}/version/{version}/entity/{entityid}/subjectof
- GET — Get 'targetOf' associations — /codesystem/{codesystem}/version/{version}/entity/{entityid}/targetof

Value Sets

List Methods | Expand Methods

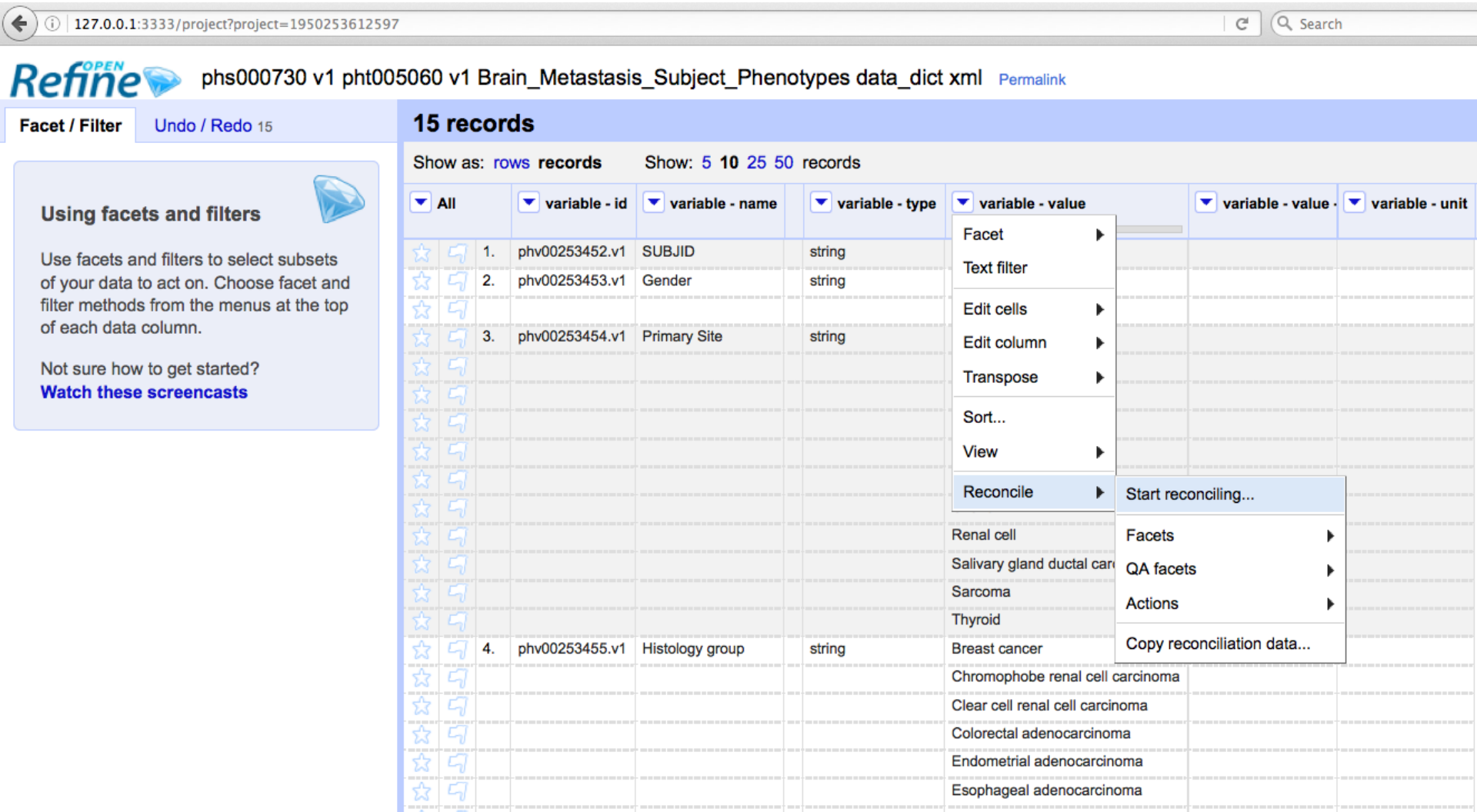
- GET — Search Value Sets — /valuesets



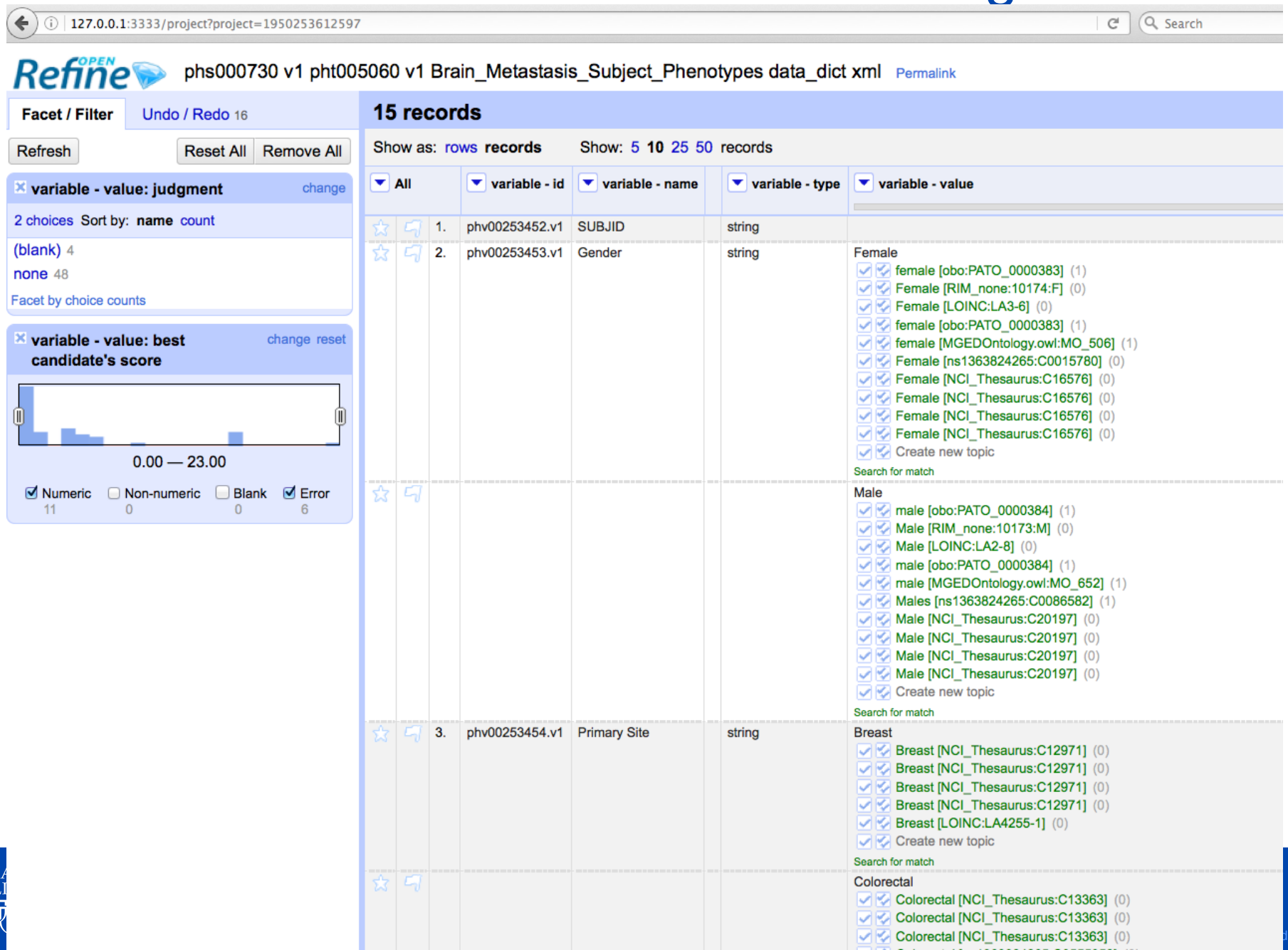

```
– <EntityDirectory xsi:schemaLocation="http://www.omg.org/spec/CTS2/1.1/Entity http://informatics.mayo.edu/cts2/spec/CTS2/1.1/entity/Entity.xsd" complete="PARTIAL" numEntries="10"
  next="http://lexevs63cts2.nci.nih.gov/lexevscts2/entities?page=1&filtercomponent=resourceSynopsis&bypass=1&matchvalue=Female&format=xml&maxtoreturn=10">
  – <core:heading>
      <core:resourceRoot>http://lexevs63cts2.nci.nih.gov/lexevscts2/</core:resourceRoot>
      <core:resourceURI>entities</core:resourceURI>
    – <core:parameter arg="matchvalue">
        <core:val>Female</core:val>
      </core:parameter>
    – <core:parameter arg="format">
        <core:val>xml</core:val>
      </core:parameter>
    – <core:parameter arg="filtercomponent">
        <core:val>resourceSynopsis</core:val>
      </core:parameter>
    – <core:parameter arg="maxtoreturn">
        <core:val>10</core:val>
      </core:parameter>
    – <core:parameter arg="bypass">
        <core:val>1</core:val>
      </core:parameter>
      <core:accessDate>2016-06-02T16:49:25.649-04:00</core:accessDate>
    </core:heading>
  – <entry about="http://purl.obolibrary.org/obo/PATO_0000383">
