UC Riverside
Cliodynamics

Title
The Polaris2025 Seshat Data Release

Permalink
https://escholarship.org/uc/item/9wj6z91m

Journal
Cliodynamics, 16(2)

Authors
François, Pieter
Chalstrey, Ed
Reddish, Jenny
et al.

Publication Date
2025-12-25

DOI
10.21237/C7clio.50838

Data Availability
Associated data will be made available after this publication is published.

Copyright Information
This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

eScholarship.org Powered by the California Digital Library
University of California

Cliodynamics: The Journal of Quantitative History and Cultural Evolution

The Polaris2025 Seshat Data Release: An Expanded Dataset for Cliodynamic Research

Pieter François[1,2], Ed Chalstrey[2], Jenny Reddish[3], Enrico Cioni[1], Samantha Holder, Jakob Hauser[3], Jakob Zsambok[3], Matilda Peruzzo[2], Dániel Kondor[3], Majid Benam[3], Helena Miton[5], Maria del Rio-Chanona[6], James S. Bennett[3,7], Daniel Hoyer[3], Harvey Whitehouse[1] and Peter Turchin[3]

[1]*University of Oxford*
[2]*The Alan Turing Institute*
[3]*Complexity Science Hub*
[6]*Graduate School of Business, Stanford University*
[7]*University College London*
[8]*University of Washington*

This report describes the current canonical dataset named Polaris2025, a subset of Seshat: Global History Databank data. This dataset focuses on a well curated list of polities and variables that have been already used in publications or will be used in forthcoming ones. The report provides an introduction to the methods and procedures of the Seshat project relating to the curation and release of the Polaris2025 dataset. The Polaris2025 dataset builds upon an earlier release, the Equinox2020 dataset, and is now the most current and complete set of publicly available Seshat data. The Equinox2020 dataset remains valuable as a replication dataset. This report also describes important changes in the Seshat Databank itself, notably the new SQL database, the community tools, and the GIS portal.

Introduction

This report describes the current canonical Polaris2025 dataset, which is a subset of Seshat: Global History Databank data. Polaris2025 focuses on a well curated list of polities and variables that have either have been used for past publications by the Seshat team or are intended to form the basis of forthcoming publications. The data included in Polaris2025 are extracted from the main SQL Seshat database,

Corresponding author's e-mail: reddish.jenny@gmail.com

Citation: Pieter François, Ed Chalstrey, Jenny Reddish, Enrico Cioni, Samantha Holder, Jakob Hauser, Jakob Zsambok, Matilda Peruzzo, Dániel Kondor, Majid Benam, Helena Miton, Maria del Rio-Chanona, James S. Bennett, Daniel Hoyer, Harvey Whitehouse, Peter Turchin. 2015. The Polaris2025 Seshat Data Release: An Expanded Dataset for Cliodynamic Research. *Cliodynamics* 16: 117–129.

which stores the much larger and continuously growing Seshat dataset. This larger evolving dataset also includes variables that are currently still under construction or at the early stages of development, data that are being augmented continuously by the engagement of historians and other scholars. Such in-progress data remain inaccessible to the public until they are presented in a peer-reviewed publication, at which point they become open-access. The benefit of periodically releasing a well curated snapshot is that it offers analysts a stable dataset to work with that can be used for replication purposes. Future releases of snapshots can be expected when new variables and values have been curated (see below) and come into play for analysis.

The Polaris2025 data are made available as a comma-delimited spreadsheet, which is suitable for statistical analysis. To keep it concise, the spreadsheet excludes the narrative paragraphs that explain data codes, as well as the references upon which the coding decisions are based. However, this important contextual information can be found easily for each datapoint on the main Seshat SQL database. Minor updates to Polaris2025 will be released as numbered versions of the Polaris2025 dataset. It is recommended to always use the most recent version, which can be accessed here.

This report also provides a brief 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. This report also outlines improvements and novel features that have been implemented in the Seshat: Global History Databank since the release of the Equinox2020 dataset. The rest of this report is structured as follows. First we give an overview of the Seshat Databank. We then discuss the temporal and geographic scope of the data, followed by a discussion of the data gathering strategy. Last, we briefly discuss new developments of the Databank.

