Challenges in Using a Graph Database to Represent and Analyze Mappings of Cancer Study Data Standards

Robinette Renner, MS MHI[1], Guoqian Jiang, MD PhD[2]
[1]University of Minnesota, Minneapolis, MN, USA [2]Mayo Clinic, Rochester, MN, USA

Abstract

While using data standards can facilitate research by making it easier to share data, manually mapping to data standards creates an obstacle to their adoption. Semi-automated mapping strategies can reduce the manual mapping burden. Machine learning approaches, such as artificial neural networks, can predict mappings between clinical data standards but are limited by the need for training data. We developed a graph database that incorporates the Biomedical Research Integrated Domain Group (BRIDG) model, Common Data Elements (CDEs) from the National Cancer Institute's (NCI) cancer Data Standards Registry and Repository, and the NCI Thesaurus. We then used a shortest path algorithm to predict mappings from CDEs to classes in the BRIDG model. The resulting graph database provides a robust semantic framework for analysis and quality assurance testing. Using the graph database to predict CDE to BRIDG class mappings was limited by the subjective nature of mapping and data quality issues.

Introduction

Sharing clinical data can foster innovation[1, 2], reduce the time from bench to bedside[2, 3], improve patient outcomes[2], reduce research costs[2], and increase transparency[2]. Unfortunately, the heterogeneous nature of data stored in local systems creates data silos that make data sharing nearly impossible[4]. While the use of data standards can facilitate the sharing of clinical data[4], sometimes one must support multiple standards. For example, regulatory submissions to the Food and Drug Administration (FDA) must be submitted using Clinical Data Interchange Standards Consortium (CDISC) standards such as the Study Data Tabulation Model (SDTM)[5]. Electronic Medical Record (EMR) systems, on the other hand, are beginning to use HL7's Fast Healthcare Interoperability Resources for electronic data transmission[6]. If researchers want to consume data from an EMR and use it for studies requiring FDA submission, they must support both standards. This landscape is complicated by the need to annotate clinical terms with concepts from multiple clinical vocabularies such as the use of Medical Dictionary for Regulatory Activities (MedDRA) concepts for adverse event report and Logical Observation Identifiers Names and Codes (LOINC) concepts for reporting laboratory values[7]. Most often mapping between standards is done manually which is time-consuming, expensive, and error-prone[8]. The need to support multiple standards has resulted in an urgent need for tools to map between the standards.

The experience of the Center for International Blood and Marrow Transplant Research (CIBMTR) is an excellent case study of the problems with manual mapping. The CIBMTR collects outcomes data for cellular therapy research[9]. To facilitate data collection, the CIBMTR offers electronic data submission directly from a transplantation center's database[9]. Common Data Elements (CDEs) obtained from the National Cancer Institute's (NCI) cancer Data Standards Registry and Repository (caDSR) provide the foundation for data transmission. These CDEs have been mapped to the Biomedical Research Integrated Domain Group (BRIDG) model[10]. The BRIDG model captures the semantics of data used for clinical research and regulatory submission[11]. The BRIDG model has been harmonized and mapped to a variety of data standards and data models. For example, BRIDG has been mapped to CDISC's Clinical Data Acquisition Standards Harmonization (CDASH) and SDTM standards[12-14]. The US Food and Drug Administration (FDA) in collaboration with Health Level 7 (HL7) created the Common Data Model Harmonization (CDMH) project. This project has mapped BRIDG to four clinical data models[15]: Sentinel[16], Patient-Centered Outcomes Research Network (PCORNET) Common Data Model[17], Informatics for Integrating Biology & the Bedside (i2b2)[18], and Observational Medical Outcomes Partnership (OMOP)[19]. The CDMH project also mapped BRIDG to HL7's latest standard, Fast Healthcare Interoperability Resources (FHIR)[20]. The goal of the CDMH project is to provide observational data to researchers, which is a typical clinical use case that demonstrates the value of mappings of clinical study data standards. Unfortunately, while mapping CDEs to the BRIDG model can facilitate the adoption of other clinical data standards, mapping to the BRIDG model itself is a labor-intensive process. A semi-automated mapping strategy could help reduce the mapping burden.

A semi-automated application that maps CDEs to the BRIDG model[21, 22] was created using the Ontology Alignment by Artificial Neural Network (OAANN) developed by Huang et al.[23, 24]. The new algorithm predicts CDE to BRIDG class mappings using key attributes of the ISO 11179 metamodel for CDEs and a robust training set of nearly 1,200 CDEs that have been manually mapped to an appropriate BRIDG class. It returns a list of 10 potential BRIDG classes for a CDE of interest. A subject matter expert then reviews the list and selects the most appropriate BRIDG class. The algorithm was able to predict BRIDG class mappings with up to 94% accuracy. Accuracy was calculated by dividing the number of the predicted mappings that are correct by the number of the manually mappings.

While this approach is effective, it has some limitations. First, the algorithm uses pattern recognition only. The underlying semantics of the CDE and potential synonyms are not considered. Second, if sufficient training data does not exist for a particular BRIDG class, then the algorithm will never map a CDE to it. Graph databases have the potential to overcome these limitations.

