

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

Deepak K. Sharma, MS¹, Harold R. Solbrig, MS¹, Eric Prud'hommeaux², Kate Lee, MPH³, Jyotishman Pathak, PhD³, Guoqian Jiang, MD, PhD¹

¹ Mayo Clinic College of Medicine, Rochester, MN; ² W3C/MIT, Boston, MA; ³ Weill Cornell Medicine, New York, NY

Abstract

In this paper, we present a platform known as D2Refine for facilitating clinical research study data element harmonization and standardization. D2Refine is developed on top of OpenRefine (formerly Google Refine) and leverages simple interface and extensible architecture of OpenRefine. D2Refine empowers the tabular representation of clinical research study data element definitions by allowing it to be easily organized and standardized using reconciliation services. D2Refine builds on valuable built-in data transformation features of OpenRefine to bring source data sets to a finer state quickly. We implemented the reconciliation services and search capabilities based on the standard Common Terminology Services 2 (CTS2) and the serialization of clinical research study data element definitions into standard representation using clinical information modeling technology for semantic interoperability. We demonstrate that D2Refine is a useful and promising platform that would help address the emergent needs for clinical research study data element harmonization and standardization.

1 Introduction

Present frameworks and tools (such as OpenEHR ADL Workbench¹, Eclipse Model Driven Health Tools², XML Schema Editors, UML Editors) to create and manage clinical models for semantic interoperability are complex and have steep learning curve. The tedious process of organizing and transforming the user-defined model definitions to a conformant format for these tools is another hindrance. Many integration solutions to achieve standardization and harmonization are proprietary and hence produced knowledge content (i.e., clinical models) is inherently dependent on the environment they are created in.

In this study, we present a platform known as D2Refine Workbench (D2Refine for short)³ aimed at addressing these issues. D2Refine provides a platform to help for clinical study data element standardization and harmonization in much simpler way with significantly reduced learning curve. D2Refine is built on top of an open-source solution called OpenRefine (formerly known as Google Refine)⁴, which offers a simpler, spreadsheet like interface for the loaded content. The existing capabilities of OpenRefine to bi-directionally serialize the contents in some of the easiest, popular and non-preparatory formats like Microsoft Excel, Comma/Tab Separated Values, XML and JSON, and establish the basis for easier handling of content that is more likely available in these formats already. D2Refine extends OpenRefine's import and export capabilities to transparently examine, validate and transform the content with almost no additional steps for the user. Large projects like the Database of Genotypes and Phenotypes (dbGaP)^{5, 6}, the Phenotype Knowledge Base (PheKB)^{7, 8} and The Cancer Genome Atlas (TCGA)⁹ have published their data dictionaries in spreadsheet-like interface using either delimited values variants (CSV - comma-separated values) or Extended Markup Language (XML) schemas. The D2Refine platform can import models in a variety of formats supported by OpenRefine for data element harmonization and standardization.

D2Refine can also be used to load clinical study data models, data dictionaries and present them in the tabular format, which is the simplest format for humans to review as well as machines to process. D2Refine extends the feature of the OpenRefine reconciliation services by implementing the Common Terminology Services 2 (CTS2) services^{10, 11}. This makes it much easier to manage models for their 'term bindings' – for standardization and their harmonization. D2Refine extends this feature to enable harmonizing local data elements with the common data elements (CDEs) as recorded in standard metadata repositories like NCI caDSR¹². The data models in D2Refine can also be exported to standard representations like OpenEHR's Archetype Definition Language (ADL)¹³, OMG Archetype Modeling Language (AML)¹⁴ or W3C Shape Expressions (ShEx)¹⁵. D2Refine can serialize data models into HL7 FHIR Profiles.

The objective of this paper is to describe our efforts in developing the D2Refine platform and demonstrate how the platform addresses the emergent needs to have a better platform for clinical data model creation, serialization, data element standardization and harmonization.

2 Materials and Methods

2.1 Materials

2.1.1 OpenRefine - The Base Platform

OpenRefine⁴ (formerly Google Refine), is a popular and widely used tool to clean up and manage data. We use OpenRefine as a base platform to develop D2Refine. OpenRefine has an extensible framework (developed by extending MIT's Butterfly Framework) that allows users to programmatically extend it. The built-in reconciliation feature of OpenRefine allows users to leverage online services to validate, transform or bind data values to their intended meaning. The export/import extensions provide a way to serialize content in OpenRefine to a desirable standard or customized format. D2Refine attempts to reduce the learning curve for its users to implement suitable import, export and reconciliation services of OpenRefine. There are many open-source and free extension libraries available for OpenRefine platform to enhance the capabilities of OpenRefine and hence D2Refine.

