

Operationalizing OCL-Driven CIEL Refresh in OHDSI Standardized Vocabularies: a Repeatable, Version-Aware Pipeline for OpenMRS-to-OMOP Harmonization

Polina Talapova, MD, PhD^{a,b}, Denys Kaduk, MD^{a,b,c}, Daryna Ivakhnenko, MD,^{a,b} Andrew S. Kanter, MD, MPH^d,

^a SciForce, Ukraine/Estonia

^b Thesaurus Health, Germany

^c V. N. Karazin Kharkiv National University, Ukraine

^d Columbia University, Department of Biomedical Informatics, USA

Background

The Columbia International eHealth Laboratory concept dictionary (CIEL) is a clinically rich, multilingual interface terminology used by OpenMRS (www.openmrs.org) and other health information systems, particularly in low-resource implementation settings where routine care, public-health programs, questionnaires, laboratory observations, medications, and locally specific clinical concepts must be captured at the point of care.¹ Because CIEL concepts are mapped to major international terminologies, they represent a valuable entry point for bringing OpenMRS-derived real-world data into standardized observational research. However, CIEL refreshes in the OHDSI Standardized Vocabularies have historically been difficult to execute as a predictable, auditable, and repeatable vocabulary lifecycle. This creates a semantic vulnerability: the same CIEL source code may resolve differently across vocabulary releases, producing mapping drift during ETL and reducing comparability across OMOP CDM instances. We therefore framed CIEL refresh not as a periodic vocabulary update, but as a community-governed semantic engineering problem linking OpenMRS clinical documentation to reproducible OHDSI real-world evidence.

Methods

We implemented the CIEL refresh as a complex OHDSI community contribution, specifically – release-oriented semantic engineering pathway (Fig. 1) composed of five controlled layers: upstream acquisition, deterministic mapping selection, bounded expert override, vocabulary-build execution, and audit metadata generation.² CIEL content was programmatically retrieved in JSON format from Open Concept Lab (OCL) through REST APIs and loaded into structured source tables, including `ciel_concepts`, `ciel_concept_names`, `ciel_mappings`, version tables.³ This eliminated manual file handling and made upstream acquisition repeatable across releases (and potentially across other data sources in the future). The canonical source-to-target mapping set was generated by `maps_for_load_stage.sql`, which applies ranked, rule-labeled selection logic to OCL-derived mappings from CIEL concepts. Only OCL map types with direct OMOP relevance – SAME-AS, NARROWER-THAN, and BROADER-THAN – were consumed as semantic signals for downstream relationship generation. Each selected mapping retained an explicit rule label and rank number, making the automated decision path inspectable rather than implicit.⁴

Rank 1 resolves mappings directly to OMOP Standard concepts using rule-specific SAME-AS, NARROWER-THAN, and BROADER-THAN logic; Rank 2 resolves otherwise eligible non-Standard targets through active OMOP “Maps to” or “Maps to value” crosswalks. Residual one-to-many cases with Standard candidates are explicitly assigned Rank 8 for manual review, whereas cases without a defensible Standard resolution are assigned Rank 9; only Rank 1-2 results proceed automatically to relationship staging. Clinically sensitive and ambiguous mappings were reviewed by four medical doctors using selected blinded assessments, candidate-remapping proposals, and consensus discussion with the mapping author; mappings were retained when reviewer consensus was achieved. For mappings finalized through this review pathway, acceptance required 100% agreement among participating reviewers and the mapping author.

Conflict resolution is rule- and context-specific rather than a simple global ordering of map types: vocabulary, source class, target domain, multiplicity, hierarchy, and selected clinical exceptions determine the preferred representation. For example, rule-defined BROADER-THAN mappings may be retained for radiology and laboratory concepts when they provide the more defensible OMOP representation. The rank therefore records a deterministic resolution pathway, not a probabilistic confidence score.

Manual curation was implemented as a bounded exception layer, not as a parallel mapping system. In `load_stage.sql`, automated relationships are first staged from Rank 1-2 results; for any CIEL source code represented in `concept_relationship_manual`, those automated staged relationships are removed before validated manual relationships are applied. Thus, reviewed manual content has explicit source-code-level authority over automated resolution.⁵ The full run sequence reset the development schema, rebuilt canonical mappings, applied validated manual exceptions, executed `load_stage.sql` and `genericupdate()`, and then ran structured QA. The current repository records execution times of approximately 1 min 43 s for `maps_for_load_stage.sql` and 3 min 31 s for `load_stage.sql` in the reference development runs.

