



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

Guoqian Jiang, MD, PhD
Mayo Clinic

NCI ITCR Annual Meeting
May 30, 2017

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

Challenges = Heterogeneities

Coding List

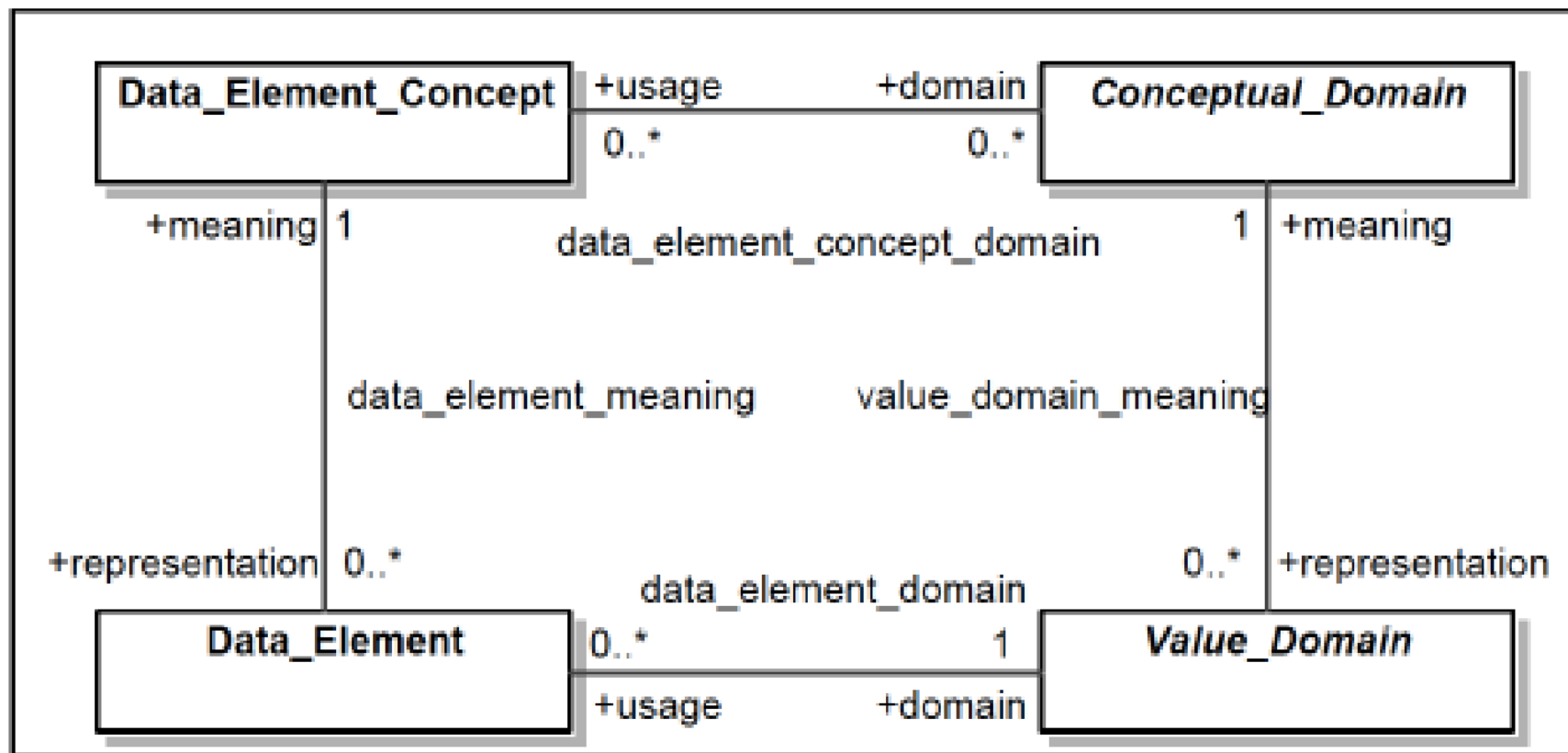
Study A

Yes = 1 No = 0
Event = 1 Censoring = 0

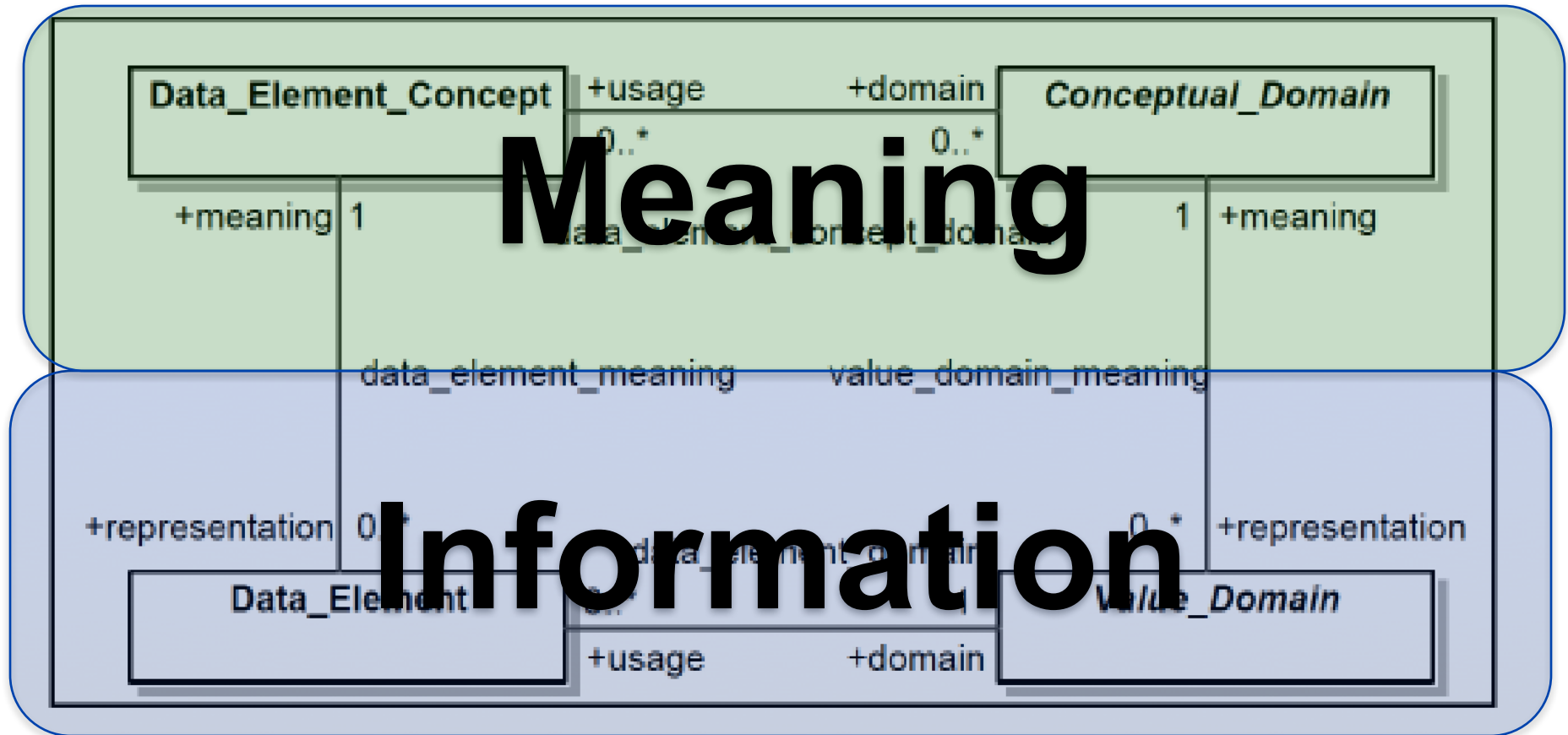
Study B

Yes = 1 No = 2
Event = 0 Censoring = 1

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

caDSR CDE: Assessment Method Type

https://cdebrowser.nci.nih.gov/CDEBrowser/search?dataElementConceptDetails=9&tabClicked=1&PageId=ValidValuesGroup

National Cancer Institute U.S. National Institutes of Health | www.cancer.gov

CDE Browser

Back Help

Data Element Data Element Concept Value Domain Classifications Usage Data Element Derivation Admin Info

Object Class [More Details](#)

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

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

Meaning of the CDE
Defined in NCI Thesaurus Concept Codes

caDSR CDE: Assessment Method Type

Value Domain

https://cdebrowser.nci.nih.gov/CDEBrowser/search?listValidValuesForDataElements=9&tabClicked=2&PageId=DataElementCc

National Cancer Institute

U.S. National Institutes of Health | www.cancer.gov

CDE Browser

Back Help

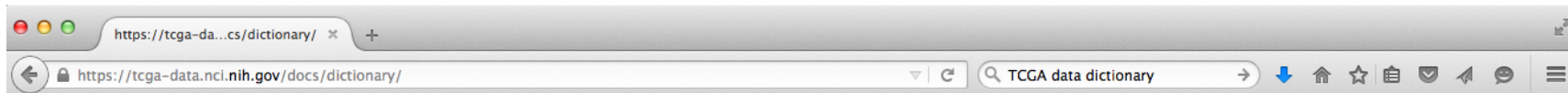
Data Element Data Element Concept Value Domain Classifications Usage Data Element Derivation Admin Info

Permissible Values

Meaning of the Permissible Values
Defined in NCI Thesaurus Concept Codes

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 attached, 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 1p/19q co-del 1p/19q intact (non-del) Only 19q del 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 enrollment 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 enrollment 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 Unknown Yes	
3225706v1	Day of Ablation	Ablation Performed Day Number	Numeric value that represents the day the ablation was performed.	22 21 20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1	
3225709v1	Were Ablation Techniques Utilized?	Ablation Performed Indicator	Text indicator to signify whether or not ablation techniques were used.	Unknown Yes 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 11 10 9 8 7 6 5 4 3 2 1	
3225702v1	Ablation Type	Ablation Performed Type	Text term to describe the type of ablation therapy.	Other Radiofrequency Ablation Microwave Ethanol Injection Radiosurgical Ablation	

Data Dictionary Viewer

[Dictionary Viewer](#) > Diagnosis

 Print

Diagnosis


 Download Template

TSV



Summary

Type	diagnosis
Category	clinical
Description	Data from the investigation, analysis and recognition of the presence and nature of disease, condition, or injury from expressed signs and symptoms; also, the scientific determination of any kind; the concise results of such an investigation.
Unique Keys	<ul style="list-style-type: none">idproject_id, submitter_id

Links

Links to Entity	Link Name	Relationship	Required?
 Case	cases	Diagnoses Describes Case	Yes