A Brief 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. In addition, Seshat also allows for browsing and for situating one or more polities in the context of the wider region, or the same temporal period, or even within a group of polities with a similar social structure.

Currently Seshat focuses on the time period between the Neolithic Revolution and the twentieth century. 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. The Databank already covers well over half of all humans who lived between 10,000 BCE and 1800 CE. Our main 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 (but see our discussion below on quasi-polities on how we deal with the smallest scales). However, over the next few years other units of analysis, e.g., settlements and trade networks, will be added. For each polity, we code variables relating to 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 at intervals around 100–200 years) are also recorded.

The Codebook includes over a thousand variables in various stages of development. These variables are coded (where possible) for over 800 polities between roughly 10,000 BCE to 1950 CE, subject to data availability and limitations. As of July 2025, the Databank contains around 80,000 coded values ("Seshat records," see below). Polaris2025, however, publishes only a well curated subset (57,560 records) from 625 polities and 172 variables. For comparison, the previous release (Equinox) comprised 44,477 records from 373 polities and 135 variables. Differences between the two datasets can be viewed in the table below. New variables include luxury good variables, religious demography, human sacrifice and crisis consequences. A python notebook that constructs the full Polaris dataset from the API can be found on GitHub.

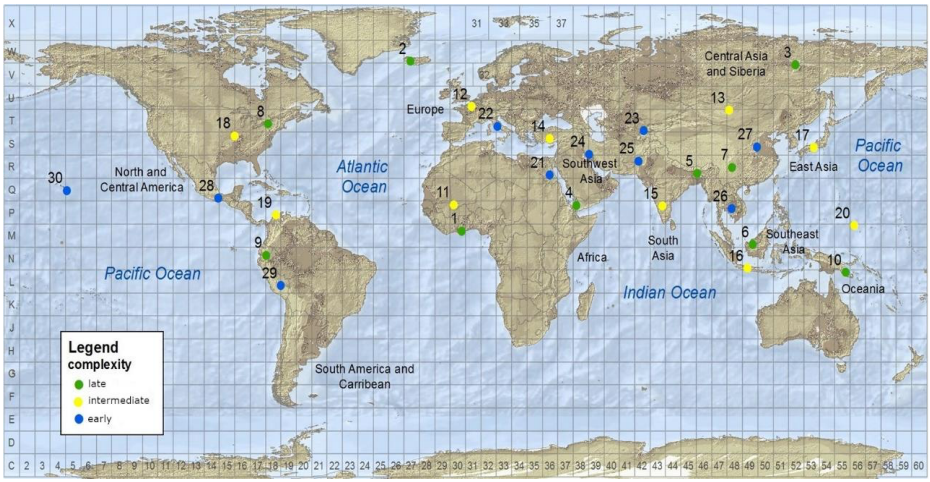
Dataset	**N. records**	**N. variables**	**N. polities**
Equinox	47,477	135	373
Polaris	57,338	172	625

Temporal and Geographic Scope

In order to assess whether different societies show commonalities in the way they have evolved, we developed a geotemporal, stratified sampling scheme that aimed (1) to include as much variation among the sampled societies as possible in terms 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. There have been two distinct phases in growing the sample of polities in Seshat. In a first phase we assembled a core set of polities aimed at overcoming the above-mentioned issues. Data collection in a highly systematic manner was key in this phase. To achieve this, we divided the world into ten major regions (named in the table of Figure 1, e.g., South Asia, Europe). Within each region we initially selected three natural geographic areas (NGAs). NGAs were, however, never intended to be a unit of analysis. Their only purpose was to organize the sampling of polities and these polities were our sole original unit of analysis. We derived our sample of polities by identifying all polities whose territory encompassed each NGA over time, listing the resulting polities in temporal sequence to construct time series of polities. In this way we were able to gather data on moments of transition both within and between polities within a given area over very long time spans. As polities were our unit of analysis, we then collected data pertaining to entire polities (and not just to the area of the NGA).