Graph databases represent data as a collection of nodes (classes) and edges (relationships). In contrast to relational databases, they provide an efficient way to represent highly interconnected data[25]. Several researchers have used graph databases for a variety of mapping and data integration problems. For example, Alqahtani et al. used a graph database to integrate business data from two heterogeneous data sources[26]. Johnson et al. combined disparate tumor-related models using a Neo4j graph database and annotated the semantic terms with concepts from the NCI Thesaurus[27]. Campbell et al. used a graph database to represent the SNOMED-CT terminology[28]

The work of Ulrich et al. is particularly interesting[29]. They used a graph database to map data elements in disparate metadata repositories and tested the application using more than 600 cancer-related data elements[29]. Their algorithm used the five-gram algorithm and the metric Longest Common Subsequence to determine the similarity of two CDE's name. If the CDE was associated with a list of allowed values, they supplemented this analysis with a comparison of the allowed values. While this work is promising, it has several limitations. First, their analysis is based only on two CDE attributes: long name and allowed values. It is not clear if the CDEs used in the work of Ulrich et al. were based on the ISO 11179 metamodel. Therefore, those two attributes may have been the only ones available for analysis. Second, their analysis is based strictly on pattern recognition. It does not take into consideration the meaning of the terms (i.e. semantics) in the CDE.

Mapping CDEs to the BRIDG model lends itself well to a graph database approach. The BRIDG model has already been implemented as a Neo4j graph database[30] and other researchers have incorporated the NCI Thesaurus into their graph database work[27]. CDEs based on the ISO 11179 metamodel, such as those found in the caDSR, consist of interrelated components. Such relationships could be represented using a graph model. Also, caDSR CDEs are annotated with concepts from the NCI Thesaurus[31]. BRIDG classes are annotated with definitions that are associated with NCI Thesaurus concepts that explicitly document BRIDG as the definition's source. Therefore, the NCI Thesaurus serves as a common point of reference between CDEs and the BRIDG model. A graph database could leverage this relationship to facilitate mapping CDEs to BRIDG classes.

In this paper, we present our work to develop a graph database that incorporates the BRIDG model, caDSR CDEs, and NCI Thesaurus concepts and to use graph-based algorithms to predict potential BRIDG class matches for the CDEs. Finally, we discuss the feasibility of using a graph-based mapping approach.

This work is significant because it is the first comprehensive representation of the BRIDG model, caDSR CDES, and NCI Thesaurus concepts in a graph database that we are aware of. It also highlights the subjective nature of mapping data standards and the challenges in developing a semi-automated mapping strategy.

Methods

Graph Model Development

BRIDG Model: The foundation of the BRIDG portion of the graph model was the Neo4j implementation of the BRIDG model provided by Jane Pollack[30]. She based her graph model on the UML model of the BRIDG model available on the BRIDG website[11]. In her graph model, classes and attributes are represented as nodes. Relationships between classes and attributes or classes and classes are represented as edges. Each node and edge has properties that contain additional information. We then supplemented Pollack's model with edges from the BRIDG class node to the appropriate NCI Thesaurus concept node. We determined the appropriate concept by mapping the BRIDG class name to an NCI Thesaurus concept that had both the BRIDG class name listed as a synonym and a matching definition. Since the focus of this work was on matching CDEs to BRIDG classes, BRIDG attributes were not associated with NCI Thesaurus concepts. Figure 1 shows a portion of the BRIDG UML[12] and the corresponding graph model.

