3.09 Collaborative Historical Information Analysis

Patrick Manning, University of Pittsburgh, Pittsburgh, PA, United States
Pieter François, University of Hertfordshire, Hatfield, United Kingdom; and University of Oxford, Oxford, United Kingdom
Daniel Hoyer, University of Toronto, Toronto, ON, Canada
Vladimir Zadorozhny, University of Pittsburgh, Pittsburgh, PA, United States

3.09.1 Introduction: World-Historical Information Resources

3.09.1.1 Big Data in Human History: The Mission

Human society is undergoing transformations, both positive and negative. But we are short on information about those transformations. We almost always lack information on how local shocks and transformations add up at the global level (Manning, 2015a). Whether it is war, drought, massacre, economic crisis, or aging of the human population, we rarely have information on the global implications of local events. Even when we can collect information on changes in the world of today, it is difficult to know whether they are the consequence of long-forming developments or something entirely new. The researchers in this group have identified themselves as working on "Big Data in Human History." We share pursuit of an underlying mission, as follows:

Big Data in Human History (BDHH). This association of historical social scientists explores big theoretical and empirical questions on the human past. Its mission lies in pursuing big questions at multiple scales of data and analysis, relying on advanced methods, and assuming interactions throughout the human social system.

Big questions. We apply theory and test hypotheses on dynamics in human society: work, community, exchange, well-being, sustainability, social evolution, social complexity, knowledge, inequality, conflict, rupture, and environmental interactions.

Multiple scales. We work with data and analysis at multiple scales, comparing and linking scales. In time, these different scales cluster around: the contemporary era, the industrial era, the post-1500 era, the era of cities and states, the era of agriculture, and the era of human diaspora. In space, we work from the local to the regional or global, terrestrial or maritime. In populations, we investigate individuals, communities, large-scale societies, and humanity. In topics, we range across fields of study from society to economics, government, culture, knowledge, health, genetics, and environment. In scales of data availability, we range from the huge but inconsistent quantities available on contemporary and recent times to the very scarce quantities and limited topics of data available on earliest human history.

Methods informed by information science. We work with digitized data supported by advanced metadata, ontologies, archiving, and curation; our analysis relies on large-scale computation and linked open data or succeeding methods. Our mission is to pursue big questions about interactions within human social systems, at multiple scales of data and analysis, using methods from information science.

3.09.1.2 Participants in "Big Data in Human History"

The collaborative of social science research groups, as identified above, is Big Data in Human History. The collaborative includes four participating groups and a fifth which, while quite active on its own, is an observer of the Big Data in Human History group. In addition, other relevant groups of researchers addressing large-scale social science exist, and may link up with BDHH in time.

Collaborative for Historical Information and Analysis (CHIA), a unit of the World History Center (WHC), Pittsburgh, PA
Seshat—Evolution Institute (Seshat-Evolution), San Antonio, FL.
International Institute of Social History (IISH), Amsterdam.
Institute for Research on World History (IROWS), Riverside, CA.
Minnesota Population Center (MPC), Minneapolis, Observer of BDHH.
Other groups of social-science researchers working on large-scale issues may find BDHH to be of interest, and we welcome new members.

3.09.1.3 Context: Big Research Projects in Social Science and Natural Science

Previous large-scale projects. During the past two centuries, large-scale research projects in social science and natural science have brought substantial advances in knowledge and substantial changes in the world of today. In social science, national censuses, beginning in the 19th century, brought immense expansion in knowledge through government efforts to compile population data. The creation of national income analysis in the early 20th century became the basis of national and international economic policies, utilized by organizations such as the Organization for Economic Cooperation and Development. In the natural sciences, large-scale research projects created atomic science (including bombs), European collaboration created the atom-splitting CERN institute near Geneva, and the United States and USSR built independent programs for space exploration. Worldwide collaboration brought the understanding of continental drift in the 1960s and the successful modeling of climate change over short and long time periods in the 1990s. Another collaborative, international program achieved the complete documentation of the genome of humans and other species. From this perspective, the development of world-historical data on human society is the natural next step in such large projects to expand knowledge. But, as with the other projects, this step will not be taken automatically—it will come only if dedicated researchers can gain substantial funding to work collaboratively on world-historical data resources over the course of years. The disciplines to participate in this shared work include history, anthropology, sociology, human biology, ecology, economics, psychology, and computer and information science.

Problems in the third stage of global studies. One may identify three stages or generations in the definition of social science analysis of global issues in cross-disciplinary terms. The first generation of global studies opened in the 1940s with postwar advances in social sciences; the second generation opened in the 1980s with new computational techniques and perceptions of contemporary globalization; and the third generation of global studies is opening now with advances in historical and cross-disciplinary depth (Manning, 2013).

The emerging third generation of global studies begins with an effort to extend the value of second-generation materials. Recent work has included steps that are necessary but not sufficient to the creation of global historical data. Work with GIS has created georeferenced documents that give little evidence of change over time; time-series data (mostly economic) have been prepared without georeferencing; and population censuses—detailed but addressing widely separated points in time—tend not to be georeferenced. Global historical datasets require simultaneous documentation of values and variations in space, time, content, and scale. In addition, the questions and obstacles that have limited previous large-scale projects in social science data remain before us. How are we to work across disciplines? How are we to keep up with new developments? What type of social networking is necessary to facilitate submission of data? We will discuss below the expected benefits of crowdsourcing and data mining, but what new technical advances or new organizations will appear within the next few years? How do we prepare the appropriate mixture of adoption, collaboration, or other responses to these changes? How are we to work across institutions—and across the boundaries of nation and language?

Hesitations in global social analysis. World-historical consciousness and world-historical interpretation have been advancing rapidly, but world-historical research is advancing much more slowly. The analysts of world history and global social sciences, to the degree that they work with datasets, use comparative data on nations or similar regions, with little reference to patterns at the supranational regional or global levels. Admirable work has been completed on small scales in collecting data on economics, social structure, and human-natural interaction, but there is yet no large-scale project to document transformation of human society over the past several centuries. The start-up problems in developing cross-disciplinary collaboration in large-scale social analysis, while serious, are parallel to the problems that slowed initial work on climate, genetics, and national income analysis: they can be overcome with patience, determination, and inclusive open dialogue. One project of the 1990s, Electronic Cultural Atlas Initiative (ECAI), did excellent work in developing metadata for many social and cultural variables, but did not succeed in obtaining datasets from those who held them nor in developing a fully global perspective.