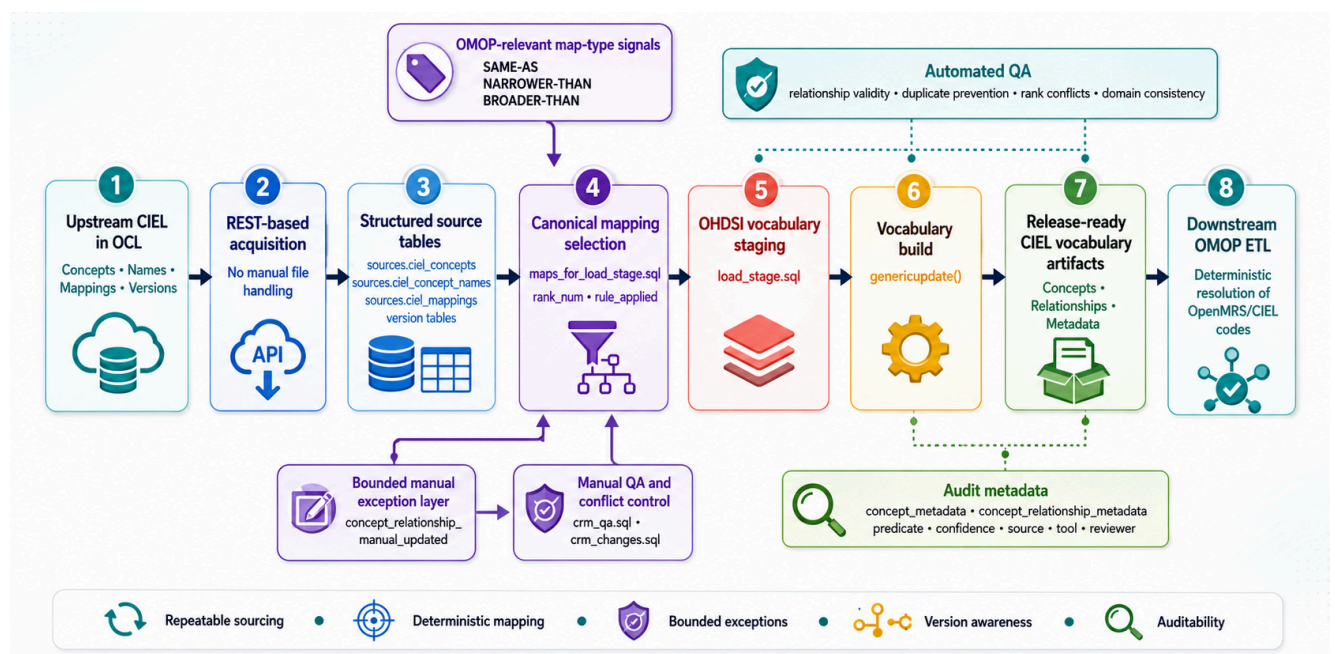


Figure 1. CIEL-to-OMOP refresh as an auditable semantic supply chain

Finally, metadata.sql populated concept_metadata and concept_relationship_metadata, including SSSOM predicate-like fields aligned with mapping directionality, confidence/provenance fields, and reviewer/tool/source attributes where available. This design made each refresh traceable from upstream CIEL/OCL release to OHDSI vocabulary staging and downstream OMOP ETL behavior.

Results

All reported percentages use CIEL source codes as the unit of analysis, with the denominator stated as either all represented CIEL codes or mapped CIEL codes. The refreshed CIEL-to-OMOP footprint covered 58,349 CIEL codes represented in the OHDSI Standardized Vocabularies, with most mapped to OMOP Standard concepts (57,455; 98.5% of all codes). Mappings were concentrated in clinically expected domains, primarily Condition (36,950; 64.3% of mapped codes), Drug (10,021; 17.4%), and Observation (5,974; 10.4%).⁶ The dominant pattern was 1-to-1 equivalence, with 49,228 mappings represented as direct semantic correspondence (85.7% of mapped codes). The pipeline also preserved complex clinical meaning when simple equivalence was insufficient. Post-coordinated OMOP representations were generated using combined “Maps to” and “Maps to value” logic for 343 mappings (0.6%), while 4,224 mappings required multi-target “Maps to” sets (7.4%). Ambiguous CIEL concepts with “AND/OR” naming patterns were deliberately generalized upward where needed, avoiding false precision and preserving analytic safety. This rule is methodologically important: the refresh does not simply maximize specificity, but selects the most defensible OMOP representation under source ambiguity. The refresh pathway also separated automatable semantic resolution from expert-governed exceptions. Automated rank-based selection carried the main mapping burden, while the manual layer remained thin, reviewable, conflict-checked, and release-delta oriented. This reduced the risk that manual curation would become an opaque alternative vocabulary, while still preserving expert control over clinically sensitive edge cases.

