

# POI Data Dictionary

Worldwide Point of Interest Dataset

~97 million records • 240+ countries • 27-field unified schema

**POIData.xyz**

## DATA DICTIONARY

Point of Interest (POI) Dataset

Worldwide Coverage

Licensing Documentation

Version 1.0 | February 2026

~97 million records | 240+ countries and territories

Classification: Public

---

# 1. Overview

This document defines the schema, field-level specifications, and data quality attributes for the Point of Interest (POI) dataset as distributed by [poidata.xyz](#). It is intended for use by data licensees, integration engineers, and product teams to ensure correct interpretation, loading, and governance of the delivered data.

The dataset contains approximately 97 million structured records spanning 240+ countries and territories, describing physical business locations, public amenities, and transport infrastructure. Each record represents a single POI and includes geographic coordinates, categorization, contact details, operating hours, source provenance, and quality scoring.

This document is country-agnostic.

---

# 2. General Specifications

File Format	CSV (comma-delimited, UTF-8 encoding)
Delivery	One CSV file per country (ISO 3166-1 alpha-2 code)
Global Record Count	~97,001,036 records (as of February 2026)
Total Fields	27 (identical schema across all country files)
Character Encoding	UTF-8 (supports all scripts including CJK, Arabic, Cyrillic, etc.)
Coordinate System	WGS 84 (EPSG:4326)
Null Representation	Empty string (no value between delimiters)

<b>File Format</b>	<b>CSV (comma-delimited, UTF-8 encoding)</b>
Multi-Value Delimiter	Pipe with spaces: '   '
Timestamp Format	YYYY-MM-DD HH:MM:SS (UTC)
Geographic Scope	Worldwide: 240+ countries and territories

### 3. Field Reference

The table below provides a complete specification for every field. The schema is identical across all country files. Data types reflect recommended database column types.

Field Name	Data Type	Null?	Description	Notes
unique_id	Integer	No	Globally unique identifier for each Point of Interest (POI) record.	Primary key. Assigned at ingestion; immutable. Unique across the entire worldwide dataset.
poi_name	String (VARCHAR 255)	Yes*	Primary business or location name of the POI.	*Extremely rare nulls (<0.001%). May contain diacritics, non-Latin scripts (CJK, Arabic, Cyrillic, etc.), and special punctuation.
brand	String (VARCHAR 100)	Yes	Corporate or franchise brand associated with the POI, if applicable.	Empty when independent/unbranded. ~98% null globally. Values reflect international and local brand names.

Field Name	Data Type	Null?	Description	Notes
poi_tel	String (VARCHAR 30)	Yes	Telephone number. Stored as string to preserve leading zeros and country codes.	No standardized formatting. Typically begins with country calling code (no + prefix). ~21% null. Lengths vary by country (5–19 digits observed).
super_category	String (VARCHAR 100)	No	Top-level industry or sector classification.	Controlled vocabulary. ~138 distinct values globally (e.g., Restaurants, Retail, Construction, Medical Practices, Transportation).
main_category	String (VARCHAR 100)	No	Primary business category within the super category.	Controlled vocabulary. One value per record. Thousands of distinct values. Max observed length: 39 chars.
second_category	String (VARCHAR 100)	Yes	Secondary category providing additional classification detail.	Optional refinement of main_category. ~46% null. Max observed length: 69 chars.

Field Name	Data Type	Null?	Description	Notes
all_categories	String (TEXT)	Yes	Pipe-delimited list of all applicable category labels across taxonomies.	Delimiter: '   ' (space-pipe-space). Typically 2–6 entries per record; up to 18+. Max observed: 909 chars. ~4% null.
location_id	Bigint (VARCHAR 20)	No	Unique identifier for the precise geographic location, calculated from latitude/longitude via a Hilbert space-filling curve. Two neighboring IDs (differing by <10) may be as close as 10 metres.	Fixed length: 13 chars. Generally unique per record. Co-location possible in multi-tenant buildings.
area_id	Integer (VARCHAR 10)	No	Broad geographic area identifier, roughly corresponding to a large city metro area.	Fixed length: 6 chars. Used for coarse regional aggregation.
small_area_id	Bigint (VARCHAR 20)	No	Granular sub-area within the parent area, roughly corresponding to a city block.	Fixed length: 9 chars. Enables fine-grained geographic filtering and spatial joins.
latitude	Decimal (10,6)	No	WGS 84 latitude coordinate of the POI.	Range: –90 to +90. Precision varies: ~70% at 4 decimal places, ~17% at 6. Not fixed-precision.