3.09.1.4 Collection and Analysis of Comprehensive, Big Historical Data on the Human System

This title phrase conveys the essential nature of world-historical data resources and their contents. Here we take the phrase apart, word by word, to provide a brief description of the context, problems, and resources for creating world-historical data.

"*Data*." The form of data to be included in the world-historical resource will begin with machine-readable, linked open data (LOD)—that is, either structured tabular data or graph-based semantically structured data—but will expand by stages to include text, images, and data in other media. The initial focus on machine-readable data stems from the availability and relevance of this data type. Here are some of the categories:

Data in print and manuscript form, in known and accessible repositories, including textual, archaeological, and visual material.
Data in repositories that are not accessible.
Digitized data.
Digitized data that are documented and/or linked.
Data created by theoretical and analytical projections and simulations.
Conceptual data—all the data one could imagine.

"Historical." The term "historical data" has multiple meanings. It ranges from handwritten medieval documents to the terabytes of digital information concerning past societies that will ultimately by held in linked, global repositories. Any bit of evidence that sheds light on some aspect of our shared past qualifies as historical data. There exists an immense amount of information from every historical society that has ever existed covering a wide array of topics and themes.

"Big." Big Historical Data are big in three senses. First, the total amount of existing, recorded historical data (accessible or not) clearly reaches terabytes in volume. Second, the number of dimensions of interactions among historical variables is such that, even before the volume of data reaches gigabytes, large-scale computational resources are necessary to handle the necessary analytical calculations. Third, in addition to recorded data, many new data can be created through manipulation of existing data. For instance, census data on population permit calculation of birth rates, death rates, and expectation of life. So we will create, collect, and repurpose large quantities of historical data on the human system.

"Human System." For our purposes, we treat human society as a system, in which the elements are connected and mutually dependent. In it, the terms "human" and "humanity" have relevance at multiple scales: from single individuals through families and communities, including nations, empires, and diasporas, up to the human community as a whole. These differences confirm why world-historical data must address multiple scales of aggregation, from local to global, and interactions among processes centered in various scales.

"Comprehensive." A world-historical data resource must be comprehensive. By "comprehensive world-historical data," we mean well-documented and linked data, conveying information on interconnected topics of human existence during the last several centuries. But "comprehensiveness" also means "selectivity." Because the range of a world-historical data resource must be so multidimensional, only so much data can be selected for each segment. For these reasons of scale, it would be best to begin with a prototype—a relatively small-scale version of the data resource that includes all the main functions and dimensions that the full-scale version will have, in order to confirm that the full data resource will produce valuable results.

"Collection and Analysis." Only small proportions of existing world-historical data are now accessible for analysis by researchers, and only small portions of the available data have been digitized. This information has the potential to address any number of Big Questions, but needs to be translated into usable, combinable, and analyzable forms; accessing Big Historical Data offers the

best path to achieve this. Only when the heterogeneous data from wide-ranging sources have been documented (as to their source, dimensions, and transformations) can they be linked into larger units.

3.09.1.5 Need, Objectives, and Challenges

Need. Do we need separate research projects to create world-historical data? Why is it not feasible simply to calculate worldwide data by taking the sum of national data for all countries? The answer, most directly, is because national data do not now go back very far for most nations, and also because records and recording units are kept differently in different nations. More generally, understanding of global society in present and past times requires explicit attention to multiple scales of aggregation (not just the national level) and also to the many exchanges and migrations across national frontiers. We need global historical data, a comprehensive repository, and globally interactive analysis. The most essential need is for a federalized set of linked repositories—distributed but coherent—containing comprehensive world-historical data, supported by an extensive infrastructure and staff, with powerful systems of analysis and visualization (Manning, 2013). These repositories must be open to all, with facilities designed for researchers at all levels as well as for students and the general public. More broadly, the repositories will function through collaboration among groups building, contributing to, and using the resources. These repositories will not be centralized. Instead, although distributed across several world-historical research initiatives, they will be federated and linked through overarching conceptual and technical standards and through powerful search and analysis tools. Here too, flexibility is paramount, so several strategies for combining, collaborating, and linking datasets and research projects must be employed. Structuring datasets as linked open data is a particularly powerful means to achieve this (Bizer et al., 2009; Brennan et al., 2013). Further, given the scale of the task a key feature of these repositories will be their crowdsourcing facilities, which enables qualified users to participate in data input, documentation, and editing. These resources will provide data on a range of units of analysis, link data across different units of analysis, and test research hypotheses on historical and social change at all levels.

Objectives. Creation of world-historical data resources will require collaborative groups, working on well-chosen, specific objectives. Yet the work must be flexible, able to change course based on new technologies and new discoveries about the human social system, and based on changing social priorities as to what are the most important issues to study. Six related goals stand out as critical first priorities:

BDHH emphasizes collaboration among social science research groups working on large-scale social issues.

A campaign for global collection and systematic documentation of global historical data.

Achieve gold standards, described in robust documentation, on data and metadata standards and on interoperability between the different ontologies of the constituent repositories.

Integration of data to create aggregate-level data, for which disaggregated elements also remain available.

Estimation of missing data at multiple levels of aggregation.

Coordination in analysis at multiple scales.

Challenges in the scientific and technical work of social science and information science: Certain recurring areas of difficulty continue to inhibit the work of creating world-historical data resources.

Global frameworks. Conceptually, global frameworks remain unfamiliar to many researchers: they have grown up thinking of the world as a collection of independent units rather than as a broad and interactive system. As a result, elementary errors in designing global research are not uncommon: for instance, designating modern countries as analytical units and "transposing" these back in time for historical comparison, rather than exploring consistent geographical regions, such as exploring the long-run economic development of "England" or "China" over thousands of years overlooking changing borders (Maddison, 2007).

Documentation: Metadata. To be linked and aggregated effectively, each constituent dataset needs a set of standardized metadata (including dataset name, source and rights, variables, scope [topic, space, time, scale], and previous transformations). Expanding this minimum list, for example by including metadata on the data quality, on levels of uncertainty, and on levels of disagreement, is the next challenge as this allows to engage with the data in more complex ways.