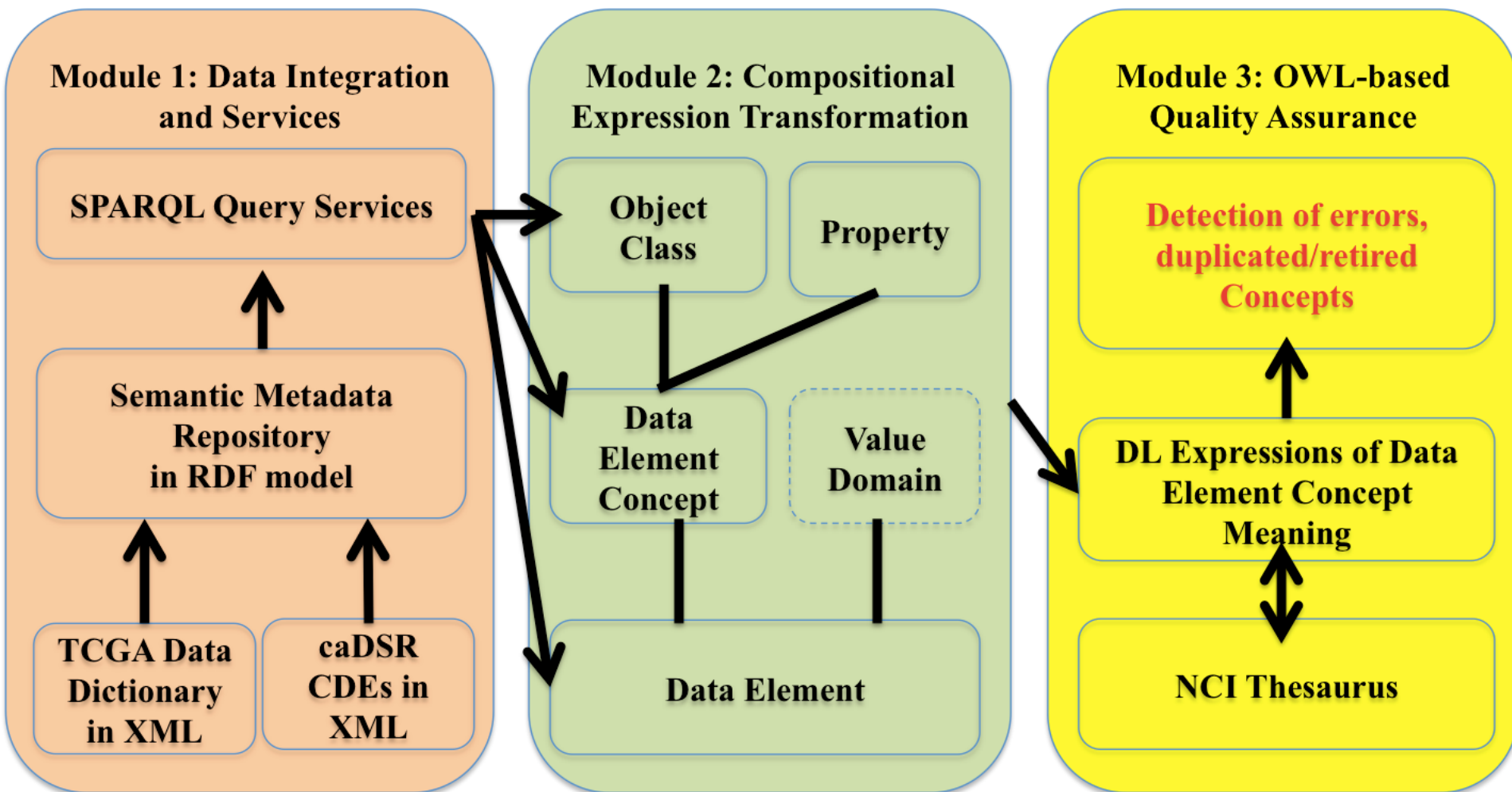
Properties

Property	Description	Acceptable Types or Values	Required?	CDE
 age_at_diagnosis	Age at the time of diagnosis expressed in number of days since birth.	number, null	Yes	3225640 - caDSR
		<ul style="list-style-type: none">Enumeration:<ul style="list-style-type: none">primarymetastasisrecurrenceotherUnknown		
 classification_of_tumor	Text that describes the kind of disease present in the tumor specimen as related to a specific timepoint.		Yes	3288124 - caDSR

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
 - FHIR Data Validation
- Aim 3: To deploy and evaluate a QA web-portal for collaborative CDE review and harmonization.
 - D2Refine: A platform for metadata standardization and harmonization

Semantic Web OWL-based QA Tools



version (http://rdf.cadsr.org/cde/version/20150207) : [/Users/m005994/Documents/icd11project/dec-ontology-02.ttl]

version (http://rdf.cadsr.org/cde/version/20150207)

Classes Object Properties Data Properties Annotation Properties Individuals OWLViz DL Query OntoGraf QDM Workbench SPARQL Query

Class hierarchy: 'Karnofsky Performance Status Score'

Class hierarchy (inferred)

Annotations: 'Karnofsky Performance Status Score'

Annotations

label
Karnofsky Performance Status Score

prefLabel
Karnofsky Performance Status Score

altLabel
Performance Status Assessment

Description: 'Karnofsky Performance Status Score'

Equivalent To

'Observable Entity' and ('is about' some (C25217 and ('inherits 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'

CDEs Violating Constraints

Equivalent CDEs

The screenshot displays the OntoGraf web interface for the 'Karnofsky Performance Status Score' class. The left pane shows the class hierarchy, with 'Observable Entity' as the parent. The right pane shows the annotations and description. Red arrows highlight specific classes in the hierarchy that violate constraints and their equivalent classes in the description.

- CDEs Violating Constraints:** This text is positioned in the center-right of the image. Red arrows point from the following classes in the hierarchy to it:
 - '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 Number'
 - 'Case Report Form Tissue Source Location Complete Year Number'
 - 'Case Report Form Tissue Source Location Completion Day Number'
 - '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 1'
 - '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 Co'
 - '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'
- Equivalent CDEs:** This text is positioned in the center-right of the image. Red arrows point from the following classes in the description to it:
 - 'Performance Status Assessment Timepoint Category'
 - 'Performance Status Assessment Eastern Cooperative Oncology Group Scale'

Reasoner state out of sync with active ontology ☒ Show Inferences

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
 - **FHIR Data Validation**
- Aim 3: To deploy and evaluate a QA web-portal for collaborative CDE review and harmonization.
 - D2Refine: A platform for metadata standardization and harmonization

ShEx, RDF and FHIR

Eric Prud'hommeaux – W3C/MIT (presenter)
Harold Solbrig – Mayo Clinic
Guoqian Jiang – Mayo Clinic
And the caCDE-QA Team

NCI ITCR 2017 Annual Meeting
Santa Cruz, CA
May 31 - June 1, 2017

ShEx as part of FHIR specification in STU3

Browser address bar: <https://www.hl7.org/fhir/rdf.html>

FHIR® FHIR Release 3 (STU)

Home Getting Started Documentation Resources Profiles Extensions Operations Terminologies

Linked Data > RDF

Formats XML JSON **RDF**

3.1 Resource Description Framework (RDF) Representation

FHIR Infrastructure Work Group	Maturity Level: 2	Ballot Status: Trial Use
--	-------------------	--------------------------

This page and the RDF forms are jointly maintained by the HL7 FHIR project and the [W3C Semantic Web Health Care and Life Sciences Interest Group](#).

W3C

FHIR resources can be represented as an [RDF graph](#) serialised in the [Turtle format](#), or as [JSON-LD](#). The RDF format is defined to assist the process of bridging between operational data exchange and formal knowledge processing systems. While the RDF form offers a fully functional representation of FHIR resources, it has different operational characteristics to the [JSON](#) and [XML](#) representations, and would be implemented for different reasons. Systems focused on operational exchange of data would not generally use choose to use RDF.

3.1.1.6 Schema

FHIR uses ShEx for representing the turtle schema. See [fhir.shex](#) for definitions.

- Logical Model**
 - Structure
 - Definitions
 - Examples
- Mappings**
 - Summary
 - Mapping Script
- FHIR Resources**
 - Profile Summary
 - Examples
- HL7 V2.4 Messaging**
 - Examples

FHIR Profiles for Structured Cancer Reports

Colorectal Report

This implementation guide is developed from the Royal College of Pathologists of Australasia's (RCPA) "Colorectal Cancer - Structured Reporting Protocol" (2nd edition 2012). The colorectal cancer protocol is not intended to apply to tumours of the appendix, small bowel and anus. Local excisions of colorectal carcinomas will be dealt with in a subsequent protocol. Synchronous primary tumours should have separate protocols recorded for each tumour.

This is the [Colorectal Logical Model](#). For context, see the [explanation of how this guide works](#).