Limitations

The clinical review was targeted to selected edge cases rather than designed as a formal validation study of the full vocabulary. No predefined gold-standard sample, chance-corrected inter-rater reliability analysis, precision/recall assessment, or statistically evaluated comparison with the previous refresh process was performed. Accordingly, the reported 98.5% mapping coverage reflects coverage, not an estimate of semantic accuracy across all CIEL mappings.

Operational performance was only partially characterized. Although execution times were recorded for the principal SQL scripts, hardware requirements, total end-to-end runtime, curator effort, and turnaround-time reduction were not systematically benchmarked. Downstream effects on ETL outputs, cohort definitions, or phenotype stability were likewise not evaluated experimentally.

The implementation is reproducible within the OHDSI vocabulary build environment using open-source SQL, but it is not yet packaged as a self-contained container or standalone workflow; portability and ease of adoption outside that environment therefore remain to be assessed.

Conclusions

This work converts CIEL refresh from a high-touch, release-dependent mapping activity into a repeatable, OCL-sourced, version-aware OHDSI vocabulary contribution process. The contribution is methodological as much as operational: CIEL-to-OMOP harmonization is treated as an auditable semantic supply chain in which upstream terminology change, ranked mapping selection, expert exception handling, vocabulary build execution, QA, and downstream ETL stability are governed as one lifecycle.

For OpenMRS-derived real-world data, this enables more deterministic resolution of CIEL codes to OMOP Standard Concepts and reduces avoidable mapping drift across vocabulary releases. Importantly, the contribution demonstrated here is reproducibility and explainability of the resolution process; improved semantic accuracy relative to previous releases remains a hypothesis requiring independent validation. The same model can be generalized to other non-Standard vocabulary refreshes when a stable upstream distribution exists and community governance requires explainable mapping choices rather than opaque heuristics. By explicitly distinguishing automatable resolutions, expert-governed exceptions, and unresolved mappings, the pipeline provides a reusable governance pattern in which uncertainty is surfaced rather than silently converted into false precision. In addition, where CIEL to Standard vocabulary maps are not SAME-AS (1:1), and where there is research value to CIEL users (particularly in low- and middle-income countries), the OHDSI Africa WG is exploring nominating appropriate CIEL concepts from non-Standard to Standard. This enhancement would then be incorporated into the refresh process described in this paper.

References

1. Kanter A. CIEL: Bridging Terminology for the Global Community [Internet]. Open Concept Lab; 2022 Oct 19 [cited 2026 Jun 1]. Available from: <https://openconceptlab.org/2022/10/19/ciel-bridging-terminology-for-the-global-community/>
2. Observational Health Data Sciences and Informatics. Vocabulary-v5.0: CIEL [Internet]. GitHub; [cited 2026 Jun 4]. Available from: <https://github.com/OHDSI/Vocabulary-v5.0/tree/master/CIEL>
3. Observational Health Data Sciences and Informatics. Vocabulary-v5.0: CIEL load_source_tables [Internet]. GitHub; [cited 2026 Jun 1]. Available from: https://github.com/OHDSI/Vocabulary-v5.0/tree/master/CIEL/load_source_tables
4. Observational Health Data Sciences and Informatics. Vocabulary-v5.0: CIEL maps_for_load_stage README [Internet]. GitHub; [cited 2026 Jun 2]. Available from: https://github.com/OHDSI/Vocabulary-v5.0/blob/master/CIEL/maps_for_load_stage_README.md
5. Observational Health Data Sciences and Informatics. Vocabulary-v5.0: CIEL manual_work [Internet]. GitHub; [cited 2026 Jun 2]. Available from: https://github.com/OHDSI/Vocabulary-v5.0/tree/master/CIEL/manual_work
6. Observational Health Data Sciences and Informatics. Vocab. CIEL [Internet]. GitHub Wiki; [cited 2026 Jun 4]. Available from: <https://github.com/OHDSI/Vocabulary-v5.0/wiki/Vocab.-CIEL>