Cliodynamics: The Journal of Quantitative History and Cultural Evolution

The Equinox2020 Seshat Data Release

Peter Turchin*[3,4], Daniel Hoyer*[2,5], James Bennett*[6], Kiran Basava[1], Enrico Cioni[1], Pieter François[1], Samantha Holder, Jill Levine[2], Selin Nugent[7], Jenny Reddish[4], Chelsea Thorpe[8], Sal Wiltshire[8], Harvey Whitehouse[1]

[1] *University of Oxford*
[2] *The Evolution Institute*
[3] *University of Connecticut*
[4] *Complexity Science Hub, Vienna*
[5] *George Brown College*
[6] *University of Washington*
[7] *Oxford Brookes University*
[8] *Independent researcher*

This report describes the current canonical time-series dataset named "Equinox2020," a subset of Seshat: Global History Databank data for a well-curated list of polities and variables available on the Seshat Data Browser. The report provides an introduction to the methods and procedures of the Seshat project relating to the curation and release of the Equinox2020 dataset.

Introduction

This report describes the current canonical time-series dataset named "Equinox-2020," a subset of Seshat: Global History Databank data for a well-curated list of polities and variables. The variables selected for inclusion are intended to form the basis of a number of forthcoming publications. The data are from the March 2020 snapshot of the Seshat Wiki, with minor updates in May 2020 that fixed a few errors. The data are published on the Seshat Data Browser in two ways.

- First, the Browser itself is a visual interface for browsing the data, including narrative paragraphs that explain data codes, as well as references.
- Second, the data are available as a comma-delimited spreadsheet, suitable for statistical analyses (note that for the sake of concision the spreadsheet does not include descriptive paragraphs or references).

Because Seshat is a dynamic databank that evolves as new variables are added and errors are corrected, you should always use the latest data and specify the version you have used in any analysis. The current dataset (as of May 20, 2020) is

Corresponding author's e-mail: peter.turchin@uconn.edu

Citation: Turchin, Peter, et al. 2020. "The Equinox2020 Seshat Data Release." *Cliodynamics* 11 (1): 41–50.

Equinox2020: version 2020-05-18. Data on which past papers were based are also made available for replication purposes at http://seshatdatabank.info/datasets/.

This document also provides an introduction to the Seshat Databank and explains how data is collected and curated. Separate articles (Turchin et al. 2018, 2019a, 2019b) describe different approaches for analyzing the data, dealing with, for example, various ways of aggregating and scaling data as well as handling the explicitly unknown (missing), uncertain or disputed data that are an important feature of the underlying Seshat data.

A Quick Introduction to Seshat: Global History Databank

Founded in 2011, Seshat: Global History Databank systematically collects what is currently known about the social and political organization of human societies and how they have evolved over time (François et al. 2016; Turchin et al. 2015). The overall goal of Seshat is to enable researchers to conduct comparative analyses of human societies and rigorously test different hypotheses about the social and cultural evolution of societies across the globe, spanning long periods of human history. Currently Seshat focuses on the time period between the Neolithic and Industrial Revolutions. The spatial reach is global, and eventually we plan to include in the databank information on all past societies, up to the present, for which historical or archaeological data are available.

Our unit of analysis is a *polity*, an independent political unit that ranges in scale from villages (independent local communities) through simple and complex chiefdoms to states and empires. For each polity, we code variables on social complexity, warfare, religion and rituals, agriculture and resources, institutions, well-being, and the production of public goods; changes in variable values during the period covered by the polity (generally around 100–200 years) are also recorded. Overall, the current codebook includes over 1500 variables. These variables are coded for any past polity that occupied one of our sampling locations (see below) between the Neolithic and Industrial Revolutions, roughly 10,000 BCE to 1900 CE, subject to data availability and limitations. Currently there are over 400 such polities from 35 sampling regions in Seshat. As of May 2020, the databank contains nearly 400,000 coded values ("Seshat records," see below). Equinox2020, however, publishes only a well-curated subset (47,400 records) from 374 polities and 136 variables.