    – <core:name>
        <core:namespace>obo</core:namespace>
        <core:name>PATO_0000383</core:name>
      </core:name>
    – <core:knownEntityDescription href="http://lexevs63cts2.nci.nih.gov/lexevscts2/codesystem/OBI/version/April2015/entity/obo:PATO_0000383">
      – <core:describingCodeSystemVersion>
          <core:version href="http://lexevs63cts2.nci.nih.gov/lexevscts2/codesystem/OBI/version/April2015">OBI-April2015</core:version>
          <core:codeSystem uri="http://purl.obolibrary.org/obo/obi.owl">OBI</core:codeSystem>
        </core:describingCodeSystemVersion>
        <core:designation>female</core:designation>
      </core:knownEntityDescription>
    </entry>
  – <entry about="http://www.hl7.org/Library/data-model/RIM/10174:F">
    – <core:name>
        <core:namespace>RIM_none</core:namespace>
        <core:name>10174:F</core:name>
      </core:name>
    – <core:knownEntityDescription href="http://lexevs63cts2.nci.nih.gov/lexevscts2/codesystem/HL7/version/V3 R2.36/entity/RIM_none:10174%3AF">
      – <core:describingCodeSystemVersion>
          <core:version href="http://lexevs63cts2.nci.nih.gov/lexevscts2/codesystem/HL7/version/V3 R2.36">HL7-V3 R2.36</core:version>
          <core:codeSystem uri="http://www.hl7.org/Library/data-model/RIM">HL7</core:codeSystem>
        </core:describingCodeSystemVersion>
        <core:designation>Female</core:designation>
      </core:knownEntityDescription>
```


