GPH2022_vs_COW2016

Understanding the Construction of the GeoPolHist (GPH) Database and Its Differences from the Correlates of War (COW) Database

July 2025

To provide users with a detailed understanding of how the *GeoPolHist* (GPH) database was developed from the *Correlates of War* (COW) database – and to facilitate future enhancements of the GPH tool – the comparative table *GPH2022_vs_COW2016* documents all modifications introduced in the GPH geopolitical entities database (latest version: September 2024) relative to the COW political units database (latest version: 2016).

As outlined in Dedinger and Girard (2021, Appendix), the GPH database builds on the COW lists of political units by correcting various identified errors and expanding the coverage to offer the most comprehensive account possible of the world's geopolitical entities and their political status over the past two centuries. Wikipedia serves as the primary source used in constructing the GPH database. In keeping with its collaborative ethos, GeoPolHist invites users to contribute by reporting errors or suggesting improvements to help enrich and refine the tool.

The *GPH2022_vs_COW2016* file is a summary table comparing the entity lists published in the COW files *States2016* and *Entities* (available on the COW website) with the GPH list in the file *GeoPolHist_entities_status_over_time*. Each row corresponds to one entity and a specific period during which it held a particular political status.

For each COW and/or GPH entity-period combination, six variables are encoded:

- COW (GPH) code: the entity's identification code
- COW (GPH) name: the entity's name
- COW (GPH) status: the entity's political status
- Start year: beginning of the status period
- End year: end of the status period
- Sovereign COW (GPH) code: for non-sovereign entities, the code of the corresponding sovereign entity

The final column, Change in GPH vs. COW, indicates any modifications made to the COW entry in GPH. These changes fall into several categories: alterations to codes, names, sovereign codes, statuses, and/or dates. Some COW entries are omitted in GPH; conversely, new entries are added.

Changes in code

GPH modifies certain COW codes for the following reasons:

- Error correction: For instance, the COW code 374 (assigned to both the Chechen-Ingush Republic and Transcaucasia) was corrected to 376 for the Chechen-Ingush Republic in GPH.
- Entity disaggregation: The COW entity "510/Tanzania (Tanganyika) (German East Africa)" was split in GPH into "512/Tanzania (Tanganyika)" and "510/German East Africa." Tanganyika, a German protectorate from 1885 to 1891, was part of German East Africa from 1891 to 1916. German East Africa, a German colony from 1885 to 1919, was composed of Tanganyika (1891-1916), Rwanda (1885-1916), and Burundi (1890-1916).
- New GPH entities and recoding: Example: "3316/Carpatho-Ukraine" becomes 3154 in GPH. This follows a broader recoding of post-Austro-Hungarian entities (codes 3150–3154). Carpatho-Ukraine was one of the entities that emerged from the disintegration of the Austro-Hungarian Empire and would become part of Czechoslovakia. Similar recoding was applied to Chinese and Dutch East Indies entities, reflecting the creation of new codes (e.g., GPH codes 7117–7137 Chinese concessions; 8501–8574 for Indonesian regions). Example: recoding of "7121/Inner Mongolia," "7131/Diaoyu (Senkaku) Is.," and "7220/Kiauchau." Similarly, the entities "8541/Bali" and "8547/Tapanuli" and "8551/Indrapuna" were recoded, and the Indonesian entities were completed with entities such Pontianak or Ternate, which are found in the RICardo database (https://ricardo.medialab.sciences-po.fr/data/RICardo_RICentities.csv)

Changes in Name

As explained in Dedinger and Girard (2021), changes in the names of entities over time present challenges for creating a historical list of political entities. The authors of the COW project addressed this issue by adopting a specific rule: when an entity's name changes over time, they use the most recent name and, in cases where confusion may arise, add the earlier name in brackets (Russett, Singer, and Small 1968, 951). The GPH database adopts this approach, systematically applying it to all relevant entities and correcting certain anomalies found in the COW lists. Approximately one hundred COW names are modified in the GPH database.

- Corrections: The COW entity lists in the "Entities" and "States" files occasionally assign different names to the same code/entity. For example, entity 110 is labeled "Guyana" in the "States" file and "Guyana (Br. Guyana)" in the "Entities" file; entity 370 appears as "Belarus" in the "States" file and "Byelorussia (Belarus)" in the "Entities" file. The GPH list standardizes all such cases. Another example is the COW entity "Portuguese India (incl. Goa, Diu, Daman)," which is renamed "Portuguese India" in GPH. In addition, GPH creates five new entities considered "part of" Portuguese India: "7591/Goa," "7592/Diu," "7593/Daman," "7594/Dadra," and "7595/Nagar Haveli." Finally, the COW lists contain a case of homonymy with entities "372/Georgia" and "1010/Georgia." In the GPH database, entity 1010 has been renamed "Georgia (U.S.)" to resolve the ambiguity.
- Disambiguation and expansion: In numerous cases, COW entities are labeled solely with their contemporary names. The GPH database systematically applies the rule defined by the COW authors and constructs new names that incorporate the various historical appellations of each entity over the past two