Colorectal Report

Name	Flags	Card.	Type	Description & Constraints
Colorectal				
subject	Σ	1..1	Reference(Australian Patient (Human))	
requester	Σ	0..1	Reference(Practitioner)	
performer	Σ	0..1	Reference(Practitioner), Reference(Organization)	
preAnalytic	I	0..1		01: If there is a perforation, the perforation indicated.
clinicalInformation		0..1	string	
operatingSurgeonDetails	Σ	0..1	Reference(Practitioner)	
perforation	I	0..1	code	Binding: PresentAbsentNotstated (required)
natureOfPerforation		0..*	code	Binding: PerforationType (required)
clinicalObstruction		0..1	code	Binding: PresentAbsentNotstated (required)
tumourLocation	I	0..*	code	Binding: ColorectalCancerTumourLocation (required) 02: If the tumour location is rectum, then the tumour is from the a
synchronousTumours		0..1	string	
distanceAnalVerge		0..1	Quantity	
typeOfOperation		0..1		
code	I	0..1	code	Binding: ColorectalCancerOperationType (required) 03: If other procedure(s) is selected, then record the type of procedure. 04: If anterior resection is selected, record the resection type.
other		0..1	string	
anteriorResectionType		0..1	code	Binding: HiLowUltralow (required)

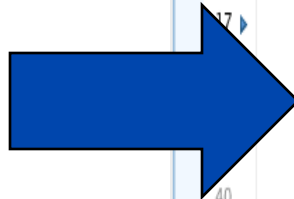
This value set contains 9 concepts

All codes from system <http://snomed.info/sct>

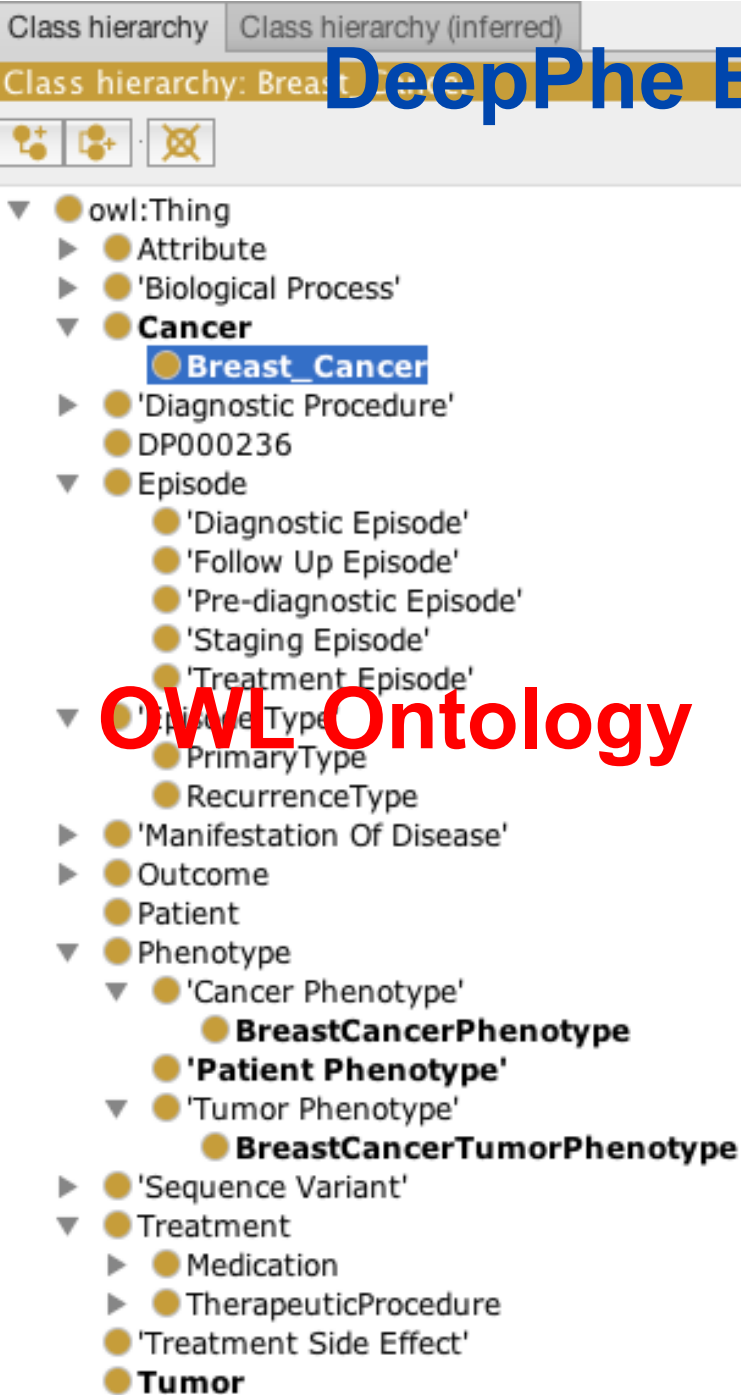
Code	Display	Definition
32713005	Caecum	
9040008	Ascending colon	
48338005	Hepatic flexure	
485005	Transverse colon	
72592005	Splenic flexure	
32622004	Descending colon	
60184004	Sigmoid colon	
49832006	Rectosigmoid junction	
34402009	Rectum	

DeepPhe Breast Cancer Model

OWL Ontology



FHIR Profile



breastcancermodel.xml

```
<?xml version="1.0" encoding="UTF-8" standalone="yes" ?>
<StructureDefinition xmlns="http://hl7.org/fhir" xmlns:ns2="http://www.w3.org/1999/xhtml">
  <id value="Breast Cancer"/>
  <meta>
    <lastUpdated value="2016-04-28T00:42:15.124-05:00"/>
  </meta>
  <url value="http://hl7.org/fhir/StructureDefinition/BreastCancer"/>
  <status value="draft"/>
  <publisher value="NCI ITCR caCDE-QA and DeepPhe Project Team (Breast Cancer Profile)"/>
  <date/>
  <description value="Base StructureDefinition for Breast Cancer Profile"/>
  <fhirVersion value="1.0.2"/>
  <kind value="resource"/>
  <abstract value="false"/>
  <base value="http://hl7.org/fhir/StructureDefinition/DomainResource"/>
  <snapshot>
    <element> [9 lines]
    <element> [9 lines]
    <element>
      <path value="BreastCancer.hasBodySite"/>
      <short value="has BodySite"/>
      <definition value="has BodySite"/>
      <min value="0"/>
      <max value="*/>
      <type>
        <code value="BodySite"/>
      </type>
    </element>
    <element>
      <path value="BreastCancer.hasCourse"/>
      <short value="has Course"/>
      <definition value="has Course"/>
      <min value="0"/>
      <max value="*/>
      <type>
        <code value="Course"/>
      </type>
    </element>
  </snapshot>
</StructureDefinition>
```

Modeling and validating HL7 FHIR profiles using semantic web Shape Expressions (ShEx)



Harold R. Solbrig^a, Eric Prud'hommeaux^{b,c}, Grahame Grieve^d, Lloyd McKenzie^e, Joshua C. Mandel^f,
Deepak K. Sharma^a, Guoqian Jiang^{a,*}

^a Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA

^b Computer Science and Artificial Intelligence Laboratory, Massachusetts Institutes of Technology, Cambridge, MA, USA

^c World Wide Web Consortium (W3C), Cambridge, MA, USA

^d Health Intersections Pty Ltd, Victoria, Australia

^e Gevity Consulting Inc, Vancouver, BC, Canada

^f Harvard Medical School Department of Biomedical Informatics, Boston, MA, USA

ARTICLE INFO

Article history:

Received 18 November 2016

Revised 10 February 2017

Accepted 12 February 2017

Available online 16 February 2017

Keywords:

HL7 Fast Healthcare Interoperability

Resources (FHIR)

Shape Expressions (ShEx)

Resource Description Framework (RDF)

Semantic web technology

Quality assurance

ABSTRACT

Background: HL7 Fast Healthcare Interoperability Resources (FHIR) is an emerging open standard for the exchange of electronic healthcare information. FHIR resources are defined in a specialized modeling language. FHIR instances can currently be represented in either XML or JSON. The FHIR and Semantic Web communities are developing a third FHIR instance representation format in Resource Description Framework (RDF). Shape Expressions (ShEx), a formal RDF data constraint language, is a candidate for describing and validating the FHIR RDF representation.

