CLIO World Tables:  
A Global Historical Database

John Gerring, Professor, Department of Political Science, Boston University  
<jgerring@bu.edu>

James Mahoney, Professor, Departments of Political Science and Sociology, Northwestern University  
<james-mahoney@northwestern.edu>

August 10, 2011
Abstract

We hereby announce the creation of a global historical database, the CLIO World Tables. CLIO is intended for use by anyone who seeks historical information about demographic, economic, political, and socio-cultural phenomena. It encompasses data from a large range of sources – published and unpublished, hard copy and digital (including extant datasets), archival and non-archival, English and foreign language – and serves as a user-friendly clearinghouse for historical information gathered on a global scale. In this report, we describe the rationale for CLIO, the process of data collection, and the product – as well as the possible future of CLIO.
With this publication, we announce the creation of a global historical database, CLIO World Tables. CLIO is intended for use by anyone who seeks historical information about demographic, economic, political, and socio-cultural phenomena. It encompasses data from a large range of sources – published and unpublished, hard copy and digital (including extant datasets), archival and non-archival, English and foreign language – and will serve as a user-friendly clearinghouse for historical information gathered on a global scale.

Currently, the database includes nearly fifty variables, each with varying temporal and spatial coverage. Longitudinal coverage for each variable extends back at least a century, and in some instances for millennia. Spatial coverage is global, which is to say that each variable encompasses more than one region of the world. Principal political units include States (understood as Empires or Nation-states), Semi-sovereign units (e.g., colonies or territories), and Non-sovereign units (e.g., regions or municipalities). Where possible, we employ GIS polygons or points to represent these units in space.

In this report, we describe the rationale for CLIO, the process of data collection, the organization of the database, and the shape of the data. Appendices discuss key sources (Appendix A), database architecture (Appendix B), and the imagined future of CLIO (Appendix C).

Rationale

Until recently it might have been argued that there was little need for a public data resource focused on historical evidence since social scientists were largely absorbed with contemporary issues and historians were focused on narrow slices of time and space for which they could readily collect their own data. However, the rise of interest in global history in recent years, and the increasing importance of history in the social sciences, means that it is important that we start to think seriously about how to overcome our collective data deficit.

Many subjects of contemporary concern demand an expansive spatial and temporal domain. A short list would begin with colonialism/imperialism, intra- and inter-state conflict, democracy, development, the diffusion of cultures, globalization and trade, nation and state formation, religion, revolution, and slavery. These subjects, like many others, are not neatly bounded in one region or time-period.

Of course, these phenomena may play out quite differently at different times and in different places. However, in order to get a clear view of the commonalities and the disjunctions we require a comprehensive view of a subject. One cannot determine whether things are different in site A without looking at sites B, C, and D.

We do not wish to imply that drawing comparisons across time and space is easy. Global history presents many challenges, especially if causal inference (as opposed to descriptive inference) is the object of analysis. However, our attention here will focus on a more basic challenge: the problem of finding relevant information.

1 CLIO is a collaborative project with many participants. Key among them were Natalie Lam (project manager), Erzen Oncel (project manager), and Jennifer Rosen (research assistant), Karim Nagib (database manager), Michael Trosen (web designer). Assistance was also received from Abel Amado, Ceren Ergenc, Jitka Hiscox, Ben Lewis, Andreea Maierean, Joseph Robinson, Humphrey Southall, and Johan van Gorp. CLIO was funded by the National Science Foundation (grants 0648292 and 0647921), with supplemental funding from the Clinton Global Initiative at Boston University and the Frederick S. Pardee Center for the Study of the Longer-Range Future.

2 Unclaimed, disputed, or internationally governed territories are a residual category with little data content.

Historical data is problematic in four respects. It is *scarce* (there is not much of it), *inaccessible* (hard to locate and contained in a variety of data formats or in hard copies), *incommensurable* (across varying sources), and *ambiguous* (uncertainty is difficult to assess). These problems persist not simply because of the nature of the evidence. They persist because individual incentives in the academy work against the provision of good data and there is no public or private body that is prepared to bear the burden of providing this public good. Let us explore these issues in greater detail.

Evidence about the past – which, for present purposes, is defined as prior to World War II – tends to be scattered across archives, published works, and web sites, and may be embedded in original-language sources that are inaccessible to researchers. Consequently, those who wish to tackle historical subjects are faced with the prospect of assembling their own data, more or less from scratch – a laborious process. Each dataset is the product of an enormous investment of time and resources. Not surprisingly, these datasets tend to remain proprietarial.

We recognize that any original data collection effort requires a period of “copyright protection” in order to compensate the collector for her investment. However, from the perspective of scholarly progress this proprietarial period ought to be of short duration and should be lifted at the moment a dataset is employed in a publication. Although this standard is generally subscribed to it is not always adhered to. Consequently, historical data tends to be submerged for long periods and released in a piecemeal fashion. *After* going public, datasets tend to remain in a static form. They are not regularly updated, revised to incorporate the suggestions of scholars, or augmented from their original form. Slowly, they fall into desuetude.

Likewise, because the collection of data is undertaken in a piecemeal, uncoordinated fashion – sometimes with a single scholar’s very specific research goal in mind – we may discover a lot about topic X in site A from Dataset 1, and topic Y in site B from Dataset 2. But we have no way to make explicit comparisons across topics and sites. Extant data is highly disorganized.

Finally, the location of historical data is mysterious – hidden away in countless archives, newspapers, web sites, and published sources. Historians spend a great deal of time acquainting themselves with sources pertinent to their chosen topic, typically a narrow slice of the past. Those with a broader agenda – and therefore focused on a broader range of topics, places, and time-periods – are at pains to solve their information problem.

