A Semantic Web-based Quality Assurance Tool for Cancer Study Common Data Elements

Guoqian Jiang, MD, PhD[1], Harold R. Solbrig[1], Eric Prud'hommeaux[2], Cui Tao, PhD[3], Chunhua Weng, PhD[4], Christopher G. Chute, MD. Dr. PH[5]
[1] Department of Health Sciences Research, Mayo Clinic, Rochester, MN; [2] W3C/MIT, Boston, MA; [3]University of Texas Health Science Center at Houston, Houston, TX; [4]Columbia University, New York City, NY; [5]Johns Hopkins University, Baltimore, MN
Email to: Jiang.Guoqian@mayo.edu

Abstract

Background: Domain-specific common data elements (CDEs) are emerging as an effective approach to standards-based clinical research data storage and retrieval. A limiting factor, however, is the lack of robust automated quality assurance (QA) tools for the CDEs in clinical study domains.

Objective: The objectives of the present study are to prototype and evaluate a Semantic Web-based QA tool for the study of cancer CDEs using a post-coordination approach.

Methods: The study starts by integrating the NCI caDSR CDEs and The Cancer Genome Atlas (TCGA) data dictionaries in a single Resource Description Framework (RDF) data store.

We designed a compositional expression pattern based on the Data Element Concept model structure informed by ISO/IEC 11179, and developed a transformation tool that converts the pattern-based compositional expressions into the Web Ontology Language (OWL) syntax.

Invoking reasoning and explanation services, we tested the system utilizing the CDEs extracted from two TCGA clinical cancer study domains.

Results: In total, TCGA data dictionary contains 775 CDEs for 38 clinical cancer domains, which cover 21 cancer types. In the present study, we performed a case study of two clinical cancer domains: Clinical Pharmaceutical and Clinical Shared, which contain 18 and 98 CDEs respectively.

The reasoning services identified 6 CDEs with equivalent CDEs from the domain Clinical Pharmaceutical and 29 CDEs with equivalent CDEs from the domain Clinical Shared. In total, there are 12 groups of equivalent CDEs. Human-based review shows that among 12 groups of equivalent CDEs identified, the CDEs in 2 groups had modeling errors.

In total, there are 19 CDEs (out of 116 CDEs) identified with constraint violations. Human-based review shows that all 19 CDEs had modeling errors in their asserted primary properties.

Conclusion: Compositional expressions not only enable reuse of existing ontology codes to define new domain concepts, but also provide an automated mechanism for QA of terminological annotations for CDEs.

Objectives

Background:
National Cancer Institute (NCI) has implemented the Cancer Data Standards repository (caDSR) that adopted the ISO/IEC 11179 Metadata Registry (MDR) standard.

Part 3 of the ISO/IEC 11179 model describes a model for formally associating data model elements with their intended meaning.

Earlier studies [1-2] have uncovered serious issues with at least some of the caDSR definitions and have highlighted a need for robust, principled and automated quality assurance (QA) tools for the CDEs in cancer study domains.

Date element meanings, as recorded in the caDSR frequently use a simple form of "post-coordination", where a primary (focus) concept is identified along with one or more secondary identifiers that modify or qualify the intended target meaning. When taken on its own, this approach does not lend itself to automated validation and consistency checking.

Description Logic (DL)-based mechanisms allow ontology curators to formally and unambiguously represent concept meanings and relationships, and to use off the shelf reasoning tools such as HermiT to automate the computation of the relationship between two class expressions and consistency checks.

Objective:
The objective of the present study to design, develop and evaluate a quality evaluation tool for cancer study CDEs using a post-coordination approach.

Figure 1

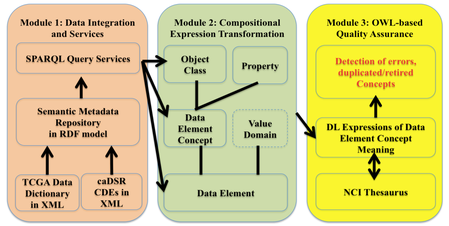

Figure 1. System Architecture

Methods

Figure 1 shows the architecture used in our evaluation.

- Module 1, Data Integration and Services combines the information from the caDSR Common Data Elements and the TCGA data dictionary as a cohesive unit, which allows the SPARQL query services to access the contents of both resources as a single unit.
- Module 2, Compositional Expression Transformation converts the data element meaning definitions recorded in the caDSR elements into DL expressions which become the inputs to Module 3.
- Module 3. OWL-based Quality Assurance which uses the combination of the NCI Thesaurus and additional disjointness axioms to detect potential errors and duplications in the data element definitions.

