Translating Knowledge about Past Societies into Seshat Data[1]

Peter Turchin
University of Connecticut
Complexity Science Hub, Vienna

The *Seshat: Global History Databank* was founded in 2011 with the goal of systematically collecting data about social, political, and economic organization of human societies and how they have evolved over time. From the beginning the first guiding principle of the Seshat project was to reflect the current state of knowledge about past societies as accurately as possible within practical constraints (I'll discuss practical limitations later on). Second, and equally important, our aim for the database was to reflect not only what is known, but what is unknown, or poorly known.

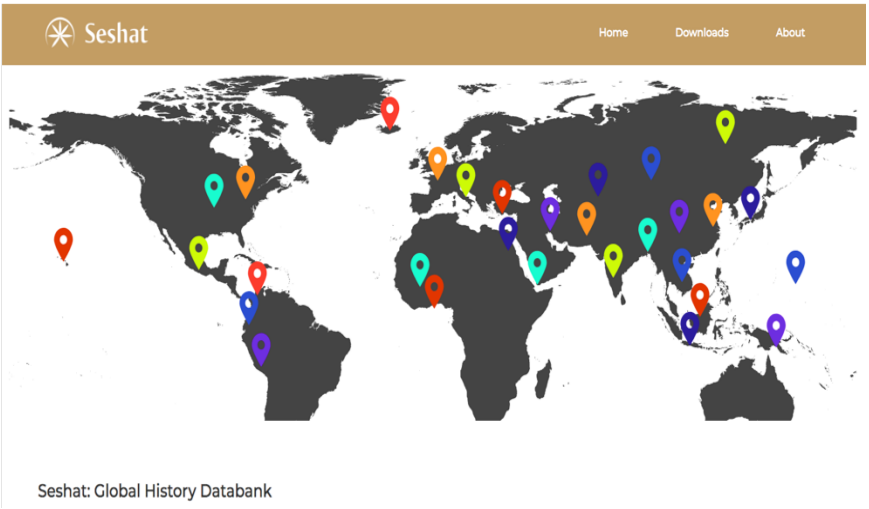

Figure 1. Screenshot of Seshat's data page (seshatdatabank.info/data)

Earlier this year we published our first article using large amounts of data from Seshat (Figure 1). The data underlying this paper were published together with

[1] This piece is adapted from a Seshat blog post first published May 10, 2018.

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

Citation: Turchin, Peter. 2018. Translating Knowledge about Past Societies into Seshat Data. *Cliodynamics* 9: 143–147.

the article in the form of a spreadsheet. We are now able to make these data available in a visual interface based on the Seshat Knowledge Graph.[2] This interface allows all to browse the data, read the explanations of why a particular variable was coded with a particular value, and—most importantly—suggest improvements, alert us to errors, and help us fill in the gaps. Because this is a "beta release", we also want to know about any problems with the interface, so that we can fix them. You can find the data on the Seshat Project Website.

During the past seven years the general principles guiding the Seshat project have not changed, but specific approaches to how we expand the Seshat data have evolved in several significant ways. Initially our idea was that all data in Seshat would be collected by *experts*—academic historians, archaeologists, and other scholars of the past. However, we quickly discovered that this approach has serious drawbacks, even with historians who were very enthusiastic about the Seshat project. For example, asking experts to fill in hundreds of boxes is a horrible misuse of their expertise. For many variables, once you have established an effective coding scheme, 80–90 percent of data can be accurately entered by well-trained research assistants (RAs) working with standard texts. The time and effort of experts is a highly valuable resource, and it should be deployed strategically, where it is truly needed—resolving difficult coding issues and locating elusive information. Furthermore, only an expert can make the judgement that the field doesn't know about a particular variable— that it is a true knowledge gap.

At the beginning of the project we experimented with different types of RAs. We discovered that using temporary undergraduate student labor was not a viable approach. It simply didn't make sense to invest several months into training RAs, determining how accurate and efficient they are, and then losing them for good. As a result, we shifted resources to hiring long-term RAs who work for the project for at least a year, and usually for many years. All our RAs have at least an equivalent of Bachelor's degree, many have Master's, and some even a PhD.

The third crucial element in our data gathering process is the close supervision of RAs by PhD-level social scientists, who include Seshat postdocs, regional editors, variables coordinators (who focus on a particular set of Seshat variables), and Seshat directors (see Who We Are on the Seshat web page). Their role is to train the RAs, check their coding decisions, and to ensure consistent application of the coding schemes. It would not be possible to have generated so much high-quality historical data, as we have done, without the incredible hard work of our RAs and extreme generosity of our expert collaborators, who

[2] For a non-technical explanation of the software engine underlying Seshat Databank, see Peregrine et al (2018).

donated their time and knowledge to help our project. I also need to stress the importance of a Horizon2020 grant awarded by the European Union, which helped to fund work on the technical infrastructure that underlies this our new Knowledge Graph and user interface led by Drs. Kevin Feeney and Gavin Mendel-Gleason.