Heretofore, historical data has been treated primarily as a private good. This would not be problematic if the incentives offered by the intellectual marketplace favored efficient collection efforts (non-duplicating), comparable data (variable definitions shared across eras and sites), unification in a single database (or small number of functionally defined databases across which data could be easily merged), and periodic updating and upgrading. Unfortunately, the incentive structure of academics generally favors free-riding on others’ data, hoarding one’s own, and single-use datasets (datasets constructed for solving a single analytic purpose, which may or may not be useful for other questions). It is not that academics are ungenerous relative to other humans. It is, rather, that little recognition follows from data collection and organization. Most important, there exists no hegemon with the capacity to solve this collective action problem. The social sciences are poorly funded, and historical work especially so.

A contrast with *contemporary* social science data is informative. Here, a large number of governmental and non-governmental bodies work to collect, aggregate, and disseminate data. A (very) partial list includes the Center for Global Policy at George Mason University (Political Instability Task Force), the Cline Center at the University of Illinois (The Social, Political and Economic Event Database Project [SPEED], The Societal Infrastructures and Development Project [SID]), the Food and Agriculture Organization (food, agriculture, malnutrition), Freedom House

---

(democracy and freedom), Gapminder (various), the International Monetary Fund (financial and trade statistics), the Minnesota Population Center at the University of Minnesota (International Integrated Public Use Microdata Series [IPUMS]), the Organization of Economic Cooperation and Development (economics), the Quality of Government Institute (democracy and governance), Transparency International (the Corruption Perceptions Index), USAID (the Demographic and Health Survey, among others), the UNICEF & United Nations Development Programme (GDP formerly, human development currently), World Bank (governance indicators, the Database of Political Institutions, the World Development Indicators), and the World Health Organization (health outcomes). Most importantly, governments collect data on their activities and on their citizens – often aided by the foregoing institutions. Consultancy firms also collect a wide variety of economic and political data which may be released after several years and often prove useful to scholars. Yet, all of these efforts are focused primarily on current years or on the recent past.

Only a few historical databases exist, and these tend to be focused on highly specialized topics and generally do not extend further than 1800 or 1815. The Cross-National Time-Series Data Archive, includes a variety of political, economic, and social data sources, but is proprietary and therefore cannot be fully integrated with other data sources (Banks 1994). Correlates of War pertains to intra- and inter-country conflict (Singer, Diehl 1990). Polity includes a number of variables measuring different aspects of polities, mostly centered on the concept of democracy (Marshall, Jaggers 2007). All three databases treat sovereign countries as the principal unit of analysis, with little or no attention to non-sovereign entities or subnational levels. All are maintained at nominal cost by scholars, supplemented with minimal government funding, and thus operate on a shoe-string.

In addition, there are a number of data repositories, one of which is explicitly historical in focus. These include the Dataverse Network Project, the World-Historical Dataverse, and the Inter-University Consortium for Political and Social Research. However, these are little more than storage facilities. If widely used (as we hope they will be), and if consolidated (so that only one or two come to dominate the field), they solve one problem associated with historical data: obscurity. But they do little to stimulate data collection, dissemination, or unification (such that variables gathered by different researchers can be easily and fruitfully merged).

Finally, we must consider the form and format of the data that finds its way into various historical datasets. Extant datasets often suffer from ambiguity about sources, coding procedures, and other elements that mitigate the accuracy and reliability of the recorded data. While some of these faults are inherent to the enterprise – collecting data globally is, after all, a daunting task – others may be mitigated through careful data-collection and data-preservation procedures, as outlined in subsequent sections of this report.

Data Collection

Although originally focused on the topic of colonialism, we quickly came to terms with the fact that the causes and consequences of colonialism are intertwined with every other element of long-term historical development throughout the world. Consequently, no a priori boundaries were placed around this data-collection effort except those mandated by limited resources (i.e., time and money).

Topics were considered eligible for the database if (a) they had some bearing on long-term patterns of development or on other topics of concern to the social sciences and humanities, (b)

5 www.umich.edu/~cowproj/
6 URLs are as follows: Dataverse Network Project (http://thedata.org/home), World-Historical Dataverse (www.dataverse.pitt.edu/), Inter-University Consortium for Political and Social Research (www.icpsr.umich.edu/icpsrweb/ICPSR/).
data was available for more than one region of the world, (c) considerable data was available for periods prior to 1950, (d) data incommensurabilities (across sources, regions, and time-periods) were not insuperable, and (e) data was fairly easy to assemble. Our strategy was to focus on the low-hanging fruit, leaving the higher branches for later iterations of the project.

With these criteria in mind, we canvassed an extraordinarily wide variety of sources — published and unpublished, hard copy and digital, archival and non-archival, English and foreign language. This canvas was conducted in correspondence with scholars and archivists around the world, who played key roles in the search.

Several hundred sources are currently integrated into the database. Among these, several serials and compilations loom large, most importantly, the Almanach de Gotha, the British Colonial Office Lists, the British Colonial Office Reports, the CHAT dataset, the International Historical Statistics series, the Statesman’s Yearbook, and the Union Postale Universelle Statistique Generale. Further information about these key sources is contained in Appendix A. It should be noted that each source collates data obtained from other sources, who may in turn have collated data from other sources. Thus, we participate in a multi-layered undertaking, the product of many hands and many years’ labor.

