Please note: This is a preprint of a manuscript that has been accepted for publication in the *Journal of Cognitive Historiography*. The final published version may contain small differences such as formatting and copy editing. Please direct correspondence to peter.turchin@uconn.edu.


Turchin, Peter. University of Connecticut. peter.turchin@uconn.edu
Whitehouse, Harvey. University of Oxford. harvey.whitehouse@anthro.ox.ac.uk
François, Pieter. University of Oxford. pieter.francois@stb.ox.ac.uk
Hoyer, Daniel. Evolution Institute. dhoyer@evolution-institute.org
Alves, Abel. Ball State University. aalves@bsu.edu
Baines, John. University of Oxford. john.baines@orinst.ox.ac.uk
Baker, David. Macquarie University. david.baker@mq.edu.au
Bartkowiak, Marta. Adam Mickiewicz University in Poznań. martabartkowiak86@wp.pl
Bates, Jennifer. Brown University. jennifer_bates@brown.edu
Bennett, James. University of Washington. jsb11@uw.edu
Bidmead, Julie. Chapman University. bidmead@chapman.edu
Bol, Peter. Harvard University. pkbol@fas.harvard.edu
Ceccarelli, Alessandro. University of Cambridge. ac2045@cam.ac.uk
Christakis, Kostis. British School at Athens. knossoscurator@bsa.ac.uk
Christian, David. Macquarie University. director.bighistory@mq.edu.au
Covey, Alan. University of Texas. r.alan.covey@austin.utexas.edu
De Angelis, Franco. University of British Columbia. franco.de angelis@ubc.ca
Earle, Timothy K. Northwestern University. tke299@northwestern.edu
Edwards, Neil R. Open University. neil.edwards@open.ac.uk
Feinman, Gary. Field Museum. gefinman@fieldmuseum.org
Grohmann, Stephanie. University of Edinborough. steph.grohmann@ed.ac.uk
Holden, Philip B. Open University. philip.holden@open.ac.uk
Júlíusson, Ární. University of Iceland. arnidan@akademia.is
Korotayev, Andrey. National Research University Higher School of Economics. akorotayev@gmail.com
Kradin, Nikolai. Russian Academy of Sciences. kradin@mail.ru
Kristinsson, Axel. Independent Scholar. axel@akademia.is
 Larson, Jennifer. Kent State University. jlarson@kent.edu
Litwin, Oren. George Mason University. ollitwin@masonlive.gmu.edu
Mair, Victor. University of Pennsylvania. vmair@sas.upenn.edu
Manning, Joseph G. Yale University. joseph.manning@yale.edu
Manning, Patrick. University of Pittsburgh. pmanning@pitt.edu
Marciniak, Arkadiusz. Adam Mickiewicz University in Poznan. arekmar@amu.edu.pl
McMahon, Gregory. University of New Hampshire. Gregory.McMahon@unh.edu
Miksic, John. National University of Singapore. seajnm@nus.edu.sg
Moreno Garcia, Juan Carlos. French National Centre for Scientific Research. jcmorenogarcia@hotmail.com
Morris, Ian. Stanford University. jmorris@stanford.edu
Mostern, Ruth. University of Pittsburgh. rmostern@pitt.edu
Mullins, Daniel. University College London. d.mullins@bbk.ac.uk
Currently, Seshat focuses on the period between the Neolithic and Industrial Revolutions. Our unit of analysis is a *polity*, which we define pragmatically as an independent political unit ranging 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, institutions, well-being, the production of public goods, and various technologies. The current codebook includes over 1500 variables, of which c.500 are the focus of active coding. As of May 2019, the Databank contains nearly 300,000 records linked to over 400 polities.

To address potential problems of selection bias and non-independence of data points, we use a sampling scheme based on 30 “Natural Geographic Areas” (or NGAs) across the globe. NGAs are simply spatial units that enable us to sample the diversity of past societies. First, ten major world macro-regions were identified (Figure 1a). To maximize potential variation, for each world region we selected one NGA in which social complexity—as defined in (Turchin et al. 2018)—emerged early, another where it emerged late, and a third in between. In different world regions, the time points at which social complexity emerged may be very different. For example, complex societies in Susiana, the early complexity NGA in Southwest Asia, go much further back in time than in Hawaii, the early complexity NGA for Oceania-Australia. NGAs are not the units for which we collect data, rather they help us select polities that become part of the Seshat sample. This approach yields a stratified sample of polities driven by geography and antiquity of social complexity, allowing analysts to construct spatially anchored time-series, while recognizing that the spatial extent of sampled societies fluctuates with time (as polities rise, expand, go into decline, and disappear).

Figure 1. (a) Locations of NGAs sampling global variation in cultural evolution. (b) Frequency distribution of the starting dates for data sequences in Seshat Databank. Adapted from (Turchin 2018).