2.1.2 The Source Content

OpenRefine supports a variety of formats like CSV, XML, JSON and can even load content from databases easily. In addition to regular user defined datasets, we loaded the data dictionaries from dbGaP, PheKB and TCGA in their released formats (XML and MS Excel Spreadsheets) for their standardization and export using D2Refine. Each of the data dictionaries contains a collection of metadata including variable definition, constraints and links to standard terminologies. Some of these data dictionaries include references to standard terminologies like NCI Thesaurus¹⁶ and Common Data Elements from Cancer Data Standard Registry and Repository (caDSR)¹² already embedded in most of the publicly shared models they define.

2.1.3 The Standards

The process of transformation of data dictionaries into standard formats in D2Refine requires implementation of model standards. The transformation employs 'constraint-based modeling' approach to transform data dictionary definition into clinical archetypes. An *archetype* is a set of constraints on a target class in a target reference model. The OpenEHR's Archetype Definition Language (ADL)¹³, the Object Management Group (OMG) Archetype Modeling Language (AML)¹⁴, W3C RDF Shape Expressions (ShEx)¹⁵ and HL7 FHIR Profiles¹⁷ are some of the standards D2Refine implements (or embeds their available implementations) to achieve that.

OpenEHR ADL is a standard for writing clinical models. OpenEHR's ADL Workbench (ADW)¹ is a platform that allows users to create and manage archetypes and archetype templates in ADL and using ADL Object Model (AOM). The in-memory incarnation of the model in AOM is serialized into ADL using Object Data Instance Notation (ODIN) format¹⁸. The ADL format is proprietary and requires ADW to effectively view and utilize it for model-driven template generation for clinical applications.

OMG's AML standard is composed of three UML Profiles - Reference Model Profile (RMP), Terminology Profile (TP) and Constraint Profile (CP) for addressing *constraint-based modeling* aspects dealing with the reference model, terminology binding and constraint specifications respectively. An AML archetype is represented in OMG Unified Modeling Language (UML). Just by having archetypes in non-proprietary UML format, it opens the door to utilize existing UML tools and editors to work with these models – reducing learning curve for the modelers. The UML format enables models for Model-Driven Development environment, while other proprietary formats require additional transformations.

The World Wide Web Consortium (W3C) Resource Description Framework (RDF Shape Expressions (ShEx)¹⁵ language is developed to describe RDF graphs. The Shape Expressions to RDF is similar to XML Schema to XML instance data. Once models are represented in the Shape Expressions, they could be transformed to various formats like RDF, XML, JSON, CSV or any custom format using dynamic customization of *semantic action* feature of Shape Expressions. The simple constraint grammar of Shape Expressions depicts constraint definitions that can be easily read and understood by humans and processed by machine. There are already some implementations of Shape Expressions data validation tools available.

The emerging popular clinical data standard - Health Level 7 (HL7) Fast Healthcare Interoperability Resources (FHIR)¹⁷ is for digitally exchanging healthcare information while preserving clinical semantics of the information

being exchanged. HL7 FHIR has defined a collection of core resources that can be put together to describe data models in clinical contexts. HL7 FHIR is relatively newer effort and to leverage benefits of FHIR resources requires additional step of mapping and transforming existing data and models into FHIR. D2Refine is intended to enable a mechanism to map data dictionary elements to FHIR resources, and map data dictionary constraints into FHIR Profiles.

2.1.4 The Reference Models

The Reference Model (RM)¹⁹ is an important component of *constraint-based modeling* approach for creating *archetypes*. The definition of an archetype is based on RM class upon which it defines constraints. Although with future D2Refine development efforts, we plan to include user-defined reference models, currently the system works with only two reference models – the OpenEHR Reference Model and the CIMI Core Reference Model. OpenEHR RM model is an evolving model that includes basic classes, data types and data-value type classes that form the basis for modeling most clinical scenario. CIMI Core Reference Model is inspired from OpenEHR RM and is bit more general to specify virtually any model clinical or non-clinical. CIMI also publishes a set of *reference archetypes*, the CIMI Core Reference Archetypes. These archetypes constrain the CIMI Core Reference Model and form the basis of all other archetypes published by the CIMI group.