Linkage. Existing social science datasets are not easily linked or combined. The ICPSR and Dataverse repositories hold many thousands of social science datasets (in Excel, SPSS, or SAS format). But there is yet no easy way to search data across these discrete datasets, nor to combine them into a larger scale dataset, thus inhibiting global analysis. (Text-based datasets, in contrast, are much more easily searched, though less amenable to quantitative analysis.) A third bottleneck is the difficulty of documenting social science data in systematic and comprehensive terms.

Ontologies. To enable comprehensive documentation, comprehensive ontologies must be maintained on scope (topic, space, time, and scale) and on data transformations. Documenting what precisely each datum held in repositories is, what time/space/topic it details, who created it and when, and similar information is critical to collate information across datasets.

Missing data. Another critical bottleneck is the problem of missing data: at any stage, social science datasets commonly have large proportions of missing data. In some cases, the missing data exist but are not accessible; in other cases, they do not exist but can be estimated or simulated. Various methods, from multiple imputation to establishing upper and lower bounds, can be used to fill in gaps and allow the data that does exist to be analyzed without running into issues of over- or underfitting.

Collaboration. The practice of scientific collaboration needs to advance among historians and social scientists. For instance, how can we arrange for regular meetings—perhaps biennial—of most of the groups mentioned above, to help advance their work? As large-scale collaborative projects involve many scholars from very different backgrounds, ground rules need to be established to ensure that credit is shared appropriately and in ways that dovetail well with established traditions of giving credit for all involved disciplines.

High-speed computation. World-historical research requires access to the hardware, software, and staff assistance with high-speed computation that will make it possible to assemble and analyze global historical datasets.

Challenges in the social application of scientific research. In the interface between social science researchers and the public, the most immediate bottleneck is the difficulty of arranging research funding—at local levels and global levels.

For research of global significance, one must think of global organizations (such as the UN, UNESCO, or World Bank) and of international collaborations, private or public. Yet here it is worth noting that, of the large-scale research efforts noted above, whether in natural science or social science, the US Government and, at least until recently, the European Union have been the main supporters of really big research. Funding for social scientific research, however, remains low relative to other areas and is in danger of dwindling further, especially in the United States (the budget for the US National Science Foundation over the past 10 years has been approximately 8 billion USD, of which only about 17% went to social science research (AAAS)). In Europe, the EU's Horizon 2020 program provides almost €80 billion of total funding between 2014 and 2020, of which 24.4 billion (about 30%) is designated for research addressing "major social concerns" (EC_Research).

The second bottleneck to large-scale historical research is social opposition. In recent years, with the destruction of priceless historical monuments, we have been reminded that some social interests wish to destroy the past, rather than preserving it. At a less obviously destructive level, many social interests are fearful of any change to the social order (or of potential loss of their privileged positions), and therefore oppose research and analysis that appears oriented toward change.

The third bottleneck is human reluctance to join in large-scale social collaboration. So far, it is only in war that whole societies can be brought to work as one. Thus, despite widespread recognition of the dual threats of social inequality and environmental degradation, it is difficult to achieve consensus or concerted action to address either crisis (McNeill, 2000). In particular, social-scientific scholars are still slow to examine how the various aspects of social science analysis interact with one another. For instance, we believe that the crisis in inequality interacts with the contemporary crisis in climate, at least in the sense that social disruptions caused by climate change amplify inequality—and vice versa. As a result, the paired crises in climate and social inequality pose challenges in both research and policy (Manning, 2017). Our effort is to put these aspects of human history in contact with one another. In our opinion, the general and specific needs for world-historical analysis interact with each other. We need to develop an overall system for documenting historical patterns of human society, focusing specifically on patterns of inequality. In particular, we seek to focus on the global crisis in inequality and its interaction with the contemporary crisis in climate, as social disruptions caused by climate change amplify inequality—and vice versa. As a result, the paired crises in climate and social inequality pose challenges in both research and policy.

Natural science research on climate change has advanced greatly, while social science research on inequality has advanced only a little. Yet there remain parallels between the two fields of study. In each case, there has been the question of whether new knowledge is seen as of interest and worth investing in. For inequality and climate change, each is at once a scientific issue and a policy issue. For climate change, after verified results and a clear scientific consensus took hold, powerful public interests rejected the analysis and refused to take policy action. For inequality, if future studies were to lead to a scientific consensus and policy recommendations, one can be certain that powerful interests will refuse to implement research findings or take up an active policy. In each case, however, the underlying crisis advances and, with time, forces additional people to recognize it as a threat requiring study and action.

3.09.2 Organizations and Collaborative Infrastructure

"Big Data in Human History" is the collaborative node of the participating organizations. It was established in July 2016 after years of informal and smaller scale collaboration among the organizations. Creation of its mission statement (section "Big Data in Human History: The Mission," above) clarified many issues in the plans for collaboration, especially in developing a language for discussing research projects set at different time frames. The various groups have built information infrastructure within their research groups. The advances are roughly parallel, and contacts among groups have been fruitful. We expect that there is a need for building an overall, shared information architecture while allowing for variety and innovation in its constituent parts.

3.09.2.1 Collaborative for Historical Information and Analysis (CHIA)

The Collaborative for Historical Information and Analysis was formed under the leadership of the World History Center to undertake the project of building a world-historical data resource for the period after 1500. Five US universities joined in the project and shared a grant for the creation of the information infrastructure of such a data resource (Zadorozhny et al., 2013).

Research collaborative. The CHIA Research Collaborative functions at the levels of research institutions and individual researchers, to establish close contacts among research institutions in various disciplines and various regions, to share data, and to collaborate in research projects (Manning, 2013). Further, the Collaborative is to facilitate the sharing of data by individual researchers through crowdsourcing. The Collaborative is thus to exist at multiple levels. The principal objective of the Research Collaborative is the demonstration of a continuously growing accumulation of historical data. Additional activities of the Research Collaborative are to be collection and delivery of several other sorts of historical data, structuring of an efficient, transdisciplinary social network, and annual meetings to review priorities and achievements of the collaborative (**Table 1**).

Table 1 List of 30 NGAs

World region	*Low complexity*	*Medium complexity*	*High complexity*
Africa	1. Ghanian Coast	11. Niger Inland 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, PNG	20. Chuuk Islands	30. Big Island Hawaii

Source: Turchin P, Brennan R, Currie TE, Feeney K, Francois P, Hoyer D, Manning J, et al. (2015) Seshat: The global history databank. *Cliodynamics: The Journal of Quantitative History and Cultural Evolution* 6(1): 77–107.