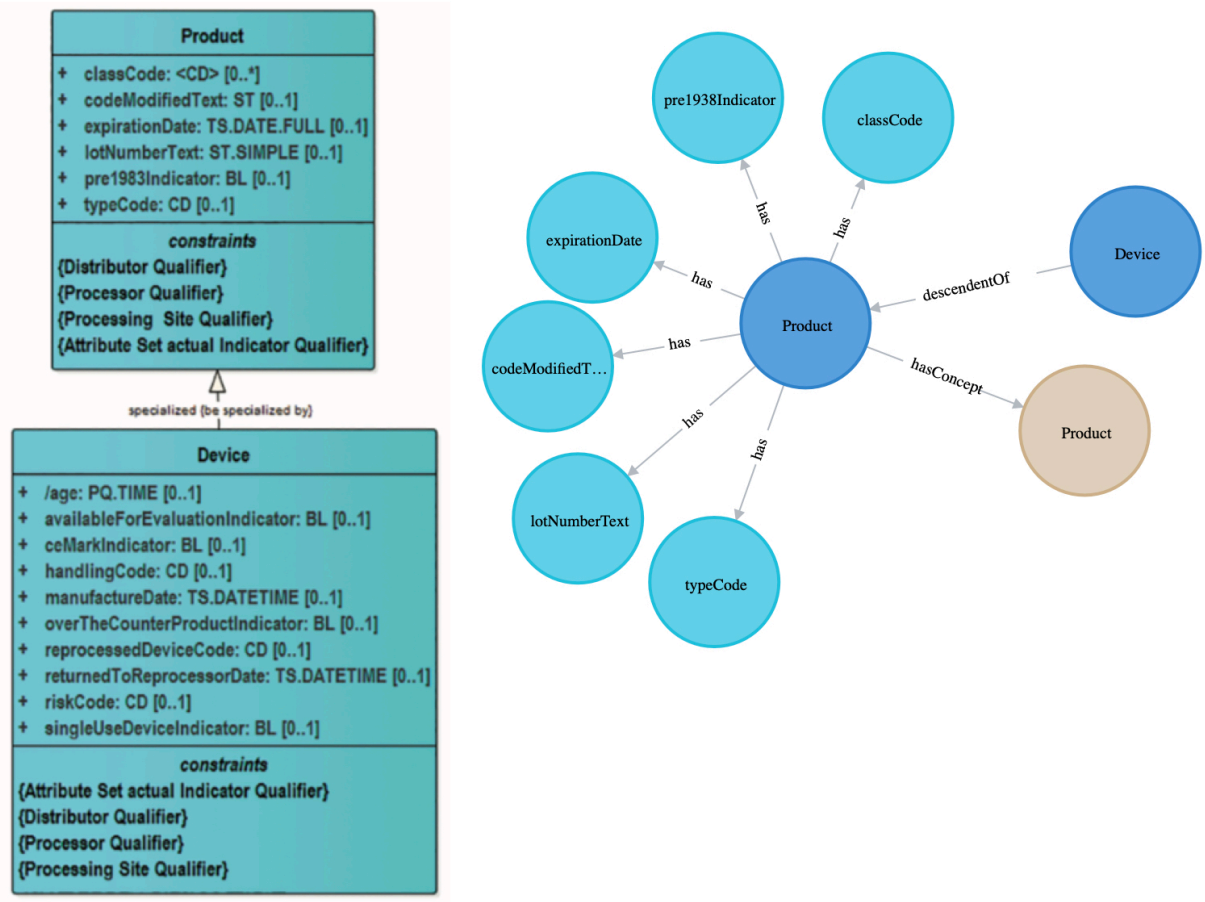

Figure 1. UML representation of a section of the BRIDG model[12] (left) and its corresponding graph model representation (right).

NCI Thesaurus: The concept attributes available within the NCI Thesaurus[32] served as the foundation for the NCI Thesaurus portion of the graph model. In addition to creating a node for Concept, we created nodes for concept attributes such as Semantic Type, Synonym, and Definition. Representing these attributes as nodes allowed for associating multiple instances of an attribute to a concept. Additional properties could also be associated with an attribute. We captured parent-child relationships between concepts as an edge between two concept nodes.

caDSR CDEs: The ISO 11179 metamodel[33, 34] served as the foundation for the caDSR Common Data Element (CDE) portion of the graph model. A CDE consists of two parts: a Data Element Concept (DEC) and a Value Domain. The DEC is the conceptual representation of the CDE and should contain the bulk of the semantic meaning for the CDE[31, 34]. The DEC consists of two parts: the Object Class and the Property. The Object Class is equivalent to a UML class and the Property to a UML attribute[34]. The Object Class and Property are both defined using concepts from the NCI Thesaurus. The Value Domain is associated with a Representation Term that describes the meaning of the data being captured. The Representation Term is also defined using concepts from the NCI Thesaurus[31]. We created nodes for the CDE and its main components, such as the Data Element Concept (DEC) and Value Domain (VD). Each node has properties that contain additional attributes about each component. Edges were created to document the relationships between the various nodes. Also, edges to the associated NCI Thesaurus concepts were added. Figure 2 shows the Data Element Concept portion of the ISO 11179 metamodel[34] and the corresponding section of the graph model.

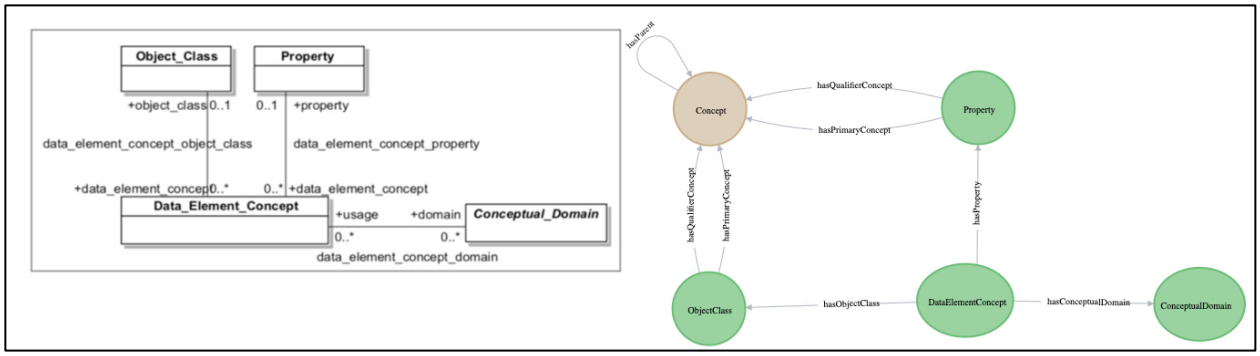

Figure 2. Data Element Concept section of the ISO 11179 metamodel [34] (left) and its corresponding graph model representation (right).

