Harmonizing bioCADDIE Metadata Schemas for Indexing Clinical Research Datasets Using Semantic Web Technologies

Harold R. Solbrig[1], Guoqian Jiang[1]

[1] Mayo Clinic College of Medicine, Rochester, MN [solbrig.harold, jiang.guoqian]@mayo.edu

Abstract. An important role of the NIH Big Data to Knowledge (BD2K) biomedical and healthCAre Data Discovery Index Ecosystem (bioCADDIE) is to promote data integration through the adoption of content standards and alignment to common data elements and high-level schema. The objective of this study was to investigate how a combination of Semantic Web technologies and the ISO/IEC 11179 data element model could be used in the alignment of a biomedical study database and the bioCADDIE indexing schema. Using the database of Genotypes and Phenotypes (dbGaP) as a representative example, we were able to demonstrate the viability of the general approach and propose a number of promising next steps.

Keywords. Metadata Standards; The Biomedical and Healthcare Data Discovery Index (bioCADDIE); The database of Genotypes and Phenotypes (dbGaP); ISO/IEC 11179; Semantic Web Technologies

Introduction

The biomedical and healthCAre Data Discovery Index Ecosystem (bioCADDIE)[1] project has been funded by the NIH Big Data to Knowledge (BD2K) initiative to develop a data discovery index (DDI) prototype which will provide a searchable index of biomedical study data. An important role of the BD2K bioCADDIE is to promote data integration through the adoption of content standards and the alignment of common data elements with a high-level schema. As a harvester pilot project funded by bioCADDIE[2], we are creating a prototype of clinical research data discovery index (crDDI) using the HL7 Fast Healthcare Interoperability Resources (FHIR) standard[3]. The crDDI can be used to index datasets from NIH pilot data commons such as the database of Genotypes and Phenotypes (dbGaP)[4, 5] and The Cancer Genome Atlas (TCGA)[6]. One of the deliverables in this project was an investigation of the applicability of OWL and ISO/IEC 11179 metamodel[7] in the alignment of study metadata and the model of the bioCADDIE index.

Part 3 of the ISO/IEC 11179-3 defines a formal model of a data element registry and its basic attributes. The model provides a structure to represent data elements, their types, units of measure, possible values, etc. It also specifies how each of these components can be associated with their intended meaning -- the real world objects properties that these data elements represent. In this study, we transform the dbGaP and bioCADDIE models from their native XML Schema and JSON Schema representations into their corresponding OWL equivalents. We then align the results with an

OWL[8] representation of the ISO/IEC 11179-3 model, which serves the role of an Upper Level Ontology (ULO). We demonstrate that the result of this process, when used in combination with a description logic (DL) reasoner, can be used to discover, validate, and uncover issues with possible alignments between dbGaP and bioCADDIE model components.

Materials and Methods

This project utilized three resources - the XML Schema representation of the dbGaP data dictionary, the JSON Schema representation of the bioCADDIE metadata schema files, and an OWL representation the model of ISO/IEC 11179 Edition 3 Part 3.[9]

dbGap

dbGaP[4] is an NIH pilot data commons charged to archive, curate and distribute information produced by studies in investigating the interaction of genotype and phenotype. The dbGaP database structure is defined in XML Schema. **Figure 1** shows a diagram illustrating a portion of the dbGaP schema for the Study resource.