Headquarters. The Headquarters houses elements including an archive of global historical data (reaching across time, space, scales from local to global, and data from various disciplines); a clearing house to facilitate collaborative development of consistent data and metadata; and an intellectual center to develop global connections in social science theory and to make key decisions in developing the overall project. The Headquarters too exists at multiple levels. The principal deliverable of the Headquarters group is an archival system able to hold all the incoming data in a form enabling analysis throughout the repository. Additional activities and deliverables include overall project design, upgrade metadata, analysis and visualization, and project communications through meetings, website, and journal. The Headquarters function is that of overall leadership of CHIA. Within Headquarters are located the Executive Committee of CHIA, administration and finance, and the *Journal of World-Historical Information*. Headquarters will also conduct annual meetings for CHIA as a whole, will coordinate the development and submission of grant proposals, and will coordinate the recruitment of new affiliates. One more function is to maintain (or certainly not inhibit) feedback and interplay of projects, data layers, and affiliates.

Interplay of collaborative and headquarters. This interplay may be illustrated through the issue of creating the global archive: Is one to construct it by assembling existing pieces and moving incrementally, or should one begin with an all-at-once design? Reliance on both the Collaborative and Headquarters seems to be the best choice.

3.09.2.2 Seshat: Global Historical Databank (Seshat Databank)

Seshat: Global History Databank is a large, international research initiative, housed by the innovative think tank, the Evolution Institute (Seshat). Seshat seeks to create digital infrastructure built upon a vast repository for structured data to facilitate systematic testing of social science theories and exploration of Big Questions. Our time frame is the last 10,000 years of human history (i.e., from the Neolithic to the present).

Cliodynamic approach to history. The overall mission is to gather information on historical polities in order to rigorously test different hypotheses pertaining to the rise, the staying power, and the fall of these societies across the globe and human history. Our approach is essentially cliodynamic, namely the marriage of "scientific" methods and analytical techniques with historical investigation (Turchin, 2008).

Seshat is at heart a collaborative venture. The core team consist of a multidisciplinary group of scholars, including evolutionary scientists, historians, anthropologists, archaeologists, economists, and other social scientists. An Editorial Board facilities the integration of individual research projects with the overall Seshat project and identifies new themes and theories to be tested. Research Associates and Assistants lead day-to-day tasks, contributing to project design, analysis, and publication.

Collaborating with experts. Critically, Seshat involves collaboration in two key ways. First, scholars in various fields—from sociology to economics to evolutionary theory—consult on every Seshat research project. Their participation is essential in terms of identifying the main theories, explicating hypotheses, generating a list of variables needed to test these hypotheses, and performing statistical analyses. Second, all Seshat data is vetted by domain experts, namely scholars specializing in the topics and historical periods we are exploring.

Collaborating with data. Seshat is designed to be flexible, scalable, and compatible with other repositories of historical information (Turchin et al., 2015; François et al., 2016). Our scope is constantly expanding as new research projects are formed, leading to new swaths of data. Further, all Seshat data is made freely and openly accessible to the public, with the hope that this will facilitate future research. Seshat stores data as a graph-based, RDF linked-data format. This allows for easy and flexible storage of metadata as well as facilitates collaboration with other Web-based data projects, particularly ones likewise employing an RDF-based data structure (e.g., the IISH projects noted here). Importantly, continued collaboration in the future will benefit from the spread of further well-structured, linked digital data, the ability to combine the data in Seshat with information from projects like CHIA, IISH, and IROWS. To accomplish this, data across projects need to be compatible and interoperable (see "Data and Metadata" and "Ontology" below).

Associated projects. The Seshat initiative pursues several related, though distinct, research projects on various Big Questions (Seshat). By 2016, these include the evolution and principal components of social complexity, causes and nature of warfare, dynamics of moralizing religious thought and egalitarianism in the human system, and deep roots of modern economic and political development. These projects are funded by several large research grants from various sponsors, including Tricoastal Foundation, the European Commission's Horizon 2020 program, the Economic and Social Research Council, the European Research Council, the John Templeton Foundation, the Irish Research Council, James Bennett, and Bernard Winograd.

3.09.2.3 International Institute of Social History (IISH)

The IISH has four major sections to its research activities. Each of them supports several autonomous research projects with various sources of funding. In the discussion here, key projects are emphasized within each section.

Global labor history. This is the largest section of IISH research, addressing all aspects of labor in the modern world—not only wage laborers, but also chattel slaves, sharecroppers, housewives, self-employed, and many other groups. The section has as many as 11 ongoing projects, of which attention is given here to the most fully developed of them, the Global Collaboratory on the History of Labor Relations, 1500–2000. The project is intended to collect primary data on labor relations at selected cross-section years: 1500, 1650, 1800, 1900 [Africa also 1950], and 2000. Researchers, specialized on local regions and time periods, work with an elaborate common typology to develop a comprehensive inventory of people and types of labor relations at these moments. Each research group's results are reviewed for consistency; periodic meetings of researchers by continental region develop a large-scale view of patterns of work as they change over time.

Global migration history. Global Migration History, while closely related to the research projects in labor history and in economic history, takes account of the specific characteristics of migration. In a series of conferences and books on migration organized mostly by continental regions, this research is developing a general measure of migrations among regions in the modern world, the "cross-cultural migration rate (Lucassen and Lucassen, 2014)."

Global economic history. The IISH supports six current projects in economic history, both regional and global, of which the CLIO-INFRA project is the most developed. CLIO-INFRA collects a wide range of data on wages, prices, output, and demographic measures. It has published a widely consulted survey of global well-being since 1820 (Van Zanden et al., 2014).

Structured social economic data projects. This research section includes IISH participation in CLARIAH, a multidisciplinary network linking researchers in media studies, socioeconomic history, and computational linguistics. The project is aimed at creating a common infrastructure for the arts and humanities.

3.09.2.4 Institute for Research on World-Systems (IROWS)

IROWS conducts research on a wide range of topics, gathered into three major sections. Two of the three sections focus especially on analyses of interacting societies from the era of cities forward to the modern period, but also linking to the modern period. Research results are made available through the posting of over 100 working papers on the IROWS website. Recent research is as follows:

Settlement and polity upsweeps and global state formation. IROWS is conducting research on the growth/decline phases of polities and settlements and on the evolution of global governance and global state formation. These projects address the relationships between climate change, epidemic diseases, and rise and fall of cities and empires since the Neolithic and Bronze Ages.