Database Implementation

BRIDG: We populated the BRIDG portion of the Neo4j graph database using the CSV load files and Cypher queries developed by Jane Pollack[30]. To allow for consistent comparison between the graph database mappings and the mappings provided by the Artificial Neural Network algorithm[21 22], we used version 3.2 of the BRIDG model. A CSV load file was created with the relationships between each BRIDG class and the associated NC Thesaurus concepts. A Cypher query was developed to import the CSV load file.

NCI Thesaurus: To populate the NCI Thesaurus section of the Neo4j graph, we first downloaded the Web Ontology Language (OWL) file from the NCI Thesaurus download website[35]. We created a Python script that leveraged the owlready2[36, 37] package to parse the OWL file. The script then generated separate CSV files for each type of NCI Thesaurus node and relationship. For example, separate import files were created for the Concept and Semantic Type nodes. Separate Cypher queries were created that imported each CSV import file.

caDSR CDEs: To populate the CDE portion of the Neo4j graph database, we first downloaded from the CDE Browser website[38] an XML file that contained the information for the 1,689 CDEs used in the ANN mapping project [21 22]. To import the information, we used an approach that was similar to the approach for importing the NCI Thesaurus information. We created a Python script that parsed the XML file and created CSV files for each type of node and edge. We then created a set of Cypher queries that imported each CSV file.

Quality Assurance Testing

BRIDG: To ensure that the Neo4j database was populated correctly, we performed basic quality assurance testing. We tested the BRIDG content by developing a Python script that queried the Neo4j database and created an Excel file with the BRIDG class name and definition. The XlsxWriter Python package[39] was used to create the Excel file, and the Neo4j Bolt Driver Python package[40] was used to query the database. We supplemented the Excel file with the BRIDG class names and definitions obtained from the BRIDG website[11]. The information from both sources was then compared to ensure consistency.

caDSR CDEs: We tested the quality of the CDE information the Neo4j graph database by creating a Python script that queried the Neo4j database [40], the XML source document downloaded from the CDE Browser [38], and generated an Excel file [39] containing the CDE information from both sources. The Excel file was then analyzed to ensure that the CDE information in the Neo4j database was accurate.

NCI Thesaurus: Testing of portions of the NCI Thesaurus content in the Neo4j occurred during the testing of both the BRIDG and CDE content. Both of those testing approaches contained NCI Thesaurus concept attributes such as concept preferred name, concept unique identifier, and definition. Robust quality assurance testing of all concept attributes such as synonyms and semantic types is in progress.

Algorithm Implementation

We leveraged the Shortest Path algorithm in the Neo4j library[41] to determine the shortest path between a CDE and a BRIDG class. We developed a Python script that iteratively ran a Cypher query that found the unweighted shortest path between a CDE and each BRIDG class and returned the distance. Distance was measured as the number of edges between the CDE and the BRIDG class. We tested the algorithm by running it against all of the CDEs used in the

ANN mapping project[21, 22]. Since those CDEs have been manually mapped to the BRIDG model by a team of subject matter experts, they represent a gold standard by which we could determine the accuracy of the shortest path algorithm.

The Python script returned an Excel file that contained the CDE public ID, CDE Long Name, the target BRIDG class name, the distance between the CDE and the target BRIDG class, and a text description of the path in terms of the nodes used. The manually mapped BRDIG class was added to the Excel spreadsheet to make it easier to determine the accuracy of the algorithm's prediction. We also developed a Cypher query that visualized specific mapping paths.

Results

Graph Model

Figure 3 shows the graph model of the BRIDG model, caDSR CDEs, and NCI Thesaurus concepts. This model shows how the NCI Thesaurus represents a nexus point connecting BRIDG classes to caDSR CDEs and providing them with. rich semantic annotations. Table 1 summarizes the number of key nodes in the graph database. In addition, the BRIDG classes are associated with 205 distinct NCI Thesaurus concepts, and the CDEs are associated with 1,143.