2.1.5 The Standard Template

All data dictionaries have their own schema to define data-set definitions. The alignment of schema elements with D2Refine requirements for validation and transformation routine is an additional step. We have developed a *Standard Template* which is based on OMG AML specifications¹⁴. The *Standard Template* eliminates repeat custom validation and transformation implementation for each type of data dictionary. All we have to do is to map the incoming data dictionary schema to the Standard Template, which can be done relatively easier (and only once) using existing OpenRefine data management features. Once the incoming schema is *aligned* with the Standard Template, the subsequent validation and transformation routines remain the same.

2.1.6 The Libraries

D2Refine requires (and comes embedded with) the implementation libraries for 1) reference model 2) AML Profiles 3) UML Serialization 4) ADL Serialization. OpenEHR provides the OpenEHR reference model implementation libraries. We have implemented CIMI RM libraries by using Eclipse Modeling Framework (EMF)²⁰ from its original release format in UML. The Standard Template is part of the core implementation and is being developed based on AML Profile implementation received from Model-Driven Healthcare Technologies (MDHT)². The AML serialization of data dictionaries is achieved with EMF UML2 libraries and similarly the ADL serialization is performed using OpenEHR's ADL Core open-source libraries.

2.2 Methods

The steps listed below, not necessarily executed in the sequence they are listed here, describe the strategy adopted to develop the D2Refine Workbench:

2.2.1 Design and Implement Standard Template

The Standard Template defines minimum set of requirements to be filled in by data dictionary definition. The Standard Template helps capture the archetype requirements and serves as a map between tabular metadata and their AML equivalents. The Standard template would eliminate the need for any additional custom transforms for an incoming data dictionary. OpenRefine allows recording the steps to create such a map, which could be reused and applied to any data dictionary that belongs to that schema. For example the mapping of a dbGAP data dictionary to the Standard Template is done once and it could be applied automatically to any subsequent dbGAP dictionary in D2Refine. Figure 1 shows the dbGAP data dictionary template elements and their intended meaning.

| variable-id | variable-value | variable-name | variable-type | variable-description | variable-logical | variable-logical | variable-unit |
|-------------------|----------------|---------------|---------------|--------------------------|------------------|------------------|---------------|
| 1. phv00225190.v1 | Male | SEX | integer | Sex of participant | 1 | 2 | |
| 2. phv00225191.v1 | Female | sex | encoded value | Sex of participant | 1 | 2 | |
| 3. phv00225192.v1 | Missing | agepcat | encoded value | Age in 5 year categories | 1 | 14 | Years |
| | 15-29 | | | | | | |
| | 30-34 | | | | | | |
| | 35-39 | | | | | | |
| | 40-44 | | | | | | |
| | 45-49 | | | | | | |
| | 50-54 | | | | | | |
| | 55-59 | | | | | | |
| | 60-64 | | | | | | |
| | 65-69 | | | | | | |

Figure 1: The dbGAP Data Dictionary Template

The elements in the Standard Template are derived from the AML specification profiles and are manually composed, as a set of three types of requirements. These requirements are listed in Table 1.

Table 1: The Standard Template requirements

| | |
|-----------------------|--|
| Archetype Metadata | Archetype library, identification, constrained RM class, archetype specialization, organization, copyright, version and author information |
| Constraint Definition | Constraint identification, constrained type, values, multiplicity, value set, archetype reference |
| Terminology Bindings | Archetype term definitions, references to standardized terminology resources like code systems, concepts, value set and value-set members |

Once the Standard Template requirements are identified, we implement the backend extension to OpenRefine to be able to execute it when the data dictionary is loaded into D2Refine Workbench. This will be transparent to the user because the map for the alignment of data dictionary schema to the Standard Template would have been applied programmatically. The implementation of the map would notify of any missing requirement of the Standard Template.

2.2.2 Create Reconciliation Service Extensions

OpenRefine allows implementation of custom reconciliation services as extensions. D2Refine leverages this feature to mainly standardize the source contents. The reconciliation services could be extended to provide validation and identify gaps in data dictionary design. There would be mainly two types of reconciliation services - Terminology Reconciliation and Data Element Reconciliation. Terminology Reconciliation provides terminology binding for the cell values and links them to their intended meaning. The Standard Template's Terminology Bindings maps (which derives from AML's Terminology Binding Profile) will register the references to the target terminology.

