Matching and Conflation of Open Government Data with OpenStreetMap Data

Matching and Conflation of AllThePlaces Data with OpenStreetMap Data

Graduate

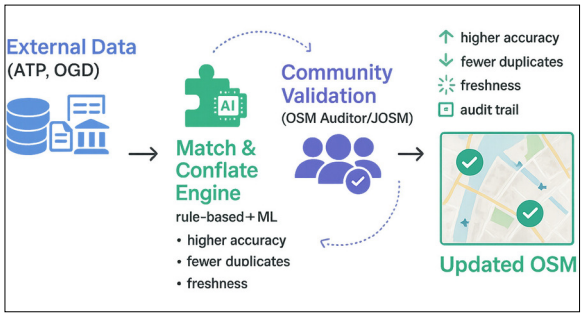
Claudio Bertozzi

Introduction: Points of interest (POIs) are important, but they change relatively quickly. This is why OpenStreetMap (OSM), the largest crowdsourcing geodata project, reconciles with publicly available, authoritative brand and retailer feeds, thereby reducing the workload of manual curation. The rule-based matcher scores candidates by distance (decay), lexical similarity, address consistency and overall tag agreement, with deterministic tie-breakers. This project delivers a repeatable workflow that ingests snapshots from AllThePlaces (ATP) covering Switzerland alongside current OSM extracts, harmonises tagging into a taxonomy, and prepares high-recall candidate pairs through joint spatial and semantic blocking. The approach emphasises explainability (structured similarity breakdowns), governance (immutable provenance metadata, deterministic reruns), and incremental adoption (drop-in matcher profiles without retooling downstream audit). By modelling both spatial proximity and semantic compatibility early, the workflow reduces noisy pair proliferation and focuses human review where residual ambiguity genuinely exists.

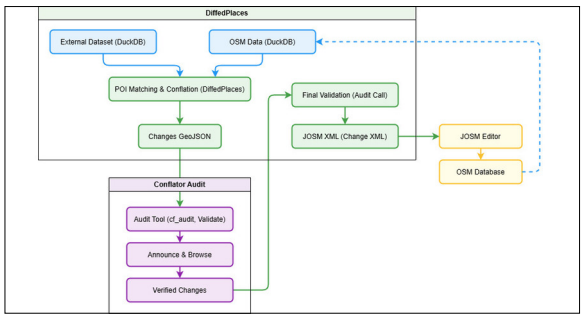
Approach / Technology: The workflow operationalises two complementary matchers inside DiffedPlaces, a web-based tool developed as part of this thesis: (i) a deterministic rules engine producing interpretable similarity decompositions and (ii) a supervised RandomForestClassifier (engineered lexical, spatial, semantic, structural features) trained with a staged regimen combining synthetic perturbations, rule-based training data generation, and high-precision pseudo-labels. Unified GeoJSON diffs progress through a multi-layer human-in-the-loop audit (command-line inspection, web-based collaborative review, quorum confirmation) before generating a ChangeXML batch for curator validation and upload. Idempotent ingestion, provenance metadata and reproducible profiles minimise hidden state; semantic blocking reduces implausible pair evaluations and lowers false positives upstream of any ML decision boundary. Feature design balances robustness (token-set and character similarities, distance decay, tag agreement) with parsimony to support later optimisation and potential model distillation. The architecture isolates concerns (ingest, match, audit, publish) so that performance tuning or classifier upgrades do not ripple unpredictably into auditing or upload procedures.

Conclusion: Empirical validation on two golden datasets - a brand-focused Aldi Süd Switzerland subset (246 outlets) and a stratified multi-category sample (200 POIs) - shows the machine learning matcher attains near-zero false positives on the brand subset (Precision 1.0000, Recall 0.9957, F1 0.9978) and improves balanced performance on heterogeneous data (F1 0.8814 vs. 0.8587 for the tuned rule-based approach) while lowering False Positive Rate. These gains cut auditor review volume in dense urban clusters and provide a defensible precision baseline for integrating further feeds. Practical impact includes shorter publication lead times, higher mapper confidence in suggested merges, and clearer escalation paths for borderline cases. Remaining limitations include elevated inference latency relative to the rule-based path, limited golden coverage of rare categories. Proposed mitigations - feature pruning, alias expansion, adaptive thresholds, probability calibration, lightweight model distillation, and periodic drift health reports - outline a path to production hardening while preserving reproducibility and transparency.

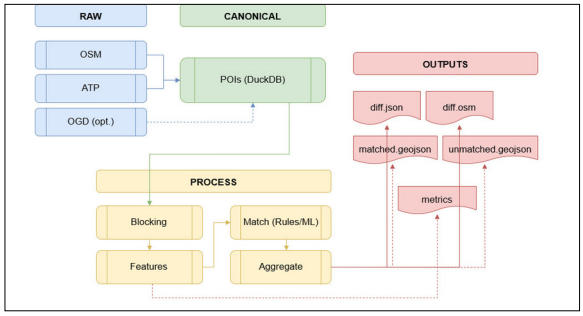
Early ideation diagram illustrating envisioned ingestion, matching, audit, and feedback loops of DiffedPlaces.
Own presentment based on ChatGPT



Operational pipeline: ingestion (blue), matching (green), auditing (purple), human upload (yellow).
Own presentment



Data flow emphasising feedback loops from ingestion through audit back to refreshed snap-shots.
Own presentment



Advisor
Prof. Stefan F. Keller

Co-Examiner
Sascha Brawer, Bern, BE

Subject Area
Data Science, Computer Science