End-to-End Implementation of a Workflow for Validating Semantic Mappings and Constructing Ontology Extensions

Jared Houghtaling[1], Polina Talapova[1], Soojin Park[2], J. Harry Caufield[3], Andrew Williams[1]

1. *Tufts Medicine - Institute for Clinical Research and Health Policy Studies (ICRHPS)*
2. *Columbia University - Vagelos College of Physicians and Surgeons*
3. *Lawrence Berkeley National Laboratory - Berkeley Bioinformatics Open-Source Projects (BBOP)*

Background

The Bridge2AI for Clinical Care (B2AI For CC) research consortium aims to capture and consolidate rich multimodal data from fifteen data contributing sites in order to support complex analytic processes in machine learning (ML) and artificial intelligence (AI); such consolidation and analytic support is nontrivial and requires a diversity of expertise and consortium-specific OMOP concepts for interacting with multimodal (e.g. images, waveforms) files alongside OMOP-shaped datasets. In this work, we demonstrate a novel, cross-platform approach (**Figure 1**) that provides a user-friendly entrypoint (i.e. Google Sheets) for clinical experts to evaluate mapping representations using *A Simple Standard for Sharing Ontology Mappings* (SSSOM) format[1]. We use multiple Google Apps Scripts to protect the data entry processes dynamically and to commit the resulting spreadsheets to a central GitHub repository. Once in GitHub, we execute a vocabulary integration workflow via GitHub Action in which the SSSOM-format mappings are processed and transformed into the shape and relational structure of the OMOP vocabulary tables. Once transformed, the tables are then stored back on the repository where they can be referenced by data contributing sites and can support analytic workflows both locally at sites and centrally in the collaborative cloud. The methodology is generalizable and allows for drafting, sharing, searching, and validating semantic mappings to standard OMOP concepts that will be made available to the entire OHDSI community to facilitate reuse of validated mappings, automation of ETLs, and adherence to ETL conventions.

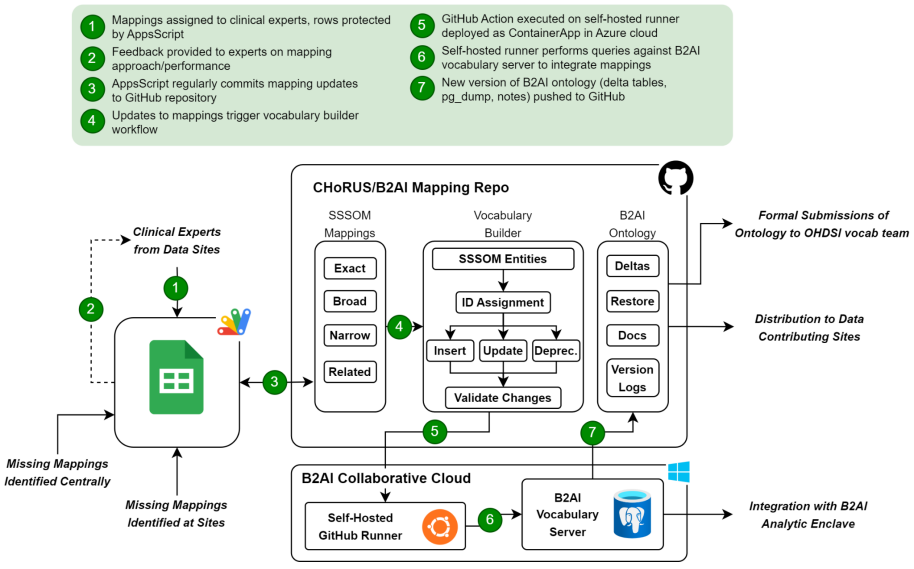

Figure 1. Process flow of B2AI ontology construction, beginning with SSSOM-oriented data entry and resulting in OMOP-shaped

Methods

a. Data entry in SSSOM-specific collaborative spreadsheets

Instead of engineering a custom application to support consortium-wide mapping and validation efforts, we chose to design and create a simple Google Sheet environment due to its familiarity, ease-of-use, collaborative features, and Apps Script integrations[2]. We shared the sheet with 14 designated clinicians across the 15 participating sites, and implemented automated protections to prevent those users from (1) inadvertently modifying each other's work, and (2) changing existing mappings that had been confirmed and validated. The environment currently supports mapping efforts stemming from a context-specific DelPhi process[3] wherein particular flowsheet descriptions were identified and prioritized based on relevance to observational health research in critical care environments. We have also established processes for sites to contribute unmapped codes to the environment, as well as processes for consortium leadership to identify and prioritize OHDSI Standardized Vocabularies[4] gaps spanning various data sites and assign validation efforts appropriately. We split validation efforts into four SSSOM categories - namely exact, narrow, broad, or related matches - and assigned clinical experts to different subsets of codes depending on their specialty. Users reviewing mappings were able to reference a consortium-specific version of Athena that searches a version of the OMOP vocabularies containing previously validated terms in the B2AI ontology.