Modeling sociocultural evolution since the emergence of sedentism. Research in this area develops computational models of the processes of demographic growth, resource usage, environmental degradation, population pressure, warfare, population cycles, and rise of sociocultural complexity in regional world-systems.

Transnational social movements. IROWS studies participants in the World Social Forum, relationships among the transnational movements, and global party formation.

3.09.2.5 Minnesota Population Center (MPC)

The Minnesota Population Center is an interdisciplinary cooperative for demographic research. It focuses especially on population data science and on census and survey methodology. In addition, MPC conducts research on population mobility, reproductive and sexual health, and work, family, and time. All of this research is conducted within the era of population censuses.

IPUMS international and IPUMS-US. The Integrated Public Use Microdata Series (IPUMS), with international and US versions of this service, provides samples of individual census records that can be defined by place and time, enabling researchers to conduct individual-level research on populations that would otherwise be unavailable to them.

Terra Populus (TerraPopulus). This project emphasizes integration of data on population and environment—data from censuses, climate data, and data on land cover and land use.

3.09.3 Scale and Theory

This section deals with two main issues: (1) How we will handle linkage and interplay of different temporal scales, from contemporary to early times as discrete periods, and including long-term analysis. (2) What are the main theories? Note that theories are most commonly set at a given scale—so how does one link theories across scales? Work at multiple scales is not new, but needs to be advanced and strengthened. Within scales, we'll consider temporal, spatial, topical, and aggregative scales. These scales, though

fundamental to analysis, are not necessarily inherent characteristics of historical materials: scales are labels given by the analyst in modeling the past. We design the scales to be as objective and as reproducible as possible, but ultimately they depend on the perspective of the observer, the analyst.

3.09.3.1 Scales

The various scales of a global framework are of different orders. The *temporal scale* is fundamentally ordinal—one can clearly rank early times and later times in the order of their occurrence. In some but not all cases, time can be analyzed in interval terms as well as ordinal terms. The *topical scale* is fundamentally categorical: it refers to the many subjects that can be analyzed in historical terms—for instance, social, cultural, and environmental phenomena. These can be treated in ordinal terms only if there is a clear ranking within the topical scale. The *spatial scale* is also fundamentally categorical. While people tend to prioritize the space they are in as the most important space, the differences of perspective among people cancel out this priority. Larger spaces can be ranked over smaller spaces, but spaces cannot generally be ranked unless they are liked to additional criteria, such as the number of inhabitants in a space. The *topical scale* is fundamentally categorical rather than ordinal. That is, while one can rank the quantities of material analyzed within a cultural framework, distinguishing the broader from the narrower analysis, such ranking can only go so far. One can distinguish social, cultural, and environmental phenomena, but attempts to rank them are largely arbitrary. Finally, the overall *aggregative scale*, while difficult to specify, is intended to be ordinal.

Spatial scales. Spatial scales range from the local to the global. The device of the gazetteer has been developed to summarize names, locations, time frames, and other data on places. For long-term and global analysis, additional problems arise that complicate the labeling of space and its use in slavery. Natural spaces (rivers, mountains, deserts) change only modestly within human history, but are labeled in many languages. Social spaces, created by humans, vary greatly over time, as do their labels. Spaces overlap in important ways, as with contested borders and multiple identities existing simultaneously at different scales; any given building is at once part of a neighborhood, city, province, country, region, etc. Space can also be highly subjective, and its boundaries often lack precision.

Temporal scales. Immediate concerns bring short-term research, documenting the problem and hoping that the solution is right at hand (organize by periods, subperiods, aggregations of periods). Research is best conducted at multiple scales.

Short-term theory. Theory may trace the pace of change over short periods of time, for instance at annual and subannual levels in economics. It can also address change from generation to generation or from century to century, with variables adjusting as appropriate for different time frames.

Theory at the scale of social eras. Theory at the level of human history must account for social eras and their interactions. Eras commonly listed include the hominid era, the era of *Homo sapiens* diaspora, the era of agriculture and other productive activities, the era of cities and states, the era of rapid global connection, the industrial era, and the contemporary era.

Theory linking major social eras. Our long-term, multiscale approach is based on the reasoning that human society of today depends on continuities as well as on changes, and on practices set in place at and across each stage of human development. Assuming that a problem appearing in a given era has its origins only within the frontiers of that era leaves untested the possibility that the problem may in contrast be newly unfolding implications of earlier changes.

Topical scales. Topical ontologies have been created for certain specialized areas: in medicine and in areas of marketing, for instance. Library catalog systems (such as the Library of Congress system) are more general, permitting classification of most published works. For the interactive, historical analysis of human society, a still more complex ontology will need to be created. Meanwhile, individual scholars are able to fashion descriptions of topical scales to assess the narrower and broader approaches in social, political, economic, cultural, environment, biomedical, and technical history.

Aggregative scales. Aggregative scales refer to the overall breadth of a study or an interpretive issue. One measure is the total amount of information in data files, in bytes or megabytes. Other measures might seek to account separately for the scale in space, time, and topic.

3.09.3.2 Theories, General and Specific

The purpose of this section is to identify major theories and paradigms applied in historical social science analysis, and specify the times, spaces, topics, and levels of aggregation at which they are most often applied. A thorough job of such mapping of theory would show where theories are in competition with one another, and where historical phenomena go untheorized. Theories need to be explicitly tested side by side against the historical record; the "less supported" theories are rejected, while the "more supported" theories continue to be refined and tested further (Turchin, 2008; Turchin et al., 2012; Hoyer and Manning, forthcoming). Theories are organized especially by the topic of their analysis, and theoretical topics are typically organized by discipline. For human society, the disciplines are commonly listed within the broad topical confines of the political, social, economic, and cultural. A growing area of theory and analysis is in human-natural interaction: specifically studies in health and in ecology or environmental studies. Social theory posed at a global level, however, remains vague for lack of comprehensive data to explore. Rather than waiting for a gradual accretion of localized projects to bring about large-scale analysis, the work of our collaborative is intended to conduct a large-scale initiative to speed the transition into global social analysis. While the initial section of this review of theory addresses some general questions in the nature of theory, the later portions of this section address theories applied at the global level, notably by researchers in our group, rather than attempting to survey social science theory in general.