Objective: Create a FHIR to ShEx model transformation and assess its ability to describe and validate FHIR RDF data.

Methods: We created the methods and tools that generate the ShEx schemas modeling the FHIR to RDF specification being developed by HL7 ITS/W3C RDF Task Force, and evaluated the applicability of ShEx in the description and validation of FHIR to RDF transformations.

Results: The ShEx models contributed significantly to workgroup consensus. Algorithmic transformations from the FHIR model to ShEx schemas and FHIR example data to RDF transformations were incorporated into the FHIR build process. ShEx schemas representing 109 FHIR resources were used to validate 511 FHIR RDF data examples from the Standards for Trial Use (STU 3) Ballot version. We were able to uncover unresolved issues in the FHIR to RDF specification and detect 10 types of errors and root causes in the actual implementation. The FHIR ShEx representations have been included in the official FHIR web pages for the STU 3 Ballot version since September 2016.

Discussion: ShEx can be used to define and validate the syntax of a FHIR resource, which is complementary to the use of RDF Schema (RDFS) and Web Ontology Language (OWL) for semantic validation.

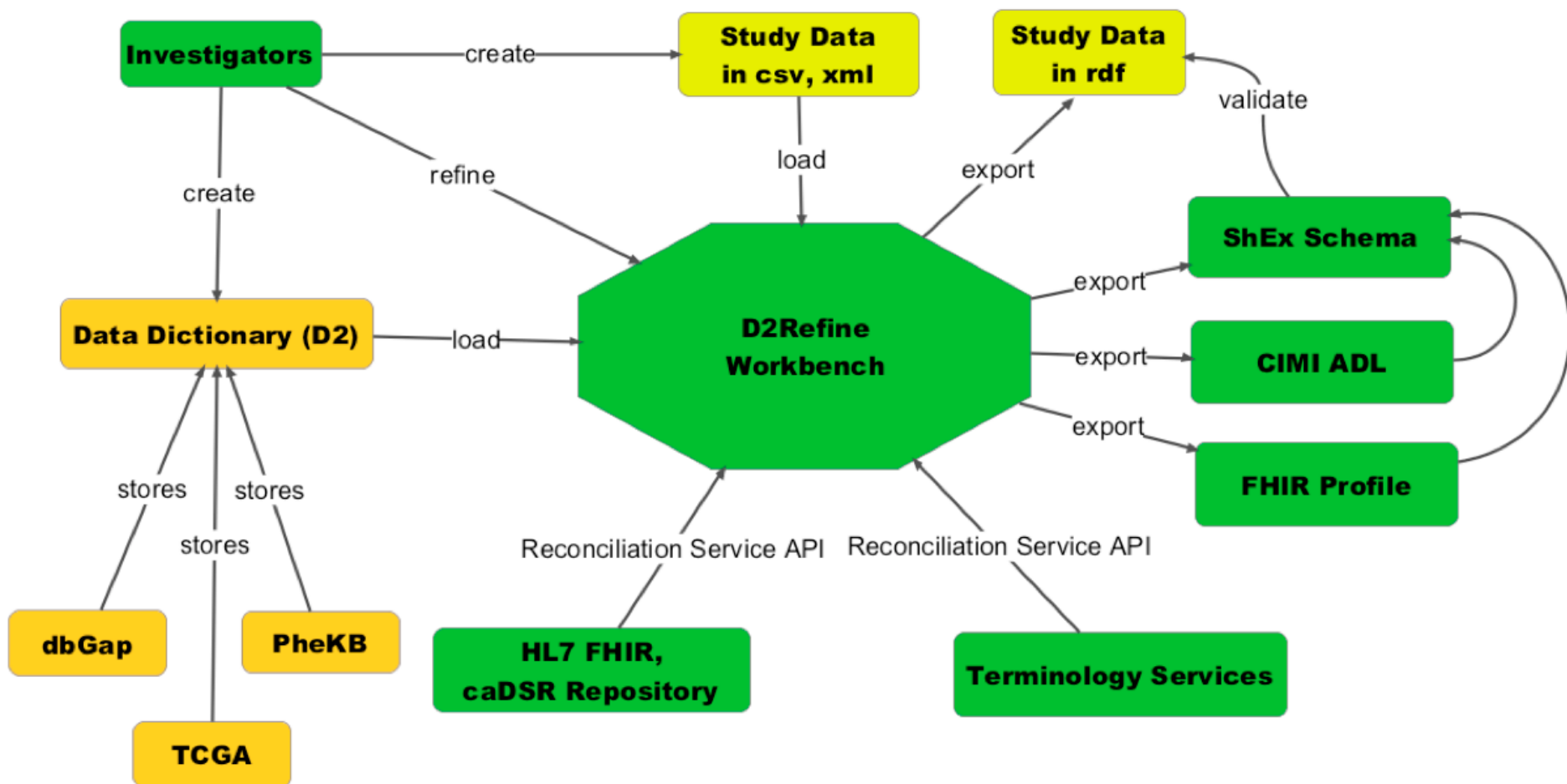
Conclusion: ShEx proved useful for describing a standard model of FHIR RDF data. The combination of a formal model and a succinct format enabled comprehensive review and automated validation.

© 2017 Elsevier Inc. All rights reserved.