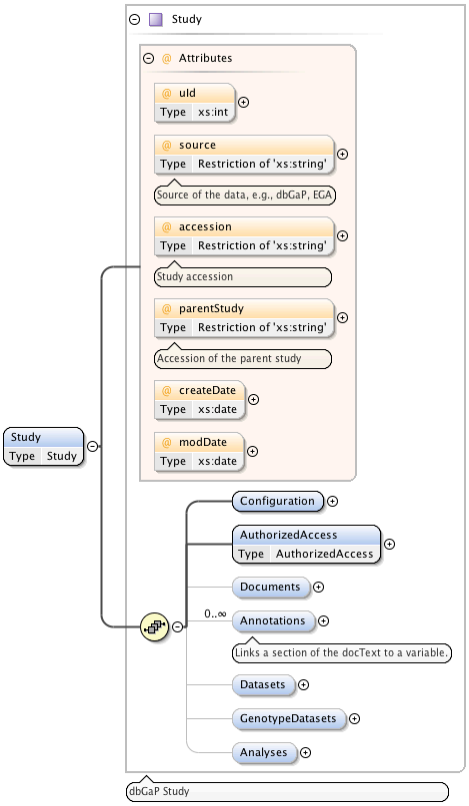


Figure 1. The diagram illustrating the dbGaP study and data dictionary XML schemas.

bioCADDIE

bioCADDIE[1] is a data discovery index (DDI) being developed to index resources such as dbGaP. The bioCADDIE metadata schema files represent a collection of descriptive metadata and structure for datasets being developed by the bioCADDIE Metadata Specification Working Group 3.[10] Figure 2 shows an overview of the structure of the bioCADDIE metadata schemas.

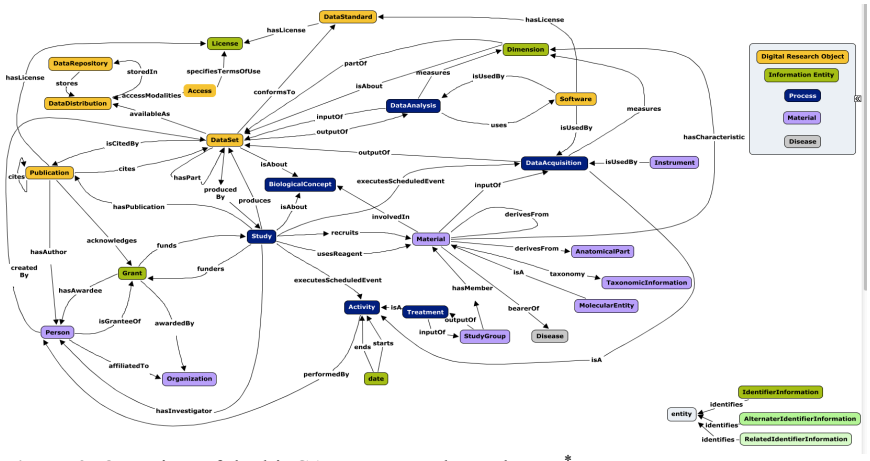


Figure 2. Overview of the bioCADDIE metadata schemas*

ISO/IEC 11179-3

ISO/IEC 11179 Edition 3 Part 3[7] defines a conceptual model of a metadata registry - a registry of the contents and semantics of data elements such as those defined in dbGaP and bioCADDIE. An OWL representation of this conceptual model was developed by the eXtended MetaData Repository (XMDR) group[11] for use with RDF based metadata repositories. Figure 3 shows a high level view of the ISO/IEC 11179 metamodel.

Methods

We converted the dbGap and bioCADDIE schemas into their equivalent OWL representation, using the ISO/IEC 11179 model as a common Upper Level Ontology (ULO), and demonstrated how an OWL DL reasoner could be used to evaluate proposed similarities between elements in the two models. We then prototyped approaches to using a reference ontology such as Ontology of Clinical Research (OCRe)† to *discover* common or similar elements between the two models.

The dbGaP XML Schema was transformed into the OWL equivalent by representing the complex types as OWL Classes, properties referencing complex types as Object Properties and properties referencing simple types and their restrictions as Data Properties. Data properties were classified as sub properties of ISO 11179 Data Element, their target types as 11179 Value Domain with the Boolean and Enumeration types as

* Source: https://github.com/biocaddie/WG3-MetadataSpecifications

† https://bioportal.bioontology.org/ontologies/OCRE

Enumerated Value Domain and everything else as Described Domain. RDF domains and ranges were defined for the properties. A prefix, "GAP ", was added to all labels to allow dbGaP elements to be easily distinguished in the combined environment.