centuries. For example, “Hawaii” becomes “Hawaii (Sandwich Is.),” “Virgin Islands” becomes “Virgin Islands (Danish West Indies),” “Central African Republic” becomes “Central African Republic (Ubangi Shari),” and “Zaire (Kinshasa) (Belgian Congo)” is expanded to “Democratic Republic of the Congo (Zaire) (Kinshasa) (Belgian Congo) (Congo Free State).”

Changes in Sovereign Code

- Creation of new GPH entities and changes in sovereign coding: This applies, for example, to entities that were part of Prussia. The COW lists treats “255/Germany/Prussia” as a single entity, whereas GPH distinguishes between two separate entities: “255/Germany (Zollverein)” and “239/Prussia.” As a result, Hanover, Frankfurt, and Nassau—coded as part of “Germany/Prussia” in COW—are recoded in GPH as part of “Prussia,” which itself becomes part of “Germany (Zollverein)” from 1871 onward.

New GPH statuses

As explained in Dedinger and Girard (2021), the COW database defines ten political status types, covering both sovereign and non-sovereign entities. To allow for more precise coding across the full historical period and to reflect the evolving nature of political statuses over time, the GPH database introduces ten additional status categories and corrects certain entries found in the COW database. Below are examples illustrating the application of these new GPH status types:

- Associated state of: In the COW database, Mauritania, Niger, and Chad are listed as having the status “colony of” France until 1960. The GPH database updates this coding by assigning the status “associated state of” France for the period 1958–1960. French West Africa and French Equatorial Africa were dissolved in 1958, and a new form of political association with France was established, including Mauritania, Niger, and Chad. This association ended in 1960 when these countries gained independence.
- Dependency: This status was introduced to acknowledge that the term “colony” was no longer used in Chapter XI of the United Nations Charter and to avoid a proliferation of non-sovereign status types after 1945. The GPH status “dependency of” replaces the COW status “colony” for the Comoros from 1947 to 1975, the status “part of” China for Hong Kong from 1997, and the status “part of” Indonesia for West Irian (Dutch New Guinea) from 1969.
- International: The COW database does not assign any political status to the entities “0/League of Nations” and “1/United Nations.” The GPH database addresses this gap by introducing the status “International.”
- Sovereign (limited): This status is used by GPH to reflect cases of constrained sovereignty. For example, Norway is coded as “part of” Sweden from 1816 to 1905 in the COW database, but GPH assigns it the status “Sovereign (limited).” After being ceded by Denmark to Sweden in 1814, Norway entered a personal union with Sweden, in which the Swedish king also ruled Norway and controlled its foreign policy—although Norway retained a degree of autonomy. A similar revision is made for Luxembourg, which is recoded as “Sovereign (limited)” from 1816 to 1892, rather than “part of” the Netherlands as in COW. These reclassifications are, of course, open to interpretation.

- Sovereign (unrecognized): This status was created to reflect entities that have declared independence but are not universally recognized. For instance, Kosovo is coded as "sovereign" in the COW database from 2008 onward. In contrast, the GPH database classifies it as "Sovereign (unrecognized)" to account for the fact that not all UN member states recognize Kosovo's independence.

Date Modifications in the GPH Database

Nearly 15% of the dates recorded in the COW database have been modified in the GPH database. These changes were made for various reasons, too numerous to list exhaustively. In most cases, the revisions reflect updated or more reliable historical information—often drawn from Wikipedia—or result from the introduction of new political status categories in GPH. Users are encouraged to report any errors or inconsistencies they may encounter.