The NGAs themselves were chosen to try and maximize variation in this initial sample of polities. 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. This allowed 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, such as "small-scale societies," "egalitarian tribes," "acephalous societies" (see below). Beginning in 2017, we added 5 new NGAs (Southern Mesopotamia, Basin of Mexico, Middle Ganga, Galilee, Crete) to the original 30 and added their associated polities.



World Region	Late Complexity		Intermediate Complexity		Early Complexity	
Africa	1	Ghanaian Coast	11	Inland Niger Delta	21	Upper Egypt
Europe	2	Iceland	12	Paris Basin	22	Latium
Central Eurasia	3	Lena River Valley	13	Orkhon Valley	23	Sogdiana
Southwest Asia	4	Yemeni Coastal Plain	14	Konya Plain	24	Susiana
South Asia	5	Garo Hills	15	Deccan	25	Kachi Plain
Southeast Asia	6	Kapuasi Basin	16	Central Java	26	Cambodian Basin
East Asia	7	Southern China Hills	17	Kansai	27	Middle Yellow River Valley
North America	8	Finger Lakes	18	Cahokia	28	Valley of Oaxaca
South America	9	Lowland Andes	19	North Colombia	29	Cuzco
Oceania-Australia	10	Oro, Papua New Guinea	20	Chuuk Islands	30	Big Island Hawaii

Figure 1. Map and table of the original 30 sampling points developed during the first phase of data gathering.

After the release of Equinox2020, we began our second phase of sample development. This left the concept of NGA resolutely behind, as they had served

their purpose by allowing us to start building a stratified sample. By 2020 we had reached critical mass in the number of polities in the sense that most polities we now add either have a predecessor or successor polity already in the Databank. Especially for some areas of the world (e.g., the Sahel or Eastern Europe) we now have almost full coverage of all historical polities for specific clusters of variables, e.g., the social complexity variables. This "fuller" history of regions is key to obtain a more complete picture of dynamic changes.

Polaris has removed NGAs (Natural Geographic Areas) as a method to string together spatially overlapping polities over time. NGAs were never more than one specific entry into the dataset and analysts were always free to devise different methods to analyze our data. Instead, we are now devising a number of ways to engage with our data and so cater for a wider variety of analysts' needs. Currently this is the area of database construction that is being most rapidly developed and major advances will be announced on our website. We want to give access to our data in four ways. First, we are working towards a set-up whereby you can click on any specific point of the Seshat maps in the GIS interface and generate a time series specifically for this geographic point. Currently, the maps only link the shapefile of the Seshat polities with the data of these polities. See below for a more detailed description of the work on the maps and the Cliopatria project. Second, we are working towards placing all of Seshat in a genealogical tree. For this we are currently linking all polities with the associated, and often multiple, successor and predecessor polities. Third, we are providing analysts with a number of polity threads centered around a set of geographical points. If there is a gap of more than 100 years in our data coverage, a new thread is started. These threads can be used for dynamic regression analysis. Creating accurate and useful polity threads is an ongoing process that will be continuously revised and expanded. The latest set of polity threads can be found on the GitHub page. Fourth, we are providing our data in table format (Figure 2) structured via geographic regions. This is mostly for browsing or for downloading regional data.

To organize and sample data about societies typically known only archaeologically, we use the concept of "quasi-polity." This applies especially to societies for which precise polity boundaries are not known or are not appropriate. A "quasi-polity" 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. In this way we can integrate (often patchy) data from different sites and different polities within this area of cultural homogeneity to estimate what a "generic" polity was like. 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.