We discovered that our best data are collected when all three groups (RAs, expert scholars, and social scientists) work together. At the beginning of coding a particular society (or a Seshat *polity*—a politically independent society bracketed by a start and an end dates) we get expert help in suggesting a set of standard texts and answers to some general questions (e.g., about periodization). RAs then are instructed to get as much data coded from these standard sources, using a "low-hanging fruit" approach. In other words, if they do not quickly discover an answer, they stop researching this question and add it to a list of issues to resolve later with expert help. Once this phase is over, we go back to the experts with a list of questions addressing data gaps and difficult coding issues.

As much as is possible, we also run specialized workshops that bring together members of the Seshat project with the experts, focusing on either world regions (e.g., Egypt, Southeast Asia), or specific variables (ritual and religion, agricultural productivity) (Figure 2).

Figure 2. Participants listen to a presentation at the Seshat workshop *Testing the Axial Age*, Oxford, December 2017. Pictured are (clockwise, from bottom): Jennifer Larson, Julye Bidmead, John Baines, Jill Levine, Patrick Savage, Barend ter Haar, Prof. Vesna Wallace, Christina Collins, Harvey Whitehouse, Thomas Currie, Daniel Mullins, and Agathe Dupeyron. Photo by author.

In summary, expanding the Seshat Databank, and especially finding data for difficult-to-code variables, is a result of collaboration between experts and the Seshat staff. This process combines expert's specialized knowledge of a particular historical society with our experience on translating historical knowledge into a coded form.

As mentioned above, establishing an effective coding scheme is a key feature of making the Seshat project work (the full Codebook is available online). If the variables are too vague, too abstract, or require too much interpretation then they become difficult to code and the chances that there will be disagreements between coders increases. For example, when gathering data into Seshat, we 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"). Prior to collecting the data, we run a workshop, usually involving experts, that develops an understanding on how to code a particular aspect 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 variables into multiple simple variables that can be coded in a binary fashion (absent/present). The initial coding scheme is then tested by Seshat RAs who apply it to several test cases, working in consultation with the experts. The coding scheme is then refined based on the suggestions from both experts and RAs and is applied to the whole sample. Sometimes we discover that we must adjust the coding scheme after a substantial number of polities have already been coded using the old one. Switching to a better definition results in a certain amount of inefficiency, because RAs have to go back to already coded polities and recode them using the new scheme. This process takes time, and such old codes sometimes linger in the databank until eventually encountered and corrected. Before using Seshat data in statistical analyses they undergo a systematic Data Quality check. Every datapoint is checked by a different RA than the one who entered it.

Seshat is a massive, complex, "living" entity that constantly evolves. In a project as large and multi-faceted as Seshat and with an underlying database as vast as this one is, 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 there is new information that changes the coded value, of which we are not yet aware. We cannot wait until this "cleaning" process is over—because it is never over. Our approach, thus, is to address such remaining problems as we discover them, gradually making the Databank better, while understanding that there are always will be some errors in the data. The suggestions and critiques of other scholars are very useful 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. Additionally, I want to stress that the databank is designed to be iterative in nature. We are constantly rechecking the data we've already coded to make sure we have the most up-to-date information which reflects the relevant ambiguities and nuance. As new historical and archaeological knowledge becomes available, we aim to include it in Seshat.

In closing, I'd like to reiterate that what you will see on the Seshat Data page is a "beta" release. Undoubtedly there are bugs in how the interface works, and in the data there are errors and missing values, which can be filled in. Please use the instructions on the Data page explaining how you can alert us to errors and suggest additional data. We need your help to make Seshat better!

Acknowledgements

Development of the software underlying Seshat Databank was supported by the Horizon 2020 grant agreement No 644055 [ALIGNED, www.aligned-project.eu]. Full details on Seshat project funding is available on the project site's Acknowledgements page. The author thanks Harvey Whitehouse, Pieter François, Thomas Currie, and Daniel Hoyer for comments on the manuscript.

References

Peregrine, Peter N., Rob Brennan, Thomas Currie, Kevin Feeney, Pieter François, Peter Turchin, and Harvey Whitehouse. 2018. "Dacura: A New Solution to Data Harvesting and Knowledge Extraction for the Historical Sciences." *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 0(0): 1–10. doi: 10.1080/01615440.2018.1443863.