It should be clear that our aim is not to create an immaculate database — free from error, perfectly precise, and entirely comprehensive. It is, rather, to reach a level of accuracy, precision, and comprehensiveness comparable to that displayed in the work of specialists on these subjects. It is important to bear in mind that this database represents a compilation of existing knowledge about the past. As such, it must suffer the same shortcomings that extant historical knowledge presents.

However, insofar as traditional historical research has illuminated parts of the human past — insofar, that is, as we know anything at all about history — CLIO should help to unify that knowledge, making it more accessible, and providing mechanisms for estimating error in a systematic fashion.

To that end, several concrete steps were taken. First, we are scrupulous in tracking the source — and where available the primary source — of each data point. Likewise, we strongly urge end-users to cite the source — and, where possible, the primary source(s) — of all data obtained from the database. Correct attribution is essential in any scholarly endeavor, and particularly in a venture that combines data from so many diverse sources. Second, multiple data sources are employed wherever available, and especially wherever we have reason to doubt the veracity of a source. Users of CLIO can pick and choose among multiple sources, aggregate across them, impute missing data, or construct uncertainty estimates based on the frequency and reliability of scores for the same time and place. Third, coders register uncertainty about a data point explicitly wherever this is noted in the source — by checking “estimate,” “uncertain,” or “circa” (for dates). Fourth, two Remarks fields — associated with variables and observations, respectively — allow coders to include remarks on any facet of the data that might affect its viability or interpretation. Finally, PDF copies of hard-copy data sources and digitized copies of dataset sources are preserved, wherever possible, and linked directly to the database so that users may easily reference the source.

This information system should make the task of any future re-coding easier and allows for a full reporting of the procedures employed. Replicability must be maintained: users of the database should be able to reconstruct the origin of each “primary” piece of data as well as the construction of all “secondary” data.

We trust that an additional audit for faulty or incomplete data will be provided by the community of users. Once the data is published, we hope to facilitate user input with a “wiki” format that allows anyone to comment on particular features of the data, or provide additional data.

Organization of the Database
CLIO is best explained by considering a single row (record or observation) in the database. A row may contain (1) the location of interest, (2) full name of location (including superordinate entities), (3) alternate names for the location (if any), (4) GIS coordinates (polygon or lat/long points), (5) the continent within which the location is situated, (6) the sovereign state (empire or nation-state) within which the location is situated (if any), (7) the date range of the data (e.g., Jan 1-Dec 31, 1889), (8) circa (an indication of uncertainty about the date), (9) a variable (e.g., population), (10) attributes describing that variable (e.g., sex: male/female), (11) the source of the data, and (12) miscellaneous remarks pertaining to the data, source, or its interpretation.

Each appears as a column in the query output (either HTML or CSV). Note that only the location (1) is necessary, though (9) and (10) would also be necessary for most sorts of queries. Others cells may be empty. Indeed, they may not even be relevant to the query. (The system allows users to choose which columns to include in the output.) Let us explore these features in more detail.

Each location is recorded in English, if available, and in alternate spellings and languages, as needed. A variety of numeric codes are available based on systems devised by Arthur Banks, COW, GeoNames, IMF, ISO, Polity, World Bank, and so forth, which the user may specify.

In addition, GIS coordinates are noted wherever possible. Polygons (GIS shape files) are employed to represent the borders of sizeable units (at a low or medium resolution) such as nation-states, empires, and colonies, as well as some of the larger regional units of empires and nation-states over the past two centuries (from 1815 to the present). All other locations are represented with point coordinates, drawn from GeoNames. This includes micro-states and micro-colonies, the smaller regional units of nation-states and empires, as well as all municipalities. (Large units do not receive any GIS representation prior to 1815.)

A primary source for polygon geometries representing nation-states, empires and colonies is the Digital Chart of the World. For sub-national units, the primary source is the Global Administrative Unit Layers (GAUL). Point geometries were generated from point coordinates gathered through internet search engines. If point coordinates cannot be located through a simple search, GeoNames serves as a fall-back source. (GeoNames should not be employed as a primary source because the database often lists multiple entries for one location, creating confusions that are difficult to untangle.) Where necessary, we relied on auxiliary sources such as Wikipedia entries to help resolve ambiguity associated with locations.

All locations are classified into a simple taxonomy (ontology), as follows:

A. State Any political unit enjoying formal (legal-juridical) sovereignty according to international (European) norms – not necessarily de facto sovereignty.

1) Empire A sovereign unit (a) that is multiethnic, (b) where a clear and fairly consistent distinction exists between a stable territorial core or metropole (presumed to be superior, both culturally and politically) and a periphery, (c) where the core rules in an autocratic fashion over the periphery, (d) where the periphery has no effective or constitutional option to secede.

2) Nation-state A sovereign unit that (a) has relatively fixed (de jure or de facto) borders, (b) violates one or more of the defining attributes of an empire. Not defined by size, either territorial or demographic, and thus includes city-states where they enjoy sovereignty. May be multi-ethnic (the national idea is an aspiration, not always achieved).

Sub-state units Parts of an empire or nation-state.

a) Semi-sovereign units Peripheral area of a sovereign state where formal sovereignty is held by larger unit but the smaller unit is not formally
integrated on an equal footing with the core. E.g., possession, colony, territory, dependency, protectorate.

b) Non-sovereign units Area of a sovereign unit that is formally integrated on an equal footing with the core. E.g., region, district, state, province, municipality, village.

B. Unclaimed, disputed, or internationally governed territory A residual category including all areas lying outside the purview of sovereign units (as defined above). E.g., Antarctica, Western Sahara.

In addition, we record continents, as follows: Asia, sub-Saharan Africa, Middle East, Europe, Latin America, North America, and Oceania.