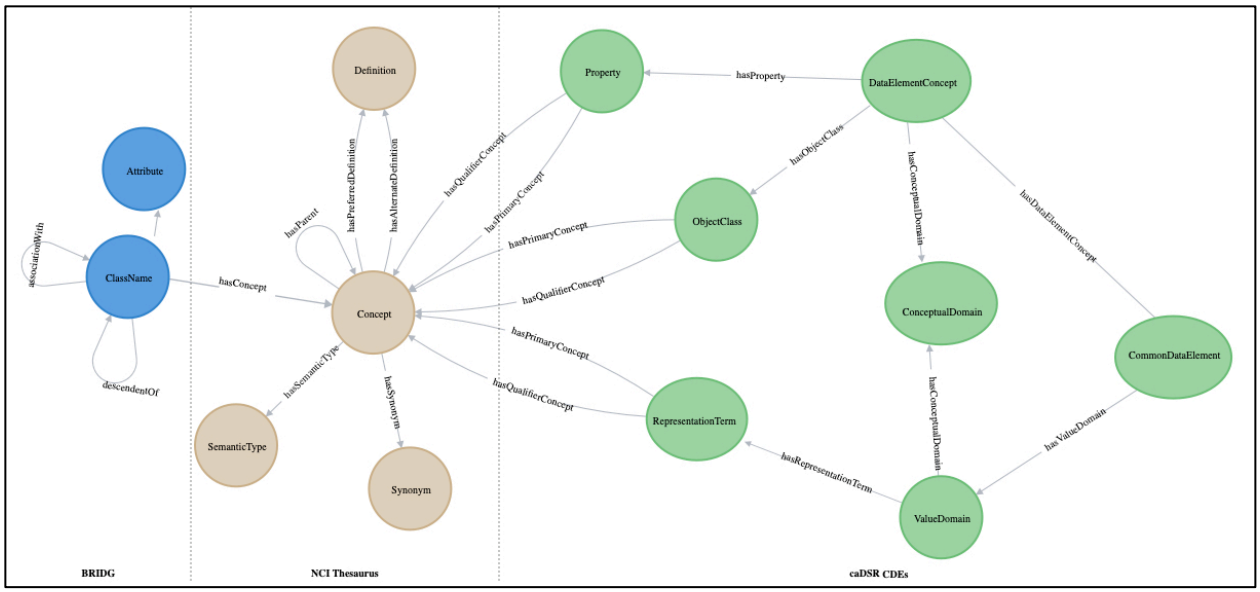

Figure 3. Representation of the graph-based data model.

Table 1. Summary of the number of key nodes in the database.

Section	Node	Count
BRIDG	Class	233
	Attribute	639
caDSR	Common Data Element	1689
	Data Element Concept	1193
	Object Class	622
	Property	638
	Value Domain	588
	Representation Term	524
NCI Thesaurus	Concept	147,010
	Semantic Type	128
	Synonym	426,150

Quality Assurance Testing

Quality assurance testing of the BRIDG section of the model verified that BRIDG classes were associated with the correct NCI Thesaurus concept where possible. 12% (28 out of 233) of the BRIDG classes are not associated with an NCI Thesaurus concept. This gap impacted 5 of the 1,689 CDEs that we analyzed.

Quality assurance testing of the caDSR CDEs detected issues when the CDEs were constructed using concepts from the NCI Metathesaurus instead of the NCI Thesaurus. This occurred in 4 CDEs out of the 1,689. caDSR best practice states that CDEs should be constructed using concepts from NCI Thesaurus [34]. Therefore, these CDEs were incorrectly constructed. The graph database implementation can facilitate the detection of such data quality errors.

Shortest Path Analysis

We hypothesized that the path from a CDE to a BRIDG class that had the shortest distance should match the BRIDG class to which the CDE was manually mapped by subject matter experts. We calculated the match rate as the percent of CDEs for which the shortest path led to the manually mapped BRIDG class. The match rate produced by the shortest path algorithm was 16.6% (280 out of 1,689 CDEs). This is much lower than the match rate produced by the ANN algorithm. The ANN algorithm produced a match rate of between 34 - 94%[22]. For the ANN algorithm, the lowest match rate of 34% corresponded to a set of CDEs that were semantically different from the training set and contained many CDEs that were manually mapped to BRIDG classes for which there was insufficient training data. To determine why the algorithm was returning such a low match rate, we looked at those CDEs that had a path distance of 4 and a path that went through the Object Class's primary concept.

Analysis of Paths with a Distance of Four

A path with a distance of 4 indicates that the CDE and the BRIDG class directly share an NCI Thesaurus concept. Since the Object Class is equivalent to a UML class[34], a path from a CDE to a BRIDG class that goes through the Object Class's primary concept and has a distance of 4 should represent a correct match. Figure 4 shows such a path.

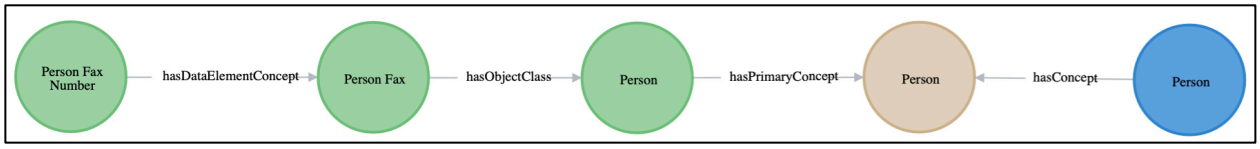

Figure 4. Example path with a distance of 4.

64 CDEs had a path to a BRIDG class that had a distance of four and went through the Object Class primary concept. However, the match rate for these CDEs was only 33%. A closer analysis of the incorrectly matched CDEs shows that the mappings were more complex than the shortest path results initially indicate. For the majority of the CDEs analyzed, they were manually mapped to a more specific BRIDG class. For example, the shortest path algorithm associated with the CDE Product Collection Date with the BRIDG class Product. However, the CDE was manually mapped to the BRIDG class PerformedSubstanceExtraction. The manually mapped BRIDG class is correct. A subject matter expert is able to evaluate the meaning of the entire CDE, while the shortest path algorithm focuses on one concept. Table 2 presents some example CDEs along with the manually mapped BRIDG class, predicted BRIDG class, and a comment explaining the discrepancy. Even this analysis is subjective. Another subject matter expert may have a different interpretation of the mappings.

Table 2. Example of discrepancies between manually mapped and predicted BRIDG classes.

