CORE

Lab 7: Build a Task-Specific Data Pipeline

Find, evaluate, curate, harmonize, and merge datasets with natural language

Chapter Chapter 7	**Level** Intermediate	**Time** ~2.5-4 hr
Mode Individual or pair	**Technology** Browser spreadsheet	**Deliverable** Map mission brief

Why This Lab Matters

Most maps inherit their truth claims from data long before design begins. This lab treats dataset discovery, reliability assessment, cleaning, and merging as design work. Your final product is not a polished map; it is a traceable task-specific dataset and the pipeline record that explains why it deserves to be mapped.

> **Essential question**
>
> What cartographic decision should you be able to defend by the end of this lab?

Learning Objectives

- Write dataset search prompts that combine technical requirements with map purpose.
- Compare keyword search and semantic search for the same mapping task.
- Evaluate candidate datasets for recency, completeness, authority, granularity, bias, coverage, license, and field suitability.
- Create a reliability ledger that preserves provenance and human judgment.
- Design a merge plan across at least three datasets using a common geography, join key, or task-specific schema.

At a glance

Inputs: Lab scenario, prompts, data notes, or map concept supplied in the activity.
Outputs: Map mission brief; Candidate inventory with at least six datasets; Reliability and fitness review.
Evaluation: Clear purpose, defensible cartographic judgment, concise prompting, and evidence of human verification where needed.

LAB WORK

Setup: Pick a Map Mission ~10 min

Choose a mission that requires more than one dataset. The final merged table should have 12-30 rows so the exercise stays manageable.

- **Heat resilience**: cooling centers, older adult population, transit access, and heat-risk indicators.
- **Safe routes to school**: schools, crashes, sidewalks or bike lanes, and demographic context.
- **Food access**: grocery stores, population, vehicle access, and transit stops.
- **Flood readiness**: flood zones, shelters, social vulnerability, and road access.
- **Invasive species response**: observation records, protected lands, roads or trails, and habitat context.

```
Map mission brief
Topic:
Place or study area:
Audience:
Decision or understanding the map should support:
Spatial unit for the final dataset, if known:
Minimum time period:
Required fields or variables:
Sensitive data or ethical guardrails:
What the final map must not imply:
```

Setup deliverable

Create a Data Pipeline Dossier. Every prompt, source URL, rejected dataset, cleaning rule, and human decision should be saved there.

PROCEDURE

Part 1 — Run a Keyword Search and a Semantic Search

~25 min

Start by showing yourself how much the wording matters. Use a normal web search or data portal search first, then use a structured LLM prompt.

```
Keyword-style search phrase:
[Example: heat risk data Phoenix]

Semantic dataset search prompt:
Locate open datasets for [PLACE] that can support a map about [PURPOSE] for [AUDIENCE]. I need
datasets that can be combined into a task-specific table at [GEOGRAPHY OR UNIT]. Return
candidate datasets with:
- dataset name
- source organization
- URL
- format or access method
- geometry or spatial fields
- temporal coverage
- key attributes
- license or reuse notes
- likely join key or common geography
- reliability score from 1 to 5
- one reason the dataset might be inappropriate
Prioritize sources from government, academic, tribal, nonprofit, or clearly documented
institutional sources.
```

Open every promising link in the browser. Do not trust a link only because the LLM listed it.

Mini-deliverable

A candidate inventory with at least six datasets, including at least two you will probably reject.

PROCEDURE

Part 2 — Reliability and Fitness Review ~30 min

Use the LLM as a reviewer, but make the final judgment yourself. Paste only dataset descriptions, metadata excerpts, field lists, or source summaries that you can verify in the browser.

```
Act as a dataset reliability reviewer for a map project. Evaluate the candidate datasets below
for this map mission. Score each from 1 to 5 on:
- recency
- completeness
- authority
- granularity
- bias and coverage
- license clarity
- field suitability
- merge usefulness

For each dataset, identify one claim it can support, one claim it cannot support, one likely
source of bias, and one verification task a human must complete.

Map mission:
[PASTE]

Candidate datasets:
[PASTE INVENTORY OR METADATA EXCERPTS]
```

Dataset	Keep / Maybe / Reject	Best Use	Main Risk	Human Verification Needed
Example: Short note	Example: Short note	Example: Short note	Example: Could overstate certainty	Example: Short note
Example: Short note	Example: Short note	Example: Short note	Example: Could overstate certainty	Example: Short note
Example: Short note	Example: Short note	Example: Short note	Example: Could overstate certainty	Example: Short note

Starter examples only; expand this in your own notes or submission document.

Mini-deliverable

Select three datasets to keep. Write one paragraph explaining why each belongs in the final task-specific dataset.

