Pawzz Community Verified Directory

Data Workflow and Verification System Design Submission

Answer to Question 1: Tools and System Design for Community Verified Directory

1. Project Objective

The objective is to design a scalable, accurate and auditable community verified directory for unorganised and local animal care businesses including veterinary clinics, NGOs, shelters, feeders, boarding centres, trainers, groomers and pet shops. The system must ensure data accuracy, reduce duplication, standardise inconsistent inputs and minimise dependency on the founder.

2. Technology Architecture

Data Collection will be managed through Google Forms for community submissions, structured intern research and controlled scraping from public platforms. Data Storage will use PostgreSQL as the primary backend with Airtable for workflow tracking. Data Cleaning will be handled using Python with Pandas, regular expressions and fuzzy matching logic. Automations such as Zapier or Make will sync submissions into the database. Power BI will be used for monitoring operational performance and quality.

3. Data Pipeline Workflow

The workflow will follow defined stages: Submitted, Needs Fix, Ready for Review, Verified, Published and Flagged. Raw data enters at Submitted stage. Incomplete or invalid data moves to Needs Fix. Cleaned entries move to Ready for Review. Verified listings have confirmed proof and assigned confidence scores. Published listings are visible publicly. Flagged entries require re evaluation due to disputes or inconsistencies.

4. Required Fields and Proof Standards

Mandatory fields include business name, structured address, phone number, category, source channel, geo coordinates, proof type and verification status. At least two independent proofs are required such as Google Maps listing, digital presence, reviews or direct phone confirmation.

5. Handling Unstructured Inputs

WhatsApp submissions will be exported and structured using automated extraction techniques. Screenshots will be processed using OCR before cleaning. Standardisation rules will ensure consistent formatting of phone numbers, addresses and names before database insertion.

6. Duplicate Detection System

Duplicate prevention will use exact matching on phone numbers and geo coordinates combined with fuzzy matching on business names and addresses. A weighted confidence score will identify probable duplicates before publication.

Answer to Question 2: Red Flags, Verification and Duplicate Prevention

7. Verification System

Verification includes phone confirmation, Google listing validation and documentation of supporting proof. A structured scoring model assigns credibility based on direct confirmation, digital footprint and reviews.

8. Red Flag Detection

Listings are flagged if phone numbers are unreachable, addresses are vague, digital presence is absent, reviews are suspicious or conflicting details are identified across sources.

9. Conflict Resolution

In case of conflicting information such as multiple phone numbers or addresses, cross verification is conducted using digital listings and direct calls. All verified information is updated after confirmation and documented in the audit log.

10. Audit Trail and Governance

A structured change log captures listing identifier, field changes, previous value, updated value, verifier name, timestamp and reason for modification. This ensures accountability and traceability.

11. Escalation Framework

Listings are rejected if fraud or false information is confirmed. Minor inconsistencies trigger re verification. Listings with limited proof may be published with caution notes if supported by credible community references.

12. Dashboard and KPIs

Operational dashboards track listing status distribution, verification turnaround time, rejection rate, duplicate rate and backlog size. Quality dashboards monitor proof completeness and dispute frequency.

13. Scalability Roadmap

The system will evolve from spreadsheet based management to a structured SQL backend with automation and eventually machine learning based duplicate detection and advanced confidence scoring models.