Title

IMI-CDE: an interactive interface for collaborative mapping of study variables to common data elements

Authors

Tao, Shiqiang
Chou, Wei–Chun
Li, Jianfu
et al.

Publication Date

2022-06-14

DOI

10.1109/ichi54592.2022.00070

IMI-CDE: an interactive interface for collaborative mapping of study variables to common data elements

Shiqiang Tao
Department of Neurology
The University of Texas Health Science Center at Houston
Houston, USA
shiqiang.tao@uth.tmc.edu

Wei-Chun Chou
Department of Neurology
The University of Texas Health Science Center at Houston
Houston, USA
wei-chun.chou@uth.tmc.edu

Jianfu Li
School of Biomedical Informatics
The University of Texas Health Science Center at Houston
Houston, USA
jianfu.li@uth.tmc.edu

Jingcheng Du
School of Biomedical Informatics
The University of Texas Health Science Center at Houston
Houston, USA
jingcheng.du@uth.tmc.edu

Pritham Ram
School of Biomedical Informatics
The University of Texas Health Science Center at Houston
Houston, USA
Pritham.M.Ram@uth.tmc.edu

Rashmie Abeysinghe
Department of Neurology
The University of Texas Health Science Center at Houston
Houston, USA
Rashmie.Abeysinghe@uth.tmc.edu

Hua Xu
School of Biomedical Informatics
The University of Texas Health Science Center at Houston
Houston, USA
hua.xu@uth.tmc.edu

Xiaoqian Jiang
School of Biomedical Informatics
The University of Texas Health Science Center at Houston
Houston, USA
xiaoqian.jiang@uth.tmc.edu

Peter W Rose
San Diego Supercomputer Center
University of California San Diego
La Jolla, USA
pwrose@ucsd.edu

Lucila Ohno-Machado
Department of Biomedical Informatics
University of California San Diego
La Jolla, USA
lohnomachado@health.ucsd.edu

Guo-Qiang Zhang
Department of Neurology
The University of Texas Health Science Center at Houston
Houston, USA
guo-qiang.zhang@uth.tmc.edu

***Abstract*—The National Institute of Health (NIH) launches the RADx Radical research collaboratives (RADx-rad) to advance new, non-traditional approaches for COVID-19 testing. RADx-rad projects are required to adopt common data elements (CDEs) to collect data to increase data interoperability. To overcome the challenges in finding appropriate CDEs for a wide range of study variables, we create a web application - IMI-CDE to ease the burden of mapping study variables to CDEs from researchers. IMI-CDE can automatically recommend CDE candidates for a study variable based on its name and description. Together with interactive mapping interfaces, IMI-CDE allows researchers to perform variable-CDE mapping with one mouse click. In addition, the IMI-CDE application supports users with multiple roles to work collaboratively on the mapping tasks. We have piloted the IMI-CDE with RADx-rad projects. 22 researchers from 8 different projects have started to use the IMI-CDE system for variable-CDE mappings. The beta-testing evaluators reported the system is intuitive, effective, and easy to use.**

***Index Terms*—COVID-19, Common Data Element, CDE, Mapping, Data Dictionary, Study Variable, CDE Recommendation**

2022 IEEE 10th International Conference on Healthcare Informatics (ICHI) | 978-1-6654-6845-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICHI54592.2022.00070

2575-2634/22/$31.00 ©2022 IEEE
DOI 10.1109/ICHI54592.2022.00070

I. Introduction

A Common Data Element (CDE) is a combination of a precisely defined question or data variable together with a specified set of responses or value options that is common to multiple datasets or used across different studies [1]. The National Institutes of Health (NIH) encourages and facilitates the use of CDEs across the research community. A NIH CDE resource portal was launched in January 2013. Currently, the portal provides access to 28, 945 CDEs in 17 collections [2].

The COVID-19 pandemic has driven numerous scientific research on epidemiology. The multitude of research raises the concern of data fragmentation and urges the need for better data interoperation. The National Institute of Health (NIH) launches the RADx Radical research collaboratives (RADx-rad) [3] to advance new, non-traditional approaches for COVID-19 testing. RADx-rad supports 50 awardees from 61 different institutes from over 20 states in the US (and through them from 20 different countries). Given the focus on new and non-traditional approaches for technology development, the projects use a very wide variety of variables and potentially value sets. In order to facilitate data integration across different research topics, NIH requires project awardees to adopt CDEs when collecting data. However, mapping study variables to CDEs is often challenging given the wide range of study variables and the large size of CDEs. There is no existing tool that can conveniently map study variables to CDEs. To address this challenge, we developed IMI-CDE – an Interactive Mapping Interface for collaborative variable-CDE mapping. IMI-CDE, as a web application, provides user-friendly interfaces for researchers to upload their data dictionaries, find CDE recommendations for study variables and perform mapping from data variable to CDE with a single button click. IMI-CDE supports multiple user roles to collaborate on variable-CDE mapping tasks. User feedback can be submitted to CDE experts through IMI-CDE comment and messaging systems.