PROCEDURE

Part 3 — Assign Dataset Roles ~20 min

A merged dataset should not be a pile of everything you found. Each source should have a job.

```
Help me assign roles to my kept datasets. Use these categories if useful:
- anchor dataset: defines the rows or map features
- context dataset: adds background or explanatory variables
- risk or exposure dataset: adds vulnerability, hazard, or urgency
- access dataset: adds services, routes, facilities, or distance context
- verification dataset: checks or qualifies another source

For each dataset, explain why it belongs, what field or concept it contributes, and what would
happen if I removed it.
```

Choose one dataset to be the anchor. The anchor usually determines the final rows, such as counties, census tracts, schools, facilities, shelters, or observation records.

Reminder

Fit matters more than fidelity alone. A precise dataset can still be wrong for your map's purpose.

PROCEDURE

Part 4 — Build a Merge Plan Before Merging ~30 min

Create a merge plan that names the final table structure before you touch the spreadsheet.

```
Design a merge plan for these three kept datasets. Return:
1. final table grain: one row per [what]
2. proposed join key or matching method
3. fields to keep from each dataset
4. fields to rename for clarity
5. fields to aggregate, summarize, or classify
6. rows likely to fail matching
7. ethical or interpretive risks introduced by the merge
8. a final canonical schema with field name, definition, source dataset, and cleaning rule

Datasets:
[PASTE ROLES, FIELD LISTS, AND SOURCE NOTES]
```

Copy the canonical schema into your dossier. Do not merge until the schema is clear.

Mini-deliverable

A canonical schema with 8-15 final fields. Each field must have a source and a definition.

PROCEDURE

Part 5 — Create the Task-Specific Dataset ~45-75 min

Build a small merged dataset in a browser spreadsheet. Use actual records if manageable. If a dataset is large, select a defensible subset such as one county, ten facilities, one neighborhood, or the most recent year.

```
Act as a data engineering assistant. Given my canonical schema and source fields, create step-by-step browser spreadsheet instructions to build the merged table. Include formulas where appropriate for XLOOKUP, VLOOKUP, FILTER, text cleanup, date parsing, category recoding, or simple scoring. Explain how to check unmatched rows and duplicate keys.

Canonical schema:
[PASTE]

Source field notes:
[PASTE]

Spreadsheet tool:
[Google Sheets / Excel web / other]
```

- Create one tab for each source dataset or source excerpt.
- Create one tab called `canonical_dataset`.
- Use the anchor dataset to define the rows.
- Bring in fields from the other datasets using a join key, shared geography, manual crosswalk, or clearly documented lookup.
- If you cannot merge automatically, create a manual crosswalk tab and document the rules.

Mini-deliverable

A merged task-specific dataset with 12-30 rows and 8-15 fields, plus a source column or source notes for each derived field.

PROCEDURE

Part 6 — Audit the Merged Dataset ~30 min

Now critique the dataset you created. If you cannot upload a file to a free LLM, paste the first 12 rows, the field list, and summary notes instead.

```
Audit this merged task-specific dataset for map readiness. Evaluate:
- missing values
- duplicate keys
- inconsistent categories
- questionable joins
- temporal mismatches
- spatial mismatches
- fields that imply more certainty than the sources support
- sensitive or ethically risky fields
- whether the dataset can support the map mission

Return a problem log with severity, affected fields, likely cause, and recommended fix. Then
give a publishability score from 1 to 5, but explain the score as a judgment, not a verdict.
```

Fix at least three issues or explain why you are choosing not to fix them.

Human judgment

The LLM can identify possible problems. You decide which problems matter for the map's purpose.

PROCEDURE

Part 7 — Write the Data Lineage Note ~25 min

Close the pipeline with a lineage note that a reviewer, instructor, or future collaborator could inspect.

```
Using my data pipeline dossier, help me draft a data lineage note with these sections:
1. map mission
2. datasets kept and why
3. datasets rejected and why
4. merge grain and join logic
5. cleaning and recoding decisions
6. reliability concerns that remain
7. claims the final dataset can support
8. claims the final dataset cannot support
9. attribution and license notes
10. next verification step before mapping

Keep it concise, transparent, and suitable for inclusion with a student project or professional
prototype.
```

Core deliverable

A data pipeline dossier and merged task-specific dataset that another person could inspect and partially reproduce.

DELIVERABLE

Submission ~15 min

- Map mission brief
- Candidate inventory with at least six datasets
- Reliability and fitness review
- Three kept datasets and at least two rejected datasets
- Dataset role assignment
- Canonical schema
- Merged task-specific dataset with 12-30 rows
- Problem log and fixes
- Data lineage note
- A 300-500 word reflection on how natural language changed what counted as data engineering