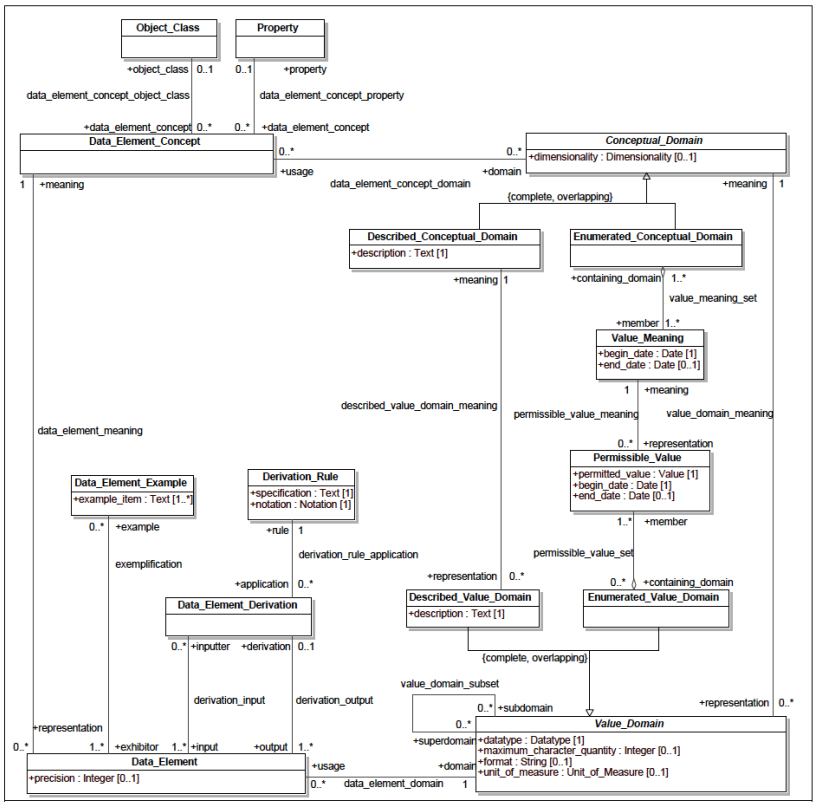


Figure 3. ISO/IEC 11179-3 Consolidated Data Description metamodel

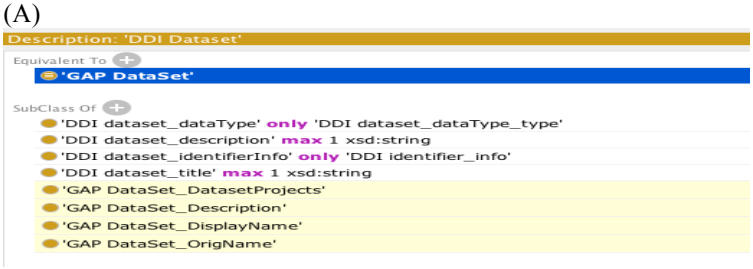
We repeated this transformation process with the bioCADDIE JSON Schema, mapping objects to OWL Classes, containments and associations to OWL Object Properties and data types to OWL Data Elements and then anchoring this transformation to the ISO 11179 model as described above, prepending the labels with "DDI".
We then created a mapping ontology that imported the dbGaP, bioCADDIE and ISO 11179 ontologies. This mapping ontology allowed us to (attempt to) make assertions about potential alignments between dbGaP and bioCADDIE data elements and properties and to understand the ramifications of these decisions.

