ETLing from your OMOP CDM to your OMOP CDM? An efficient solution to vocabulary migration and data quality enhancement

Clair Blacketer[1,2], Anton Ivanov[1], Evanette Burrows[1], Dmitry Dymshyts[1], Frank DeFalco[1]

[1]Janssen Research & Development, Raritan, NJ [2]Department of Medical Informatics, Erasmus MC, Rotterdam, NL

Background

In the realm of data management, particularly within healthcare and research databases, the Extract-Transform-Load (ETL) process plays a critical role in standardizing disparate data sources into a unified Common Data Model (CDM) like the Observational Medical Outcomes Partnership CDM [1,2]. However, updating the vocabulary of an existing CDM instance traditionally necessitates a complete rerun of the ETL process, a task that is both challenging and time-consuming. This becomes especially problematic when access to the native data is restricted, as is often the case with vendor-licensed standardized data. To address these challenges, we present a novel method that allows for vocabulary updates in an existing CDM instance without reverting to the native data. This method proves invaluable when a full ETL is cost-prohibitive or impossible due to data access limitations, and it also offers an opportunity to address outstanding data quality issues.

Methods

Our approach involves creating an ETL specification that treats the existing CDM as if it were a native data source. Key steps in our methodology include:

1. **Source Concept ID Mapping**: We extract source concept IDs from the old CDM, identifying the corresponding source codes and source vocabularies using the old vocabulary, and then map these to the current source concept IDs and standard concept IDs in the new vocabulary. This method accounts for date-specific source concept IDs. For example, in the US some drug source codes are reused over time as drugs enter and leave the market. Coupling the source code with the valid start and end date guarantees the proper source concept IDs are assigned, ensuring accurate mappings over time

2. **Data Quality Logic**: We implement logic to correct data quality issues. These are issues that were identified prior by the Data Quality Dashboard (DQD)[3].
    1. Persons with an unknown year of birth are removed
    2. Persons with a year of birth in the future are removed
    3. Persons with a year of birth prior to 1875 are removed
    4. Observation periods with a start date in the future are removed
        1. If a person only has one observation period that starts in the future, that person is also removed
    5. Observation periods with a start date that occurs after the end date are removed
        1. If a person only has one observation period that meets this criterion, that person is also removed
    6. If an observation period end date is a future date it is truncated to the end date of the database
    7. Any death records with a death date that occurs in the future are removed
    8. Any visit occurrence or visit detail records with a start date that that occurs in the future are removed
3. **Focus on Event Tables**: Our primary focus was on event tables (e.g., CONDITION_OCCURRENCE, PROCEDURE_OCCURRENCE, DRUG_EXPOSURE, OBSERVATION, MEASUREMENT, and DEVICE_EXPOSURE), where the majority of vocabulary mappings occur. Non-event tables (e.g., PERSON, OBSERVATION_PERIOD, LOCATION) were checked to ensure that any used concepts remained standard in the new vocabulary.
4. **Operationalizing ETL Logic**: We developed code to operationalize the ETL logic, preparing the new vocabulary by transferring any custom vocabularies or concepts from the old to the new vocabulary. We then generated the new CDM using this new vocabulary and CDM-to-CDM logic.
5. **Quality Assurance**: Post-generation, we used the DQD to identify any overlooked issues or errors in the vocabulary mapping by comparing the results to previously generated DQD output.

Once we completed these steps and addressed any issues identified by the DQD we applied the logic to four OMOP CDM standardized datasets: IQVIA® Adjudicated Health Plan Claims Data, IQVIA® Disease Analyzer France, IQVIA® Disease Analyzer Germany, IQVIA® Longitudinal Patient Database Australia.

Results

The full ETL logic employed can be found at: ETL-LambdaBuilder Documentation.

Our process revealed several key considerations for mapping to a new vocabulary without accessing native data:

1. **Domain Changes:** If the standard concept has changed domains, the corresponding record must also move to the appropriate domain.

2. **Multi-Domain Mappings:** Some source concepts previously mapped to multiple standard concepts across different domains. Failing to identify these could result in record duplication under the new mappings.
3. **Primary Key Re-issuance:** Domain movements and mapping changes necessitate re-issuing primary keys for event tables to maintain data integrity.
4. **Database-Specific Vocabulary Integration:** Incorporating database-specific vocabularies, custom concepts, and locally managed source-to-concept map files is critical. Omitting these can lead to a significant number of records being mapped to a standard concept id of zero.
5. **Data Quality Dashboard:** The Data Quality Dashboard proved essential in identifying and resolving errors throughout the vocabulary mapping process.

Discussion

We introduce a novel approach to updating the vocabulary of a CDM instance without reverting to native data. This method significantly reduces the resources and effort required to adopt a new vocabulary version. Additionally, it can be utilized to correct data quality issues, promoting consistency across federated data networks, enhancing interoperability, and fostering better evidence generation. By addressing both technological and practical challenges in vocabulary updates, this approach provides a scalable and efficient solution for maintaining up-to-date and high-quality CDM instances in environments where access to native data is limited or unavailable.

Conclusion

The methodology presented offers a viable and resource-efficient alternative to traditional ETL processes for updating CDM vocabularies. Our approach not only simplifies the transition to newer vocabulary versions but also ensures data quality and integrity, making it a valuable tool for researchers and data managers working within the constraints of vendor-licensed data and federated data networks.

References

1 Quiroz JC, Chard T, Sa Z, *et al.* Extract, transform, load framework for the conversion of health databases to OMOP. *PLoS One*. 2022;17:e0266911.
2 Janssen ETL Documentation. Janssen ETL Documentation. https://ohdsi.github.io/ETL-LambdaBuilder/ (accessed 6 June 2023)
3 Blacketer C, Defalco FJ, Ryan PB, *et al.* Increasing trust in real-world evidence through evaluation of observational data quality. *Journal of the American Medical Informatics Association : JAMIA*. Published Online First: 27 July 2021. doi: 10.1093/jamia/ocab132