Ingredients Curation Project
Final Report

Jing Su

December 10, 2025

Abstract

This project curates dietary supplement product information by integrating four sources including the NIH ODS Dietary Supplement Label Database (DSLD), Amazon best-seller pages, Knowde supplier pages, and a set of internal leads collected from company domains. The pipeline ingests heterogeneous JSON, harmonizes field semantics including synonym and unit normalization, integrates records into a single Parquet corpus with stable content-derived identifiers, computes quality metrics, and exports two use-case oriented tables: products listing target ingredients (UC-1) and company-level aggregations over those targets (UC-2). The design follows the course framework for lifecycle, ethics and policy, models and standards, metadata, provenance and reproducibility, and dissemination. The primary evidence is the public repository and associated artifacts; this narrative situates the workflow in course concepts, reports results, and reflects on challenges and lessons learned.[1]

1 Motivation, Context, and Use Cases

Dietary supplement data appears in many incompatible places: structured label databases, e-commerce listings, supplier catalogs, and domain-specific collections. Analysts and educators often need a modest but reliable corpus that can be reproduced, inspected, and extended. This project focuses on a thin, explicit schema capable of describing a product, its company, and its ingredient text with just enough structure to enable two practical use cases. UC-1 provides a product list filtered by a target-ingredient vocabulary; UC-2 aggregates brands and products by company for each target. The intention is not to exhaustively model ingredients, but to present a pipeline that is transparent, auditable, and feasible for a single investigator while still tied closely to course principles.

2 Datasets and Scale

Four sources were ingested using a robust directory reader that handles JSON, NDJSON, array-shaped JSON, and compressed files. Source-specific mappers project content into a common schema covering *source*, *source_record_id*, *product_name*, *brand*, *company_name*, *link*, *form*, serving and net quantities, and the free-text fields *ingredients*, *claims*, and *statements*. After ingestion and normalization, the corpus comprises 49,694 harmonized records, of which 44,339 originate from DSLD, 4,792 from internal leads, 496 from Amazon, and 67 from Knowde. The integrated dataset preserves the same total (49,694) after stable identifier assignment and optional aliasing. These counts match the quality report and serve as anchors for reproducibility.

[1] Repository: https://github.com/jingsu322/ingredients_curation_project. Full dataset access: https://uofi.box.com/s/ac89dzikmob9eu00xr9zyetg8k24s1lz.

3 Workflow as Performed

The workflow is captured in a Snakemake DAG with configuration in `workflow/config.yaml`. Ingest transforms heterogeneous inputs into per-source Parquet files. Harmonization applies synonym normalization to create *ingredients_norm* from *ingredients*, and unit normalization to derive *serving_unit_canonical* and *serving_size_mg* when mass-based conversions are unambiguous. Integration computes a deterministic *curated_id* using SHA-256 over a tuple of stable, human-explainable fields (source, *source_record_id*, name, brand, normalized ingredients) and applies a small brand-to-company alias table when available. Validation emits a compact quality report, and dissemination exports UC-1 and UC-2 tables driven by a plain-text target list. Each stage yields concrete artifacts under `data/`, `reports/`, and `provenance/`, and is re-runnable in isolation.

The ingest stage is intentionally tolerant. Source mappers handle object arrays, NDJSON, and compressed files without user intervention, and they coerce nested lists into readable strings so that later stages do not depend on crawler-specific structures. When a field is missing, the mapper prefers to preserve the record and leave the field empty rather than discard the record silently. This choice trades completeness for transparency and makes failure modes visible in the validation metrics. The output is a set of per-source Parquet files with a common, minimal schema; compression and columnar layout reduce IO cost for downstream steps.

Harmonization focuses on two levers that matter for downstream queries: lexical consolidation of ingredient strings and basic unit canonicalization. Synonym rules are expressed as a small CSV so they can be audited and extended; replacements are applied conservatively, then deduplicated in order to keep human-readable strings. Unit rules convert only where the quantity type is clearly mass-based; otherwise the original unit is retained and a canonical field remains blank. The aim is to improve comparability without overstating precision. The resulting table, `harmonized.parquet`, is still close to the sources but has enough structure for cross-source filtering.