CDE Long Name	Manually Mapped BRIDG Class	Predicted BRIDG Class	Comment
Person Fax Number	StudySitePersonnel	Person	Manual mapping is technically incorrect because it introduces semantics not contained within the CDE.
Cellular Therapy Product Unique Identifier	Subject	Product	Manual mapping is technically incorrect because it introduces semantics not contained within the CDE.
Product Batch Number Unique Identifier	ProcessedProduct	Product	Manual mapping is potentially correct; ProcessedProduct has an association with Product.
Person Sex Type	BiologicEntity	Person	Manual mapping is correct; BRIDG is designed to accommodate non-human subjects and CDEs are generally created from a human perspective.
Product Tissue Donor Relationship Type	SubstanceExtraction AdministrationRelationship	Product	Manual mapping is correct; the CDE was mapped to a more specific BRIDG class that better captured the semantics.

Discussion

Reasons for the Low Match Rate

The low match rate returned by the shortest path algorithm was unexpected but brought to light the subjective nature of mapping and data quality issues. Defining metadata and mapping data standards is a subjective process. Each individual involved in the process has use cases and perspectives that influence their work. In many ways, the BRIDG model, the caDSR CDEs, and the NCI Thesaurus are like the parable of the three blind sages trying to describe an elephant. Each is describing the same thing but doing it from a slightly different perspective. For example, the CDEs in the caDSR are primarily created to capture patient-related data such as outcomes data. As a result, they are patient-focused. The BRIDG model was developed to represent both pre-clinical and clinical data[11]. As such, it needs to capture information about non-human research subjects. The CDE Person Birth Date illustrates this situation well. The CDE uses the concept for Person, which is also used by the BRIDG class Person. This resulted in a shortest path with a distance of four. However, the BRIDG model reserves the Person class for human-specific attributes such as occupation. Attributes such as date of birth and gender are in the class BiologicEntity. The path between the CDE Person Date of Birth and the BRIDG class of BiologicEntity has a distance of five. The CDE, the manual mapping to the BRIDG class BiologicEntity, and the associated concepts are fundamentally correct. However, the shortest path algorithm is purely objective. It looks at the underlying data and matches the CDE to the BRIDG class of Person.

A similar difference of perspective is seen with how concepts are defined in the NCI Thesaurus and used by BRIDG and the CDEs. This is exemplified by CDE Surgical Procedure Performed Date, which was manually mapped to the BRIDG class "PerformedProcedure." The mapping distance between the CDE and BRIDG class is seven. The reason for the long mapping distance lies with how the underlying concepts are defined. The CDE uses the concept "Surgical Procedure," which has a semantic type of "Health Care Activity" and a root concept of "Activity." The BRIDG class uses the concept "PerformedProcedure," which has a semantic type of "Research Activity" and a root concept of "Conceptual Entity." The representations of the two concepts in the NCI Thesaurus are different but technically correct. A surgical procedure can be both a health care activity and a research activity. Unfortunately, the different representations have created semantic silos, which the shortest path algorithm struggles to overcome.

The results of the shortest path analysis also revealed potential data quality issues with both the choice of BRIDG classes selected during the manual mapping process and with how the CDEs were defined. For example, the algorithm found a path with a distance of four between the CDE Person Fax Number and the BRIDG class Person. During the manual mapping process, the CDE was mapped to the BRIDG class StudySitePersonnel. The mapping is technically

incorrect because it introduces information not contained in the semantic definition of the CDE. Most likely, the individual who performed the manual mapping based the mapping on the context in which the CDE was used and not strictly on the semantics of the CDE. Manually mapping the CDE to StudySitePersonnel limits the reusability of the mapping to other contexts.

The shortest path algorithm also highlighted instances where the CDE was constructed using incorrect concepts. For example, the CDE Disease Involvement Site Other Specify has a path length of seven to the manually mapped BRIDG class of TargetAnatomicSite. Analysis of the CDE revealed that it had incorrectly used the generic NCI Thesaurus concept for Location instead of the correct concept for Anatomic Site. This error made it difficult for the shortest path algorithm to find the path to the correct BRIDG path.

Limitations

There are several limitations to our work. First, we used an unweighted implementation of the shortest path algorithm. According to caDSR best practice, the bulk of a CDE's semantic meaning should be contained in the Data Element Concept[31, 34]. Adjusting the algorithm's parameters so that paths going through the Value Domain have a higher weight, may have produced a better match rate. Second, the algorithm only considers one concept. The complete meaning of a CDE cannot be understood by looking at one concept in isolation. This limits the ability of the algorithm to predict an appropriate BRIDG class. Finally, 12% of the BRIDG classes were not associated with an NCI Thesaurus concept. While this had little impact on this project, to extend this work to other use cases, this gap should be closed.

Next Steps

We plan to enhance to shortest path algorithm implementation to include weighting and to better handle relationships between BRIDG classes. The weighting will prioritize those paths that go through the Data Element Concept which should contain the bulk of the semantic meaning for the CDE[31, 34]. Also, sometimes the manually mapped class was either a descendent of or associated with the predicted BRIDG class. Adjusting the algorithm to include such classes along with the BRIDG class associated with the shortest path may improve the effectiveness of using a graph database to semi-automate the mapping process.