II. Methods

A. IMI-CDE System Architecture

IMI-CDE is a web application developed using Ruby on Rails framework adapted from our previous work to match data dictionaries to ontologies [4]. It has MySQL as its backend database and internet browser as its frontend. As illustrated in Fig. 1, IMI-CDE consists of seven function modules: (1) data dictionary import interface; (2) CDE recommendation system; (3) variable-CDE mapping interface; (4) mapping review; (5) mapping result export; (6) role-based access control; and (7) messaging system. Four user roles are created in IMI-CDE: system administrator, data dictionary manager, data dictionary member, and CDE expert. Data dictionary manager upload data dictionaries and manage its versioning. Data dictionary manager and member can perform the actual variable-CDE mapping tasks with the help of the CDE recommendation system. The mapping results are then reviewed by CDE experts. The communication among users are supported by the messaging system through emails.

B. CDE Recommendation Algorithm

The CDE recommendation system automatically assigns top 5 best-matched CDEs by default in three NIH CDE repositories, i.e., NIH Minimum, NIH All, and NLM COVID, respectively, to a study variable requested by the user. The system also identifies un-matched study variables that would need to be further discussed (e.g., create new CDE and update current CDE). The matching algorithm for the recommendation system is based on Elasticsearch BM25 (Best Match 25) [5] which tries to calculate and rank the relevance scores between the query text and the repository documents-CDE records in the three NIH CDE repositories. The BM25 algorithm improves upon TF-IDF (Term Frequency-Inverse Document Frequency) [6] by overcoming its two shortcomings: (1) not taking document length into account; (2) term frequency not saturated. For a query q, with query tokens q1, q2, ...qM, the BM25 relevancy score for document Di (i from 1 to N) in repository D is displayed in Fig. 2, where M is the tokens number in query q, N is the documents number in the repository D, $N(q_j)$ is the number of documents in the repository that contain token q_j, $f(q_j, D_i)$ is the number of times token q_j occurs in document D_i, $|D_i|$ is the number of tokens in document D_i, d_avg is the average number of words per document, and b and k are hyper-parameters for BM25.

C. Variable-CDE Mapping Workflow

First of all, a data dictionary manager uploads a data dictionary. The user will be asked to choose a name column and a description column for study variables from the data dictionary. With the text content of these two columns, the IMI-CDE recommendation system runs at the background to fetch candidates for every study variable in the data dictionary. This process may take several minutes and a notification email will be sent to user once it is completed. Following that, users can open the variable-CDE mapping interface and select the target CDE for each study variable based the system recommendations. After all study variables are mapped, users can request CDE experts to review the mapping result. At last, the mapping result can be exported and downloaded into CSV format.

D. Evaluation

To evaluate the effectiveness of the IMI-CDE application, we invite researchers from RADx-rad projects to perform the beta-testing. The evaluation consists of two sessions. In the first session, the IMI-CDE developers demonstrated the variable-CDE mapping workflow. Then evaluators try to map their own data dictionaries independently. In the second session, all evaluators show their mapping results and provide their feedback.

III. Results

A. Variable-CDE Mapping Interface

Fig. 3 shows the IMI-CDE mapping interface for a project called "CoCreate Aim 1 Survey". Three areas are highlighted and labeled by number in the figure. Area 1 lists all the study