Temporal and Geographic Scope

In order to assess whether different societies show commonalities in the way they have evolved, we developed a geo-temporal, stratified sampling scheme that aimed (1) to include as much variation among the sampled societies as possible in terms

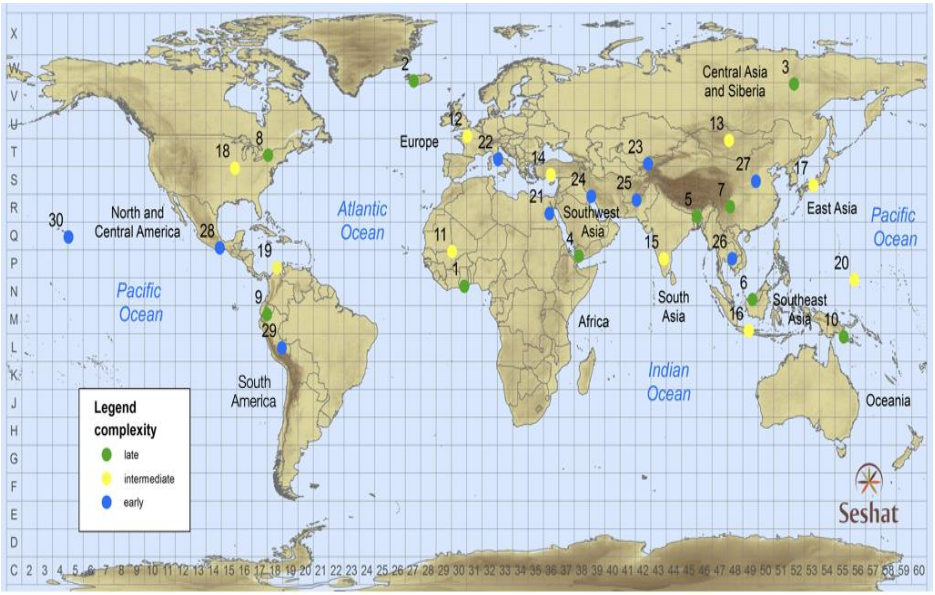

Figure 1. Locations of Natural Geographic Areas (NGAs) in the original World Sample-30. For the current list of NGAs, see http://seshatdatabank.info/databrowser/.

of social organization, and (2) to ensure representation of different parts of the world (Turchin et al. 2018). This issue is challenging as societies can expand or contract in geographical space, appear or disappear in the historical and archaeological records, and show varying degrees of continuity with earlier or later societies.

To overcome these issues and ensure that we collected data in a systematic manner, we divided the world into ten major regions (named in Figure 1, e.g., South Asia, Europe). Within each region we initially selected three natural geographic areas (NGAs), our basic geographical sampling units. Each NGA is defined spatially by a boundary drawn on the world map that encloses an area delimited by naturally occurring geographical features (for example, river basins, coastal plains, valleys, and islands). The extent of the NGAs does not change over time, and NGAs thus act as our fixed points that determine for which societies we collected data. The data themselves, however, are collected not for an NGA, but for the entire society, or polity, that happened to occupy the NGA at a given time. Each NGA, then, serves as geographic "anchor" from which we generate a list of all the polities that occupied it over the course of history. Such a sampling approach allows us to be

consistent and methodical about designating societies for which we gather data. It also allows us to construct spatially anchored time series, as long as it is understood that the spatial extent of sampled societies fluctuates with time (as polities rise, expand, go into decline, and collapse). Note that some polities, like the Roman Empire, can be present in several NGAs at the same time; the data for the polity, however, is represented only once in the downloadable data sheet. Likewise, each polity that occupied an NGA has dates associated with the period during which it occupied the NGA in question, though the polity itself will have a different set of dates indicating its own duration (see Table 1 for an example).

Table 1. Partial list of polities associated with the Upper Egypt NGA. For a full list, see http://seshatdatabank.info/databrowser/upper-egypt.html.

NGA	Polity	Duration in NGA	Polity duration
Upper Egypt	Egypt – New Kingdom Thutmosid	1550–1294 BCE	1550–1294 BCE
Upper Egypt	Egypt – New Kingdom Ramesside	1293–1071 BCE	1293–1071 BCE
Upper Egypt	Egypt – Thebes-Libyan Period	1070–761 BCE	1070–747 BCE
Upper Egypt	Kushite Empire	760–656 BCE	760–656 BCE
Upper Egypt	Egypt – Saite Period	655–526 BCE	664–526 BCE
Upper Egypt	Achaemenid Empire	525–405 BCE	550–331 BCE
Upper Egypt	Egypt – Inter-Occupation Dynasties	404–343 BCE	404–343 BCE
Upper Egypt	Ptolemaic Kingdom I	305–218 BCE	305–218 BCE
Upper Egypt	Ptolemaic Kingdom II	217–32 BCE	217–32 BCE
Upper Egypt	Roman Empire – Principate	31 BCE–283 CE	31 BCE–283 CE

Within each world region we identified NGAs that would allow us to cover as wide a range of forms of social organization as possible, ensuring that we captured information about the kinds of societies that researchers have previously discussed in relation to social complexity ("states," "chiefdoms," "stratified societies," "empires," etc.) without using typological definitions of such societies or employing strong, limiting definitions about what features such societies should have. We also wanted to make sure that we captured information about societies that are not traditionally thought of as complex ("small-scale societies," "egalitarian tribes," "acephalous societies").