Imperfections in this taxonomy will be readily apparent to readers. The distinction between empire and nation-state is strained, e.g., with respect to Russia, the Soviet Union, or Great Britain in previous eras. (Those who wish to ignore this distinction may simply resort to the superordinate category, state. However, here one faces the difficulty of defining sovereignty in the case of contested regions (e.g., Western Sahara). Some locations (e.g., Vatican city) are inherently difficult to classify. Naturally, the taxonomy becomes more difficult to apply as one moves back in time to historical eras when most political units did not have a fixed identities or locations and sovereignty was not a well-established concept.

Over time, we expect that this taxonomy will be developed and further subclassifications added. For example, it may be useful to distinguish between regions and municipalities. We may also allow for additional taxonomies, cross-cutting the above. No taxonomy is perfect, and no taxonomy will serve all goals. However, for purposes of organization it was necessary to impose an initial taxonomy on the thousands of locations that populate the database.

Locations have a historical existence which stretches back in time. Any information associated with that location – e.g., its population, land area, superordinate units, GIS coordinates – is understood to refer to that unit as it existed at a specific point in time. For example, any data associated with the United States in 1813 refers to the United States as it existed in 1813.

There is one exception. In some instances it is desirable to create a longitudinal (panel) data with units having consistent boundaries through time. A fixed-boundary variable refers to the population of all those residing within the boundaries of the US at some designated point in time (say, 2005). So, the population (fixed) of the US in 1813 would consist of all those residing within the 2005 boundaries of the US in 1813 (i.e., including parts of the Spanish and British empires in North America).

Principal sources for coding locations and time-periods within this taxonomy include the Correlates of War (COW) Territorial Change dataset and WorldStatesmen.org. COW provides the default coding for names and dates, with WorldStatesmen as a secondary resource. Additional materials were consulted as necessary (i.e., for obscure locations).

Dates are recorded as intervals, i.e., Begin and End dates. Specific dates are noted, wherever available. Where no specific day within a calendar year is noted the default coding is June 30, i.e., the middle of the year. Dates are sometimes assigned to data that may not have an original date. For example, data described in a source as “precolonial” will be assigned the date of the last year prior to colonization. More or less permanent geographic features (e.g., “Latitude of capital city” or “Landlocked”) are assigned to every year for which a country or other political unit is in existence. Where the date is ambiguous but clearly applies to contemporary features of a unit, we assign the date of publication of the source. Thus, “yellow fever” is dated at 1998, the year of publication for the study from which the data is drawn. Where the data refer to a cumulative total – e.g., “total
number of slaves shipped from country between 1500 and 1900” – the begin date is 1900 and there is no end-date.

Table 1 provides an overview of the variables, the columns containing what is usually the quantity of interest. These are arranged into five broad categories: Politics, Economics, Infrastructure (transport, communications), Demography/health, and Education.

If the units for a variable are standard across all sources then this information is incorporated into the variable name or its definition. If they are different, this is treated as an attribute of the variable. For example, a government revenue statistic may be expressed in local currency (LCU), British pounds, or some other currency. Currency, therefore, is an essential attribute of government revenue, and appears as a column next to the latter in the output table.

Unless otherwise specified, variables are recorded in “raw” numbers, using whatever format was employed by the original source and employing the lowest level of aggregation possible, so as to minimize error and maximize flexibility for the end-user. Of course, sometimes a higher level of aggregation is required in order to achieve comparabilities across space and time. But, usually, clarity and agreement is higher at a lower level. In any case, low-level variables can easily be aggregated by the end-user to form higher-order concepts.

For all data, we record notations from the original source that indicate a data point is an estimate, is uncertain, or – for dates – circa. These codings are based on what the source says, not the coder’s intuitive sense of accuracy and precision. Thus, if the field is left blank for an observation this should not be taken as an indication that the data point is secure. By the same token, if a data point is accompanied by estimate, is uncertain, or circa it is reasonable to assume that this data point is less secure than other of the same type.

For each datum, a source is listed. This refers to whatever source we employed to code the cell, i.e., our immediate source. Each source is associated with a Remarks field, where general comments pertaining to that source may be entered. This includes comments regarding primary sources used to compile that source, where indicated (not all secondary sources are fastidious in this respect). References are abbreviated, but full references – including page numbers (where relevant) and the usual bibliographic information – may be obtained by clicking on the abbreviated name. Where the source is saved in PDF format (as most published material is) the user may click through to observe that document.

Each observation (row, record) is also associated with a Remarks field, where all additional notes pertaining to the observation may be entered. This includes commentary on the viability of the source, disagreements apparent across sources, special coding rules, or any other facet of the data that might be relevant to its interpretation.
Table 1: Variables

**Government revenue**  Government receipts of tax and non-tax revenue.
  - Currency: LCU, USD, ...
  - Type of government: All/Central(federal)/Sub-central/Provincial/Local.
  - Type of revenue: Total/Total from taxation/Consumption taxes/Customs taxes/Direct taxes/Excise taxes/Income and property taxes/Industrial and turnover taxes/Petrol and Tobacco taxes

**Government expenditure**  Government spending on goods and services.
  - Currency: LCU, USD, ...
  - Type of government: Central(federal)/Sub-central/Provincial/Local.
  - Category: Total/Agriculture/Consumers/Edcucation/Justice/Marine/Ministry of Finance/Ministry of the Interior/Pensions/Police/Political Administration/Railway/Telegraph/Telephone/War/Other

**Public debt**  Government liabilities to both domestic and foreign lenders.
  - Currency: LCU, USD, ...

**Public debt, interest service on**  Interest accrued on unpaid government debt.
  - Currency: LCU, USD, ...

**Public debt, virtual service on**  The unpaid part of foreign service for nations defaulting on debt.
  - Currency: LCU, USD, ...