Figure 2

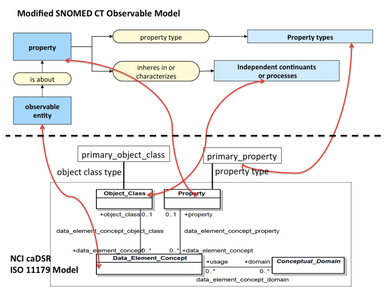

Figure 2. Mappings between SNOMED CT Observable Model and ISO/IEC 11197

Table 1: Compositional Expression

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

Figure 3

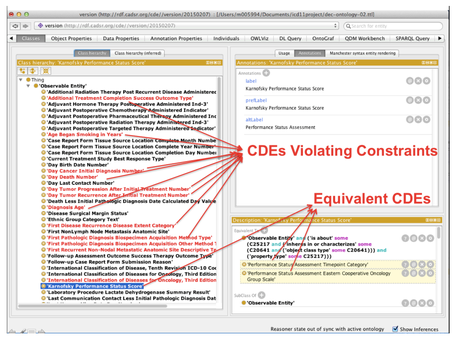

Figure 3. A screenshot of Protégé 5 illustrating CDEs violating constraints and equivalent CDEs are detected by DL reasoning services.

Table 2: Equivalent CDEs

Domain	Equivalent CDEs	Errors
Clinical Shared	2181650 Patient Smoking History Category 2228604 Started Smoking Year	Yes
Clinical Shared	88 Performance Status Assessment Eastern Cooperative Oncology Group Scale 2792763 Performance Status Assessment Timepoint Category 2003853 Karnofsky Performance Status Score	Yes

Table 3: Modeling Errors

Domain	CDEs Violating Constraints	Asserted Primary Property	Semantic Type	Our Suggested Primary Property	Semantic Type
Clinical Pharmaceutical	2975232 Prior Therapy Regimen Text	C1708|Agent	Chemical Viewed Functionally	C25365|Description	Intellectual Product
Clinical Shared	2791104 First Disease Recurrence Disease Extent Category	C13717|Anatomic Site	Body Location or Region	C25372|Category	Classification
Clinical Shared	3108203 Neoplasm Anatomic Subdivision Name	C13717|Anatomic Site	Body Location or Region	C42614|Name	Conceptual Entity
Clinical Shared	3124503 First Recurrent Non-Nodal Metastatic Anatomic Site Descriptive Text	C13717|Anatomic Site	Body Location or Region	C25365|Description	Intellectual Product
Clinical Shared	3427536 Tumor Disease Anatomic Site	C13717|Anatomic Site	Body Location or Region	C25365|Description	Intellectual Product
Clinical Shared	2006657 Diagnosis Age	C15220|Diagnosis	Diagnostic Procedure	C2515|Age	Organism Attribute
Clinical Shared	2896956 Month Cancer Initial Diagnosis Number	C15220|Diagnosis	Diagnostic Procedure	C25164|Date	Temporal Concept
Clinical Shared	2896958 Day Cancer Initial Diagnosis Number	C15220|Diagnosis	Diagnostic Procedure	C25164|Date	Temporal Concept
Clinical Shared	2896960 Year Cancer Initial Diagnosis Number	C15220|Diagnosis	Diagnostic Procedure	C25164|Date	Temporal Concept

Discussion

- The transformation allows us to take advantage of the built-in feature of OWL in expressing the disjointness and domain/range restrictions among a set of OWL classes, and subsequently invoke the reasoning services provided in existing OWL-DL reasoning tools (e.g., the HermiT reasoner to check the inconsistencies and violations .
- Leveraging the upper level ontologies such as UMLS Semantic Network, basic formal ontology (BFO) or BioTop ontology (a top-domain ontology for the life science) would potentially provide a formal approach to define the constraints for supporting the CDE modeling applications.

Conclusions

- We designed a compositional expression pattern based on a version of SNOMED CT observable model, which is used to represent the data structure of a data element concept (i.e., the meaning of data element) informed by the ISO/IEC 11179 metadata standard.
- Leveraging the existing Semantic Web tools, we demonstrated that the post-coordination approach could enable an effective and automated mechanism in detecting potential CDE modeling errors and duplicate CDEs.

Acknowledgements

The study is supported in part by a **NCI U01 Project – caCDE-QA** (1U01CA180940-01A1).

References

1. Jiang G, Solbrig HR, Chute CG. Quality evaluation of cancer study Common Data Elements using the UMLS Semantic Network. Journal of biomedical informatics. 2011;44 Suppl 1:S78-85. .
2. Jiang G, Solbrig HR, Chute CG. Quality evaluation of value sets from cancer study common data elements using the UMLS semantic groups. Journal of the American Medical Informatics Association : JAMIA. 2012;19(e1):e129-36 .