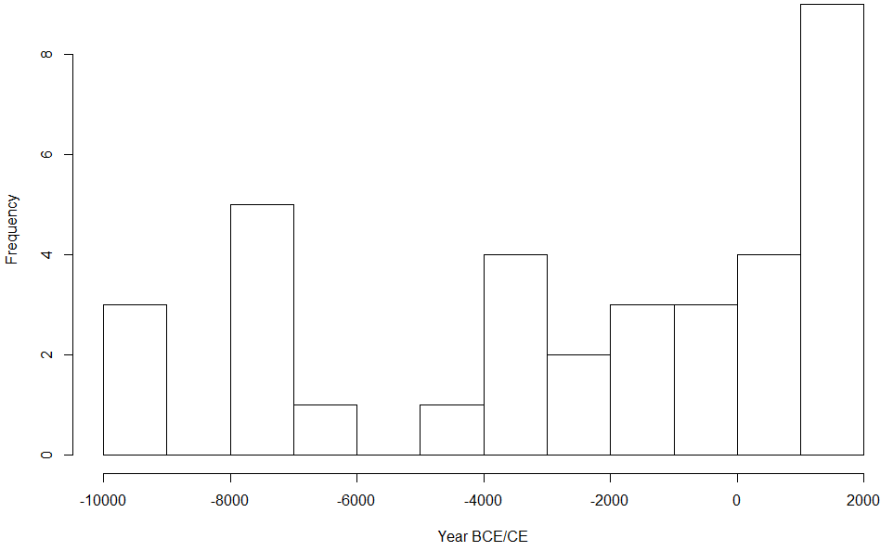

Figure 2. Frequency distribution of the starting dates for data sequences in Equinox2020. For "late-complexity" NGAs, data series are short, often starting only when European explorers reached the area. For "early-complexity" locations, data sequences extend back in time between 4,000 and 10,000 years ago. "Intermediate-complexity" cases are usually located between these two extremes.

Accordingly, within each world region, one NGA was selected that saw the earliest developments of centralized, stratified societies. We also chose a second NGA that was the opposite; ideally, it was free of centralized societies until the early modern period (after 1500 CE). Finally, the third NGA was intermediate in terms of the time when political centralization emerged within the world region. Europe, for instance, has Latium as the early-complexity NGA, Iceland for late complexity, and the Paris Basin for the intermediate NGA. Because different world regions acquired centralized societies at different times, there can be substantial variation across "early-complexity" NGAs both in the time at which our measures of social complexity start increasing and the degree of social complexity that is eventually reached at the end of our sampling period. For example, Susiana, the early-complexity NGA in Southwest Asia, has a much longer history of large societies than the Big Island of Hawai'i, the early-complexity NGA in the Pacific region.

The distribution of earliest polity starting dates for all NGAs currently in Seshat is shown in Figure 2.

Beginning in 2017, we expanded data coverage beyond the original sample of 30 NGAs. Currently, we have good data on five additional NGAs, which are included in the current canonical dataset (Equinox2020).

Data Collection

To populate the databank, for each NGA we consult the literature and chronologically list all polities that were located in the NGA, or encompassed it (see Table 1 for an example). We chose a temporal sampling rate of one hundred years, and we only included polities that span a century mark (for example, 300 CE, 400 CE, 500 CE, and so on) while omitting any polities of short duration that only inhabited the target NGA between these points. One century is short enough to capture meaningful changes in the social complexity of historical societies, but not too short to lead to *oversampled* data ("oversampling" results when the succeeding point in time contains the same data as the preceding one, thus not adding to the overall information content of the dataset). Polities are coded for the duration of the polity itself as well as the length of time it "occupied" or laid claim to the target NGA. Usually these dates are the same, but often they differ, as with the Achaemenid Empire, which was founded in 550 BCE but only claimed Egypt as its territory from 525 BCE (Table 1).

Likewise, we generally divide very long-lived polities, like the Egyptian New Kingdom and the Ptolemaic Kingdom, into different periods (Table 1). This allows us to more easily capture and represent changes during a polity's lifespan, helping to avoid reifying one particular moment within that polity's development. Our rule of thumb is to not have a single polity phase that lasts more than three centuries. This also works to limit oversampling issues, ensuring that no more than three data points will be generated by a single polity phase; though in most cases, we record changes within the polity phase as well (see below).