Field Name	Data Type	Null?	Description	Notes
longitude	Decimal (10,6)	No	WGS 84 longitude coordinate of the POI.	Range: -180 to +180. Precision varies: ~70% at 4 decimal places, ~16% at 6. Not fixed-precision.
adr	String (VARCHAR 255)	Yes	Street address or location descriptor of the POI.	Free-text in the local language of the POI. May include zone names, building identifiers, or descriptive text rather than formal street addresses. <1% null.
city	String (VARCHAR 100)	No*	City or municipality name where the POI is located.	*Effectively non-nullable. Local-language spelling. Thousands of distinct values per country.
neighborhood	String (VARCHAR 100)	Yes	Sub-city locality, neighborhood, or hamlet name.	Sparsely populated (~95% null). Hundreds of distinct values where populated. Max observed: 41 chars.

Field Name	Data Type	Null?	Description	Notes
prov	String (VARCHAR 100)	No*	Province, state, or top-level administrative region.	*Fully populated where available but may include 'UNKNOWN' sentinel values. Naming conventions and language vary by country (e.g., Flemish names in Belgium, state names in the US). Licensees may need to maintain mapping tables for localization.
postal	String (VARCHAR 20)	Yes	Postal or ZIP code for the POI location.	String type to support alphanumeric formats (e.g., UK, Canada). <1% null. Format varies by country.
country	String (VARCHAR 60)	No	Full country name in English.	240+ countries and territories supported. Single value per record. English-l anguage name (e.g., Belgium, United States of America, Japan).

Field Name	Data Type	Null?	Description	Notes
formatted_address	String (TEXT)	No	Pre-formatted, human-readable full address string.	Always populated. Concatenation of address components; may include ISO country code. May occasionally contain operational notes. Not suitable as a parsing source for structured address components.
source_url	String (TEXT)	Yes	URL of the POI's official website or primary online listing.	~98% null. Primarily populated for chain/franchise locations with centralized websites. Mix of http and https. Single URL per record.
email	String (VARCHAR 255)	Yes	Contact email address of the POI.	~70% null. Max observed: 88 chars. Validated format where available.

Field Name	Data Type	Null?	Description	Notes
opening_hours	String (TEXT)	Yes	Structured or semi-structured opening hours of the POI.	~65% null. Format varies: day abbreviations (M o,Tu,We,Th,Fr,S a,Su), mixed 12h/24h time notation, semicolon-separated day groups. Max observed: 345 chars. Requires normalization for machine processing.
crawl_time	Datetime (ISO 8601)	No	Timestamp of the most recent data collection event for this record.	Format: YYYY-MM-DD HH:MM:SS (UTC). Indicates data freshness. Used to assess currency and schedule update cadences.
alternate_names	String (TEXT)	Yes	Known aliases, trade names, or alternate spellings with associated confidence scores.	Pipe-delimited entries ('   '). Each entry format: name,score (score range 0–1). ~35% null. May include transliterations and multi-script variants. Max observed: 1,236 chars.

Field Name	Data Type	Null?	Description	Notes
top_poi_score	Decimal (15,15)	No	Composite quality/confidence score for the POI record.	Theoretical range: 0 to 1. Rare edge cases slightly exceeding 1.0 have been observed. Higher values indicate higher confidence and data quality. Implement tolerant parsing.
top_source_urls	String (TEXT)	Yes	Pipe-delimited list of the highest-quality source URLs used to compile this record.	Delimiter: '   '. Typically 1–3 URLs per record (~75% have 3). ~3% null. Ordered by source quality/relevance. Max observed: 608 chars.

## 4. Geographic Hierarchy

The dataset employs a three-tier geographic hierarchy for spatial aggregation and filtering. All three ID fields have fixed character lengths across the entire global dataset: `area_id` (6 chars), `small_area_id` (9 chars), and `location_id` (13 chars).

`area_id` represents the broadest regional grouping, roughly corresponding to a large city metro area. `small_area_id` subdivides each area into finer zones roughly corresponding to a city block. `location_id` pinpoints the precise building or site using a Hilbert space-filling curve computed from latitude/longitude. Two neighboring `location_ids` (differing by a small number <10) could be as close as 10 metres to each other. Multiple POI records may share the same `location_id` when co-located (e.g., a hotel and its on-site restaurant).

Coordinate precision varies globally: approximately 70% of values have 4 decimal places (~11m accuracy), 17% have 6 decimal places (~0.11m), with the remainder at 1–5 decimals. Licensees should not assume fixed precision.

## 5. Category Taxonomy

POI categorization follows a three-level hierarchy that is consistent across all countries. `super_category` provides broad industry classification with approximately 138 distinct values (e.g., Restaurants, Retail, Construction, Medical Practices, Transportation/Logistics). `main_category` specifies the primary business type with thousands of distinct values globally, one per record. `second_category` offers an optional further refinement (~46% null).

The `all_categories` field aggregates all applicable labels as a pipe-delimited list, enabling flexible search and cross-referencing without strict hierarchy constraints. Records typically carry 2–6 category labels, with some reaching 18+. Category labels are in English regardless of the POI's country of origin.