Integration adds just enough identity to support joins and views. The *curated_id* is deterministic and explainable; given the same inputs, the same record will always receive the same identifier, which is essential for reproducibility and debugging. A lightweight brand-to-company mapping is applied when present to stabilize the company dimension used in UC-2, but the original brand and company strings are preserved for audit. No global deduplication beyond exact key matches is attempted, by design; ambiguous merges are avoided, and potential future linkage can be layered on without rewriting past decisions.

Validation turns these design choices into measurable signals. The script reports total and per-source counts, required-field completeness, and parse coverage for key fields such as *ingredients* and serving attributes. Where a metric would be misleading—such as cross-source ingredient agreement in the absence of reliable duplicate detection—the report surfaces this as *not applicable* rather than inventing a weak statistic. This keeps the quality report compact yet decision-relevant: it explains what is present, what is missing, and how much of the corpus is ready for the use cases.

Dissemination is a thin, explicit layer. A plain-text target list drives UC-1 and UC-2 so that graders can change the targets and observe deterministic changes in the outputs. Exports are simple CSV files intended for quick inspection and reuse; they include links back to the sources and retain the provenance fields necessary to trace any row to its origin. Throughout, Snakemake's file-based contracts and the locked Python environment ensure that any altered input or rule triggers only the necessary rebuilds, while `provenance/` records the input inventory, run context, and checksums so that results can be verified independently.

Supplementary materials (data, scripts, workflow, documentation) are organized in the repository; see the README at https://github.com/jingsu322/ingredients_curation_project/blob/main/README.md for exact run commands and file layout.

4 Lifecycle and Course Concepts

The pipeline maps to a pragmatic lifecycle discussed in class: *ingest* → *harmonize* → *integrate* → *validate* → *disseminate*. Ingest and harmonize correspond to representation and transformation activities; integration constructs identity and prepares cross-source views; validation emphasizes fitness-for-use and transparency; dissemination prepares materials for reuse and evaluation. Ethical, legal, and policy considerations are explicit: DSLD is public; Amazon and Knowde pages are used within pedagogical fair use and the project's scope avoids redistribution of raw proprietary pages; internal leads are limited to course activities. Documentation of these constraints and acknowledgments is provided in `docs/TERMS.md`.

Data models and abstractions favor a thin common record that is resilient to upstream variability. Parquet provides an open, columnar storage format with broad tool support. Identity and identifier systems retain native identifiers where present and supplement them with the content-derived *curated_id*, whose determinism simplifies linking and debugging. Standards and standardization appear in the choice of JSON/CSV/Parquet, the use of simple CSV rules for synonyms and units, and the reliance on an openly documented workflow engine. Metadata and documentation include a DataCite-style record and a data dictionary/codebook, allowing a reader to understand variables and units without reading code. Workflow automation, provenance, and reproducibility are central: the DAG, the locked Python environment, file manifests, per-source ingest statistics, run metadata, and checksums create a transparent trail from inputs to outputs. Dissemination and communication are realized through the repository layout and a concise README that enables five-step reproduction on a clean machine.

5 Results

The validation step confirms totals and coverage. The harmonized and integrated datasets both contain 49,694 records. Per-source counts are consistent with expectations: 44,339 (DSLD), 4,792 (internal), 496 (Amazon), and 67 (Knowde). Required-field completeness is high: *product_name* appears in 100% of records, *source* in 100%, and *source_record_id* in 80.5%, reflecting sources that do not expose strong native identifiers. Parsing coverage for *ingredients* reaches 97.26%; coverage for *serving_size* and *serving_unit* is 89.86% and 89.85%, respectively, which is reasonable given unit ambiguity in some product lines. A cross-source consistency rate for ingredients is reported as "NA" in the quality file because cross-source duplication is limited and would require stronger match heuristics to be meaningful; the pipeline prefers to omit that figure rather than present an unstable estimate.