For those periods when the NGA is divided up among a multitude of small-scale polities (e.g., independent villages, or small chiefdoms) it is not feasible to code each individual polity. In such instances we use the concept of "quasi-polity," which is defined as a geographic area with some degree of cultural homogeneity that is distinct from surrounding areas and approximately corresponds to an ethnological "culture" (Murdock 1967; Murdock and White 1969) or an archaeological subtradition (Peregrine 2003). We then collect data for each quasi-polity as a whole. This way we can integrate (often patchy) data from different sites and different polities within the NGA to estimate what a "generic" polity was like. Such an approach is especially useful for societies known only archaeologically, for which we usually don't know polity boundaries.

Our use of polities and quasi-polities is best understood as a means of sampling the vast literature on past human societies rather than trying to impose a rigid framework on the human past. Our data coding procedures enable us to record changes in a particular variable within the lifetime of a polity, capturing variation within a polity or quasi-polity where there is such evidence. This scheme also allows us to track the gradual emergence or disappearance of a polity, as when an empire slowly disintegrates and its constituent pieces gain an increasing degree of independence from their old imperial master. Finally, we are able to flexibly incorporate multiple lines of evidence and uncertainty, as we outline below.

When gathering data into Seshat, our approach is to avoid forcing information about a past society into an arbitrary scale (e.g., "rate the social complexity of this society on a scale from 0 to 10"). Instead, and prior to collecting the data, we run a workshop, bringing together various domain experts to develop a conceptual scheme for the particular aspect of past societies that we aim to capture in Seshat. Generally speaking, we aim to use either a quantitative variable (e.g., an estimate of the population of the coded polity) or break up complex concepts into multiple simple variables that can be coded in a binary fashion (absent/present). The initial coding scheme is then tested by Seshat research assistants (RAs) applying it to several test cases, in consultation with experts (archaeologists or historians who study the coded polities). The coding scheme is documented and then refined based on suggestions from both experts and RAs and is applied to the whole sample. The codebook underlying Equinox2020 data has been published here: http://seshatdatabank.info/browser/Equinox_Code_Book.

Once a coding scheme is defined, data collection occurs in several phases. First, under direct supervision by more established scholars (professors and PhD-level researchers), RAs search published articles and books on a particular polity (with advice from a regional or polity expert on which sources are likely to be most useful) in order to find information about each variable and enter it into the databank. Second, RAs compile lists of questions on values that cannot be coded unambiguously, or on which information is lacking in the published sources, and seek help from the experts on the polity. In the final phase we ask experts to go over the data to check coding decisions made by RAs and help us fill any remaining gaps. Experts also indicate when the value should be coded as "unknown" (RAs may use the code "suspected unknown," but only experts can definitively state that something is indeed "unknown"). The current list of Seshat experts is published here: http://seshatdatabank.info/seshat-about-us/contributor-database/.

When two or more experts disagree about the value or there is ongoing debate in the literature, all choices are entered as alternative values. For example, for the Egypt – Classic Old Kingdom polity, there is disagreement as to whether the largest settlement at the time, Memphis, housed 30,000 or 50,000 people. We thus record

the variable 'Population of the largest settlement' as {30,000; 50,000}, with the curly brackets indicating that the values represent a disagreement between experts. For quantitative variables whose values are known only approximately, coders are instructed to enter a likely range [min, max] that roughly corresponds to a 90-percent confidence interval (i.e., omitting possible but unlikely or unrepresentative values).

We refer to a coded value of a particular variable for a particular polity as a "Seshat record." Seshat records have complex internal structure. First, there is the value of the coded variable. For a numerical variable, the value can be either a point estimate, or a range approximating the 90-percent confidence interval. Binary variables can take the following values: present, absent, inferred present, inferred absent, and unknown. (Numerical variables can also be coded as unknown.) "Inferred" presence or absence indicates some degree of uncertainty: direct evidence of presence (or absence) is lacking, but the RA or expert can confidently infer it. For example, if iron smelting has been attested both for the period preceding the one that is coded and for the subsequent period, we code it as "inferred present" even if there is no direct evidence for it (assuming there are no indications that this technology was lost and then regained).

Binary variables can also have temporal uncertainty associated with them. For example, if we know that iron smelting appeared in a particular polity at some point between 300 and 600 CE, we code the period previous to 300 CE as absent, the period following 600 CE as present, and the period between 300 and 600 CE as effectively "either absent, or present" (this is different from "unknown").