http://lexevscts2.nci.nih.gov//lexevscts2/entities?matchvalue=Female&format=xml&filtercomponent=resourceSynopsis&maxtoreturn=10&bypass=1

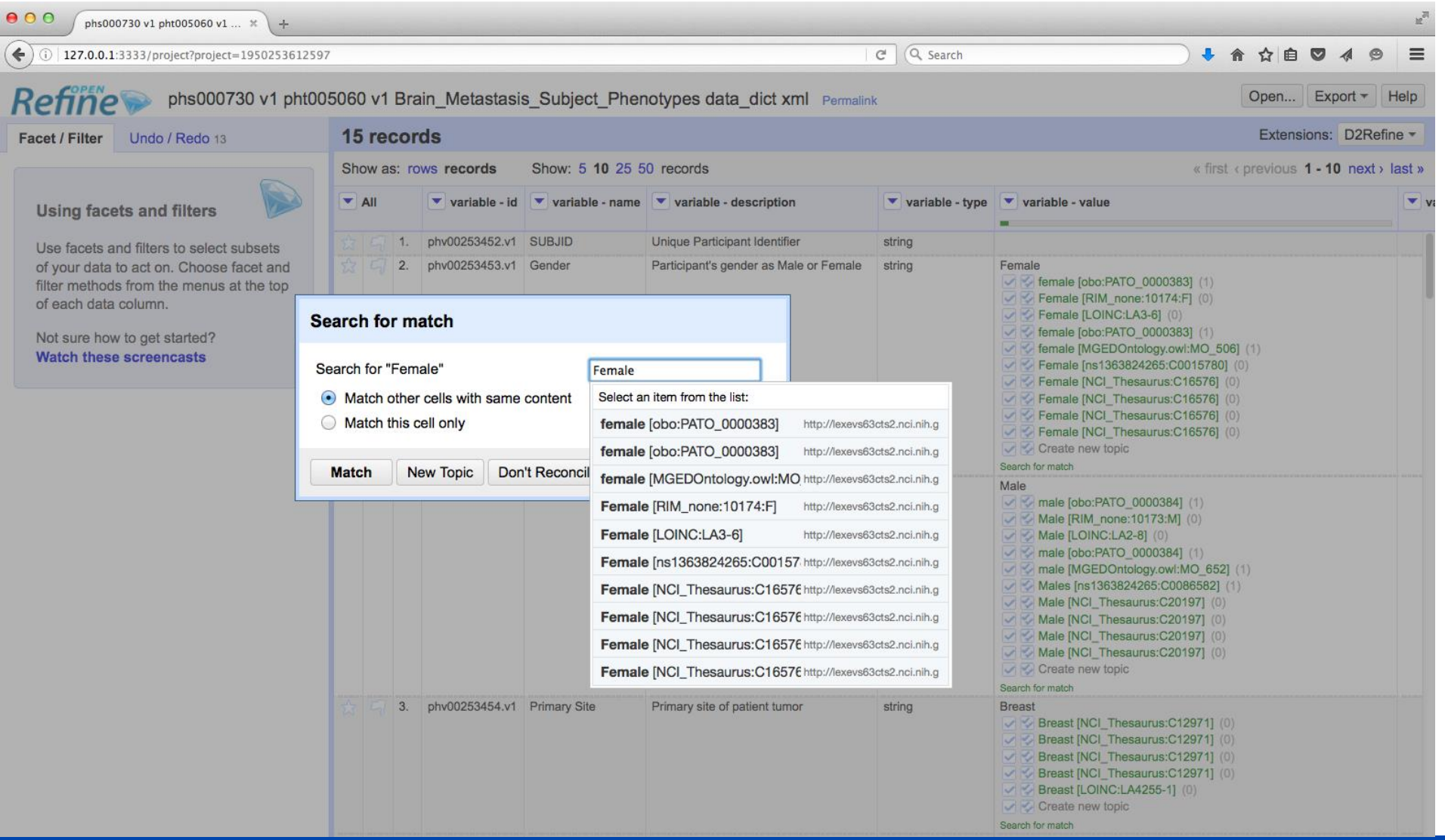
D2Refine CTS2 Reconciliation Plugin



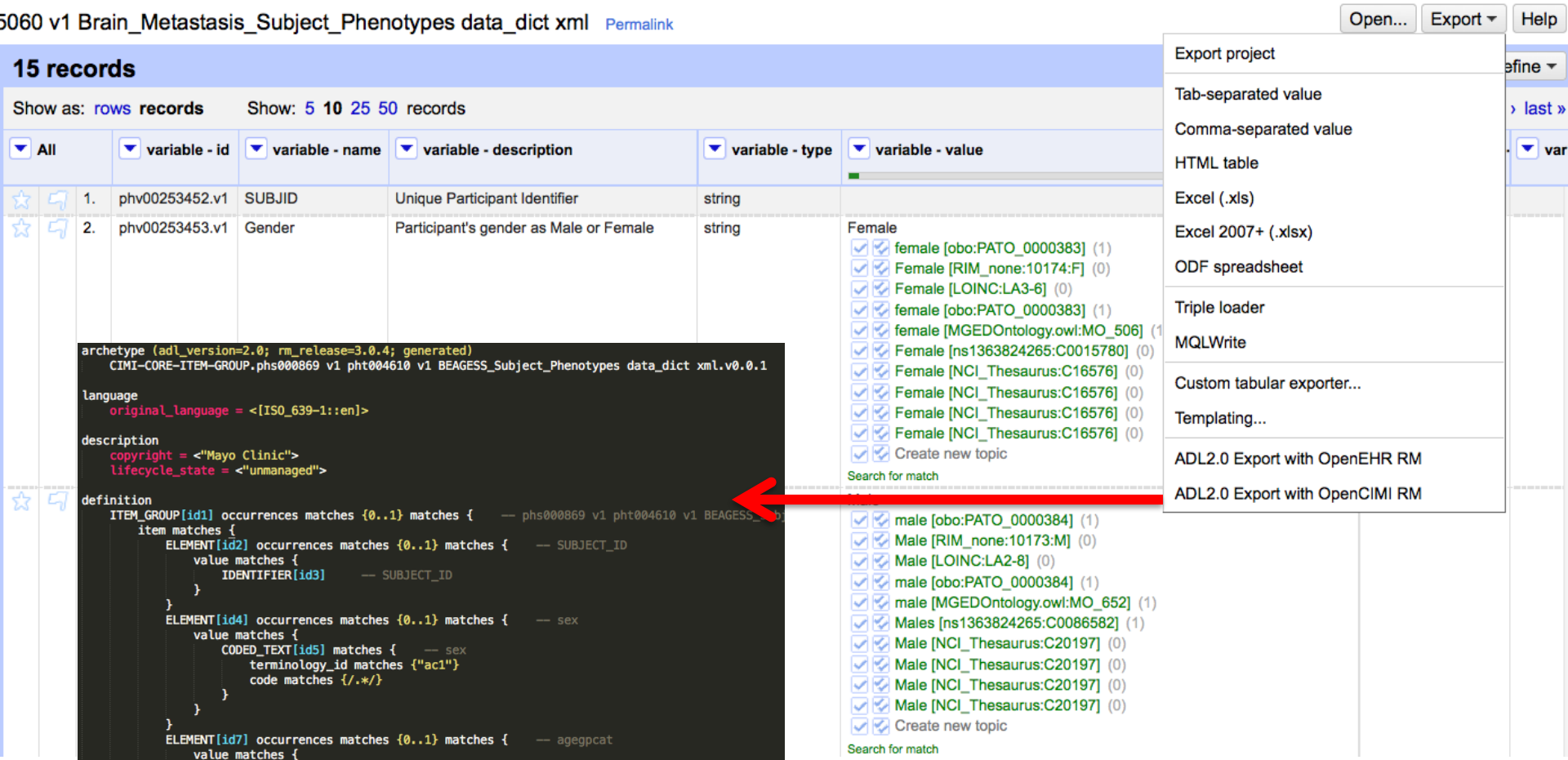
D2Refine CTS2 Reconciliation Plugin



D2Refine CTS2 Reconciliation Plugin



CIMI ADL/FHIR Exporter




```
archetype (adl_version=2.0; rm_release=3.0.4; generated)
    CIMI-CORE-ITEM-GROUP.phs000869 v1 pht004610 v1 BEAGESS_Subject_Phenotypes data_dict xml.v0.0.1

language
    original_language = <[ISO_639-1::en]>

description
    copyright = <"Mayo Clinic">
    lifecycle_state = <"unmanaged">

definition
    ITEM_GROUP[id1] occurrences matches {0..1} matches {    -- phs000869 v1 pht004610 v1 BEAGESS_
        item matches {
            ELEMENT[id2] occurrences matches {0..1} matches {    -- SUBJECT_ID
                value matches {
                    IDENTIFIER[id3]    -- SUBJECT_ID
                }
            }
            ELEMENT[id4] occurrences matches {0..1} matches {    -- sex