Results and Discussion

We successfully transformed the dbGaP and bioCADDIE metadata schemas into a common OWL / ISO 11179 syntax and have demonstrated that the DL reasoner can be used to determine and validate equivalence assertions between the resources. As

an example, the assertion that the dbGaP Dataset is equivalent to the bioCADDIE Dataset asserts that both classes include createDates, modDates, original names, data types, etc. (see **Figure 4A**). This process also uncovers underspecified types such as the dbGaP "DisplayName" element in Dataset. An assertion about the equivalence of the dbGaP Dataset "Accession ID" with the bioCADDIE Dataset "identifierInfo" shows as a typing error (see **Figure 4B**).

(A)



(B)

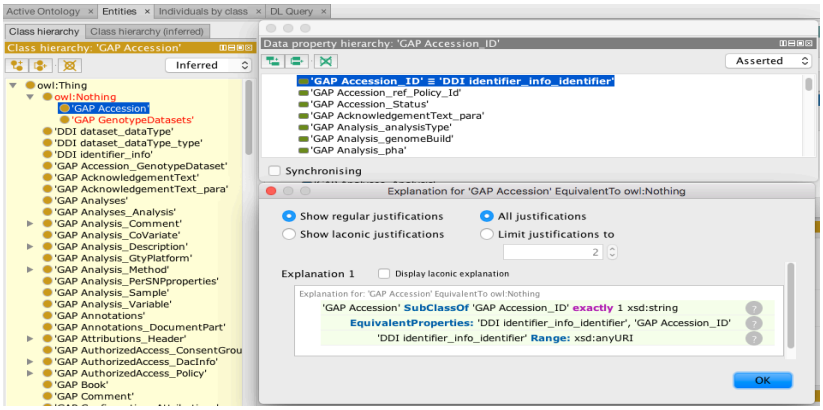


Figure 4: Errors detected by the DL reasoner

.

Conclusion

Our next step will be to propose conceptual definitions (meanings) for both sets of data elements, which should enable richer validation. We anticipate that this platform and approach, once completed, will be equally applicable to both the metamodel and model instance levels and, in combination with other Natural Language Processing (NLP) and table based alignment tools, will provide a framework for both validation and eventual dissemination through the CDISC PhUSE as well as other 11179 based tooling. The artifacts of the project are accessible at https://github.com/crDDI/ontologies.

Acknowledgement: This study is supported in part by the funding from NIH Big Data to Knowledge, Grant 1U24AI117966-01 and NCI U01 CA180940.

References

1. BioCADDIE Project. 2016 [March 10, 2016]; Available from: https://biocaddie.org/.
2. BioCADDIE Pilot Project. 2016 [March 10, 2016]; Available from: https://biocaddie.org/pilot-project-harvester-announcement.
3. HL7 FHIR DSTU 2. 2016 [March 10, 2016]; Available from: https://http://www.hl7.org/fhir/.
4. Tryka KA, Hao L, Sturcke A, Jin Y, Wang ZY, Ziyabari L, et al. NCBI's Database of Genotypes and Phenotypes: dbGaP. Nucleic acids research. 2014;42(Database issue):D975-9. Epub 2013/12/04.
5. The College of American Pathologists (CAP) eCC (electronic Cancer Checklists) 2016 [August 7, 2016]; Available from: http://www.cap.org/capecc.
6. TCGA Data Portal. 2016 [March 10, 2016]; Available from: https://tcga-data.nci.nih.gov/tcga/.
7. ISO/IEC 11179 Metadata Standard. 2016 [March 10, 2016]; Available from: http://metadata-standards.org/11179/.
8. Web Ontology Language (OWL). 2016 [March 10, 2016]; Available from: http://www.w3.org/2001/sw/wiki/OWL.
9. XMDR OWL Schema. 2016 [March 10, 2016]; Available from: https://wiki.nci.nih.gov/pages/viewpageattachments.action?pageId=10854184.
10. bioCADDIE WG3 Metadata Specifications. 2016 [March 10, 2016]; Available from: https://github.com/biocaddie/WG3-MetadataSpecifications.
11. XMDR. 2016 [March 10, 2016]; Available from: https://en.wikipedia.org/wiki/XMDR.