**Public debt, serviced in gold**

**Military personnel, number of**
  - Unit: Total/Native or sovereign/Colonial/Cavalry/Foreign or mercenary/Infantry/Naval
  - Individual/Battalion/Company/Regiment/etc...

**European officials, number of**  Refers to the unit (e.g., colony) in which they are serving.

**Leaders**  Name of leader (proper noun).
  - Terms served
  - Number of previous terms
  - Start date of term
  - End date of term
  - Birth date
  - Death date
  - Gender
  - Type of entry: regular/irregular/foreign imposition
  - Type of exit: irregular/natural death/still in office
  - Exit code (detailed coding of exit from power, particularly for irregular exits): by domestic popular protest with foreign support/by domestic popular protest without foreign support/by domestic rebel forces with foreign support/etc...
  - Post-tenure fate: No punishment/exiled/imprisoned/killed

**State antiquity value, raw**  Summary value indicating the historical duration of governments for each 50-year period between years 1 and 1950 CE, determined by three criteria: existence, local control, and congruence with present territorial borders.
  - Borders: current. Current borders projected back in time. For example, data for the United States in 1800 refers to aspects of state-ness in 1800 across the territory that is currently part of the United States.
  - Discount rates: 0%/0.1%/1%/5%/10%/50%

**State antiquity value, cumulative**  Combined raw state antiquity values for a given location from years 1-1950 CE, multiplied by rates between 0-50% in order to discount the influence of the past.
  - Borders: current. Current borders projected back in time. For example, data for the United States in 1800 refers to aspects of state-ness from 1500-1800 across the territory that is currently part of the United States.

**Land area**  Total territorial area pertaining to a location.
Units: square kilometers/square miles/ acres/ etc.

**ECONOMICS**

*Land distribution*  The distribution of total land holdings (disposable land per farm) across the total extent of agricultural land, expressed as a Gini coefficient.

*GDP*  Gross domestic product, or total market value of all goods and services produced within a country in a given year.
  
  Currency: LCU, USD,…

*Income per capita*  Income per person for a given location; total income divided by population.
  
  Currency: LCU, USD,…

*Trade, aggregate*  Total imports and exports for a nation with all trading partners.
  
  Currency: LCU, USD,…

*Trade, bilateral*  Total imports and exports between a nation and one trading partner.
  
  Currency: LCU, USD,…

*Imports, total*  Economic value of total goods and services brought into a nation through trade.
  
  Currency: LCU, USD,…

*Exports, total*  Economic value of total goods and services shipped out of a nation through trade.
  
  Currency: LCU, USD,…

**INFRASTRUCTURE**

*Railroads, length of lines open*  
Units: kilometers/miles/etc…

*Railroads, freight traffic*  
Units: cubic tons/thousand cubic tons/etc

*Railroad, passenger traffic*  Raw number of passengers per year.

*Railroad, expenditure*  Government-owned railroads except where indicated as privately owned corporation.
  
  Currency: LCU, USD,

*Railroad, revenue*  Government-owned railroads except where indicated as privately owned corporation.
  
  Currency: LCU, USD,…

*Roads, length*  
Units: kilometers/miles/etc…

Maintenance type: public/private

*Telephones in use*  

*Telegraph lines, distances covered (miles)*

*Telegraph lines, miles of wires*

*Telegraph lines, number of stations*

*Postal items, pieces of mail sent*

  *Domestic*

  *International (Receipt/Expedition/transit)*

*Telegrams, number sent*

*Merchant ships, number registered*  

  *Type: steam/motor/sail*

*Motor vehicles, number in use*  

*Motor vehicles, passenger km*

*Air cargo, ton/km*

*Radio, sets in use*  

*Television, sets in use*  

*Postal service, number of bureaux*  

*Postal service, revenue*  

  Currency: LCU, USD,…
Postal service, expenditure
  Currency: LCU, USD,…
Postal service, number of boxes
  Domestic
  International (Receipt/Expedition/transit)
Postal service, number of postal personnel

DEMOGRAPHY, HEALTH

Population
  Units: raw, 1,000s,…
  Total: Yes/No [if Total, all remaining questions except Individual/Family are irrelevant]
  Unit: Individual/Family
    74/75-79/80+ /unknown.
  Gender: Male/Female
  Ethnicity: English/Indian/Russian/etc…
  Religion: Catholic/Jewish/Muslim/etc…
  Race: Caucasian/Hispanic/African-American/etc…
  Occupation: categories to be developed as data is entered
  Origin: Indigenous/Foreign

Crude birth rate  Births per 1000 people in a given year.
  Male/Female
  Ethnicity: English/Indian/Russian/etc…
  Religion: Catholic/Jewish/Muslim/etc…
  Race: Caucasian/Hispanic/African-American/etc…

Total fertility rate  Average number of children that would be born to a woman in her life time, given age-
  specific fertility rates and reproductive period.

Infant mortality rate  Number of infant deaths per 1000 live births in a given year.
  Gender: Male/Female
  Ethnicity: English/Indian/Russian/etc…
  Religion: Catholic/Jewish/Muslim/etc…
  Race: Caucasian/Hispanic/African-American/etc…

Life expectancy  Expected longevity at birth based on current age-specific mortality rates.