The second important part of a Seshat record is a narrative paragraph explaining why a particular variable was coded in a particular way. Typically, the narrative is first written by an RA, who may quote the relevant text from a reference (a book or an article) or from a personal communication from an expert. Subsequently, experts can add to it and disagree with previously recorded estimates, which are added to the existing records to pre-serve all assessments.

The third part of a Seshat record is the references to publications or other databases. Reference can also be a "personal communication" from an expert or from several experts participating in a Seshat workshop.

Moving Forward

Seshat is a living project, meaning that our data are never "fixed" but are constantly evolving, being updated with new findings, alternate interpretations, and expanded variable lists. We are currently engaged in cleaning and analyzing data on addi-

tional variables that will appear in publications soon,[1] and following that the data will be added to the Data Browser. We are also expanding the geographical and temporal scope of our databank, gathering information on new NGAs and on archaeological cultures from the early Neolithic onwards. Lastly, our data are being utilized by other projects,[2] which we strongly encourage and support. Check our project page for updates on progress, and contact us if you have questions about accessing any of our data.

Acknowledgments

This work was supported by a John Templeton Foundation grant to the Evolution Institute, entitled "Axial-Age Religions and the Z-Curve of Human Egalitarianism"; a Tricoastal Foundation grant to the Evolution Institute, entitled "The Deep Roots of the Modern World: The Cultural Evolution of Economic Growth and Political Stability"; an ESRC Large Grant to the University of Oxford, entitled "Ritual, Community, and Conflict" (REF RES-060-25-0085); a grant from the European Union Horizon 2020 research and innovation program (grant agreement No 644055 [ALIGNED, www.aligned-project.eu]), a European Research Council Advanced Grant to the University of Oxford, entitled "Ritual Modes: Divergent Modes of Ritual, Social Cohesion, Prosociality, and Conflict"; and the program "Complexity Science," which is supported by the Austrian Research Promotion Agency FFG under grant #873927. We gratefully acknowledge the contributions of our team of research assistants, post-doctoral researchers, consultants, and experts. Additionally, we have received invaluable assistance from our collaborators. Please see the Seshat website (seshatdatabank.info) for a comprehensive list of private donors, partners, experts, and consultants and their respective areas of expertise.

References

François, P., J. G. Manning, H. Whitehouse, R. Brennan, T. Currie, K. Feeney, and P. Turchin. 2016. "A Macroscope for Global History: Seshat Global History Databank, a Methodological Overview." *Digital Humanities Quarterly* 10 (4). http://www.digitalhumanities.org/dhq/vol/10/4/000272/000272.html.

Murdock, G. P. 1967. *Ethnographic Atlas*. Pittsburgh, PA: University of Pittsburgh Press.

Murdock, G. P., and D. R. White. 1969. "Standard Cross-Cultural Sample." *Ethnology* 8 (4): 329–69. doi: 10.2307/3772907.

Peregrine, P. 2003. "Atlas of Cultural Evolution." *World Cultures* 14 (1): 2–88.

[1] We have published several preprints to preview some of these findings: Turchin et al. (2018, 2019a, 2019b).

[2] E.g., Shin et al. 2020.

Shin, J., M. Price, D. H. Wolpert, H. Shimao, B. Tracey, and T. A. Kohler. 2020. "Scale and Information-Processing Thresholds in Holocene Social Evolution." *Nature Communications* 11 (2394). doi: 10.1038/s41467-020-16035-9.

Turchin, P., R. Brennan, T. Currie, K. Feeney, P. François, D. Hoyer, J. Manning, A. Marciniak, D. Mullins, A. Palmisano, P. Peregrine, E. A. L. Turner, and H. Whitehouse. 2015. "Seshat: The Global History Databank." *Cliodynamics* 6 (1): 77-107. doi: 10.21237/C7clio6127917.

Turchin, P., T. E. Currie, C. Collins, J. Levine, O. Oyebamiji, N. R. Edwards, P. B. Holden, et al. 2019a. "Crop Productivity Estimates for Past Societies in the World Sample-30 of Seshat: Global History Databank." *SocArXiv* preprint. doi: 10.31235/osf.io/jerza.

Turchin, P., H. Whitehouse, A. Korotayev, P. François, D. Hoyer, P. Peregrine, G. Feinman, et al. 2018. "Evolutionary Pathways to Statehood: Old Theories and New Data." *SocArXiv* preprint. doi: 10.31235/osf.io/h7tr6.

Turchin, P., H. Whitehouse, P. François, D. Hoyer, S. Nugent, J. Larson, A. Covey, et al. 2019b. "Explaining the Rise of Moralizing Religions: A Test of Competing Hypotheses Using the Seshat Databank." *SocArXiv* preprint. doi: 10.31235/osf.io/2v59j.