- Date revisions based on Wikipedia sources: A notable example involves the British colony of Jamaica and its dependencies. In the COW database, Belize (British Honduras) is coded as "part of" Jamaica from 1862 to 1884, the Turks and Caicos Islands from 1848 to 1962, and the Cayman Islands from 1816 to 1962. Based on information from Wikipedia, the GPH database adjusts these dates as follows: Belize (British Honduras) is considered part of Jamaica from 1816 to 1884, and the Turks and Caicos Islands from 1873 to 1962. These changes better reflect the historical administrative arrangements of these territories.
- Date changes resulting from new GPH statuses: The GPH database introduces the status "dissolved into" to capture cases where an entity ceases to exist—either through integration into another entity or through division. This refinement leads to changes in several entries. For example, both the Federal Republic of Germany and the German Democratic Republic are coded in the COW database as "part of" Germany from 1990. In the GPH database, both entities are instead assigned the status "dissolved into" Germany (Zollverein) in 1990, with matching start and end dates.
  The case of Vietnam is more complex. In the COW database: Entity "815/Vietnam" is coded as "part of" "810/French Indochina" from 1887 to 1949; as a "colony of" France from 1949 to 1954; then as "part of" both "816/Vietnam, Democratic Republic" and "817/Vietnam, Republic of" from 1954 to 1975. Additionally, a separate entity "816/Vietnam" is coded as "sovereign" from 1954 onward—creating an overlap with "815/Vietnam."
  In contrast, the GPH database recodes this trajectory as follows: "815/Vietnam" has the status "sovereign (unrecognized)" from 1816 to 1887; in 1887, it is "dissolved into" "810/French Indochina;" the entity reemerges in 1945, but in 1954–1955 it is again "dissolved into" two new entities: "816/Vietnam, Democratic Republic" and "817/Vietnam, Republic of;" in 1975, both "816" and "817" are "dissolved into" "815/Vietnam," which from that point onward holds the status "sovereign."

Entities Not in COW

More than one-third of the differences between the GPH and COW databases correspond to new data entries added in GPH. The relatively large number of these additions highlights a key limitation of the COW database:

its list of political entities is incomplete[1] if the goal is to construct a comprehensive dataset without missing values—i.e., a database that assigns a political status to every entity for every year of the period covered.

This limitation reflects the original purpose of the Correlates of War project. Its founders aimed to build a scientific database on war, not a complete historical record of political status. As part of that effort, they compiled a list of political units, developed criteria for determining sovereignty, and identified all entities that had participated in armed conflict since 1816. The result of this extensive work was a list of approximately 1,200 sovereign and non-sovereign political units for the period 1816–2016.

The GeoPolHist (GPH) project builds upon this foundational work, but with a different primary objective: to produce a scientific database of the political status of the world's geopolitical entities over the past two centuries. As described above, in addition to correcting certain inconsistencies in the COW data, GPH introduces new political status types and identifies additional geopolitical entities. These additions allow GPH to assign a status to every entity for each year since 1816, thus filling the gaps left by the original COW framework. 1,300 geopolitical entities are listed in the current version of the GeoPolHist database.

- New GPH Statuses: The authors of the COW database defined ten political statuses, including one sovereign and independent status ("sovereign") and nine non-sovereign or non-independent statuses ("colony," "claimed," "leased," "mandated," "neutral or demilitarized," "occupied," "part of," "possession," and "protectorate"). The GPH database introduces ten additional political statuses. Three of these represent forms of sovereignty that do not meet the two COW-defined criteria for sovereignty:[2] "associated state," "sovereign_limited," and "sovereign_unrecognized." Two correspond to non-sovereign statuses ("dependency," "vassal"), two serve as catch-all categories for uncertain cases ("discovered," "unknown"), and two apply to special cases ("informal," "international").

  The "sovereign_limited" status was introduced to identify entities that were sovereign but not independent. It primarily applies to the member states of the German Confederation between 1816 and the formation of the North German Confederation in 1867, which preceded the creation of the German Empire in 1871. The previously mentioned "sovereign_unrecognized" status addresses nearly 300 missing data points across the historical period—for example, Luxembourg (1890–1914), Vatican City (1929–1964), Liberia (1847–1920), Iran (then Persia) (1816–1855), and Thailand (then Siam) (1816–1887).

  The "discovered" and "unknown" statuses help complete around 200 data entries, mostly involving non-European entities for which specific information is either unavailable or has not yet been found. Finally, the "informal" status captures entities that, while not political entities in the conventional sense, nevertheless "exist" and may exercise a significant degree of sovereignty in certain areas—such as trade policy. This status applies to five GPH entities: Australia (1816–1901), the European Union (European Economic Community) (1958–2024), Germany (Zollverein) (1816–1867), Italy (1816–1861), and Malaysia (British Malaya) (1826–1946).

1 Dedinger and Girard (2021) illustrate this point with Figure 1 (p. 213), which shows the annual evolution of the total number of entities in the COW lists. It highlights a sharp increase between 1884 and 1891. This rise can be attributed to the inclusion of a growing number of entities in the COW database following the Berlin Conference of 1884–85, which formalized the partition of Africa among European powers. Although many of these entities existed prior to colonization, they were not previously recorded in the COW lists.

2 1. A population threshold of 500,000; 2. Recognition of sovereignty by the members of the interstate system through: before 1919, the establishment of diplomatic missions from major powers; after 1919, being a member of the League of Nations or the United Nations.