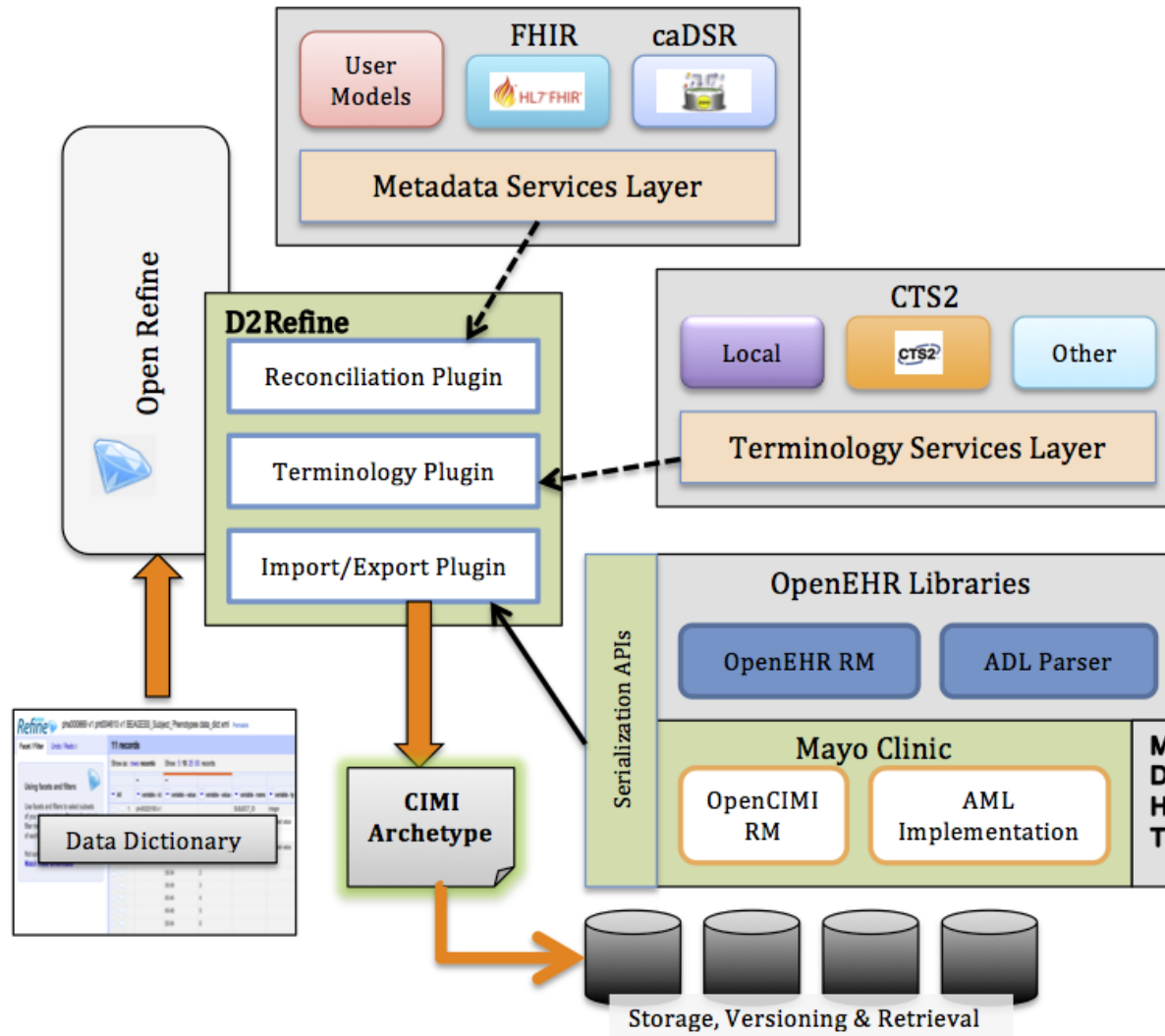
Specific Aims of caCDE-QA

- 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
 - FHIR Data Validation
- Aim 3: To deploy and evaluate a QA web-portal for collaborative CDE review and harmonization.
 - **D2Refine: A platform for metadata standardization and harmonization**

D2Refine for Metadata Harmonization and Validation



D2Refine Plugins



D2Refine CTS2 Reconciliation Plugin

The screenshot displays the D2Refine CTS2 Reconciliation Plugin interface. The browser address bar shows the URL: `127.0.0.1:3333/project?project=1950253612597`. The page title is `phs000730 v1 pht005060 v1 Brain_Metastasis_Subject_Phenotypes data_dict.xml`. The interface includes a sidebar with a "Using facets and filters" section, a main data table, and a search for match dialog.

Using facets and filters

Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started?
[Watch these screencasts](#)

15 records

Show as: **rows** records Show: 5 10 25 50 records

Extensions: D2Refine

	variable - id	variable - name	variable - description	variable - type	variable - value
1.	phv00253452.v1	SUBJID	Unique Participant Identifier	string	
2.	phv00253453.v1	Gender	Participant's gender as Male or Female	string	Female
3.	phv00253454.v1	Primary Site	Primary site of patient tumor	string	

Search for match

Search for "Female"

☒ Match other cells with same content
☐ Match this cell only

Match **New Topic** **Don't Reconcil**

Select an item from the list:

- female [obo:PATO_0000383] <http://lexevs63cts2.nci.nih.g>
- female [obo:PATO_0000383] <http://lexevs63cts2.nci.nih.g>
- female [MGEDontology.owl:MO] <http://lexevs63cts2.nci.nih.g>
- Female [RIM_none:10174:F] <http://lexevs63cts2.nci.nih.g>
- Female [LOINC:LA3-6] <http://lexevs63cts2.nci.nih.g>
- Female [ns1363824265:C00157] <http://lexevs63cts2.nci.nih.g>
- Female [NCI_Thesaurus:C16576] <http://lexevs63cts2.nci.nih.g>
- Female [NCI_Thesaurus:C16576] <http://lexevs63cts2.nci.nih.g>
- Female [NCI_Thesaurus:C16576] <http://lexevs63cts2.nci.nih.g>
- Female [NCI_Thesaurus:C16576] <http://lexevs63cts2.nci.nih.g>

Female

- ☒ female [obo:PATO_0000383] (1)
- ☒ Female [RIM_none:10174:F] (0)
- ☒ Female [LOINC:LA3-6] (0)
- ☒ female [obo:PATO_0000383] (1)
- ☒ female [MGEDontology.owl:MO_506] (1)
- ☒ Female [ns1363824265:C0015780] (0)
- ☒ Female [NCI_Thesaurus:C16576] (0)
- ☒ Female [NCI_Thesaurus:C16576] (0)
- ☒ Female [NCI_Thesaurus:C16576] (0)
- ☒ Create new topic

Search for match

Male

- ☒ male [obo:PATO_0000384] (1)
- ☒ Male [RIM_none:10173:M] (0)
- ☒ Male [LOINC:LA2-8] (0)
- ☒ male [obo:PATO_0000384] (1)
- ☒ male [MGEDontology.owl:MO_652] (1)
- ☒ Males [ns1363824265:C0086582] (1)
- ☒ Male [NCI_Thesaurus:C20197] (0)
- ☒ Male [NCI_Thesaurus:C20197] (0)
- ☒ Male [NCI_Thesaurus:C20197] (0)
- ☒ Male [NCI_Thesaurus:C20197] (0)
- ☒ Create new topic