The OMG CTS2 specification provides model for Code System, Concept, Value Set and Value Set Definition references and hence we intend to implement reconciliation services to *controlled terminologies* implemented and available through CTS2 REST services. One popular and widely used CTS2 implementation is NCI LexEVS CTS2 REST Services. Although NCI LexEVS CTS2 Service would be used as the default target service, D2Refine would allow user to add and configure any additional CTS2 compliant terminology service to reconcile with.

The Data Element Reconciliation service would be to reconcile the data dictionary elements with standard data element repositories such as caDSR. This would allow compare and identify gaps and guide modeler by offering data elements that could possibly be considered for inclusion and exclusion. This exercise would definitely make quality of data dictionaries better for their usage and harmonization with other models.

2.2.3 Implement Transformation Profiles

The Standard Template elements derived from OMG AML²¹ have been used to harmonize the three data dictionaries dbGAP, TCGA and PheKB proving that it covers at least the minimum requirements to create archetypes from these popular data dictionary schemas. The implementation of Standard Template now requires creating objects of corresponding target transformation profiles.

We create 1) *ADL Profile* (using OpenEHR AOM specifications) to transform Standard Template into OpenEHR's ADL archetype and, 2) *AML Profile* (using AML's UML Profiles) to transform into AML archetype. At present, ADL and AML are the two formalisms approved by HL7 Clinical Information Modeling Initiative (CIMI)²², so we target these two formats first. Since AML Object Model is inspired from ADL/AOM, the mappings between the Standard Template and ADL are similar to those for AML.

3 Results

A beta-version of D2Refine is available for download [<https://github.com/caCDE-QA/D2Refine>]. Figure 2 shows the system architecture illustrating the interaction among the technical components of D2Refine Workbench.

D2Refine, built on top of OpenRefine platform, adds the reconciliation plugin enabling cell values to reconcile with various internal and external services. At present D2Refine includes a configurable and fully working reconciliation services to any CTS2 Services for controlled terminology terms.

The export extensions of D2Refine offer serialization to OpenEHR's ADL, which has been completed using OpenEHR RM and CIMI RM libraries. The prototype project, which is manually created to implement and test mappings of dbGAP, TCGA, and PheKB to AML Object Model²¹ was successfully completed and would be implemented to serialize data dictionaries to the AML UML models soon. This uses the newly implemented libraries of AML's UML Profiles leveraging Eclipse Modeling Framework (EMF) and Eclipse UML2 libraries. We plan to use utility libraries from Eclipse's Model Driven Healthcare Technology (MDHT) to ease the management of UML artifacts for the AML serialization.

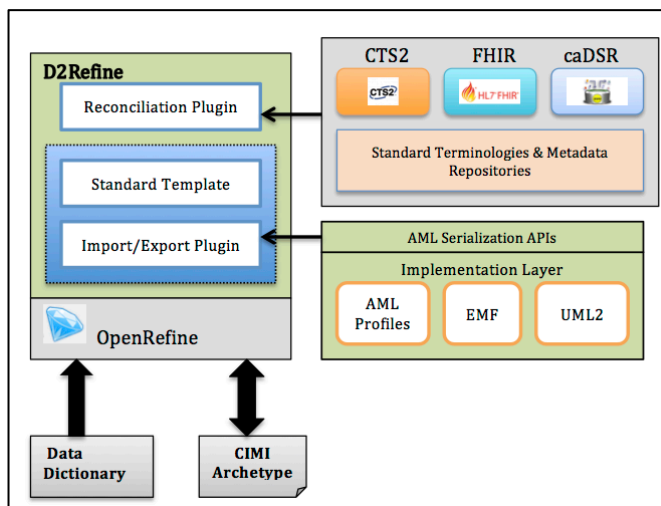


Figure 2: The D2Refine System Architecture

3.1 The Standard Template

The implementation of the Standard Template is in the development stage. The properties of the standard template would be used to identify the completeness of source content. The manual mappings of three popular data dictionaries (dbGAP, TCGA and PheKB) to Standard Template have been successfully created. We plan to utilize these mappings into scripts to triage contents of these data dictionaries and alignment with Standard Template requirements. This implementation would soon enable bidirectional serialization (data dictionaries to CIMI archetypes and vice-versa) within the D2Refine platform.