Data Collection
To populate the Databank, we consult current scholarship on each region along with expert collaborators to develop a list of all polities that sequentially occupied each NGA. In cases where NGAs encompass numerous coexisting small-scale societies, we treat these as quasi-polities, which roughly correspond to ethnological “cultures” (Murdock 1967; Murdock and White 1969) or archaeological sub-traditions (Peregrine 2003). We gather information on each variable in our codebook for each polity or quasi-polity, offering as much specificity as the evidence allows; often, a variable value is associated with the entire temporal extent of the polity, but if more granular information is known, we capture changes within the polity’s duration as well.

Prior to inputting data on a topic, we develop a conceptual scheme through Seshat workshops. The goal is to create a quantitative variable (e.g., polity population) or multiple proxy variables capturing various aspects of a more complex characteristic (e.g. well-being). Seshat research assistants (RAs) then code several test cases in consultation with experts, continually refining the variables. RAs are trained and supervised by teams of advanced (postdoctoral or professorial) social scientists and historians.
Supervisors facilitate communication between RAs and our expert collaborators, oversee coding decisions, and ensure consistent application of our coding schemes. We focus on hiring RAs with advanced qualifications and aim to retain productive researchers for lengthy periods, often several years.

Once a coding scheme is operationalized to test theories, data collection begins. First, RAs search the most up-to-date and relevant scholarship (with expert guidance and direct supervision by Seshat’s more senior researchers), sourcing both primary and secondary material, and enter preliminary data. Second, RAs compile lists of questions on values that cannot be coded unambiguously, or on which information in the published sources is lacking, and seek help from the experts on the polity. Finally, we ask experts to review the data to check coding decisions made by RAs and help us fill gaps. The coding process is never “complete,” as Seshat data are constantly checked by various stakeholders, and new relevant data may appear, or novel insights may alter the understanding of known documents and material.

When experts disagree or ongoing debate continues in the literature, multiple alternatives are coded. Likewise, when quantities are ambiguous, coders record the likely ranges [min, max] found in specialist literature. Where evidence is lacking entirely, RAs record the value as “suspected unknown”, which becomes a question to bring to our collaborators; only domain experts can verify a value as “unknown”, referencing their unique understanding of the limits of empirical evidence for a given area or topic.

We refer to a coded value of a variable for a particular polity as a “Seshat record” with a complex internal structure. Seshat records also include a narrative description explaining the background to the code and contextualizing levels of uncertainty and disagreement, along with citation to reference sources. The description may specify a quantity or range, or indicate whether a feature is present, absent, inferred present, inferred absent, suspected unknown, or unknown. “Inferred” presence or absence indicates some degree of uncertainty. For example, if iron smelting has been attested both for the period preceding the one that is coded and for the subsequent period, this could suggest a code of “inferred present” although there is no direct evidence for it (assuming general socio-technological persistence during this period and that there is no indication that this technology was lost and then regained). Variables can also have temporal uncertainty. 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 between 300 and 600 CE as effectively “either absent or present” (which is different from “unknown”).

**The Evolution of Seshat**

Seshat’s methods have evolved since its inception in 2011. Initially, all data were collected by expert collaborators, but we soon realized that asking busy colleagues to fill in hundreds of boxes was misuse of their expertise. Having established an effective coding scheme, we discovered that much information can be accurately entered by well-trained and supervised RAs from published scholarship, allowing us to deploy expertise more strategically to resolve difficult coding issues, locate elusive information and point out relevant literature, and confirm genuine gaps in the record.

Seshat is a massive, complex project, which evolves constantly. In a project as vast and multifaceted as Seshat, there will inevitably be some practical constraints on obtaining accurate or representative values or codes for specific variables because, for example, a particular bit of information has been published in an obscure source, or information of which we are not yet aware changes the coded value. We cannot wait until this “cleaning” process is over—because it never is. Our approach, thus, is to address such problems as we discover them, gradually improving the Databank, while understanding that some disagreement will always exist and that lag between our recorded data and available evidence will endure. We aim to bring as many voices as possible to bear on the information
we collect in order to utilize the most relevant scholarship in our published work. Once an article using Seshat Data is published, we present that material on our website as open-source data to be reused by other researchers. Knowing that the process of improving our data is never ending, we encourage users to offer feedback, pointing out alternate readings of evidence or directing us to previously unknown information.

The suggestions and critiques of scholars are essential in this regard. We all benefit by bringing out these issues into the open; the systematic nature of Seshat helps concentrate these discussions and identify where there are gaps in knowledge, uncertainties, and disagreements. Seshat is designed to be iterative—evolutionary—in nature. We are constantly rechecking coded data to make sure that the information we have is the most up-to-date and that it reflects relevant ambiguities and nuance. We are continually expanding both the geographic scope of data and adding new variables to cover different aspects of past societies. This is entirely a collaborative process, as Seshat functions only with the open, constructive engagement of researchers from a host of disciplines.

References


