Phishing Takedown Orchestrator – Week 1, Day 3 Report
Date: Sunday, August 17, 2025
Location: Pune, Maharashtra, India

1) Objectives
- Implement Discovery MVP to convert a CSV of candidate URLs into validated JSONL Findings.
- Integrate the discovery flow with the CLI for repeatable runs.
- Add a minimal enhancement to accept bare domains by assuming an HTTPS scheme.

2) Files Added/Updated
- examples/seed_urls.csv — comprehensive seed data for discovery.
- examples/sample_config.yaml — example config with safe defaults.
- src/app/config.py — configuration loader (no changes today, referenced by CLI).
- src/app/models.py — Pydantic models (Finding, etc.; no changes today).
- src/discovery/feeds.py — CSV ingestion, normalization, validation, JSONL writing.
- src/cli/main.py — CLI commands: show-config and discover.
- tests/test_config_and_models.py — confirms config defaults and model behavior.
- tests/test_discovery.py — validates discovery against the CSV input.
- src/phish_takedown_orchestrator.egg-info/* — packaging metadata (present in repo).

3) Discovery Flow Overview (Day 3)
- Input: examples/seed_urls.csv with columns url,source.
- Processing:
  - Trims whitespace and ignores empty URL rows.
  - Normalizes a dedupe key using a simple domain extraction from the URL.
  - Validates each entry with the Finding Pydantic model.
  - Enhancement added: bare domains (no scheme) are automatically treated as https://domain before validation.
- Output: findings.jsonl (each line is a validated Finding record serialized to JSON).

4) Key Code Notes
- src/discovery/feeds.py
  - load_findings_from_csv(csv_path):
    - Reads CSV rows with DictReader, coerces bare domains to HTTPS, deduplicates by normalized domain + source, and creates Finding objects with timezone-aware timestamps.
  - write_findings_jsonl(findings, out_path):
    - Serializes each Finding to a JSON line, converting URL and datetime fields to strings to ensure valid JSON output.
  - _normalize_domain(url):
    - Extracts and lowercases the host portion for dedupe keys.
  - Small addition on Day 3:
    - An ensure-scheme step is applied so bare domains validate as URLs.

- src/cli/main.py
  - show-config:
    - Prints the effective configuration JSON, confirming paths and defaults.
  - discover:
    - Invokes the discovery loader and writer, reports number of findings, and writes the JSONL file.

- src/app/models.py

  - Finding:
    - url (HttpUrl), discovered_at (UTC), source (default local_csv), optional risk_score.
  - Evidence, NetworkMeta, Parties, Report, Outcome:
    - Structured for later pipeline stages; unchanged today.

5) How to Run (Commands Used)

- Navigate to project:
  - cd "/Users/harshilbuch/Downloads/Harshil/Projects/Phishing Takedown Orchestrator"
- Activate environment:
  - source venv/bin/activate
- Run tests:
  - python -m pytest tests/ -v
- CLI help:
  - python -m src.cli.main --help
- Show config:
  - python -m src.cli.main show-config
- Process CSV:
  - python -m src.cli.main discover --csv examples/seed_urls.csv --out findings.jsonl
- Inspect results:
  - head -10 findings.jsonl
  - wc -l findings.jsonl
  - grep -E '"url": "https://' findings.jsonl | head -5

6) Day 3 Deliverables

- Functional discovery pipeline from CSV→JSONL with validation and deduplication.
- CLI integration for repeatable execution.
- Enhancement to accept bare domains by assuming HTTPS, increasing valid coverage of inputs.
- Tests confirming configuration defaults and discovery behavior.

7) Guardrails and Assumptions

- No external network calls; local-only processing.
- Reporting remains disabled by default.
- Artifacts (.outbox, .runs, artifacts) are created automatically and should remain git-ignored.
- Bare domains default to https://; this can be parameterized later if needed.

8) What's Next (Preview)

- Optional Day 3+ hardening (deferred unless requested):
  - Rejects log for invalid rows.
  - eTLD+1 normalization for stronger dedupe.
  - IDN/punycode normalization for dedupe while preserving original for reports.
  - Streaming writes for very large CSVs.
- Day 4 target:
  - Evidence capture MVP using Playwright with strict safety settings, writing captured artifacts and hashes.

Prepared by: HARSHIL AMIT BUCH
Project: Phishing Takedown Orchestrator
Mentor/Guide: AI-assisted development workflow