Next, we plan to explore combining the Artificial Neural Network (ANN) algorithm for facilitating the mapping of CDEs to the BRIDG model with the Neo4j graph database implementation of the BRIDG model, caDSR CDEs, and the NCI Thesaurus. Since the ANN algorithm learns from previous CDE to BRIDG mappings, it can handle the subjective nature of the mapping process. The performance of the algorithm diminishes when it has insufficient training data. The performance of the algorithm may improve if it can leverage the underlying semantics of both the CDEs and the BRIDG classes. In particular, performance may improve if the ANN algorithm can incorporate synonyms. In contrast, the Neo4j graph database and shortest path algorithm employ an objective, logical approach to mapping. It looks strictly at the semantic relationships between BRIDG classes, caDSR CDEs, and NCI Thesaurus concepts. It does not consider the context, nor does it learn from previous mappings. Combining the ANN algorithm with the graph database may result in better mapping predictions, especially when there is insufficient training data.

The incorporation of the BRIDG model, caDSR CDEs, and NCI Thesaurus concepts into one graph database creates a framework to perform robust quality assurance testing. Our analysis of the shortest path algorithm results revealed instances of incorrect mappings and poor CDE construction. We plan to develop a collection of Cypher queries to determine potential quality issues.

Another area to explore is expanding the NCI Thesaurus so that it can accommodate multiple perspectives. For example, annotating a concept such as PerformedProcedure as being both a Health Care Activity and a Research Activity will increase its interoperability when used in different contexts and use cases.

Conclusion

The process of manually defining and mapping clinical data standards combines logical, objective reasoning, along with subjective characteristics informed by the individual's experience and the particular use case. As a result, a more objective mapping approach, such as the graph-based shortest path algorithm, will not be able to replicate the manual mapping results perfectly. An approach, such as an artificial neural network-based algorithm that learns from previous manual mappings, is better able to replicate the manually mapping results but is limited by lack of training data and semantic annotations. Combining the rich semantics contained in a graph database along with the learning capabilities of an artificial neural network may provide for a more robust mapping strategy

The graph database that incorporates the BRIDG model, caDSR CDEs, and the NCI Thesaurus provides a valuable source of semantic annotations that can be leveraged for a variety of purpose. In particular, the graph database has the potential to facilitate quality assurance testing.

References

1. Tang C, Plasek JM, Bates DW. Rethinking Data Sharing at the Dawn of a Health Data Economy: A Viewpoint. *Journal of medical Internet research*. 2018;20(11):e11519.
2. Piwowar HA, Becich MJ, Bilofsky H, Crowley RS. Towards a data sharing culture: recommendations for leadership from academic health centers. *PLoS medicine*. 2008;5(9):e183.
3. Federer LM, Lu YL, Joubert DJ, Welsh J, Brandys B. Biomedical data sharing and reuse: Attitudes and practices of clinical and scientific research staff. *PloS one*. 2015;10(6):e0129506.
4. Kush R, Goldman M. Fostering responsible data sharing through standards. *New England Journal of Medicine*. 2014;370(23):2163-5.
5. Study Data Standards: What you need to know [Internet]. US Food and Drug Administration; 2017 [cited 2019 May 28]. Available from: https://www.fda.gov/downloads/Drugs/DevelopmentApprovalProcess/FormsSubmissionRequirements/ElectronicSubmissions/UCM511237.pdf.
6. Argonaut Project [Internet]. Health Level Seven International; 2019 [cited 2019 Aug 14]. Available from: http://argonautwiki.hl7.org.
7. Introduction to the Interoperability Standards Advisory [Internet]. Office of the National Coordinator for Health IT; 2019 [cited 2019 Aug 14]. Available from: https://www.healthit.gov/isa/.
8. Richesson RL, Fung KW, Krischer JP. Heterogeneous but "standard" coding systems for adverse events: Issues in achieving interoperability between apples and oranges. *Contemporary Clinical Trials*. 2008;29(5):635-45.
9. CIBMTR Progress Report 2017 [Internet]. The Medical College of Wisconsin, Inc. and the National Marrow Donor Program; 2017 [cited 2019 Aug 14]. Available from: http://www.cibmtr.org/About/AdminReports/Pages/index.aspx.
10. Renner R, Carlis J, Maiers M, Rizzo JD, O'Neill C, Horowitz M, et al. Integration of Hematopoietic Cell Transplantation Outcomes Data. 2015;9162:139-46.
11. Biomedical Research Integrated Domain Group [Internet]. [cited 2019 Aug 14]. Available from: https://bridgmodel.nci.nih.gov.
12. Becnel LB, Hastak S, Ver Hoef W, Milius RP, Slack M, Wold D, et al. BRIDG: a domain information model for translational and clinical protocol-driven research. *Journal of the American Medical Informatics Association*. 2017;24(5):882-90.
13. Kibbe W. Cancer Clinical Research: Enhancing Data Liquidity and Data Altruism. Oncology Informatics: Elsevier; 2016. p. 41-53.
14. Clinical Data Interchange Standards Consortium [Internet]. 2019 [cited 2019 Aug 14]. Available from: https://www.cdisc.org.
15. Common Data Models Harmonization FHIR Implementation Guide [Internet]. HL7 International - Biomedical Research and Regulation Work Group; 2019 [cited 2019 Aug 14]. Available from: https://build.fhir.org/ig/HL7/cdmh/.
16. FDA's Sentinel Initiative [Internet]. U.S. Food & Drug Administration; 2018 [cited 2019 Aug 14]. Available from: https://www.fda.gov/safety/fdas-sentinel-initiative.
17. The National Patient-Centered Clinical Research Network (PCORnet) [Internet]. [cited 2019 Aug 14]. Available from: https://pcornet.org.
18. Informatics for Integrating Biology and the Bedside (i2b2) [Internet]. Partners Healthcare; 2019 [cited 2019 Aug 14]. Available from: https://www.i2b2.org.
19. Observational Health Data Sciences and Informatics (OHDSI) [Internet]. 2019 [cited 2019 Aug 14]. Available from: https://www.ohdsi.org.
20. HL7 FHIR [Internet]. Health Level Seven International; 2018 [cited 2018 Mar 6]. Available from: http://hl7.org/fhir.
21. Renner R, Li S, Huang Y, Tan S, Li D, v. d. Zijp-Tan A, et al. Mapping Common Data Elements to a Domain Model Using an Artificial Neural Network. 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); 2018 3-6 Dec. 2018; Madrid.

