Trial scope: CA UCC Scraper → Sheet (4.0 h max)

SCRAPERS TARGET: https://bizfileonline.sos.ca.gov/search/ucc
SHEET TO USE: UCC Filings - California
See tab one of this sheet for example of results: ucc_filings_oregon

Target: CA SOS UCC search page. Use the search box only.
Output schema: `filing_type, debtor, file_number, secured_party, status, filing_date, lapse_date, query_prefix, scraped_at`.

Block B0 — Compliance & setup (0.3 h cap)

Tasks

- Read robots.txt and Terms. Capture 2 screenshots that show allowance or limits for automated access.
- Create a local project folder with `results/`, `logs/`, `selectors/`, `state/`.

Deliverables

- `compliance_memo.pdf` (1 page, bullet points + screenshots).
- Folder structure created.

Gate G1: If automation is clearly disallowed, **stop** and deliver B0 only.

Block B1 — Selector proof + stub harvest (1.0 h cap)

Tasks

- Open the UCC search page. Identify selectors for: search input, submit, results table rows, "Next" control, and the on-page **Results: N** text.
- Run a single search for `aaaa`. Wait for table. Harvest **first page only** into memory.

- Normalize fields to the schema above. Parse dates to `YYYY-MM-DD`. Set `query_prefix="aaaa"`, `scraped_at` UTC ISO.

Deliverables

- `selectors/01_input.png`, `selectors/02_table.png`.
- `results/CA_UCC_stub.csv` (≥10 rows, header present).
- `logs/run_stub.txt` with timestamp, detected N, row count on page 1.

Acceptance

- 100% rows have `file_number`.
- CSV headers exactly match the schema order.

Gate G2: Proceed only if a CSV was produced.

Block B2 — Pagination for `aaaa` (1.0 h cap)

Tasks

- Add reliable “Next page” loop. Stop when Next is disabled or page repeats.
- Harvest **all pages** for `aaaa`.
- Produce both `CSV` and `JSONL`. Create `manifest.json` with file names, row counts, and sha256 checksums.
- Write a **dedupe by `file_number`** within this `aaaa` run.

Deliverables

- `results/CA_UCC_aaaa.csv`, `results/CA_UCC_aaaa.jsonl`.
- `results/manifest.json`.

- `logs/run_aaaa.txt` with per-page counts and final total.

Acceptance

- CSV row count equals the on-page **Results: N** for `aaaa`.
- No duplicate `file_number` in output.

Block B3 — Sheet landing (0.7 h cap)

Tasks

- Create or use a Google Sheet with tabs: `Raw_UCC_CA` and `Sync_Log`.
- Place `results/CA_UCC_aaaa.csv` into a Drive folder `CA_UCC_Drops/tmp/`.
- **Import once** (manual import acceptable in trial) to `Raw_UCC_CA`.
- Log counts in `Sync_Log` with timestamp, source file name, inserted rows.

Deliverables

- Link to Sheet.
- Screenshot of `Raw_UCC_CA` showing headers + first 20 rows.
- Screenshot of `Sync_Log` row with counts.

Acceptance

- `Raw_UCC_CA` row count equals the `aaaa.csv` row count.
- Header names and order match the schema.

Block B4 — Proof pack (0.5 h cap)

Tasks

- Produce a short README with run steps and known limits.
- Zip evidence.

Deliverables

- `README.md` (how to run again, how to import).
- `evidence.zip` containing: `results/`, `logs/`, `selectors/`, and screenshots of the Sheet.

Stretch (only if time remains inside block caps)

- Detect “>1,000 results” for `aaa` and log **“branch required”** in `logs/run_aaa.txt`.
- Write `state/state.json` with `{query, page_number, last_file_number}` after each page.

What you must not do in the trial

- No proxies, no CAPTCHA solving, no login.
- No full alphabet crawl. Only `aaaa` end-to-end.
- No complex prefix engine beyond logging “branch required”.

Final acceptance for the 4-hour trial

- `CA_UCC_aaaa.csv` and `.jsonl` produced, checksumed, and deduped by `file_number`.
- `Raw_UCC_CA` populated with **row-count parity** to `CA_UCC_aaaa.csv`.
- Dates in ISO. Headers exact.
- Evidence pack delivered.
- If disallowed by TOS/robots, a complete **compliance memo** and selector screenshots are delivered instead.

Contractor report format (paste on completion)

- **Time used per block (B0–B4):**
- **Results:** N for `aaaa`, CSV rows, Sheet rows.
- **Screenshots:** links/filenames
- **Issues/risks:**
- **Next 4-hour plan:** pagination hardening, resume state, `abba` run, importer automation.

Inputs we provide now

- Google Sheet link (with tabs created or permission to create).
- Drive folder `CA_UCC_Drops/tmp/` write access.

This is the complete MVP trial scope.