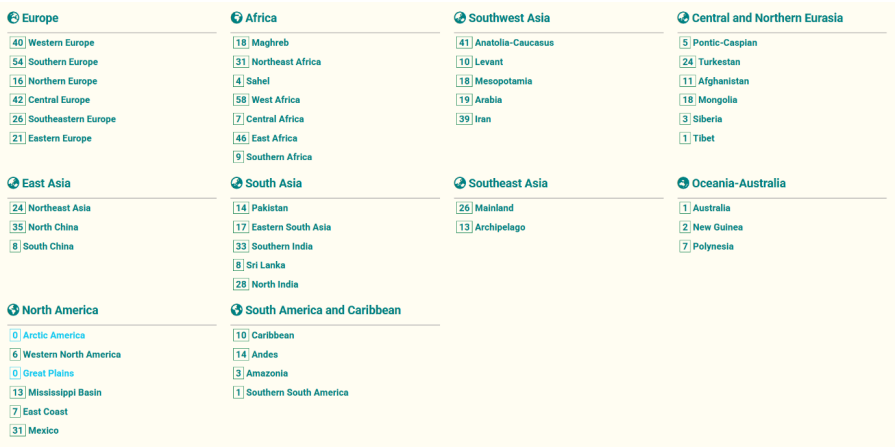
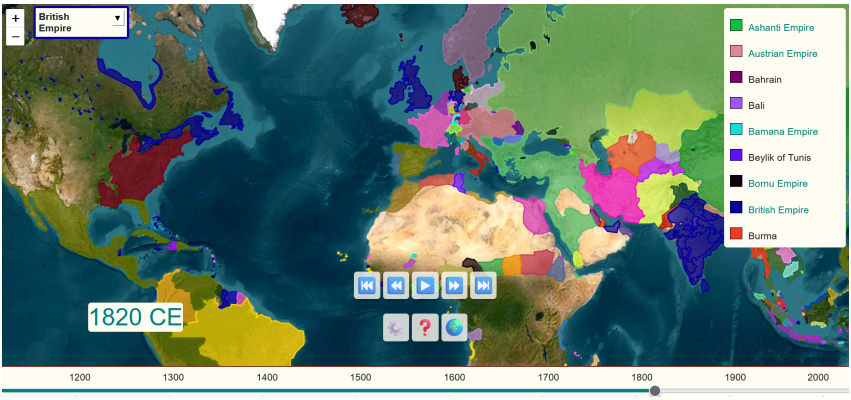


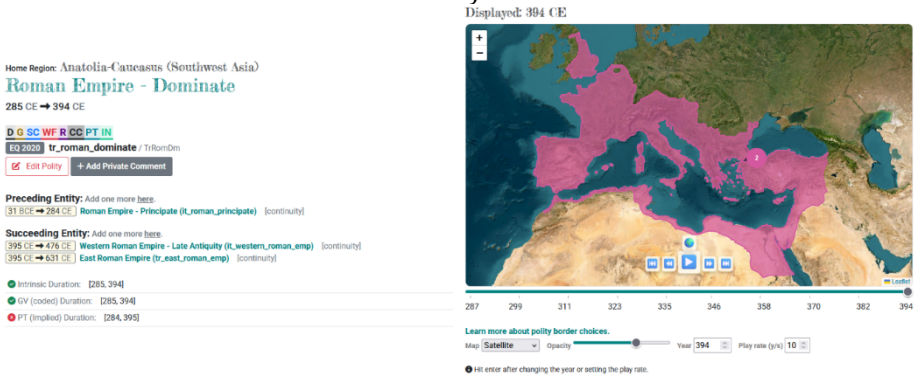
Figure 2. List of geographic areas used to group Seshat polities. Numbers indicate the total number of polities recorded for each area across the timespan covered by the database.

Data Collection and Data Quality

Debates about data quality, the appropriate mechanisms to capture the origin of the data transparently, and the process of buttressing data are evolving quickly within the fields of (quantitative) world history and the digital humanities. Being "open" no longer means simply making the dataset available, but it also entails providing information on every step of the research process and the assumptions and limitations of the data. In this fast-evolving landscape, the project team's commitment to open and transparent ways of doing research has been a key defining feature. Since the Seshat Databank was founded in 2011, our data gathering strategy and our data quality checks have followed a transparent and rigorous process taking place over several years and involving expert historians, archaeologists and anthropologists, research assistants, and (in more recent years) dedicated data review boards (DRB) for certain types of data (see http://seshatdatabank.info/methods/). The latter comprise the senior team responsible for data management on a given paper. The process below describes not only the coded data in the Polaris2025 spreadsheet, but also the entire process, and includes reviews of the explanatory paragraphs and the references that are stored on the SQL database.