---

## 6. Data Quality & Scoring

Each record carries a `top_poi_score` reflecting overall confidence in the record's accuracy and completeness. The theoretical range is 0 to 1, with higher values indicating higher quality. Rare edge cases slightly exceeding 1.0 have been observed. Scores are derived from source reliability, field completeness, cross-reference validation, and recency of `crawl_time`. Licensees should implement tolerant parsing (e.g., accept 0–1.05) rather than strict 0–1 validation.

The `alternate_names` field pairs each alias with a confidence score (format: name,score). Approximately 65% of records have at least one alternate name. These may include transliterations, abbreviations, colloquial names, and multi-script variants, which are particularly valuable for POIs in countries using non-Latin scripts.

### 6.1 Approximate Global Completeness Rates

Core identification fields (`unique_id`, `super_category`, `main_category`, `location_id`, `area_id`, `small_area_id`, `latitude`, `longitude`, `country`, `crawl_time`, `top_poi_score`) are 100% populated across all countries. Address fields are nearly complete: `city` (~100%), `formatted_address` (~100%), `adr` (~99.6%), `postal` (~99.7%). Contact fields are sparser: `poi_tel` (~79%), `email` (~30%), `opening_hours` (~35%). The `brand` field is populated for approximately 1–2% of records globally. Rates may vary significantly by country.

---

## 7. Data Provenance & Source Tracking

`source_url` contains the POI's official web presence; ~98.5% null globally, with only a small number of distinct values per country (primarily chain/franchise URLs). `top_source_urls` lists the highest-quality third-party sources consulted during data compilation; ~97% populated, typically containing 3 pipe-delimited URLs per record. The `crawl_time` timestamp records when each record was last refreshed, enabling licensees to assess data currency and schedule update cadences.

---

## 8. Administrative Geography

The `prov` field contains the top-level administrative region for each POI (province, state, region, prefecture, etc., depending on the country). Field naming conventions and language vary by country — values typically use local-language names. Examples: Belgian provinces in Flemish (e.g., 'Antwerpen', 'West-Vlaanderen'), US states as two-letter codes (e.g., 'CA', 'TX'), Japanese prefectures in romanized form.

A small number of records per country may carry the sentinel value 'UNKNOWN', indicating incomplete administrative region assignment. Licensees requiring standardized or translated region names should maintain external mapping tables (e.g., ISO 3166-2).

---

## 9. Usage Notes for Licensees

### 9.1 Telephone Numbers

Stored as unformatted digit strings with no separators. Typically begin with the country calling code (without + prefix). Lengths vary by country (5–19 digits observed). Apply country-specific parsing libraries (e.g., Google's `libphonenumber`) before display or deduplication.

### 9.2 Opening Hours

Semi-structured format with significant variation globally. Day abbreviations (Mo, Tu, We, Th, Fr, Sa, Su) are generally consistent, but time formats mix 12-hour and 24-hour notation. Day groups are semicolon-separated. ~65% null. Maximum observed length: 345 characters. Requires dedicated normalization for machine processing.

### 9.3 Addresses and Localization

All textual address fields (`adr`, `city`, `neighborhood`, `prov`, `formatted_address`) contain local-language content. For countries using non-Latin scripts, values may be in the local script, romanized, or a mix. The `formatted_address` field is always populated and provides a display-ready string, but should not be parsed for structured address components. Licensees operating across multiple countries should anticipate significant formatting variation.

### 9.4 Character Encoding

All files are UTF-8 encoded. POI names and address fields may contain characters from any Unicode block, including CJK ideographs, Arabic script, Cyrillic, Devanagari, Thai, and more. Ensure your data pipeline supports full UTF-8 processing.

### 9.5 Coordinate Validation

While each country file nominally contains POIs for that country, coordinates near borders may extend slightly beyond the country's geographic boundaries due to geocoding of cross-border areas or imprecision. Licensees requiring strict geographic filtering should validate coordinates against expected bounding boxes or administrative boundary polygons.

## 9.6 Country-Specific Considerations

Nullability rates and value distributions vary by country. Smaller or less-developed markets may have higher null rates for contact fields (email, poi\_tel, opening\_hours). The prov field naming convention follows local administrative terminology. The all\_categories and super\_category taxonomies are globally consistent (English labels) regardless of the POI's country.

---

## 10. Coverage Summary

The dataset covers 240+ countries and territories, with record counts ranging from under 100 (e.g., Antarctica, Cocos Islands) to over 12 million (India, China). Estimated coverage rates per country (reported by the data provider) range from 88% to 99%. The following are the largest country files by record count:

<b>China</b>	<b>~10.9 million records</b>
India	~12.2 million records
United States	~9.2 million records
Indonesia	~6.7 million records
Brazil	~5.1 million records
Japan	~3.7 million records
Germany	~3.0 million records
France	~2.3 million records
Italy	~2.2 million records
Argentina	~1.2 million records

A complete country listing with record counts and estimated coverage is available at [store.poidata.xyz/world](https://store.poidata.xyz/world).