b. Cross-platform data integration processes

In addition to the AppScript-based row protections mentioned above, we also created an Apps Script to transfer mappings from the Google environment to GitHub at regular intervals. This process creates a robust version history that can be used to restore prior mappings or track drift over time, and it also makes those mappings readily available to GitHub Action workflows that convert them into tables consistent with structural and relational requirements of the OMOP CDM, including vocabularies.

c. CI/CD Vocabulary Building Processes

We deployed two key pieces of infrastructure in an Azure cloud environment in support of a self-hosted GitHub runner workflow; an Azure ContainerApp that serves as a virtual machine linked to GitHub and able to execute code sets, and a Azure Flexible Postgres Server containing an indexed and constrained set of vocabulary tables to be referenced and updated based on validated mappings. The vocabulary update process is executed at regular intervals following mapping updates. We coordinated execution using a GitHub action that (1) ingests the mappings into the Postgres database, (2) performs syntactic quality control on the mappings and mapping metadata to flag potentially errant entries, (3) evaluates differences between those mapped terms and the latest OMOP vocabulary version with B2AI terms, (4) produces staging tables representing various concepts, both standard and non-standard, along with their respective relationships to existing OMOP terminologies and associated concept_id assignments greater than two billion, (5) inserts those staging tables into the constrained vocabulary tables, and (6) exports "delta" tables (i.e. B2AI-specific terms and relationships that augment the publicly available vocabularies) back to GitHub where they can undergo validation and be integrated by consortium members into local Extract, Transform, and Load (ETL) processes.

d. Ontology Validation and Dissemination

With the mappings in OMOP format, we applied validation protocols to ensure that the newly defined terms and relationships were both (1) consistent with OMOP convention, and (2) representative of the input mappings used to create them. Following the success of these processes (based on internal thresholds), we migrated the mapping output to a public repository and placed it alongside scripts that enable data contributing sites to augment their local vocabulary tables. We also established a pipeline

between validated vocabulary versions and the B2AI For CC central analytics enclave to support model building and cohort characterization efforts. We plan to submit formal versions of the B2AI For CC ontology and associated mappings to the OHDSI vocabulary team for integration into the broader OMOP vocabularies twice per year.

Results

Thus far, we have successfully applied the procedures described above to integrate more than 1000 unique concepts, representing more than 50,000 distinct rows of data across various vocabulary tables, into a B2AI-specific ontology. Now that the workflow is established, the ontology is growing organically and any mapping updates or validations are captured and consolidated daily with minimal effort. We expect to define and implement robust protocols for contributing mappings and training clinical experts in the validation process.

Conclusion

The ontology pipeline we've established here represents a complete pathway between clinical experts and interoperable mapping relationships. The work builds on prior SSSOM entity standardization and makes use of cross-platform automation strategies to increase ease-of-use, transparency, and collaboration. Because of the open-source nature of the tooling, we expect that this workflow can serve as a model for other consortia, or individual institutions, that require standardized processes to make connections or fill gaps in the OMOP vocabularies in a robust and version-controlled manner. We have recently implemented the pipeline for the Geospatial Information Systems (GIS) workgroup in OHDSI, and expect to expand and apply it to other use cases in the coming months.[5]

References

1. Matentzoglu N, Balhoff JP, Bello SM, Bizon C, Brush M, Callahan TJ, Chute CG, Duncan WD, Evelo CT, Gabriel D, Graybeal J, Gray A, Gyori BM, Haendel M, Harmse H, Harris NL, Harrow I, Hegde HB, Hoyt AL, Hoyt CT, Jiao D, Jiménez-Ruiz E, Jupp S, Kim H, Koehler S, Liener T, Long Q, Malone J, McLaughlin JA, McMurry JA, Moxon S, Munoz-Torres MC, Osumi-Sutherland D, Overton JA, Peters B, Putman T, Queralt-Rosinach N, Shefchek K, Solbrig H, Thessen A, Tudorache T, Vasilevsky N, Wagner AH, Mungall CJ. A Simple Standard for Sharing Ontological Mappings (SSSOM). Database (Oxford). 2022 May 25;2022:baac035. doi: 10.1093/database/baac035.
2. Martin A. Mastering Google Sheets for Business Analytics. Teamgate Blog [Internet]. 2024 Feb 14 [cited 2024 Jun 17]. Available from: https://www.teamgate.com/blog/google-sheets-for-analytics/
3. Nasa P, Jain R, Juneja D. Delphi methodology in healthcare research: How to decide its appropriateness. World J Methodol. 2021 Jul 20;11(4):116-129. doi: 10.5662/wjm.v11.i4.116. PMID: 34322364; PMCID: PMC8299905.
4. Reich C, Ostropolets A, Ryan P, Rijnbeek P, Schuemie M, Davydov A, Dymshyts D, Hripcsak G. OHDSI Standardized Vocabularies - a large-scale centralized reference ontology for international data harmonization. J Am Med Inform Assoc. 2024 Mar;31(3):583-590. doi: 10.1093/jamia/ocad247.
5. https://www.github.com/TuftsCTSI/CVB