3a)



3b)

Figure 3. Examples of geographic information on the Seshat website
a) Interactive world map page. The map updates depending on the selected year, and provides links to the detailed polity pages where the full data associated with an individual polity can be viewed.
b) Example of a Seshat polity page, with interactive mini-map and links to succeeding and preceding entities. Scrolling further on such pages reveals the full data associated with the polity.

The first step in building a new set of variables is usually a workshop where we discuss the theories we want to test, the variables needed to test these, and the availability of data. This results in a first-pass codebook, which is refined during the early phases of the process described below. The process of data collection for variables that are difficult to code and involve a large degree of interpretation typically involves matching each of our fully trained research assistants with one or more experts (recognized authorities on the polity in question, normally holding

a relevant doctorate and occupying a faculty position in a university). Research assistants focus on one or a small group of temporally or geographically adjacent polities. The initial input from the experts focuses on providing help with assembling initial reading lists and, where necessary, advice on how to interpret some of the key historiographical debates. The research assistants gather the information required to put forward a provisional coding recommendation, together with a condensed overview of the data used to buttress that coding, highlighting any areas of uncertainty. These codes are thus based on scholarly sources and are fully referenced. The research assistants then conduct consistency checks. Coding recommendations and the data provisionally used to buttress them are then presented to the experts for further review, often in multiple iterations. When Seshat experts point out disagreements in the literature or disagree among themselves on a particular coding, we keep a record of this so that multiple analyses can be run, taking into account the contrasting interpretations.

Finally, for a subset of variables, typically those that require especial care, we add another layer of checks. Thus, for such (groups of) variables as moralizing religion or human sacrifice, a DRB reviews the resulting coding recommendations and supporting references. At this stage, the DRB may approve codes as ready for analysis or request further review, where appropriate, involving additional experts to address any remaining points of uncertainty. The DRB is also responsible for ensuring that coding conventions are consistently applied across polities. Only when the DRB is satisfied that the rationales for the coding decisions and the associated buttressing statements are transparently and compellingly articulated, following a set of agreed-on coding conventions, are the data "frozen" and converted into the correct syntax for the analysis. As such, final responsibility for coding decisions relating to data frozen for publication is assumed by the DRB instead of the contributing expert.

Variables that are straightforward to code and for which it is relatively easy to find information (e.g., the social complexity or military technology variables) are still coded by research assistants and receive input from senior Seshat team members and sometimes experts, but not all of these data are explicitly endorsed by experts. By showcasing this data on the Databank and through community engagement tools, we present these variables for scrutiny by our users and the historical community at large, who can suggest additions or changes. Whereas our quantitative coding scheme goes a long way towards capturing nuance, uncertainty, and disagreement, and the accompanying textual paragraphs and references go one step further, we nevertheless acknowledge that converting complex arguments into a sophisticated coding scheme can be difficult. Hence, we constantly strive to improve our methods and incorporate as much context and as many alternative interpretations as possible and enable other scholars to engage with our data, and build upon it if they so desire.

New Developments and Moving Forward

Seshat is a living project – neither our data nor our methods are ever fully fixed but are constantly evolving. Our data are continuously updated with new findings and alternative variable lists. The list of variables is growing and, for example, data on religious tolerance are currently being added to the Databank. As this article shows, the number of polities, how we sample them, and how we offer entry points into studying change over time has changed a lot over the past five years and this will keep expanding over the years to come. Similarly, we have significantly improved our database architecture in three main aspects. First, we have moved on from the Wiki database which stored our data up to the release of Equinox2020. Instead, our data is now stored in a state-of-the-art SQL database, and this offers a lot more functionality in terms of linking the data and displaying it. We now have an API which allows one to programmatically query the evolving Seshat database. As a result, the Seshat: Global History Databank has taken the next step in becoming a mature building block of the open-science digital infrastructure.

Second, through the creation of user profiles, comment boxes, and engagement tools, it is now much easier for scholars and the public at large to engage with our data and suggest additions and changes. Third, we have integrated a GIS into the database. We have integrated historical geographical information from the Cliopatria project (Bennett et al. 2024) into the Seshat Databank and website. Cliopatria shapefiles and Seshat data are linked through the use of the same IDs. Cliopatria encodes, as an open-source dataset, the changing geographic extent of polities worldwide between 3400 BCE and 2024 CE. This allows for a more intuitive way to engage with and query the data. The GIS concisely visualizes change over time both in terms of the total number of polities we have data on and in terms of a growing list of individual variables. It is important to note that the GIS operates on the evolving Seshat dataset (and not on the Polaris2025 snapshot). Examples can be seen in Figure 3.

