Implementing the OMOP common data model using *dbt*

Quin Ashcroft[1], Dale Kirkwood[1], Tim Howcroft[1], Jo Knight[1,2], Stephen Dobson[1], Vishnu V Chandrabalan[1]

1 Department of Data Science, Lancashire Teaching Hospitals NHS Foundation Trust
2 Data Science Institute, Lancaster Medical School, Lancaster University

DataScience@lthtr.nhs.uk
https://www.lancsteachinghospitals.nhs.uk/

Background

Lancashire Teaching Hospitals NHS Foundation Trust (LTH) is a digitally mature secondary care provider, major trauma centre and multi-specialty tertiary referral centre in Northwest England and part of the UK National Health Service. LTH have routinely collected healthcare data for more than 1.7 million patients spanning over 15 years, covering most aspects of secondary care.

Electronic health data collected using a primary EPR and multiple disparate specialist clinical systems are held in siloed, poorly documented databases from multiple vendors with no straightforward method to create a single, linked, person-centric, semantic view. The OMOP Common Data Model was chosen for its person-centric design, rich OHDSI analytics software ecosystem, vibrant global researcher community and opportunities for national and international collaboration to accelerate research as well as to support near real-time operational and clinical intelligence to drive transformation and improvements to patient care pathways and organisational efficiency.

We describe the use of dbt (data build tool) to implement a complex extract-load-transform (ELT) workflow that transforms data from multiple sources daily and incrementally, into a single OMOP database.

Methods

dbt[1] [dbt-core with dbt-sqlserver adapter] was chosen to accelerate ELT development for several reasons:

- Collaborative development as a team with version control
- Improved code reusability, maintainability, and automation
- Multiple target architectures and parallelism
- Auto-documenting with lineage as directed acyclic graphs (DAG)
- Support for incremental refreshes, DQ tests and snapshots
- Ability to build sections of the DAG by selecting models or tags
- Integration with Prefect for workflow orchestration

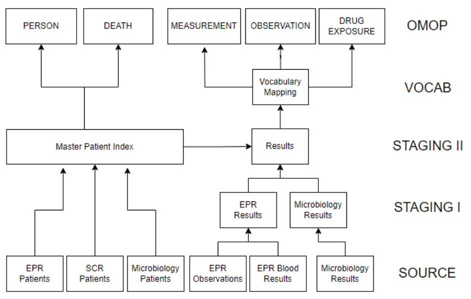


Figure 1. ELT Transformation Layers

Incremental loading for selected models using custom strategies minimised computational load on source systems, reduced build times and made daily updates feasible.

Vocabulary mapping of source concepts to standardised vocabularies (SNOMED, ICD10, OPCS4) from Athena was done using Usagi. These were exported to multiple domain/dataset-specific CSV files which were added to version control and incorporated into dbt as seeds. A single model

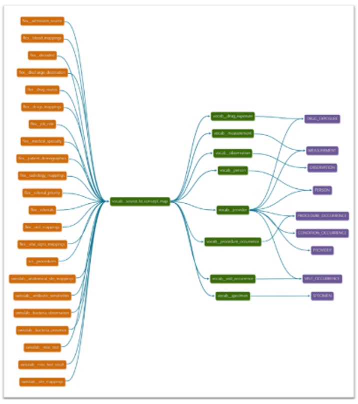


Figure 2. Vocabulary DAG Lineage

(source_to_concept_map, *Fig. 2*) was created as a union of these seed models and integrated into downstream lineage.

SQLFluff[2] with custom rules was used as a pre-commit hook for ensuring code quality, linting, fixing common issues and standardising formatting to allow improved code readability and consistency between collaborators.

GitHub was used for version control, issue-tracking, and project management. Weekly code/PR reviews, and an internal branching and merge strategy in combination with the use of dbt accelerated collaborative development.

Workflow orchestration using a custom solution built with Prefect[3] allowed fine-grained control over building selected models and their dependencies within the DAG and targeting different sources at different times of day to work in harmony with other complex ELT workloads in the organisation. Prefect server deployed on a private Azure Kubernetes cluster improved visibility of tasks, flows, logs, and failures for the entire team.