                value matches {
                    CODED_TEXT[id5] matches {    -- sex
                        terminology_id matches {"ac1"}
                        code matches {/.*/}
                    }
                }
            }
            ELEMENT[id7] occurrences matches {0..1} matches {    -- agegpcat
                value matches {
                    CODED_TEXT[id8] matches {    -- agegpcat
                        terminology_id matches {"ac2"}
                        code matches {/.*/}
                    }
                }
            }
            ELEMENT[id10] occurrences matches {0..1} matches {    -- race
                value matches {
                    PLAIN_TEXT[id11] matches {    -- race
                        value matches {"European"}
                    }
                }
            }
            ELEMENT[id12] occurrences matches {0..1} matches {    -- BMI
                value matches {
                    CODED_TEXT[id13] matches {    -- BMI
                        terminology_id matches {"ac3"}
```

Posters & Demo

- Poster I (by the U01 caCDE-QA Team)
  - **Modeling and Validating HL7 FHIR Profiles Using Semantic Web Shape Expressions**
- Poster II: (by a collaboration between the U01 caCDE-QA and U24 DeepPhe/TIE teams)
  - **Integrating TCGA Clinical Data Using Metadata-driven Tools and NLP**

Acknowledgements

- **caCDE-QA Development Team**
  - Deepak Sharma, MS, Mayo Clinic
  - Rick Kiefer, Mayo Clinic
  - Eric Prud'hommeax, W3C/MIT
  - Harold Solbrig, MS, Mayo Clinic
  - Guoqian Jiang, MD, PhD, Mayo Clinic (PI)
- **Co-Investigators**
  - Chunhua Weng, PhD, Columbia Univ.
  - Cui Tao , PhD, UTHealth
  - Christopher Chute, MD, Dr. PH, Johns Hopkins Univ.
  - Peter Szolovit, PhD, MIT
- **Consultants**
  - HL7 Advisor: Dr. Stan Huff
  - CDISC Advisor: Dr. Rebecca Kush
- **Supplement Collaborators**
  - Dr. Guergana Savova, Harvard Medical School/Boston Children's Hospital
  - Dr. Rebecca Jacobson, Univ. of Pittsburgh

Questions & Discussion

Youtube video demo link:

https://youtu.be/D93_p9QAomw