General questions in social science theory

Levels of specificity in theory. Theories are sometimes as simple as typologies of phenomena under study, such as types of family organization or scales of civilizational development. Or theories may take the form of an emphasis on a certain variable or phenomenon as of central importance in a problem under study, without advancing hypotheses of any more specificity. Further, theories have some relationship to case studies, which explore the complexities of cases but stop short of generalizing them. More generally, a theory can be found to rely on a set of philosophical assumptions and to be expressed in a framework providing the boundaries of the subject under study. A theory generally identifies key variables or factors under study and advances hypotheses on the relationships or dynamics linking them. Theories usually involve the search for causal relationships. This is often expressed by dividing variables into dependent and independent variables; it may also be expressed through the assumption of mutually interactive dynamics.

Theories by scale. The framework for a given theory includes the topical limits, and relevant spatial and temporal limits. Theories are also defined in terms of their aggregative level: typically as micro- or macrotheories.

Theories by topic and key variables. Examples of arenas of theoretical analysis: social and economic inequality, governance, economic change, social and political institutions, well-being or quality of life, evolutionary change, cultural practices, family, conflict and violence, health and healing, and environmental impact of human society.

Policy variables. Social science theories commonly give attention to "policy variables." Such variables, regardless of their significance in overall causation, are understood to be under conscious human control, and therefore present tools for explicit manipulation of social outcomes by human agents—typically, by governments.

Advances at isolated scales. Analysis in the social sciences has developed impressively in the last 50 years, with many advances at micro-, macro-, and (increasingly) mesolevels of theory and research (Gerring, 2012; Lange, 2012; Skocpol and Somers, 1980; Hoyer and Manning, forthcoming). Most of these advances, however, have taken place within sets of constraints that have made the social sciences increasingly diversified and subdivided. Rather slower to develop has been attention to linking the various subtheories in each discipline to each other. Thus, behavioral approaches have become influential in microeconomics, but it is not yet clear whether the behavioral approach has implications for macroeconomic analysis (Glimcher, 2009; Camerer et al., 2005). Sociological studies at micro- and macrostudies show considerable divergence; studies in comparative politics focus on national government, almost to the exclusion of trends and traditions in local governance. Further, general reviews tend to address social sciences in parallel silos rather than focusing on their interactions or on overall developments in the logic, philosophy, and empirical base of social science knowledge. In particular, the increasingly acute problems of social inequality have not yet led to large-scale, cross-disciplinary efforts to address the interacting dimensions of inequality in economic, social, political, cultural, and biological affairs.

Advances within disciplinary silos. The social sciences have thus responded to globalization more with intensive development of subtheories than with extensive explorations across disciplinary frontiers. For all their sophistication, they give minimal attention to change over time, global patterns, and cross-disciplinary effects. (Sociology and anthropology have little close contact, development economics has little to do with economic history, and links of demography and health are only now coming under serious study.) All in all, the current state of social science analysis accords low priority to studies that are long-term in their time frame, multiscale and especially global in their spatial scope, and cross-disciplinary in their analysis of social dynamics. Yet the current problems of globalization suggest that there is a great need for information at all of these scales, despite their relative complexity. Investing in the creation of global data will launch this wide range of discussions.

Linking theories. Distinctive theories have grown up in each of the social scientific disciplines: political science, economics, sociology, psychology, and anthropology; historians remain reluctant to embrace theory. The task of global social science, however, requires finding ways to link the various social science theories to each other and generate ways to incorporate critical historical information into theoretical frameworks. One insight is that, since all social-scientific theories focus on human individuals and groups, all the theories may be linked through various population and migration variables (Manning and Ravi, 2013).

Transcending boundaries in social science theory. A primary charge of our group is the creation of theoretical frameworks that identify mechanisms in human systems, notably mechanisms creating (or reducing) various inequalities. This fundamental work will transcend traditional boundaries among social and natural sciences. It will document multiple dimensions of inequality, link them to each other, and reveal feedbacks alternatively reinforcing or diminishing inequality across varying temporal or spatial scales and across gradients of social structures (Manning and Ravi, 2013). One of the most challenging but also most important aspects of this work is the commitment to a systematically global approach over multiple time frames. This scope is necessary to understand the global interactions driving these social structures and will require estimation of missing data through simulation and agent-based modeling, extensions of the top-down analysis of the institutional approach, and incorporation of bottom-up case studies from social history into a broad framework.

Indices of social change. Theories require measurement, quantitative or qualitative. Development and testing of theory on human society require the specification and clarification of measures of societies and their characteristics. Here are four types of measurement that are applied to the past.

Population. Of cities, states, provinces, regions, empires. Based at best on formal censuses; otherwise on enumerations or estimates of variables thought to correlate with population. Also included are estimates of expectation of life and age–sex ratios.

Production. Calculations of gross domestic product and gross domestic income, developed for contemporary analysis in the 20th century, have led to speculative estimates of GDP per capita that are applied to historical situations.

Inequality. Various indices of inequality within populations have been developed, including the Gini index, the Foster–Greer–Thorbecke index, and the Sen index. These indices each has significant weaknesses, but using multiple measures helps to clarify degrees of inequality.

Migration. Migrations are often estimated as crude numbers of migrants, as compared to base populations of regions of departure and arrival. Recent work seeks to develop estimates of rates of out-migration and in-migration for specific territories within the last few centuries.

Theory in analysis of global social inequality. Three main research themes have been selected in order to sort and discern the multiple dimensions of inequality and its causal factors:

Impact analysis. How much of each aspect of inequality results from the natural setting, from institutions, and from social processes? This aspect of the research, conducted through cross-sectional correlations across many boundaries, is to determine the relative significance of natural, institutional, and social factors in various sorts of inequality. This basic analysis is expected to confirm the multicausal nature of inequality.

Scaling up and scaling down. To what degree is inequality an emergent property of society resulting from the steadily greater scale of institutions? In contrast, to what degree do individual agency and social movements influence inequality? This second aspect of the research is to trace the patterns and relations of inequality across scales from the individual to the societal.

Scaling to the global. What are the worldwide characteristics of inequality? In what ways does the closed system of global society affect inequality by imposing limits on its open and interacting subsystems? This research theme is to characterize inequality at the global level, identifying distinctive global patterns and linking them to inequality in various subsystems.