- New GPH entities and codification: GPH extends COW's three- and four-digit coding system to five digits, enabling more granular entity identification. This aligns with recommendations from Russett, Singer, and Small (1968, 932-933), allowing researchers to track geopolitical entities over time and across various datasets with greater precision:

> "We offer here a population of those national or quasi-national political entities which should serve the needs of most comparative and/or international politics scholars concerned with such entities, or a sample thereof. Despite our belief in their importance, it does not, however, include the infinite number of sub-national and extra-national actors in world politics, territorial or otherwise. The list is so conceived as readily to accommodate most of the modifications in territory or status that can be anticipated in the decades immediately ahead. In addition to delineating such a population, we specify the general political-legal status of each entity, and the time period during which each such status was held. Finally, in order to make more efficient and more accurate the borrowing of data from one research enterprise to another, we propose a simple three-digit coding scheme keyed to the major continental regions, which others might see fit to continue. In short, we offer here what might be called a 'prominent solution'; while it may not represent the very peak of refinement, it nevertheless reflects what we consider a reasonable balance among a number of competing requirements, it is convenient for ready reference, and all of the data from both our projects are already stored in accord with this particular scheme. In order to maximize replicability and cumulativeness, we invite others to adopt it for their own investigations."

GPH adopts the 3- and 4-digit coding system developed by the Correlates of War (COW) project and extends it to a 5-digit format. In general, 3-digit codes are used for sovereign or principal entities, while 4-digit codes are assigned to non-sovereign entities or country subdivisions. By expanding this system to 5 digits, GPH enables the inclusion of entities absent from the original COW list. This finer level of granularity is essential for detailed historical analysis. Moreover, the extended format allows for the seamless addition of new geopolitical entities as historical knowledge evolves, and provides a flexible, backward-compatible framework for researchers working across multiple databases.

Continent	*3-digit Code*	*4-digit Code*	*5-digit Code*
America	002 à 199	1000 à 1999	10000 à 19999
Europe	200 à 399	2000 à 3999	20000 à 39999
Africa	400 à 699	4001 à 6999	40000 à 69999
Asia	700 à 899	7000 à 8999	70000 à 89999
Oceania	900 à 999	9000 à 9999	90000 à 99999

Not in GPH

Conversely, approximately 300 entries from the COW database have been excluded from GPH. These primarily involve cases where an entity is categorized as "part of" another—typically meaning that it is neither sovereign nor independent but belongs to a larger sovereign entity.

- Removal of COW 'part of' entries for colonial groups: In COW, certain colonial groupings are represented as being "part of" their constituent territories before their formal establishment and after their dissolution. GPH eliminates such entries, restricting the existence of these groups to their official periods of existence. This applies, for instance, to the Netherlands Antilles, the West Indies Federation, the Dutch Windward Islands, and the Leeward Islands.

  Take the case of the West Indies Federation (COW code 50), which comprised ten Caribbean islands and existed only briefly from 1958 to 1962. In the COW dataset, the Federation is encoded as "part of" each of these ten entities from 1816 to 1958 and again from 1962 to the end of the period, resulting in 20 redundant observations. GPH removes these and retains only the 1958–1962 period, during which the Federation is classified as a "colony of" the United Kingdom.

- "Russian doll" cases: These are instances where an entity is listed as "part of" another entity that is itself "part of" a third entity, and so forth. COW often resolves such situations by redundantly coding the smallest entity as "part of" every larger enclosing entity.

  For example, St. Kitts and Nevis appears in the COW database as follows: in 1882, '1075/St. Kitts' and '1076/Nevis' merge to form '60/St. Kitts-Nevis,' which is itself "part of" '1063/Leeward Islands' from 1882 to 1956. Additionally, COW codes both St. Kitts and Nevis as being directly "part of" the Leeward Islands for the same period. GPH removes these redundant entries and retains only the hierarchical relationship as it existed institutionally.

Quantitative Summary of GPH modifications

Changes in GPH vs. COW	*Number of changes*
changes in code (+ name, year and/or status)	47
changes in name (+ year, status, and/or sovereign code)	299
changes in sovereign code	25
changes in status (+ sovereign code)	80
changes in year (+ status and/or sovereign code)	686
not in COW	1765
not in GPH	286
no changes	1601
Total number of rows	*4789*

References

Béatrice Dedinger & Paul Girard (2021), "How many countries in the world? The geopolitical entities of the world and their political status from 1816 to the present," *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 54 (4), pp. 208-227.

Béatrice Dedinger & Paul Girard (2017), "Exploring Trade Globalization in the Long Run: the RICardo Project," *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 50 (1), pp. 30-48.

Bruce M. Russet, J. David Singer & Melvin Small (1968), "National political units in the twentieth century: A standardized list," *American Political Science Review,* 62 (3), pp. 932–951.