Crude death rate  Number of deaths per 1000 people in a given year.
  Gender: Male/Female
  Ethnicity: English/Indian/Russian/etc…
  Religion: Catholic/Jewish/Muslim/etc…
  Race: Caucasian/Hispanic/African-American/etc…

In-patients  Number of patients undergoing medical treatment in a hospital for a period of 24 hours or greater.
  Disease: tuberculosis/malaria/yaws/etc…
  Procedure: heart transplant/appendectomy/pacemaker/etc…

Out-patients  Number of patients undergoing medical treatment either in a clinic or surgery center, or in a
  hospital for a period of 24 hours or fewer.
  Disease: tuberculosis/malaria/yaws/etc…
  Procedure: tonsillitis/cataract/dialysis/etc…

Inoculations  Number of inoculations per year.

Hospitals
  Hospital beds
    Care type: total/acute/chronic/longterm

Deaths from epidemics
  Disease: cholera/malaria/yaws/etc…
EDUCATION

Schools, number of
  Total: Yes/No
  Level: Primary/Secondary/Tertiary(university)
  Sector: Public/Private
  Lay status: Secular/Religious

Students, number of
  Total: Yes/No
  Units: Enrollment Number/Actual Attendance
  Level: Primary/Secondary/Tertiary(university)
  Sector: Public/Private
  Lay status: Secular/Religious
  Gender: Female/Male

Teachers, number of
  Total: Yes/No (default: No)
  Level: Primary/Secondary/Tertiary(university)
  Sector: Public/Private
  Lay status: Secular/Religious
The Shape of the Data [section incomplete]

A total of ?? unique political units are contained in CLIO. Of these, ?? are empires, ?? are nation-states, ?? are semisovereign units, and ?? are nonsovereign units.

A total of ?? data points are contained in CLIO. These may be disaggregated in a variety of ways.

Subdivided by region, ??% of the data points lie in Asia, ??% in sub-Saharan Africa, ??% in % in the Middle East, ??% in % in Europe, ??% in % in Latin America, ??% in % in North America, and ??% in % in Oceania.

Subdivided by topic, ??% of the data points lie in Politics, ??% in Economics, ??% in Infrastructure (transport, communications), ??% in Demography/health, and ??% in Education. Among variables, population offers the best coverage (by far), followed by ??.

Subdivided by type of political units, ??% of the data points describe features of empires (as a whole, not locations within empires), ??% describe nation-states, ??% describe semisovereign units, and ??% describe nonsovereign units.

Historical coverage stretches from ?? [earliest data point] to ?? [latest data point]. Of course, the data is not evenly distributed. In order to visualize this distribution, Figure 1 displays a frequency histogram of the data aggregated by decade. Constr?? So far as we can discern, quantity and quality generally co-vary. Where the quantity of data is highest its quality (i.e., precision) is also probably highest.

GIS polygons are available for a total of ?? units. This includes ?? empires (??% of the total) and ?? nation-states (?? of the total). GIS points are available for ?? units. ?? geographic entities are not represented by either polygons or points. B As noted, polygons extend back only to 1815; prior to that, all GIS data takes the form of points and there is no representation for larger units.

What more can be said by way of summarizing the data??

Conclusions

Academics and policymakers recognize that every challenge facing the world today has a history – in most cases, a very long one. These histories matter, not simply because we wish to understand the causes of a current problem but also because we wish to understand its trajectory. One’s perspective on an issue is necessarily affected by one’s knowledge of its past. For example, one is apt to consider globalization quite differently if the topic is viewed in the context of current history (the past half-century), modern history (the past two centuries), written history (the past several millennia), or human history (since the advent of hominids).

We do not mean to imply that the longue duree always offers a superior vantage point. Sometimes, a present-ist perspective is justified. Our assumption, however, is that the historical perspective(s) should always be taken into consideration. Sometimes, there is valuable information to be gained by examining the long-run evolution of a topic or problem, and one cannot reasonably judge the issue until such an accounting has been accomplished. Indeed, for most topics of current interest – including poverty, disease, premature mortality, social conflict, autocracy, poor governance, environmental degradation, and globalization – a strong prima facie case can be made for the importance of history. The project envisioned here should serve to enhance the relevance of history for challenges facing the world today by making the past accessible to us in ways that it previously was not (and, prior to the computer age, could not be).

Of course, data recovery for the fifteenth century will never match the quality and quantity of data available for the twentieth century. Yet, we are equally convinced that much more can be done to gather together the data that lies out there already (in the form of primary sources, secondary sources, and specialized datasets) and to make this information more widely available to
scholars. We regard this project as an essential step in this direction, one that will greatly enhance our ability to engage history as part of our quest to understand the present and future. In much the same way that Polity IV, Correlates of War, and the World Development Indicators now function as standard references for the study of nation-states in the modern era, we anticipate that CLIO may serve as a leading source of evidence for those investigating periods prior to the twentieth century – thus serving as a jumping-off place for research on a great variety of topics in demography, economics, history, political science, sociology, and related fields.

To access the database, visit the CLIO website <CLIOWorldTables.org>. Here one will find a portal for accessing the data, which can be viewed in HTML or downloaded in a number of data formats. Prior to downloading, we ask that you register on the site. We also ask that you cite any data drawn from the site according to the original source, noting that the data was drawn from CLIO World Tables on a specific date. This will clarify provenance and give appropriate acknowledgments. We also request that you contribute your own historical data and make suggestions for sources that we might not be aware of. These will be integrated periodically to the database.