These research themes, pursued in sufficient detail, will yield theory and testable hypotheses on inequalities of varying sorts and at different scales of social aggregation. In addition, once sufficient data have been assembled, it will be possible to seek out representations of additional causal relationships in the data through application of causal Bayesian networks (Jensen and Nielsen, 2007).

3.09.4 Current Research and Collaborations

Several social science research groups have conducted research on large-scale social phenomena. While they focus on a wide range of world-historical topics, they share a great deal in methodology and philosophy; all of them encounter the inherent problems in gathering historical data and attempting to link local data and aggregate them to a global level. They work on different time periods, and these time periods commonly correspond to differences in the nature of documentation, nature of society, and questions under study. Nevertheless, the Big Data in Human History collaborative decided to seek out analytical parallels and overlaps among the temporally distinct research projects. We are developing a discourse that enables us to keep multiple and overlapping time frames in mind. We begin by dividing all of human history into seven periods and then distinguish studies conducted within individual periods, studies comparing different periods, and studies encompassing two or more periods. We describe the various time periods, starting with the most recent and working back in time, with reference to research projects focusing on each and collections of data focusing on each.

3.09.4.1 Research Within Basic Periods

Here we list seven commonly identified periods of human history, indicating their rough temporal extent and some of the historical problems and historical research projects that have focused on each. We list these distinct periods partly to emphasize that a great deal of historical research and interpretation take place within these discrete periods—but also to emphasize that research and interpretation take place at a larger level, sometimes comparing processes in the various periods and sometimes combining the initially listed periods into longer, more encompassing periods. At the maximum, we can imagine a historical interpretation of all of humanity.

Basic time periods. The following basic time periods for the conduct of large-scale social science research are here identified as follows: contemporary (since 1950), industrial (1800–1950), era of global interconnection (1500–1800), late medieval period (1000–1500), era of cities and empires (BCE 3000–CE 1000), era of agricultural development (10,000BP to 5000BP [or BCE 3000]), and era of human diaspora (80,000BP to 10,000BP). Each mutually exclusive temporal period is described in terms of general themes, specific research projects, and existing data resources.

Contemporary, since 1950. National level statistics on population, economic and social affairs, assembled by the World Bank, the International Monetary Fund, UNESCO, and OECD. IROWS conducts research on social movements, party formation, global class formation, governance, globalization, empire, and decolonization. On economic inequality: Piketty and Goldhammer (2014), Milanovic (2016), Korzeniewicz and Moran (2009). Resources include ICPSR, WTID, SEDLAC, UNU-WIDER, and UTIP.

Industrial era, 1800–1950. Social, economic, and natural changes in national and industrial era. National studies of 1918 influenza pandemic (Chandra et al., 2012) and economic inequality (O'Rourke and Williamson, 1999). Resources include a social science repository (ICPSR) and city–county data on US health statistics (Tycho).

Era of global interconnection, 1500–1800. Rapidly changing social, political, and economic systems. Migration, population movement, and demographic shifts. Impact of technology, labor mobility, and increasing productivity on development in northern Europe, contrast with developments in southern Europe, East Asia, India, Africa, and elsewhere. Resources include data on some 35,000 Atlantic slave voyages (Voyages).

Late medieval era, 1000–1500. Major research projects in this era have addressed the administration of Song China, the Medieval warm period in Europe, and the plague pandemic (Mostern, 2011; Campbell, 2016; Green, 2015). For data on this era: SONGGIS.

Era of cities and empires, BCE 3000–CE 1500. In addition to documenting the rise and fall of large states, research on this era addresses religion in the Axial Age (CHECC). Resources include digital repositories of archaeological, epigraphic, papyrological, and numismatic material, notably Pelagios Commons, OpenContext.org, The Digital Archaeological Record, papyri.info, nomisma.org, and the Epigraphic Database Heidelberg.

Era of agricultural development, 10,000BP to 5000BP [BCE 3000]. Research, especially with archaeological techniques, traces early formation of complex social systems, development of sedentary agriculture, and global population growth (eHRAF-cultures).

Era of human diaspora, 80,000BP to 10,000BP. Genetic research is currently primary for this area, supplemented by archaeology, with some work on the distribution and evolution of language groups (Language and History). Data archives have yet to be constructed for this early period, except in genetic data.

3.09.4.2 Research on Multiple or Combined Periods

Research spanning or combining the periods listed above are likely to require collaborations among research groups.

Short modern era, 1800–present. Transnational social movements, tracing efforts to engage in collective action. Industrial and contemporary eras. Institute for Research on World-Systems. Resources: For census data, the largest collection of historical data is at the Minnesota Population Center, whose data collections include IPUMS, for which US and international versions provide users with samples of individual data drawn from full-scale censuses (MPC).

Long modern era, 1500–present. The Global Collaboratory on the History of Labor Relations (GCHLR). CLIO-INFRA project on documenting economic history (CLIO-INFRA, Maddison). Correlates of War project has assembled numerous databases on topics related to politics (CorrelatesOfWar). African Population: estimation of population and migration for the African continent and 70 subunits, 1650–1950 (Manning et al., 2015a,b). From IROWS: research of globalization since 1800; global elites; IISH: global labor history; migration history; economic history. One may test the hypothesis that natural factors influenced the decisions made during early colonization and land division in North America (Bain and Brush, 2008).

The past millennium, 1000–present. Douglass North led a group proposing that development of efficient social institutions was the principal source of social change and economic growth during the past millennium (North et al., 2009). The approach was anticipated in the work of Charles Tilly; it was paralleled in some degree by the long-term economic analysis of Daron Acemoglu and James A. Robinson, and has been followed up within the arena of labor history by Leo Lucassen (Tilly, 1998; Acemoglu and Robinson, 2012; Lucassen, 2016; Turchin and Nefedev, 2009).

Era of civilizations and empires, BCE 3000–present. Seshat applies a cultural evolutionary framework to diagnose long-term and global patterns in social, cultural, political, and economic dynamics (Seshat). From IROWS: Cities, states, and world systems; long sweeps of world systems, empires, and cities; cycles in globalization, empire, and climate change; world systems at all scales; state formation (IROWS). World Cultures: indexing ethnographic collections widely covering aspects of cultural and social life (eHRAF cultures).