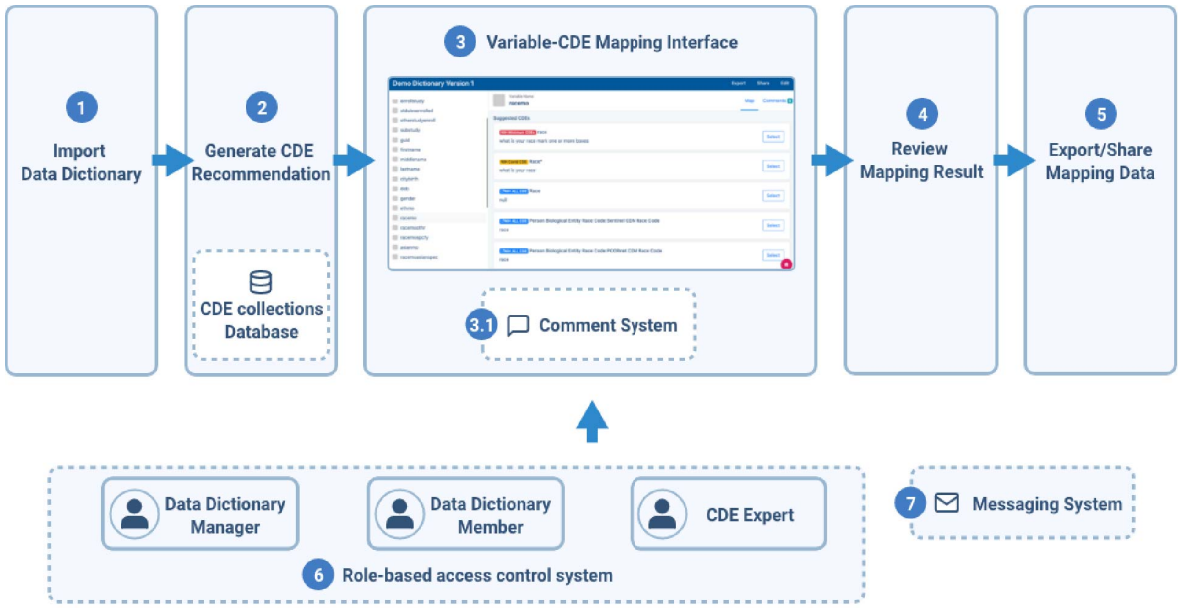

Fig. 1. IMI-CDE system architecture

$$Score(q, D_i) = \sum_{j=1}^{M} \log(1 + \frac{N - N(q_j) + 0.5}{N(q_j) + 0.5}) \cdot \frac{f(q_j, D_i) \cdot (k+1)}{f(q_j, D_i) + k \cdot (1 - b + b \cdot |D_i| / d_{avg})}$$

Fig. 2. BM25 relevancy score formula

variables in the current project. When clicking on "record_id", area 2 displays this study variable's name and description and area 3 shows the CDE candidates ranked by their BM25 relevancy scores. In this specific example, users can select the item "study_id" in the NIH Minimum CDEs classification to complete this mapping. Area 4 provides another comments tab, where users can submit feedback about variable-CDE mapping to CDE experts. There is also an "Export" for users to download the mapping results in CSV format.

B. CDE Recommendation System

The IMI-CDE recommendation system is developed and deployed as a RESTful web service using python and flask packages. It firstly creates three Elastic search indexes based on the three CDE repositories. Then, for each incoming query, the query text is preprocessed and tokenized by eliminating its non-alphanumeric characters and fetched into the matching algorithm. Finally, the top 5 rankings mapping results containing the CDEs in three CDE repositories are returned to the user.

C. Preliminary Usage

The IMI application is deployed in production [7]. As of March, 2022, there are 22 researchers from 8 different RADx-rad projects registered in the IMI application who have started to work on the mappings between study variables and CDEs.

D. Evaluations

Six researchers accepted the IMI beta-testing invitation and participated in our two test sessions. Feedback showed that they found the IMI system is intuitive and easy to use. The CDE recommendation features and the mapping interface allows the mapping work to be done with only one click. However, most evaluators reported the recommendation quality was not always desirable.

IV. Discussion

Common data elements, as standardized terms or concepts, are both human and machine readable. They allow researchers to link data across different research and therefore have an advantage in data interoperation and integration. The National Institute of Health (NIH) and National Library of Medicine(NLM) have established the importance of CDEs in their strategic plans [8], [9]. However, it is not straightforward to apply CDEs for research data collection in RADx-rad projects. First of all, it is challenging to find the best match for a study variable from more than 20,000 CDEs in the current NIH CDE repository. Some of them might be synonyms but belong to different classifications. Second, NIH has additional criteria for the use of CDE in RADx-Rad projects: Minimum required CDEs must be present, and COVID-19 CDEs collection has a higher priority than the other classifications. In addition, researchers are not necessarily familiar with CDEs compared with the traditional study variables. To ease the burden of mapping traditional study variables to CDEs, we create IMI as a dedicated tool that can facilitate the normalization of data dictionaries by mapping study variables to CDEs.