3.2 The CTS2 Reconciliation Service Extension

D2Refine now offers reconciliation with any terminology with contents available through CTS2 REST services. D2Refine extends the standard reconciliation services workflow of OpenRefine to programmatically add CTS2 Services by using its *base service URL and a human readable name*. The NCI LexEVS CTS2 Services (<https://wiki.nci.nih.gov/display/LexEVS/LexEVS+6.x+CTS2>) is one of the popular and widely accessed public CTS2 implementation, which comes configured as default target reconciliation service with D2Refine. Figure 3 shows the default CTS2 reconciliation service named *D2Refine*. The default reconciliation workflow looks for terms in any code system provided by the CTS2 Services. This lookup functionality would soon be enhanced for better configurations that could be tailored for more complex reconciliation requirements.

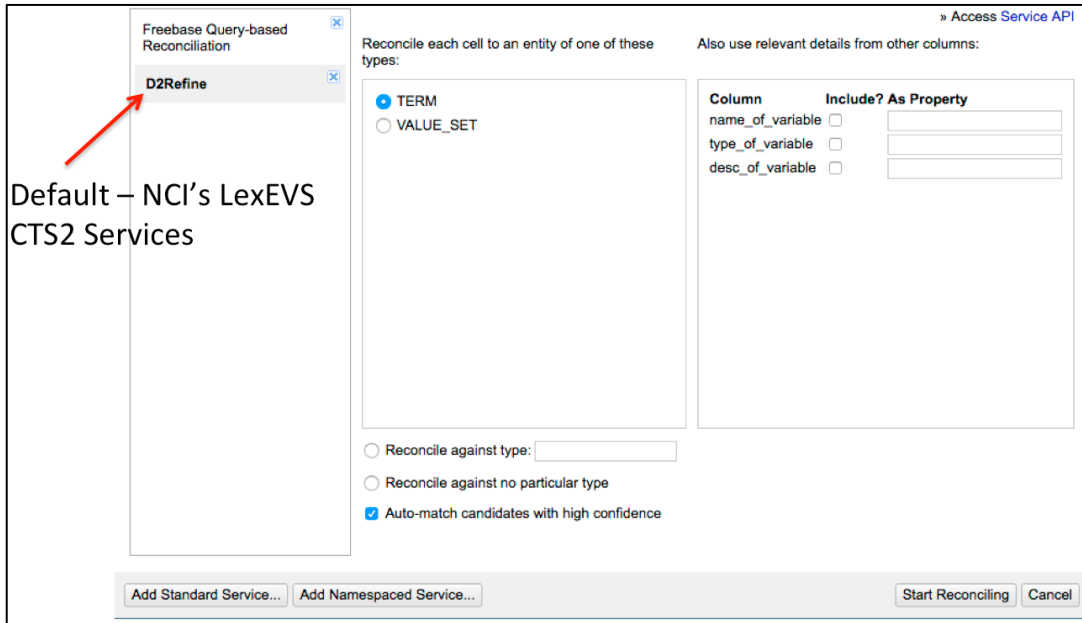


Figure 3: The *default* CTS2 Reconciliation Service of D2Refine Workbench

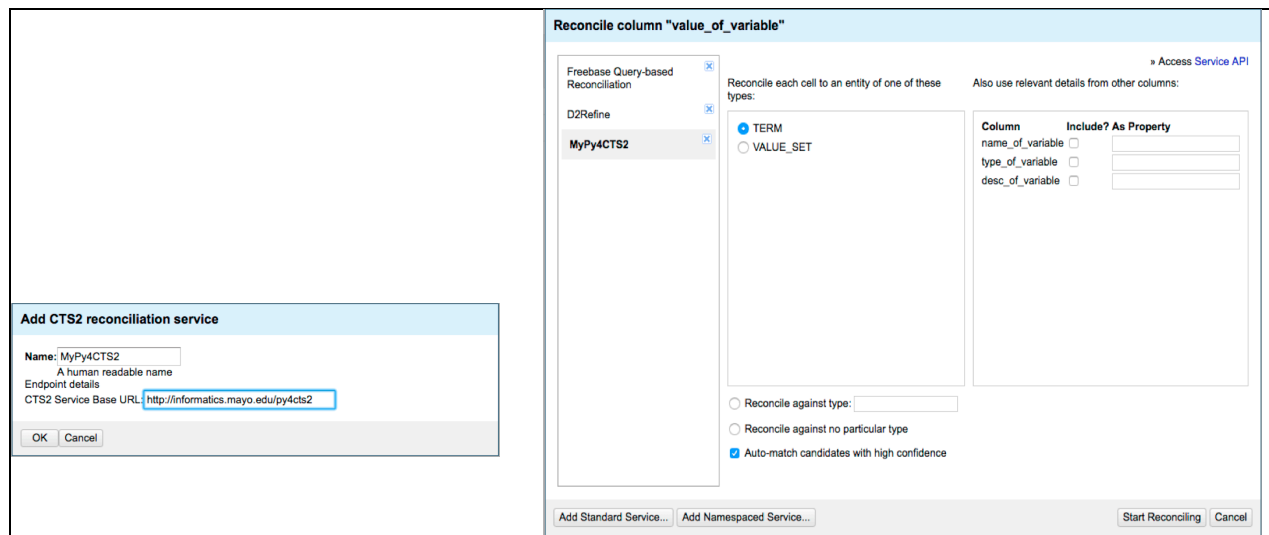


Figure 4: A CTS2 service being added as a reconciliation service

In addition to the default reconciliation service, a user is able to add any available CTS2 REST services implementation by using D2Refine menu extension to add a service. For this D2Refine needs a human readable name and a base URL of the CTS2 service. Figure 4 shows example of a CTS2 Service written in Python added as another service in addition to the default service that D2Refine provides. D2Refine preserves the workspace configuration of the user and the reconciliation service needs to be added once and it is available for any content that is processed afterwards.

