

The Regions Dataset

Reference & User Guide — v3.3 September 10 2026

1. What Is This Dataset

The Regions Dataset is a geographic reference file. It answers a question that turns out to be surprisingly hard to answer from public sources: when someone refers to a place by an geographic, informal, colloquial, or historical name — “the Great Lakes,” “the Corn Belt,” “Appalachia,” “New England,” “the Byzantine Empire” — which present-day countries, states or provinces, counties, and in some cases postal codes does that name actually correspond to?

Most geographic reference data (gazetteers, postal databases, census boundary files) work in the opposite direction: they tell you the official name of a place, not the informal or historical names people use to talk about the region it’s in. This dataset fills that gap. It has been built by hand, region by region, cross-checking colloquial usage, historical scholarship, and official administrative boundaries against each other — which is very likely why nothing quite like it turns up in an open search.

The dataset covers three broad kinds of regions:

- Geographic and cultural regions in current use — “the Lowcountry,” “the Great Lakes,” “the Eastern Seaboard,” “the Mid-Atlantic,” “the Anglosphere.”
- Former political entities — empires, colonies, short-lived republics, and other polities that no longer exist but whose historical footprint people still reference — “the Byzantine Empire,” “the Roman Empire,” “the Republic of Texas,” “New France.”
- Metropolitan and indigenous-peoples regions — named metro areas and present and historical territories of indigenous nations.

This dataset reflects editorial opinion — decisions about which regions to include, what to name them, and what geographic area each one covers. If you disagree with us, we hope you’ll take the time to tell us and explain why.

2. The Dataset is Continually Improved

This dataset is drawn from and actively maintained in support of the Historical Marker Database (HMdb.org), a continually growing global catalog of historical markers. Because this underlying data continues to serve that live project, it is under continuous review, correction, and expansion. You should not consider it static, one-time download. Licensees are provided with a link that can be repeatedly used to fetch the latest copy of the dataset.

If you discover an error or omission, we want to know. Errors and omissions will be promptly corrected. Differences of opinion will be carefully considered, as will suggestions for additional regions.

3. The Original Use of the Dataset

The Historical Marker Database uses the dataset to indicate what regions markers and memorials find themselves in. Querying the dataset with four datapoints, Country, State, County, and Postal Code, the result is a list of regions. The additional information in the dataset allows for the regions to be announced in complete sentences. Following are two examples. The dataset is also used to create lists of all markers and memorials found in each region.

Regionally, this marker is on Virginia’s Northern Neck. It is also in the American South, specifically in the Upper South, in the Tidewater, in the Chesapeake Bay Region, and on the Eastern Seaboard. Globally, it is in the North Atlantic Region, North America, the Western Hemisphere, the Western World, and the Anglosphere. Historically, it finds itself in what was once the territory of the Mississippian Culture, one of the original Thirteen Colonies, one of the Confederate States of America, and the Antebellum South.

Regionally, this marker is in England’s Midlands. It is also on the British Marches. Globally, it is in the North Atlantic Region, in Europe, and specifically in Atlantic Europe, on one of the British Isles, in the Western World, and in the Anglosphere. Historically, it finds itself in what was once the Roman Empire.

4. What's In the File

The dataset is delivered as a single flattened (“Cartesian”) CSV: one row per combination of a named region and one piece of geography that belongs to it. This trades some redundancy — a region with forty member counties appears forty times, once per county, with its name and metadata repeated on every row — for a file that opens directly and correctly in Excel, Google Sheets, or any BI tool without requiring a join, a relational database, or any setup at all.

Internally, the source system stores this as two related tables: one row per named region (its display name, category, and how it should be described in a sentence), and one row per piece of geography that belongs to that region (a country, a state/province, a county, or in some cases a postal code). The flattened file is the join of those two tables — every geography row carries its parent region's full metadata along with it.

5. Column Reference

Columns fall into two groups: fields describing the region itself (repeated identically across every row belonging to that region), and fields describing one piece of that region's geographic footprint (unique to each row).