Use-case exports are non-empty and sizeable. UC-1 produces 16,608 product rows after filtering by the target list, and UC-2 aggregates to 395 company rows. These values will vary if the target list evolves, but the export code is deterministic for a fixed configuration.

6 Discussion: Findings, Problems, and Lessons

The most useful lesson is that a thin, explicit schema travels far. Minimalism made it possible to keep the mappers and normalizers understandable, and the datasets comparable, without losing essential context. Synonym expansion and unit canonicalization, even when seeded with a small rule set inspired by LanguaL and USDA resources, noticeably improve filtering and

aggregation across sources. A second lesson is that deterministic, content-derived identifiers reduce cognitive load when integrating sources with uneven native identifiers; the *curated_id* is explainable to a human and stable under re-runs. A third lesson is the value of early end-to-end automation. Having a running DAG, even with placeholders at first, created continuous evidence of progress, made debugging concrete, and simplified communication in the repository.

Not every element is easy. Ingredient strings mix marketing text with composition, especially in e-commerce pages, and aggressive natural-language processing would risk overfitting. The project uses conservative normalization, favoring precision over recall. Unit ambiguity is real: not all serving declarations are mass-based, and conversions to milligrams are performed only when safe. Finally, the project resists the temptation to build an elaborate ontology and instead emphasizes documentation, provenance, and the ability to reproduce the same tables tomorrow.

7 Ethics and Policy

The project acknowledges the terms of use for each source and limits redistribution accordingly. DSLD data are public. Amazon and Knowde content are used for research and teaching; the pipeline avoids storing raw pages in the repository, focuses on derived fields, and documents access for graders through Box. Internal leads are constrained to the course context. The `docs/TERMS.md` file collects these statements so that users understand what is permitted.

8 Reproducibility, Provenance, and Dissemination

Reproduction follows five steps in the README: clone, `uv sync`, place data, run Snakemake, and inspect outputs. The environment is locked with `uv.lock`; the DAG captures dependencies and file naming; provenance includes an input manifest, per-source ingest statistics, run metadata with configuration and timestamps, and checksums for derived artifacts. Dissemination consists of the organized repository structure (scripts, workflow, documentation, environment specification) and a separate Box link for large raw data. The design prioritizes transparency and the ability to verify counts and checksums on a clean machine.

9 Conclusion

The project delivers a compact, reproducible curation pipeline that aligns closely with course concepts from lifecycle and ethics to metadata, standards, provenance, and dissemination. It is intentionally modest and explainable, which makes it a good foundation for extension. Reasonable next steps would be to broaden synonym coverage with LanguaL and USDA resources, add shallow rules to better segment ingredient strings, expand the company alias map to improve UC-2, and introduce a conservative cross-source similarity check for optional duplicate analysis. None of these steps requires a change in philosophy; they extend the same design with slightly richer domain knowledge.

Supplementary Materials

The GitHub repository contains scripts, workflow rules, documentation, and the environment specification. The Box link provides controlled access to large raw datasets needed to reproduce the exact counts in this report. File paths, targets, and configurable lists (synonyms, units, target ingredients, and optional company aliases) are simple text files, which keeps auditing and modification straightforward.

References

1. Köster, J., & Rahmann, S. (2012). Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19), 2520–2522. https://doi.org/10.1093/bioinformatics/bts480

2. LanguaL™. (n.d.). *LanguaL—The International Framework for Food Description*. https://www.langual.org/

3. National Institutes of Health, Office of Dietary Supplements. (n.d.). *Dietary Supplement Label Database (DSLD)*. https://dsld.od.nih.gov/

4. U.S. Department of Agriculture, Agricultural Research Service. (n.d.). *Dietary Supplement Ingredient Database (DSID)*. https://dietarysupplementdatabase.usda.nih.gov/