D2Refine allows use of multiple reconciliation services for a single data dictionary. D2Refine stores the *bound* CTS2 Term URI as a property of cell's and this is used to fulfill *term-binding* requirements of the Standard Template when data dictionary is transformed to one of the target formats. The cells contents are then replaced with suitable representation of the term using the CTS2 Services.

D2Refine searches for matching terms for a cell value and returns with top few results based on their score (based on CTS2 Service implementation, usually Levenshtein distance²³). The chosen term can be associated with one cell value or all cell values matching the same textual search criteria.

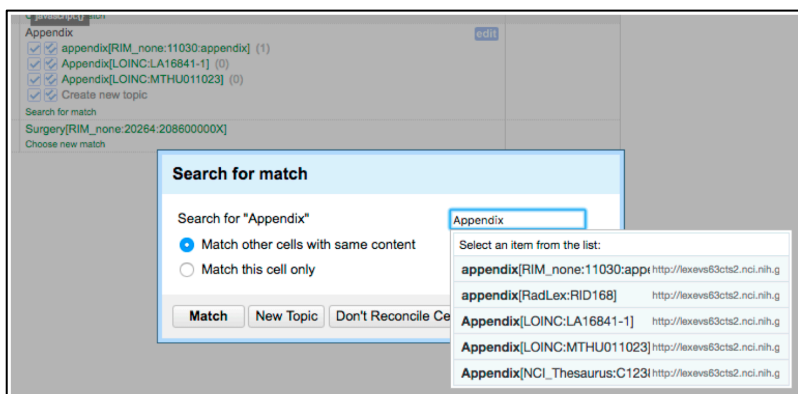


Figure 5: Searching for terms using reconciliation services

In case the returned results are not the expected candidates, user can launch search by manually looking for alternate phrase and D2Refine shows preview of the terms returned with manual search. The manual search of terms makes it easy to adjust the default term-binding behavior.

3.3 Transformation Profiles

D2Refine allows the contents of the data dictionaries to be transformed as OpenEHR's ADL archetypes. This is achieved through extending the export mechanism to include two new export options.

Figure 6 shows two new extensions added to Export Menu of OpenRefine to persist data dictionaries in ADL 2.0 format using either OpenEHR RM or CIMI Core RM. ADL Transformation Profile implementation required open-source ADL Core libraries, which are available from OpenEHR Github repository [https://github.com/openEHR/adl2-core].

