Progress Report — Ingredient-Centric Curation for Dietary Supplements

Overview and feedback incorporated

This project curates ingredient-centric knowledge for dietary supplements: given an ingredient (including synonyms), identify products that contain it and the companies that market them. The design is anchored in three lightweight use cases—UC-1 (Ingredient→Products), UC-2 (Ingredient→Companies), and UC-3 (Quality & Coverage). Proposal feedback has been integrated: synonym expansion will consult the LanguaL Thesaurus and the USDA ARS Dietary Supplement Ingredient Database, improving recall while keeping terminology controlled (USDA ARS, n.d.; LanguaL, n.d.). The workflow remains automation-first (Snakemake, pinned Python) to preserve a realistic scope (Mölder et al., 2021).

Status against the preliminary plan

Plan. Completed. A minimal abstraction spans `Ingredient`, `Product`, and `Company` with `USES` and `MARKETS` relations. The field codebook (`docs/codebook.csv`) and a Lifecycle & Compliance plan (`docs/plan.md`) are in place, aligning with USGS and DCC lifecycle guidance (USGS, n.d.; Higgins, 2008).

Acquire. Completed at classroom scale. **Amazon access has been successfully resolved** and sample data are collected. Samples for **DSLD** (on-market, 2023–2025), **Knowde**, and a de-identified **internal leads** subset are staged under `data/raw/`. Full datasets are archived in Box for evaluation: https://uofi.box.com/s/ac89dzikmob9eu00xr9zyetg8k24s1lz. Acquisition remains compliant with site terms and robots through a samples-only policy and manifests for later regeneration (Koster & Pebesma, 2022).

Process (integration & cleaning). In progress. Seed tables for **synonyms** and **units** (`rules/synonyms.csv`, `rules/units.csv`) support normalization to canonical mass units (mg) and cautious handling of IU. A profiling scaffold (`reports/profiling.csv`) summarizes missingness and basic ranges.

Integrate (identity & identifiers). In progress. Ingredient names are anchored to DSLD/INCI where possible; brand→company uses a small alias list and domain cues. Source identifiers (e.g., `asin`, `upcSku`, URLs) are retained, while integrated entities receive **UUIDv7** or content hashes, strengthening identity and traceability.

Validate. Partially implemented. A JSON Schema (`metadata/dataset.schema.json`) checks curated tables. A compact **quality report** is planned to track coverage, parsing success, and cross-source consistency.

Preserve & Disseminate. On track. The repository is public—https://github.com/jingsu322/ingredients_curation_project—using `uv` for a pinned environment, with DataCite and `schema.org/Dataset` metadata underway. Final artifacts will include CSV/Parquet (optionally SQLite), checksums, and a "Reproduce in 5 steps."

Evidence of progress (artifacts)

- Public GitHub repository with code, seeds, plan, codebook, and sample data: https://github.com/jingsu322/ingredients_curation_project
- Box archive for full datasets: https://uofi.box.com/s/ac89dzikmob9eu00xr9zyetg8k24s1lz
- Workflow scaffolding (`Snakefile` stub, `workflow/config.yaml`) and a provenance layout (`provenance/`) for manifests and checksums.
  These artifacts demonstrate concrete movement toward UC-1/UC-2 exports and the UC-3 quality summary.

Challenges and scope adjustments

- **Heterogeneity and cleaning.** Cross-source variation in naming, serving forms, and units requires careful normalization. This project introduces explicit **synonym** and **unit** dictionaries to stabilize integration while keeping rules auditable.
- **Identity and identifiers.** Ambiguity in brand/company mapping is common. The approach emphasizes transparent identity: preserve source IDs, assign **UUIDv7** for integrated entities, and flag uncertain links for later review.
- **Standards and standardization.** To avoid ad-hoc drift, the project uses JSON Schema for structure, **DataCite** for dataset-level metadata, and a `schema.org/Dataset` snippet for discovery (DataCite Metadata Working Group, 2021; Schema.org, n.d.).

- **Legal and policy constraints.** Redistribution follows a minimal, **samples-only** model with robots-aware acquisition, avoiding bulk page content and copyrighted media (Koster & Pebesma, 2022).
  No expansion of scope is required at this time; the successful Amazon access reduces risk. The project remains feasible within the remaining schedule.

Next steps (clear, actionable)

1. **Finish normalization & cleaning (M6).**
   - Implement parsing of serving size, net quantity, and per-ingredient amounts; apply `rules/units.csv` and `rules/synonyms.csv`.
   - Emit `reports/profiling.csv` and `logs/parse_failures.csv` for transparency.
2. **Complete integration & identity (M6, M9).**
   - Consolidate DSLD, Amazon, Knowde, and internal leads into a single integrated table.
   - Assign **UUIDv7** / hashes; preserve source IDs; write out `curation/company_links.csv` with confidence notes.
3. **Validation & quality (M6, M8).**
   - Finalize `metadata/dataset.schema.json`; add structural checks to the DAG.
   - Produce `reports/quality_report.csv` covering coverage, parsing success, and simple consistency tests.
4. **Export views for UC-1 and UC-2.**
   - Generate `data/curated/uc1_products.csv` and `uc2_companies.csv`; include example queries and a small inspection notebook.
5. **Standards & metadata (M11, M8).**
   - Complete **DataCite** fields and embed a `schema.org/Dataset` JSON-LD snippet in the README.
   - Expand `docs/codebook.csv` with concrete examples and cite sources for synonym lines (LanguaL/USDA ARS).
6. **Reproducibility & dissemination (M12, M15).**
   - Finish the Snakemake DAG; capture run parameters, commit hashes, and

checksums in `provenance/`.

- Tag a GitHub release and, if needed, publish a mirrored Zip for classroom submission.

References

DataCite Metadata Working Group. (2021). *DataCite metadata schema documentation for the publication and citation of research data* (Version 4.4). DataCite. https://schema.datacite.org/

Higgins, S. (2008). The DCC Curation Lifecycle Model. *International Journal of Digital Curation, 3*(1), 134–140. https://doi.org/10.2218/ijdc.v3i1.48

Koster, M., & Pebesma, E. (2022). *Robots Exclusion Protocol* (RFC 9309). IETF. https://doi.org/10.17487/RFC9309

LanguaL. (n.d.). *LanguaL™—The international framework for food description*. https://www.langual.org/

Mölder, F., Jablonski, K. P., Letcher, B., et al. (2021). Sustainable data analysis with Snakemake. *F1000Research, 10*, 33. https://doi.org/10.12688/f1000research.29032.2

Schema.org. (n.d.). *Schema.org vocabulary*. https://schema.org/

U.S. Geological Survey (USGS). (n.d.). *USGS Science Data Lifecycle*. https://www.usgs.gov/

USDA ARS. (n.d.). *Dietary Supplement Ingredient Database*. https://dietarysupplementdatabase.usda.nih.gov/