22. Renner RL, Shengyu; Huang,Yulong; van der Zijp-Tan, Ada Chaeli ; Tan, Shaobo; Li, Dongqi ; Kasukurthi, Mohan Vamsi; Benton, Ryan; Borchert, Glen; Huang, Jingshan; Jiang, Guoqian Using an Artificial Neural Network to Map Cancer Common Data Elements to the Biomedical Research Integrated Domain Group Model in a Semi-automated Manner. *BMC medical informatics and decision making*. Forthcoming 2019.
23. Huang J, Dang J, Huhns MN, Zheng WJ. Use artificial neural network to align biological ontologies. *BMC genomics*. 2008;9(2):S16.
24. Huang J, Dang J, Vidal JM, Huhns MN, editors. Ontology matching using an artificial neural network to learn weights. IJCAI workshop on semantic Web for collaborative knowledge acquisition; 2007.
25. Robinson I, Webber J, Eifrem E. Graph databases. Sebastopol (CA): O'Reilly Media, Inc.; 2013.
26. Alqahtani A, Heckel R. Model based development of data integration in graph databases using triple graph grammars. In: Mazzara M, Salaun G, Ober I, editors.: Springer Verlag; 2018. p. 399-414.
27. Johnson D, Connor AJ, McKeever S, Wang Z, Deisboeck TS, Quaiser T, et al. Semantically linking in silico cancer models. *Cancer informatics*. 2014;13(Suppl 1):133-43.
28. Campbell WS, Pedersen J, McClay JC, Rao P, Bastola D, Campbell JR. An alternative database approach for management of SNOMED CT and improved patient data queries. *Journal of biomedical informatics*. 2015;57:350-7.
29. Ulrich H, Kock-Schoppenhauer AK, Duhm-Harbeck P, Ingenerf J. Using Graph Tools on Metadata Repositories. *Stud Health Technol Inform*. 2018;253:55-9.
30. Pollack J. BRIDGModel2Graph [Internet]. GitHub; 2019 [cited 2019 Aug 14]. Available from: https://github.com/nmdp-bioinformatics/BRIDGModel2Graph.
31. caDSR Training Material - Course 1040 Creating Well-formed Metadata and Metadata Business Rules [Internet]. National Cancer Institute; [cited 2019 Aug 14]. Available from: https://wiki.nci.nih.gov/display/COREtraining/1040+Creating+Well-formed+Metadata+and+Metadata+Business+Rules.
32. NCI Thesaurus [Internet]. National Cancer Institute; [cited 2019 Aug 14]. Available from: https://ncit.nci.nih.gov/ncitbrowser/.
33. ISO 11179 Specification part 1 version 3. Switzerland: ISO/IEC 2015.
34. ISO 11179 Specification part 3 version 3. Switzerlend: ISO/IEC; 2013.
35. NCI Thesaurus Download [Internet]. National Cancer Institute Enterprise Vocabulary Services; [cited 2019 Aug 14]. Available from: https://evs.nci.nih.gov/evs-download/thesaurus-downloads.
36. Lamy J-B. Owlready2 Documentation [Internet]. [cited 2019 Aug 05]. Available from: https://pythonhosted.org/Owlready2/.
37. Lamy J-B. Owlready: Ontology-oriented programming in Python with automatic classification and high level constructs for biomedical ontologies. *Artificial intelligence in medicine*. 2017;80:11-28.
38. CDE Browser [Internet]. National Cancer Institute; [cited 2018 Aug 31]. Available from: https://cdebrowser.nci.nih.gov/CDEBrowser/.
39. McNamara J. XlsxWriter [Internet]. [cited 2019 Aug 14] Available from: https://xlsxwriter.readthedocs.io/index.html.
40. Neo4j Bolt Driver 1.7 [Internet]. Neo Technology; 2018 [cited 2019 Aug 09]. Available from: https://neo4j.com/docs/api/python-driver/current/.
41. The Neo4j Graph Algorithms User Guide v3.5 [Internet]. [cited 2019 Aug 14]. Available from: https://neo4j.com/docs/pdf/neo4j-graph-algorithms-3.5.pdf.