Search for match

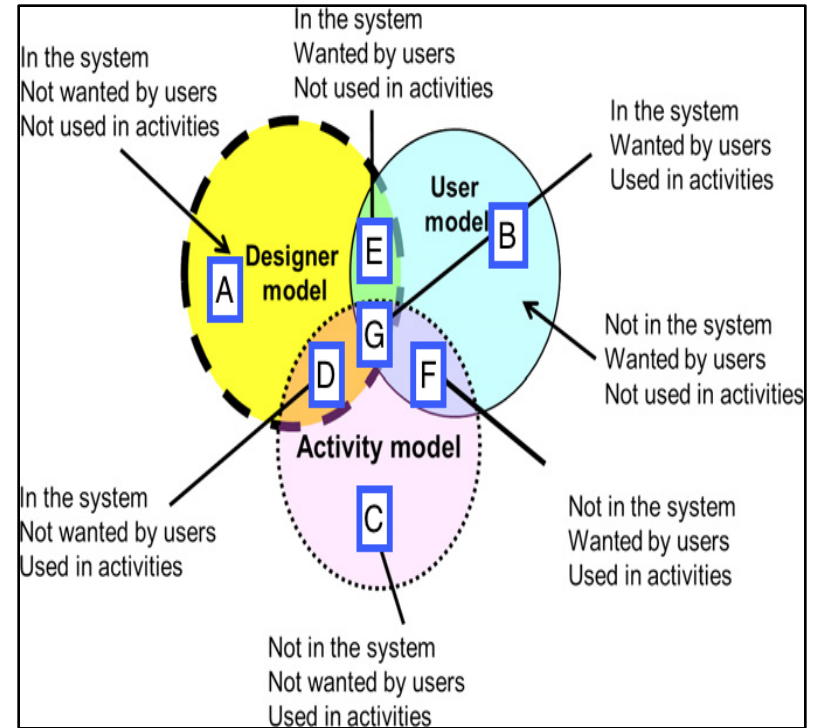
Breast

- ☒ Breast [NCI_Thesaurus:C12971] (0)
- ☒ Breast [NCI_Thesaurus:C12971] (0)
- ☒ Breast [NCI_Thesaurus:C12971] (0)
- ☒ Breast [NCI_Thesaurus:C12971] (0)
- ☒ Breast [LOINC:LA4255-1] (0)
- ☒ Create new topic

Search for match

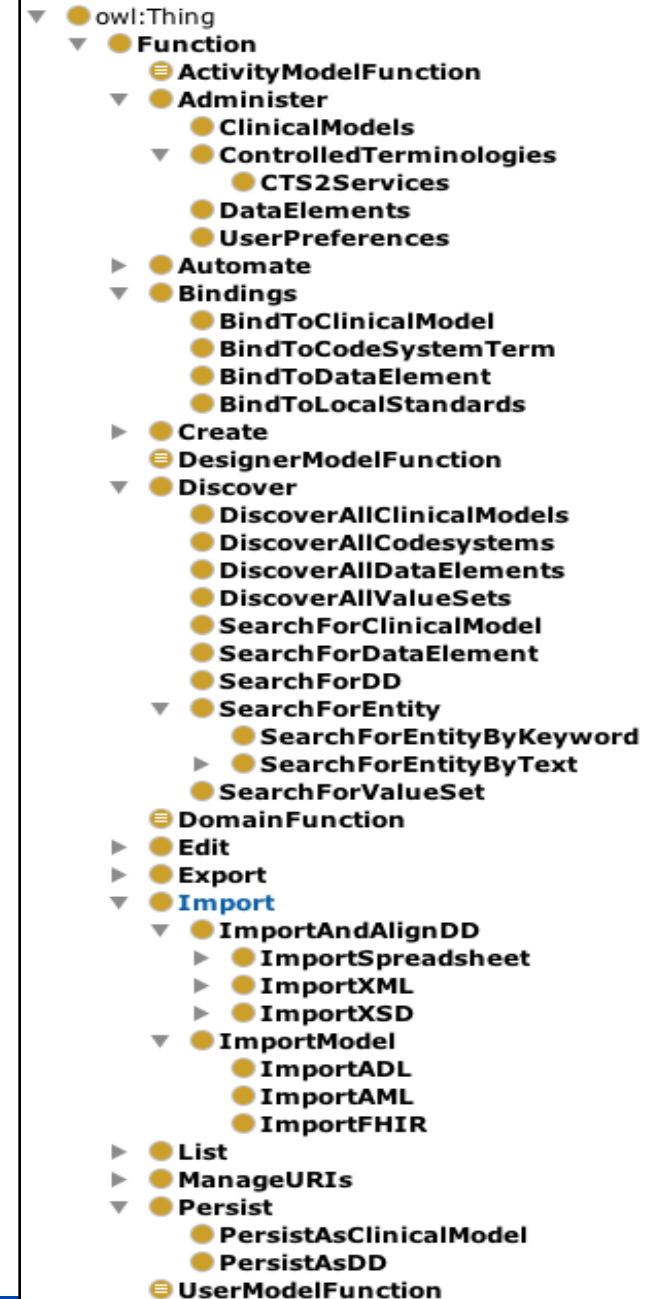
TURF Usability Framework

- **Within-Model Domain Function Saturation:** Ratio of domain functions in Designer Model to total functions in Designer Model. This is calculated as *sum of numerators of the fractions in (A, D, E, and G) divided by Sum of denominators of the fractions in (A, D, E, and G)*.
- **Across-Model Domain Function Saturation:** Ratio of domain functions in Designer Model to domain functions in all three models (Designer, User and Activity). This is calculated as *sum of numerators of the fractions in (A, D, E, and G) divided by the count of all domain functions*.




Domain Ontology for Usability Study

- The first iteration of *user analysis* and the *function analysis* of D2Refine is completed - by capturing the functions and tasks in the Domain Ontology.
- The Usability study will compare D2Refine with few other platforms, e.g. OntoMaton and eleMap, that provide similar functions and this would augment the domain ontology.
- Identified users would be invited to participate in the study and perform the tasks.



ITCR Poster Session

- D2Refine: A platform for clinical research study data element harmonization and standardization.



D2Refine: A Platform for Clinical Research Study Data Element Harmonization and Standardization

Deepak K. Sharma, MS¹, Harold R. Solbrig, MS¹, Eric Prud'hommeaux², Guoqian Jiang, PhD¹
 Department of Health Sciences Research, Mayo Clinic, Rochester, MN¹, W3C/MIT, Boston, MA²

Abstract

Introduction: The D2Refine Workbench (D2Refine for short) aimed at addressing the issues of high complexity and steep learning curve of existing tools, when it comes to create and manage clinical models for semantic interoperability. D2Refine, built on top of OpenRefine¹, offers simple, spreadsheet-like interface. D2Refine augments OpenRefine's already rich extensible platform for terminology binding with CTS2 and, serialization of clinical models into standard formats like ADL and AML.