Results

dbt enabled rapid, collaborative transformation of data from multiple, complex source databases in Oracle, Sybase, and SQL Server into OMOP in SQL Server.

Data from multiple sources were harmonised in a staging layer *(Fig. 1)* before being separated by domain into OMOP tables. *Figure 3* shows the exponential increase in volume of data created each month emphasising the importance of incremental updates.

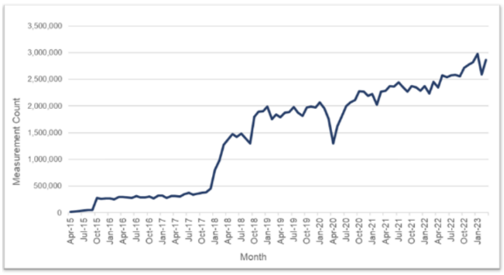
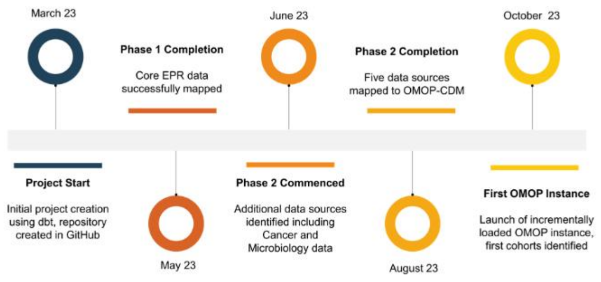


Figure 3. Number of measurements by month



March 23
Phase 1 Completion
Core EPR data successfully mapped
June 23
Phase 2 Completion
Five data sources mapped to OMOP-CDM
October 23
Project Start
Initial project creation using dbt, repository created in GitHub
May 23
Phase 2 Commenced
Additional data sources identified including Cancer and Microbiology data
August 23
First OMOP Instance
Launch of incrementally loaded OMOP instance, first cohorts identified

Figure 4. Project timeline

Alternative approaches including SQL stored procedures, Python/SQLAlchemy were tried and abandoned due to multiple reasons. These were addressed easily with dbt as described in methods.

Data engineering using dbt allowed Phase 1 to be completed in 12 weeks, and Phase 2 in 4 weeks *(Fig. 4)*. *Figure 5* demonstrates the complexity of generating a single OMOP table from multiple source models.

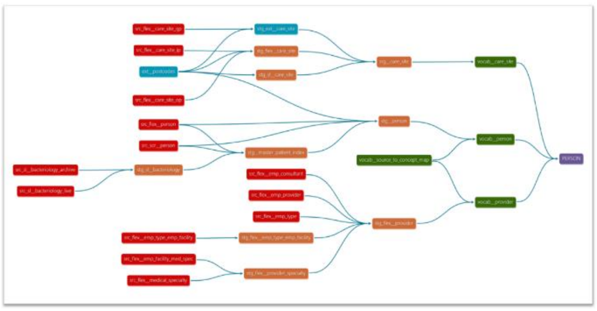


Figure 5. Person DAG lineage

The use of dbt and git allowed for additional sources to be integrated into existing lineage with minimal effort and enabled a clear path for integration of future data sources. The documentation and DAG data lineage generated by dbt were published online[4] and shared with regional development partners.

Conclusions

The use of a data transformation and orchestration workflow based around dbt allowed rapid harmonisation of multiple, high-value, healthcare data sources in a complex NHS organisation into OMOP. Incremental loading with daily updates extended the use of OMOP from research into operational intelligence and near real-time direct care uses. Similar methodology can be used at other complex sites, enabling a collaborative and open approach to OMOP transformation projects where changes may only be required at the source layer with subsequent transformations done using a shared transformation lineage.

References

1. dbt: https://www.getdbt.com/
2. SQLFluff: https://sqlfluff.com/
3. Prefect: https://www.prefect.io/
4. dbt docs: https://omop-lsc.surge.sh/

Funding

EHDEN-HDRUK 7th Data Partner Call
NIHR NW Clinical Research Network