5.1 Region-level fields

Column	Type	Description
RegionID	Integer	Unique identifier for the region. Every row belonging to the same region shares this value — it's the join key between the region-level and geography-level fields, retained in the flattened file for anyone who wants to re-derive the two-table structure.
Region	Text (≤40 chars)	The region's display name, e.g. “Great Lakes,” “Byzantine Empire,” “Northern Neck.”
RegionType	Code: 1, 2, 3, or 4	The region's tier in the classification hierarchy. 1=State or provincial, 2=Countrywide, 3=Global, 4=Historical,
RegionSubType	Code: M	The region's category: M = Metropolitan. Any other letter or empty field indicates that the region is not a Metropolitan region.
RegionTypeName	Text (≤50 chars)	A secondary label whose meaning depends on RegionType: for state/provincial-tier regions it names the parent state or province; for countrywide-tier regions it names the parent country; for global-tier regions it names a continent, ocean, or thematic grouping; for former-polity regions it names one of a small set of historical category labels (e.g. “Former Empires and Colonial Powers”). See Section 4.
Collection	Text (≤35 chars)	The named group of regions this one belongs to for maintenance purposes, e.g. “Regions of Michigan,” “Global Regions,” “Former Empires.” Multiple regions typically share a Collection.
The	Yes/No	Whether a sentence referring to this region should include the definite article — e.g. “in the Mid-Atlantic” (Yes) vs. “in New England” (No).
InOn	“in” / “on” / “at”	Which preposition idiomatically precedes the region's name — most regions take “in,” but a handful of place-like regions take “on” (e.g. “on Long Island,” “on the Eastern Seaboard”) and then there's “at the Four Corners”
OneOf	Yes/No	Whether the region is idiomatically phrased as membership in a set — “one of the original Thirteen Colonies” — rather than simple containment.
Rank	Integer	Controls display order when multiple regions of the same tier apply to the same location — used to keep auto-generated descriptions in a sensible, consistent order (e.g. naming the Northeast before the Mid-Atlantic).

Column	Type	Description
WikipediaTitle	Text	The title of a corresponding Wikipedia article, where one exists, for further reading. When this field is not populated, the value of Region will find the Wikipedia article. When the value of this field is “no article” no Wikipedia article existed the last time it was checked, and a link to Wikipedia should be omitted.

5.2 Geography-level fields

These describe one specific piece of geography belonging to the region named on that row. A region's full footprint is the union of all rows sharing its RegionID.

Column	Type	Description
Country	Text (≤50 chars)	The country or country-equivalent this row's geography sits in.
AllCountry	Yes/No	Yes means the entire country belongs to the region — State and County are blank on that row. No means only part of the country belongs, and State (and possibly County) narrows it down.
State	Text (≤50 chars)	The state, province, or equivalent first-level division. Blank when AllCountry = Yes.
AllState	Yes/No	Yes means the entire state/province belongs to the region — County is blank on that row. No means only part of it belongs, and County narrows it down further.
County	Text (≤50 chars)	The county or equivalent second-level division (parish, borough, census area, district, prefecture, etc., depending on the country). Blank when AllState = Yes or AllCountry = Yes.
Zip	Text (≤11 chars)	A postal code (complete or a prefix), used only in cases where county-level precision is too coarse — typically when a region covers only part of a single county. Zip is authoritative within a country: whenever it's populated, treat it as overriding State and County on that row — do not read state and county as adding a further restriction. In this case they are solely provided to document the location of the largest portion of the postal code area.
LastChanged	Date/time	Added to support incremental mirroring. Each row's value is the more recent of two source timestamps: the parent region's own last-changed date, and this specific geography row's own last-changed date — since either one changing means this flattened row's content has changed. A mirror can request only rows where LastChanged is newer than its last successful sync, rather than re-downloading the full file on every update.

Every row's precision level is fully described by which of AllCountry / AllState / County / Zip is populated — there's no separate “precision level” column to look up; it can be read directly off which fields are filled in on that row.

A note on the Geographic Names Used

Country names are in English. State/Province and County-equivalent names are in the local language—and may contain accented characters—when the local language uses Roman characters, and transliterated to roman characters when they are not. This page lists the exact geographic names used in this dataset: [Definitive Geographic Names](#).

6. The Classification Hierarchy (RegionType)

Every region is assigned one of four tiers. This is the single most useful column for filtering the dataset down to a relevant subset — a user who only wants currently-existing geographic regions, say, can exclude Tier 4 entirely. RegionType and RegionTypeName should be used to group regions for browsing purposes, not to limit the geography of a region.

Tier	Meaning	RegionTypeName holds	Example regions
1	State/provincial — a region defined within a single state or province	The name of that parent state or province	Lake Michigan Shore Counties (Michigan); Georgian Bay Country (Ontario)
2	Countrywide — a region defined within a single country, potentially spanning many states/provinces	The name of that parent country	Great Lakes (as currently defined); Regions of the United States
3	Global — a region that spans more than one country	A continent, ocean, or thematic grouping (e.g. “Global Regions,” “North America”)	Western World, Anglosphere, North America, Lake Superior's North Shore
4	Former polity — an empire, colony, or other political entity that no longer exists	One of a small set of historical category labels (e.g. “Former Empires and Colonial Powers,” “Former Parts of the USA”)	Byzantine Empire, Roman Empire, New France, Republic of Texas

A note on the RegionSubType field (Section 5.1)

The RegionSubType field is currently used to indicate that the region is a Metropolitan Region, typically named “(name of city) Metropolitan Area” or “Greater (name of city).” This type of region is typically classified as a State/Provincial region (RegionType=1) and placed in the state or province where the city is located. Often, a metropolitan region encompasses areas outside of the state or province where the city is located and the geography-level fields (Section 5.2) will so indicate.