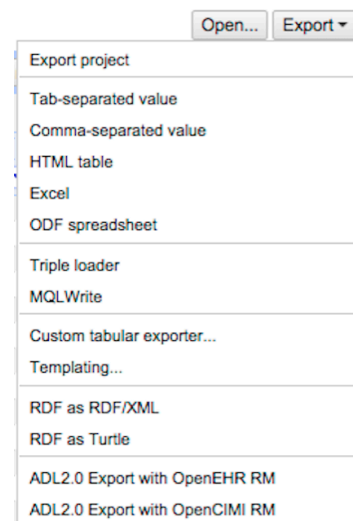


Figure 6: ADL Export Extensions

4 Discussion

The tabular format has been the first and foremost simplest way to organizing information and is being used in recording metadata for various types of models around the world. The research studies and data dictionaries developed this way need to be standardized to be interoperable outside of their local context. Existing tools like OpenEHR's ADL Workbench or UML Modeling tools have learning curve and non-trivial model management workflows, which discourages modelers and they resort to spreadsheets and create their own model creation and management processes, which are rarely interoperable outside.

D2Refine workflow addresses these issues and introduces D2Refine Workbench, which provides an environment for the modelers to define, standardize and transform their studies and data dictionaries for greater interoperability. D2Refine allows importing existing spreadsheets, XML, JSON and database tables. This reduces the learning curve for the users as they keep the original content in the same format.

The Standard Template of D2Refine Workbench not only preserves existing tabular interface, but also aids examining the content for its completeness so that it could be standardized and transformed correctly to a model in a standard format. The D2Refine Standard Template is derived from OMG AML Profiles and hence could guide the modelers to identify the missing model elements that might prevent it be serialized as a valid archetype. The Standard Template requirements have been tested manually using three popular data dictionaries and soon it would be part of D2Refine Workbench.

The standardization of elements is performed extending reconciliation features of the underlying OpenRefine platform. D2Refine introduces the CTS2 reconciliation services for terminology binding of cell values. Though the

default CTS2 reconciliation service connects to NCI LexEVS CTS2 Service, any other CTS2 REST service could be added in addition for reconciliation. Current functionality includes terminology binding to the terms in a code system but would be extended for value-sets. The implementation of reconciliation services to a metadata repository like caDSR and to HL7 FHIR resources is also planned for development which would be crucial to harmonize the models in D2Refine to external standardized models.

Manually translating an existing data dictionary or study definition into ADL or AML is not practical. As the studies and models grow, it becomes impossible to a modeler to manage them without the help of some tooling. D2Refine augments OpenRefine platform with export extensions as transformation profiles. So far we completed the ADL transformation profile implementation, which transforms D2Refine model to ADL archetypes. Since ADL archetypes are useful for any translation only in OpenEHR tooling environment (using ADL Workbench), another two potential useful transformation profiles, in addition to AML transformation (under development), would be HL7 FHIR Profiles and W3C RDF Shape Expressions (ShEx). Extending the D2Refine to work with a HL7 FHIR reconciliation services would create HL7 FHIR schema and profiles. Analogous to XML Schema, ShEx formally define RDF structures, which could be used to validate and transform RDF instance data. There are already a few of implementations available¹⁵ that could be used for implementing D2Refine Workbench's ShEx transformation profile. Another advantage of ShEx is that ShEx offers *semantic action* capability, which could be used to hook up any standard or custom translation of the instance data to RDF, XML, JSON, HTML or other formats.

Microsoft Excel remains the dominating environment when it comes to tabular format to manage information, as majority of computer users already are conversant with MS Excel. We think that making D2Refine's extensions available in Microsoft Excel would give D2Refine huge boost and enormously help vast number of Excel users to get onboard with the features that D2Refine offers. This will also help them access reconciliation services and harmonization workflow right from their Excel spreadsheets. We have successfully developed some prototype *add-ins* to Microsoft Excel and we now plan to extend Microsoft Excel environment to replicate the functionality of D2Refine.

To gauge the effectiveness of D2Refine and the utility of these mappings, we are designing and conducting a usability study. The usability study is being designed and conducted using TURF EHR Usability Framework²⁴. TURF Framework based usability study would help us measure how useful D2Refine is to accomplish tasks as compared to other solutions. The usability study includes four components of TURF Framework – Task, User, Representation and Function. The *user analysis* part of the study is to identify the user base and create a domain model. The *function analysis* focuses on what functions user would want and how many of those are implemented. Task completion time and complexity is analyzed with *task analysis*. The *representation* part of the study would evaluate how well D2Refine represents and aligns with user's expectations. Identified users would be invited to participate in the study and perform the tasks and their tasks completion steps and feedback comments are documented for statistical analysis. The usability study results would help us evaluate our existing strategy and future development of D2Refine Workbench.