---

7 This helps us maintain site security, gives us a bit of information about our end-users (which we can report to our funders), and helps us chart the future direction of CLIO. It also allows us to inform you when there are important additions or renovations to the site (if you indicate a wish to be contacted). All contact information is strictly confidential and will not be shared with anyone or used for commercial purposes.
References


Appendix A:
Key Sources

Almanach de Gotha (AdG). First published in 1763, the initial volumes primarily recorded the familial structure and government appointments of European nobility, including only cursory demographic information for the various populations of Europe. In the 1780s, the publishing house of Justus Perthes, located in the German city of Gotha, took over annual publication of the AdG. As the series expanded its geographic coverage across Europe, it expanded its statistical coverage to detailed population, infrastructure, and economic data as well. By the middle of the 19th century, AdG contained comprehensive statistics for not only European states proper, but also the colonies and territories encountered through European expansion and imperialism. The series continued to expand and refine its coverage until publication ceased in 1944, with the Soviet invasion of Germany in the Second World War. About 160 of the AdG’s original 182 volumes currently survive in print or electronic form. In employing the AdG, we have digitized and integrated the population, infrastructure, economic, and spatial area data from the majority of extant volumes.

British Colonial Office Lists (BCOL). First published by the British Colonial Office in 1861, this series provides annual demographic and financial information for all possessions held by the British Empire, as well as brief descriptions of their acquisition, history, infrastructure, and administration. The series continues until the era of major decolonization in the 1960s. In integrating the data from this source, we have digitized population, government, and infrastructure data from the majority of surviving volumes. (While the series provides some educational statistics for the British possessions, we have opted to integrate more comprehensive and consistent data found elsewhere.)

British Colonial Office Reports (BCOR). Published by the British Colonial Office, BCOR provides comprehensive information for the British-administered trusteeship and mandate territories between 1920 and 1963. Structured similarly to the British Colonial Office Lists, the BCOR includes demographic, financial, infrastructure, and educational statistics for various territories under British administration. In integrating the data from this source, we have digitized the volumes for the territories of Cameroon, Iraq, Palestine and Transjordan, British Togoland, and Tanganyika; these are the territories for which the Reports provide comprehensive, recurring coverage.

Cross-Country Historical Adoption of Technology (CHAT). This dataset contains information on 104 different types of technologies developed in the last two centuries. This dataset aims to display the effect of diverse technologies on countries’ total factor productivity, using annually observed data from the agricultural, financial, health, industrial, communications, and transportation sectors. The dataset collectors, Comin and Hobijn, have assembled mainly original source data, merging sources where necessary to create consistent variable definitions over time. We have integrated data primarily from the health, communications, and transportation variables in the CHAT dataset, citing both the dataset authors and the primary sources from which they draw.

International Historical Statistics (IHS). This source represents a remarkably comprehensive compilation of statistics from 1750 to the present century, published in three volumes: Africa, Asia, and Oceania (1993), Europe (2003a), and the Americas (2003b). (Later editions of IHS update the data but do not alter the historical data; thus, our use of different volumes affects only the final years.
in the series.) B.R. Mitchell, the supervising editor, primarily draws upon government publications, censuses, and yearbooks to present global comparative data for the following set of variables: population, vital statistics, labor, agriculture, industry, external trade, transportation, communication, prices, education, and national accounts. We have integrated data from many of IHS’s variables into our data set, taking care to cite both IHS and the primary sources from which it draws.

**Statesman’s Yearbook (SY).** Originally associated with Thomas Carlyle, SY was first published in 1864 as a comprehensive political reference for those in public or diplomatic office. Published annually, the yearbook now provides historical, political, social, economic, and demographic information for all countries of the world. In referencing sources, the SY relies heavily on official statistical publications from the United Nations, World Bank, International Monetary Fund, European Union, and other major international organizations. We have both integrated data from the SY and also employed it as a reference list in mapping the higher-level locations (e.g., modern nation-states) in our database.

**Union Postale Universelle Statistique Generale (UPU).** The UPU began publishing annual reports on the postal statistics of its member countries in the 1870s. Currently, this information is available both in report format and, for more recent years, as a publicly searchable online database through the UPU website (www.upu.int). We have digitized the postal statistics published in the UPU’s historical reports, and we have also integrated spatial area information provided in the online database for years 1980-2010.
Appendix B:

Database Architecture

The database is implemented with Django, an open source high-level Python web framework. Data models were created to represent the multiple variable tables and relationships, as well as tables and relationships for geographic and political entities. The underlying database used is PostgreSQL, an open source object-relational database system. Django allowed us to work on a higher object-oriented level, abstracted from the underlying SQL database. It also provided us with a built-in administrative website that was used for data entry, auditing and cleaning; and a geographic add-on to facilitate storing geographic points and polygons.

The Django project falls under the colonialismdb namespace, split into Django “apps.” They are common, sources, and one for each for the following categories: population, government, education, and infrastructure, as illustrated in Figure B1.

The common app contains the model classes that are used and inherited from by all other apps’ model classes. Political entities and geographic locations are derived from the base model class PoliticalUnit. Geographic locations are represented by the subclass Location. Period-specific ones are represented by the subclass TemporalLocation. Geographic features (points and multi-polygons) are separately stored in classes derived from BaseGeo to which Location classes point. Another commonly used model class is the abstract class Category, from which model classes representing such data as race, religion, languages, etc. are derived. Finally, there is the base abstract model class BaseDataEntry, from which all apps’ data entry classes are derived.

All variable apps have a MainDataEntry model class that holds all data entries by location and time period. In addition, app-specific category classes are stored in their respective apps. The sources app contains those model classes specific to data sources, data tables, and the digital data files from which they were retrieved.