A note on the geography-level fields disagreeing with RegionType

Geography-Level fields can include areas outside of the RegionType indicated for the Region. Examples are Metropolitan areas discussed in the previous note, state-wide and or province-wide areas primarily in one state or province that crosses into another state or province, or countrywide areas that incidentally stray into adjacent countries.

7. Geographic Precision

Regions are recorded at the coarsest level of precision that's geographically accurate, and only narrowed further where the region's real-world boundary genuinely requires it. In practice this produces four precision levels, cascading from country down to postal code:

- Whole country (AllCountry = Yes) — used when the entire country is part of the region.
- Whole state/province (AllState = Yes, within a named Country) — used when only part of a country is included, but a given state or province is included in full.
- Specific county (or equivalent) — used when only part of a state or province is included. This is the most common precision level in the dataset by a wide margin.
- Postal code — used in the rare case where even a full county is too coarse, typically because a region only covers part of a single city. When postal code is present, County and State should be ignored. The zip code's post office may be located in the state and county shown, but zip codes can straddle more than one county and/or state or province.

Historical (Tier 0) regions in particular are frequently precise down to the county level, because former political boundaries very often don't align with modern state or national borders — the Byzantine Empire's brief foothold in Spain, for instance, covers only two provinces of Andalusia plus the Balearic Islands and Murcia, not the whole country.

8. Known Limitations and Editorial Notes

- Coverage depth is uneven. Some regions — particularly long-established U.S. regions — are researched down to the county or postal-code level. Others, especially outside the U.S. and Canada, are currently recorded only at the country or state level and would benefit from the same treatment over time.
- Historical (Tier 4) boundaries especially reflect editorial judgment, rather than universal consensus. Ancient and medieval political boundaries are frequently disputed or only loosely attested in the historical record; where scholarship disagrees, this dataset generally favors the better-attested, more conservative reading.
- Modern administrative geography changes over time, and the dataset requires periodic revalidation against current official sources — county mergers, renamings, and reclassifications (e.g. Connecticut's 2022 replacement of its eight counties with nine planning regions) do occur and need to be tracked.
- Precision bottoms out at the postal-code level. A handful of regions are defined by informal boundaries even finer than a postal code can express (e.g. specific townships in areas where postal codes span multiple, geographically dissimilar municipalities); these are approximated at the nearest precision level the schema supports, and a small number of edge cases remain acknowledged but unresolved.

9. Suggested Use Cases

Because it maps informal region names to formal geography rather than the other way around, this dataset is best suited to use cases that start from how people actually talk about places, such as:

- Editorial or content-tagging systems that want to auto-suggest or validate regional tags (“this dateline is in the Mid-Atlantic”) from a location.
- Historical, genealogical, and heritage projects needing to describe a location's former political affiliations accurately.
- Marketing and localization tools that target campaigns by colloquial region rather than formal administrative boundary.
- Research and data-science projects that need to resolve informal region references in text to structured, queryable geography.

10. Worked Example

A few illustrative rows from the flattened file, showing how the same region produces multiple rows at different precision levels:

Region	RegionType	Country	AllCountry	State	AllState	County
Byzantine Empire	4	Spain	No	Región de Murcia	Yes	
Byzantine Empire	4	Spain	No	Andalucía	No	Málaga
Byzantine Empire	4	Greece	Yes			
Southern Plains	2	USA	No	Texas	No	Armstrong County
Southern Plains	2	USA	No	Kansas	Yes	

(Columns Rank, WikipediaTitle, The, InOn, OneOf, Zip, RegionTypeName, and Collection omitted here for width; all are present in the actual file as described in Section 5.)

11. Dataset Snapshot (September 10, 2026)

These numbers are constantly changing as the dataset is updated and improved.

Metric	Value
Total rows in dataset	15,539
Total named regions	892 (across 93 named Collections)
Tier 1 — State/provincial	694 regions
Tier 2 — Countrywide	61 regions
Tier 3 — Global	93 regions
Tier 4 — Historical	45 regions
Distinct countries with at least one geography row	220
Regions with a linked Wikipedia article	810