5 Conclusions

In our first development iteration of D2Refine, we identified mandatory and optional elements of AML Profiles and successfully designed the Standard Template specifications from it. We were able to implement CTS2 REST reconciliation services extensions for standardization and able to work with controlled terminologies at NCI and other CTS2 services. We extended the search capabilities of the platform to include CTS2 reconciliation services. The export extensions have been implemented to serialize the data dictionaries into equivalent ADL archetypes. The outcome of this work will significantly improve capabilities to manage heterogeneous clinical study data dictionaries for their standardization and harmonization thus improving semantic interoperability. In addition, we also plan to conduct a usability study to gauge utility and effectiveness of D2Refine platform to meet the expectations and requirements of real-world use cases in clinical and translational study domains.

Acknowledgement

This work was supported in part by funding from a NIH R01 GM105688 and a NCI U01 Project – caCDE-QA (U01 CA180940).

References

1. OpenEHR ADL Workbench 2016 [Available from: <http://www.openehr.org/downloads/ADLworkbench/home>.
2. Eclipse Model Driven Health Tools 2016 [Available from: <https://projects.eclipse.org/proposals/model-driven-health-tools>.
3. D2Refine Project 2016 [Available from: <https://github.com/caCDE-QA/D2Refine/wiki>.
4. OpenRefine 2016 [Available from: <http://openrefine.org/>.
5. The database of Genotypes and Phenotypes (dbGaP) 2016 [Available from: <http://www.ncbi.nlm.nih.gov/gap>.
6. 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.
7. Kirby JC, Speltz P, Rasmussen LV, Basford M, Gottesman O, Peissig PL, et al. PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. *Journal of the American Medical Informatics Association : JAMIA*. 2016.
8. PheKB 2016 [Available from: <https://phekb.org/>.
9. TCGA Data Dictionary 2015 [Available from: <https://tcga-data.nci.nih.gov/docs/dictionary/>.
10. The OMG Common Terminology Services 2 Standard 2016 [Available from: <http://www.omg.org/spec/CTS2/1.1/>.
11. CTS2 Wiki 2015 [Available from: http://informatics.mayo.edu/cts2/index.php/Main_Page.
12. NCI caDSR Repository 2015 [Available from: <https://cbiit.nci.nih.gov/ncip/biomedical-informatics-resources/interoperability-and-semantics/metadata-and-models>.
13. OpenEHR's ADL and AOM Specifications 2016 [Available from: <http://www.openehr.org/releases/AM/latest/docs/>.
14. AML Specifications in GitHub 2016 [Available from: <https://github.com/opencimi/AML/tree/master/Specification>.
15. Shape Expressions (ShEx) 2016 [Available from: <http://shex.io>.
16. NCI Thesaurus 2016 [Available from: <https://ncit.nci.nih.gov/ncitbrowser/>.
17. HL7 FHIR DSTU 2 2016 [Available from: <https://www.hl7.org/fhir/>.
18. Object Data Instance Notation 2016 [Available from: <https://github.com/openEHR/odin>.
19. CIMI Core Reference Model 2016 [Available from: https://github.com/opencimi/rm/tree/master/model/Release-3.0.5/UML/AML_RM.
20. Eclipse Modeling Framework 2016 [Available from: <https://eclipse.org/modeling/emf/>.
21. Sharma D, Solbrig HR, Prud'hommeaux E, Pathak J, Jiang G. Standardized Representation of Clinical Study Data Dictionaries with CIMI Archetypes. *AMIA Annual Symposium Proceedings 2016 (in press)2016*.
22. Clinical Information Modeling Initiative (CIMI) 2016 [Available from: <http://www.opencimi.org>.
23. Levenshtein distance 2016 [Available from: https://en.wikipedia.org/wiki/Levenshtein_distance.
24. Zhang J, Walji MF. TURF: toward a unified framework of EHR usability. *J Biomed Inform*. 2011;44(6):1056-67.