Patent Processing Pipeline Manual

Step 0: Download Patents (PatentsView API or Uploads)

- **Quick Download (7 days)** – Downloads patents from the most recent 7-day period available in the PatentsView database. Uses default settings for immediate results.
- **Configure Download** – Opens advanced options to customize your download:
    - Smart Mode: Automatically finds recent patents going back a set number of days
    - Manual Mode: Specify exact start and end dates
    - Set maximum results (100 to 50,000 patents)
    - Choose runner: PatentsView API or Alternate Extract
    - **Start Download** – Runs the selected configuration immediately after you set mode, dates, and max results.
- **Upload CSV/XLSX** – Alternative to API download. Upload your own patent data file if you have patents from another source.

Step 1: Integrate Existing Data

- **Run Integration** – Compares downloaded patents against your existing database to filter out duplicates and identify new records. Prevents reprocessing people you've already contacted.
- **Review Potential Matches (optional)** – Manual verification for borderline cases:
    - Same Name • New Address: People with identical names but different locations (may have moved)
    - Score 10-24 Matches: People with partial name/location matches requiring human judgment
- **Options for each match**
    - Mark as Existing – Skip this person (already in your system)
    - Enrich From New Address – Process using the new location data
    - Update In Database – Update your existing record with the new address fields

- Skip for Now – Leave undecided
- **New: Update All (Bulk)** – From the summary toolbar, open the bulk update overlay to apply address updates to multiple "moved" matches in one action.

Step 2: Data Enrichment (PeopleDataLabs API)

- **Test Enrichment (5 people)** – Processes only 5 records to verify API connection and preview results before running larger batches.
- **Full Enrichment** – Enriches all new people with contact information including emails, LinkedIn profiles, job titles, and company data.
- **Full Enrichment Express** – Same as Full Enrichment but skips people who previously failed to enrich, making it faster and avoiding unnecessary API costs. Recommended for regular use.
- **Rebuild CSVs (no API calls)** – Rebuilds export files from the current dataset without making any new enrichment requests.

CSV Export Options

Full CSV Exports

- **Current**
  - Contents: "New & Existing" results **plus** all people filtered out in Step 1 for already being in the SQL `existing_people` table.
  - Columns: Includes every column from `enriched_people` **and** any additional columns that exist only in `existing_people`.
  - Provenance: Adds a column indicating the source table for each row (`source_table` = `enriched_people` or `existing_people`).
  - Purpose: A single, simplified CSV replacing the need to separately handle "New & Existing" and Step-1 filtered existing records.
- **New Enrichments**
  - Contents: Only newly enriched records from the latest Step 2 run (unchanged).
- **All**

- Contents: The full `enriched_people` SQL table exported as CSV (complete historical set).

Formatted Exports

- **Current/New/All (Formatted)** – Same datasets as above but normalized for your business format.
- **New & Existing (Formatted)** – Kept for compatibility; the **Current** export already includes this logic in the non-formatted version.

Contacts and Addresses Exports

- **Contacts (Current/New)** – CSVs containing all available email addresses associated with each person in the respective set (Current or New).
- **Addresses (Current/New)** – CSVs containing all known addresses associated with each person in the respective set (Current or New).

Advanced settings

Use these options in advanced settings at: http://<url>/dev. This mode is intended for faster test cycles, alternate runners, and developer-only utilities. These features are not required for normal processing.

Step 0 (Advanced)

- **Quick Download** (7 days) – One-click quick run using default Smart Mode and a 7-day window. Intended for rapid testing.
- **Runner selection** – Switch between PatentsView API and Alternate Extract for comparison or legacy extraction testing.
- **Upload XML** – Upload raw USPTO XML files for direct extraction.

Step 1 (Advanced)

- Dev Integration – Runs integration using a developer-set Issue Date Cutoff to test filtering behavior on recent data.
- Issue Date Cutoff – A timestamp used to include/exclude patents during integration in dev mode.
- Enrich All – Forces a full enrichment trigger directly after dev integration for end-to-end testing.

Step 2 (Advanced)

- Zaba Enrich – Runs an alternate enrichment path for internal testing workflows.
- Rebuild Zaba CSVs (no scraping) – Rebuilds CSV outputs for the alternate path without performing new network requests.