Materials: D2Refine's development is based on the following crucial ingredients:

- OpenRefine: Popular platform to manage data
- Sources: Data Dictionaries in various formats from cDGM, PheKB, and TQGA
- Standard Template: Defines minimum set of requirements to be filled in by a Data Dictionary
- Standards: OpenEHR's ADL, OMG AML, W3C RDF, Shape Expressions (SHEX), HL7 FHIR
- Reference Information Model: CIML Core RM
- Libraries: UML Implementations, OpenEHR ADL

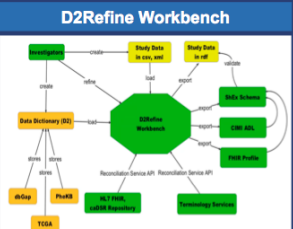
System Design: The equivalence of the data dictionaries definitions and the constraints of an archetype is the central idea at D2Refine. The standard template defines the minimum set of requirements (derived from OMG AML Specifications) to be satisfied by a data dictionary, to translated it into a clinical archetype. This enables us to represent it using standard formats of ADL, AML, SHEX and FHIR Profiles.

D2Refine introduces capability CTS2 reconciliation services to standardize the source content. This enables us create terminology bindings of data dictionary elements with CTS2 compliant controlled terminologies. The reconciliation services helps validate the model and identify gaps in data dictionary design.

D2Refine comes pre-configured reconciliation services with NCI LexEVS CTS2 REST Services interface. The reconciliation service could be extended to work not only with terminology services but also with data element repositories like NCI caDSR².

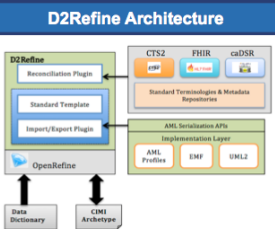
D2Refine extends export plugin of OpenRefine to serialize models in ADL and AML.

D2Refine Workbench



The diagram shows the D2Refine Workbench architecture. It features a central 'D2Refine Workbench' component. To its left is the 'Data Dictionary (SD)' which connects to 'Data Dictionary (SD)' and 'Data Dictionary (SD)'. To its right are 'Data Dictionary (SD)' and 'Data Dictionary (SD)'. Below the central component is a 'Reconciliation Service API' which connects to 'Reconciliation Service API' and 'Reconciliation Service API'. The central component also connects to 'Data Dictionary (SD)' and 'Data Dictionary (SD)'.

D2Refine Architecture

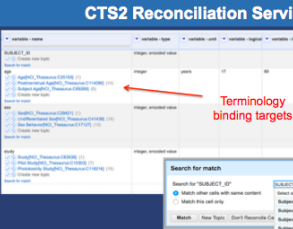


The diagram shows the D2Refine Architecture. It features a central 'D2Refine' component. To its left is the 'Standard Template' which connects to 'Standard Template' and 'Standard Template'. To its right is the 'OpenRefine' component which connects to 'OpenRefine' and 'OpenRefine'. Below the central component is the 'Data Dictionary' which connects to 'Data Dictionary' and 'Data Dictionary'. The central component also connects to 'Data Dictionary' and 'Data Dictionary'.

Results

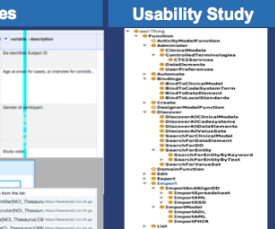
- A beta version of D2Refine is available to download from its GitHub repository¹.
- D2Refine includes a configurable and fully working reconciliation service to any CTS2 service for terminology binding.
- The export extensions of D2Refine offer serialization of a data dictionary to OpenEHR's ADL format.
- The implementation of the Standard Template, OMG AML's profiles and to support loading models in ADL, AML is in-progress.
- To gauge the effectiveness and usefulness of D2Refine, a usability study is in-progress. This Usability Study follows the analysis components of the TURF³ Framework and two saturation metrics of functionality coverage are targeted.

CTS2 Reconciliation Services



The screenshot shows the CTS2 Reconciliation Services interface. It features a search bar with the text 'Search for match'. Below the search bar is a list of results. A red arrow points to the 'Terminology binding targets' section.

Usability Study



The screenshot shows the Usability Study interface. It features a list of results. A red arrow points to the 'Terminology binding targets' section.


Conclusions

D2Refine enhances semantic interoperability of models with improved capabilities to manage heterogeneous clinical study data dictionaries for their standardization and harmonization study

- We identified elements to create minimum set of requirements for data dictionaries.
- D2Refine's CTS2 REST Reconciliation services extension implemented. The capability to search in a controlled terminology is also implemented.
- The ADL export extension is implemented to create ADL archetypes from data dictionaries.

Usability Metrics

- **Within-Model Domain Function Saturation:** Ratio of domain functions in Designer Model to Total functions in Designer Model
 - Sum of fraction numerators (A, D, E, G) divided by Sum of fraction Denominators (A, D, E, G)
- **Across-Model Domain Function Saturation:** Ration of domain functions in Designer Model to domain functions in all models (Designer, User and Activity)
 - Sum of fraction numerators (A, D, E, G) divided by count of all functions



The diagram shows the Usability Metrics. It features a central 'Activity model' component. To its left is the 'Designer Model' which connects to 'Designer Model' and 'Designer Model'. To its right is the 'User Model' which connects to 'User Model' and 'User Model'. Below the central component is the 'Activity model' which connects to 'Activity model' and 'Activity model'. The central component also connects to 'Activity model' and 'Activity model'.

Acknowledgement

This work has been supported in part by funding from NIH R01 GM105688 and a NCI U01 Project known as caCDE-QA (U01 CA180940)

References

1. D2Refine Project and Wiki
<https://github.com/Health-Sciences-Research/d2refine>
<https://github.com/Health-Sciences-Research/d2refine/wiki>
2. TURF: Toward a unified framework of EHR usability
<http://www.academichandbook.com/doi/pdf/10.1533/9780441110713.98>

Summary

- Standards-based representations and Semantic Web technologies are two enabling technologies
- 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.

Acknowledgements

- This study is supported in part by the NCI U01 project (U01 CA180940; PI: Jiang): caCDE-QA - A quality assurance platform for cancer study common data elements.
- Development Team
 - Harold Solbrig, Mayo Clinic
 - Deepak Sharma, Mayo Clinic
 - Rick Kiefer, Mayo Clinic
 - Eric Prud'hommeaux, MIT/W3C
- Co-Investigators/Collaborators
 - Dr. Cui Tao, UTHealth
 - Dr. Chunhua Weng, Columbia University
 - Dr. Christopher Chute, Johns Hopkins
 - ITCR Projects: DeepPhe(U24), UTHealth (U24), and OmniSearch (U01)

Useful Links

- The caCDE-QA Project GitHub
 - <https://github.com/caCDE-QA>
- FHIR ShEx Tutorial Materials
 - <https://github.com/caCDE-QA/ShEX>
- ShEx Web-based Validation Tools
 - <http://rawgit.com/shexSpec/shex.js/master/doc/shex-simple.html#>
 - <http://rawgit.com/shexSpec/shex.js/remote-query/doc/shex-simple.html>
- D2Refine
 - <https://github.com/caCDE-QA/D2Refine>



Questions & Discussion