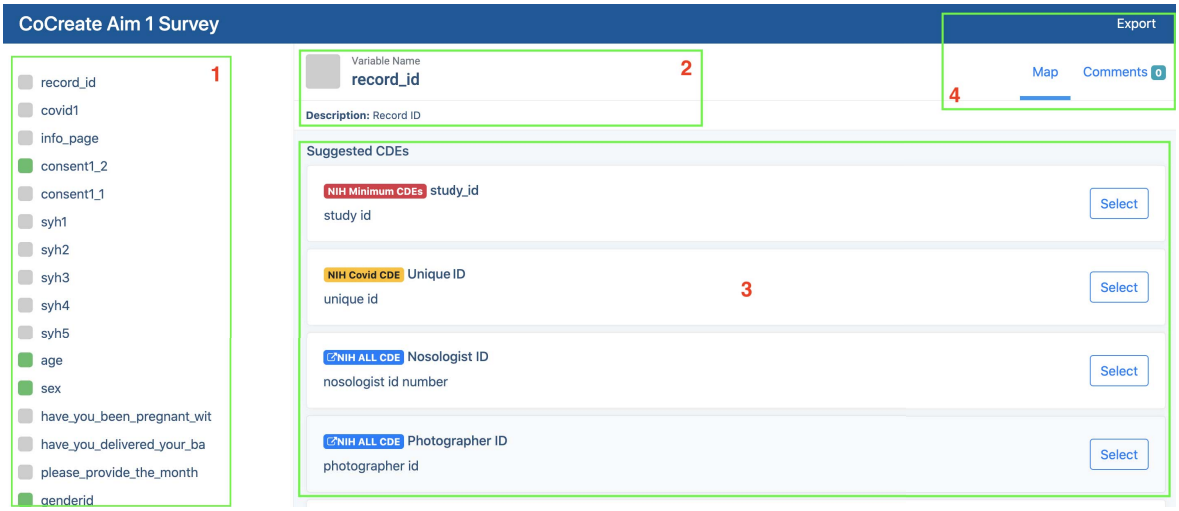

Fig. 3. Mapping interface from study variable to common data element

A. *Limitation and Future Work*

In the current IMI CDE recommendation system, the matching algorithm is solely based on the string matching of name and description of study variables with CDEs, which proved not sufficient as our evaluators reported poor quality for some CDE recommendations. More semantics such as value options and source need to be considered. A more intelligent matching algorithm based on natural language processing needs to be developed.

V. Conclusion

In this paper, we presented IMI – an interactive, intuitive, and collaborative web interface for building mappings from study variables to CDEs, so as to facilitate interoperation and integration for COVID-19 testing. IMI was successfully pilot tested with RADx-rad projects to build variable-CDE mappings.

VI. Acknowledgment

This study was supported by the National Institutes of Health - National Library of Medicine (Project Number - 4U24LM013755-01).

References

[1] Sheehan J, Hirschfeld S, Foster E, Ghitza U, Goetz K, Karpinski J, Lang L, Moser RP, Odenkirchen J, Reeves D, Rubinstein Y, Werner E, Huerta M. Improving the value of clinical research through the use of Common Data Elements. Clin Trials. 2016 Dec;13(6):671-676. doi: 10.1177/1740774516653238. Epub 2016 Jun 15. PMID: 27311638; PMCID: PMC5133155.

[2] NIH CDE Repository. https://cde.nlm.nih.gov/home. Accessed Mar 1, 2022

[3] RADxSM RAD Website. https://www.radxrad.org/. Accessed Mar 1, 2022

[4] Tao S, Zeng N, Hands I, Hurt-Mueller J, Durbin EB, Cui L, Zhang GQ. Web-based interactive mapping from data dictionaries to ontologies, with an application to cancer registry. BMC Medical Informatics and Decision Making. 2020 Dec;20(10):1-9.

[5] Robertson SE, Walker S, Jones S, Hancock-Beaulieu MM, Gatford M. Okapi at TREC-3. Nist Special Publication Sp. 1995 Nov 1;109:109.

[6] Aizawa A. An information-theoretic perspective of tf-idf measures. Information Processing & Management. 2003 Jan 1;39(1):45-65.

[7] IMI Application Site. https://imi.radxrad.org. Accessed Mar 1, 2022.

[8] Kittrie E. The US National Library of Medicine: A Platform for Biomedical Discovery & Data-Powered Health. Proceedings of the 29th on Hypertext and Social Media 2018 Jul 3 (pp. 155-155).

[9] NIH Strategic Plan for Data Science. https://datascience.nih.gov/nih-strategic-plan-data-science Accessed Mar 1, 2022