Present in all the apps are also various “admin” classes needed for the administrative website. They specify how different data is represented to the project staff, as well as manage the view/edit/activate/delete permissions available to different members. The site relies on different levels of permission to provide for some data quality control. All data entered by coders needs to be first activated by members with approval permission before it is made available to external data users. Also, the site’s data revision control keeps track of each datum’s history, e.g., additions, deletions, activations, and merges.
Appendix C:
Into the Future

A project of this nature is useful over the long run only if maintained, expanded, and improved over time. That means updating existing data, correcting errors in existing data, incorporating new datasets, digitizing data from non-digitized sources, expanding the scope and enhancing the precision of the spatial data, enhancing the usefulness and flexibility of the user interface, upgrading the hardware, and engaging in outreach to potential contributors and end-users. All of this requires time and resources, and some supervision. No system, no matter how fancy, will operate automatically.

In particular, it should be noted that a global system of anything requires a good deal of coordination. If variables, definitions for those variables, and sources are entered in a haphazard fashion confusion is apt to result and comparability will be lost. This is especially challenging as one moves beyond fairly standardized subjects to those that are not easily standardized (e.g., ethnic and racial heritage) and from global sources to more specialized sources.

Several specific goals for the next iteration of CLIO deserve mention. (We hope that readers of this report will add their own suggestions to this partial list.)

Additional variables. Over time, we would like to enhance CLIO’s collection of data in several areas. Conflict data stretching back several centuries is available from Clodfelter (2002), Harrison and Wolf (2011), and other sources. Data on global immigration is available from the British Colonial Office Lists, the British Colonial Office Reports, and Mitchell (1993, 2003a, 2003b). Data on commodities – production and land use devoted to cereals, dairy wool, cotton, cocoa, coffee, tea, livestock, mining, and so forth – is available from Mitchell (1993, 2003a, 2003b) and other sources. Changing suitability of lands for growing a variety of different crops is available from Nunn and Qian (2011). Weekly data on exchange rates, short-term interest rates, and long-term interest rates from 1870 to 1914 are available from the Neal-Weidenmier Gold Standard Database (http://ebutts05.tripod.com/nealweidenmiergsd/). These are just a few of the substantial bodies of historical data that could be incorporated into CLIO in a timely, efficient manner.

Standardized units. Although data is entered and maintained in original units, it would be helpful if queries could convert data to units of the user’s choosing. This is simple where the units are constant and interchangeable, e.g., miles and meters. It is more complicated where the terms of conversion involve assumptions about value, such as with currencies. Currently, over 40 currencies are represented in the database. It would greatly enhance the usefulness of CLIO as an on-line database if we could build widgets to allow for instantaneous conversion to common currency (perhaps USD or British pounds). An additional feature of great utility would be the capacity to convert currency values into constant (inflation-adjusted) currency. We realize that many assumptions are required for these tasks (see MeasuringWorth.com), and we do not wish to lend a false sense of precision to the enterprise. However, if properly caveated, this goal seems worthwhile. Indeed, any spatial and temporal comparison denominated in monetary terms depends upon it.

Ratio variables. Data is entered in “raw” formats. However, sometimes it is advisable to consider a variable as a ratio of something else, e.g., GDP, trade, population, or land area. We would like to build a simple widget that would allow end-users to create ratio variables of their choosing (from a fixed menu of possible denominators).

Composite variables. Often, there are multiple sources for a single variable (e.g., population). If these sources cover somewhat different periods or parts of the world it may be advisable to combine
them into a single composite variable, using the most authoritative source for each data point (if there is more than one).

**Interpolated data.** For some users, it may be helpful to provide adjunct variables with missing data interpolated through time. That is, for a single time-series, missing data within that series (in between extant data points) would be linearly interpolated.

**Spatial data.** We have begun the process of representing locations with GIS vectors. However, the collection of polygons and points is incomplete and imprecise. In some cases, the geographic entity simply cannot be represented using polygons from any of the digital sources mentioned in the text. In these instances, one may search for maps scanned from historical atlases and posted online or other digital sources. These maps can be imported into ArcGIS, georeferenced to the DCW shapefile, and digitized by tracing boundaries on the maps.

In addition, the spatial-data user interface should be enhanced. Ultimately, we would like to enable users to choose a variable and display a map showing data availability for that variable (across a specified date range). Users might also zoom to a region of interest on the map and click on the map, returning all information the system contains for the location the user clicks on. In addition to data availability, displaying information on a map helps to clarify patterns and potential causal relationships and may suggest new relationships that would not otherwise occur to a researcher.

**User comments.** There are many elements of uncertainty in an enterprise such as this, as discussed. In order to identify errors, ambiguities, and missing data we would like to make it easy for users to comment on various aspects of the data – sources, variables, data, observations, and so forth. This requires a system for allowing users to (a) identify the feature they wish to comment on and then (b) to insert comments that can be compiled in a single page of the database as well as inserted in a specific location (so that users working in that area will see a comment that is potentially relevant to their work).

**Numeric Codes.** Data merging becomes easier insofar as standardized codes can be incorporated into the dataset. Most of these pertain to sovereign units (nation-states, empires), such as those developed by Arthur Banks, Correlates of War, International Organization for Standardization (ISO), Polity, World Bank. A few, such as Federal Information Processing Standards (FIPS) and GeoNames, are more far-ranging.

**An improved taxonomy.** A number of improvements could be made to the taxonomy of locations. Principal among these is the specification of a mid-level location for municipalities and other small units. At present, we are able to direct the user to the sovereign unit (empire or nation-state) and continent, but not necessarily the region or semisovereign unit within which it resides. This is problematic for cities with common names that are likely to have multiple entries within a single sovereign unit.