Finally, recent breakthroughs in artificial intelligence present an opportunity to enhance both the quality and productivity of our data-gathering pipeline. As an initial step, we benchmarked large language models on the accuracy of their historical knowledge (Hauser et al., 2024). Earlier generations performed poorly, but our preliminary tests show that today's reasoning-oriented models are markedly more reliable (Long Phan et al. 2025). While the Polaris2025 release does not incorporate data collected with the use of AI, we are now actively exploring the extent to which AI-supported workflows for both data collection, reference identification and subsequent analysis can be helpful and complement our current workflows. We also acknowledge that there are challenges relating to the fair use of AI. Therefore we are focusing efforts on finding appropriate reference for data points to credit people's work, documenting potential biases, and always being explicit on when a data point stems from AI use. Visit our project

website for ongoing updates, and feel free to contact us with any queries about accessing the dataset.

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]), a European Research Council Advanced Grant to the University of Oxford, entitled "Ritual Modes: Divergent Modes of Ritual, Social Cohesion, Prosociality, and Conflict"; the program "Complexity Science," which is supported by the Austrian Research Promotion Agency FFG under grant #873927, grant GZ 2023-0.841.266 from the Austrian Ministry for Innovation, Mobility and Infrastructure | Bundesministerium für Innovation, Mobilität und Infrastruktur (BMIMI) to CSH's PostDoc Program for Complexity Science and Data Competence; an FCDO grant "Freedom of Religion or Belief Leadership Network"; an AHRC grant "Data/Culture. Building Sustainable Communities around Arts and Humanities datasets and software"; a Santa Fe Institute Lou Schuyler grant; and a CLARIAH-AT grant "Developing HuBERT: a Natural Language Processing algorithm for extending the Seshat Global History Databank." 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 for a comprehensive list of private donors, partners, experts, and consultants and their respective areas of expertise.

References

Bennett, J. S., E. Mutch, A. Tollefson, E. Chalstrey, M. Benam, E. Cioni, J. Reddish, et al. 2025. Cliopatria – A Geospatial Database of World-Wide Political Entities from 3400BCE to 2024CE. *Scientific Data* 12: 247. doi: 10.1038/s41597-025-04516-9.

François, P. Forthcoming. Bridging the Quantitative/Qualitative Divide: A Methodology to Study the Role of Moralizing Religion in World History. In *Seshat History of Moralizing Religion*, vol. 2: *World Survey*, edited by J. Larson, J. Reddish, and P. Turchin, 1–15. Chaplin, CT: Beresta Books.

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.

Hauser, J., D. Kondor, J. Reddish, M. Benam, E. Cioni, F. Villa, J. Bennett, et al. 2024. Large Language Models' Expert-Level Global History Knowledge Benchmark (HiST-LLM). *Proceedings of the 38th Advances in Neural Information Processing Systems (NeurIPS 2024)*. https://proceedings.neurips.cc/paper_files/paper/2024/file/38cc5cba8e513547b96bc326e25610dc-Paper-Datasets_and_Benchmarks_Track.pdf.

Phan, Long, A. Gatti, Z. Han, N. Li, J. Hu, H. Zhang, C. B. C. Zhang, et al. 2025. Humanity's Last Exam. *arXiv preprint*. arXiv:2501.14249. doi: 10.48550/arXiv.2501.14249.

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.

Turchin, P., R. Brennan, T. Currie, K. Feeney, P. François, D. Hoyer, J. Manning, et al. 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.

Turchin, Peter, Daniel Hoyer, James S. Bennett, Kiran Basava, Enrico Cioni, Kevin Feeney, Pieter Francois, et al. 2020. The Equinox2020 Seshat Data Release. *Cliodynamics* 11 (1): 41–50. doi: 10.21237/C7clio11148620.