Era of agriculture, 10,000–present. Seshat: Global History Databank. Cultural evolution as framework to uncover dynamics of human system. Large database with stratified sample of societies, tracing cultural evolution and historical causality with multiple analytical techniques (Seshat). Indexing of in-depth descriptive documents of archaeological traditions from around the world (eHRAF archaeology). Peter Richerson and Robert Boyd have led in synthesizing and analyzing processes of cultural evolution (that is, social evolution), in contrast to biological evolution (Richerson and Christiansen, 2013; Boyd and Richerson, 2005).

Precivilizational era, 80,000BP to BCE 3000.

Human era, 80,000BP–present. Genetic history. Human genome. Also the genome of other plant and animal species important to human society. Studies widely dispersed among different research groups in biology, human genetics, anthropology, etc.

3.09.4.3 Research Crossing Topical Areas as well as Time Periods

Contemporary study crossing human-natural boundaries.

Terra Populus, contemporary. Workflows for integrating and harmonizing geospatial population and environment data. Contemporary, demography, and environmental studies (TerraPopulus).

Linking variables in post-1500 study of inequality.

The CHIA on historical inequality explores interactions among economic, social, and natural factors (Manning, 2017). **Fig. 1** identifies the eight categories of a grid of numerous variables, in which cause and effect are traced in various directions. Data collection in this project includes efforts to estimate missing data; it also emphasizes extra effort to collect data in the Global South, for instance in the Caribbean (Drwenski, 2015). For a large but condensed dataset constructed in an effort to combine data of natural, demographic, and political origins (Mafrica).

Inequality before 1500.

Anthropologists and other scholars have sought to identify early developments of hierarchy in human society, as well as processes limiting the expansion of hierarchy (Flannery and Marcus, 2012; Boyd and Richerson, 2005, Seshat).

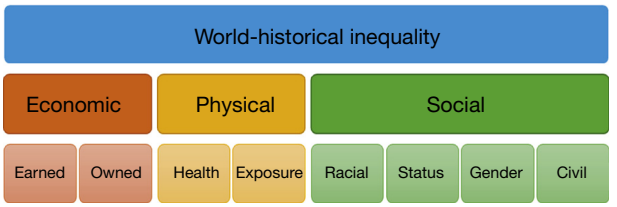

Fig. 1 Conceptual categorization of inequality. Flannery K and Marcus J (2012) *The creation of inequality: How our prehistoric ancestors set the stage for Monarchy, Slavery, and Empire*. Cambridge: Harvard University Press, Boyd 2005, Seshat.

3.09.5 Information Infrastructure

General comment on information infrastructure: Various groups have built information infrastructure within their research groups. The advances are roughly parallel, and contacts among groups have been fruitful. We state that the twin most crucial tasks to achieve in the near future are the ongoing work with shared data and metadata standards and creating powerful digital tools that link the different datasets and query the data. What follows are variegated descriptions of the different groups, but it will be clear that there are benefits in their collaboration.

3.09.5.1 Information Infrastructure at CHIA

The Collaborative for Historical Information and Analysis is a distributed working group, unifying scientists across the United States (University of Pittsburgh [headquarters], Harvard University, University of California—Merced, Michigan State University, Boston University) in the formulation of a mechanistic understanding of inequalities in human systems. The overall envelope of information addressed in the CHIA project is labeled the Human Systems Data Resource. This distributed set of overlapping repositories includes the Data Hoover (for solicitation and ingest of data), the Dataverse Archive (which serves as the Collected Data Repository (CDR) for basic documentation and preservation of ingested datasets), and additional repositories for further documentation, integration, aggregation, and analysis of data. The higher order elements of the Human Systems Data Resource are analyzed with the support of the Pittsburgh Supercomputing Center. Interaction of the elements of the Human Systems Data Resource will allow the formulation of systematic understanding of human social structure across scales, from the individual actors in social history to the institutions regulating individuals and populations and to the interactions among institutions in running the human "world." Further, this advance in the scale and sophistication of historical human understanding will be a resource for a wider variety of users from elementary and secondary students to independent scholars.

Collection and Ingest

The Human Systems Data Resource must be able to identify a wide range of heterogeneous, big, and dark data, organize it, and ensure systematic documentation as the data is incorporated into the resource. The Data Hoover project provides a systematic method for the identification and collection of relevant data via surveys of professional networks (Mostern et al., 2016; Mostern and Arksey, 2016). The Dataverse (v. 4.0, Harvard University) has a well-tested system of ingest and a high standard for documentation and preservation of contributed datasets: it is the repository for all project datasets as they are initially contributed (Dataverse) (**Fig. 2**).

Archiving and Documentation

The tension between the collection of metadata to allow data longevity and the proposed "least sets" assumed to be most likely to be reported is a persistent challenge in the construction of big datasets. Using tools available through Dataverse, project participants will engage in iterative processes examining the optimal set of metadata, informed by the dual role of collector and user. Periods of contraction designed to enhance metadata collection are expected to be followed by expansion as missing pieces become obvious. Three central dimensions of the archive, described below and shown in **Fig. 3**, are the CDR, the Integrated Data Repository (IDR), and the Derived Data Repository (DDR).

Archiving data: Headquarters. Our overall archive is composed of several levels. At the top, global level it provides access to global values of all variables by time but also with access to regional, temporal, and topical subsets of the global values at various scales. This is our objective—it is the dataset to be used for interactive global historical analysis—along with its metadata. The bottom level includes the many datasets originally contributed to CHIA, accompanied by their metadata. In between are an even larger number of datasets resulting from the transformation and aggregation of the original datasets—each with its incremental metadata. Encompassing the archive, the project as a whole requires development of an overall ontology addressing both (1) the technology of data development, storage, and visualization and (2) the domains of world history in the data incorporated and analyzed. We will explore applicability of semantic Web technologies utilizing Linked Data and Web of Data for the task of large-scale historical data integration, though the definitive ontological work may be completed after the three-year RCN stage.

CDR. CHIA infrastructure allows users to submit their datasets with core dataset-level metadata. CHIA Col*Fusion infrastructure stores heterogeneous datasets in separate databases and maintain system-wide catalog with all the required metadata that will allow system to operate. This approach is intrinsically distributed since it allows us putting the data on different machines. The dataset-level metadata includes title and description of the dataset so that other users can browse the repository. Additionally, the infrastructure stores variable-level metadata (such as name, description, type) from all datasets. This metadata enables establishing relationships between datasets in an *IDR*, as explained below.

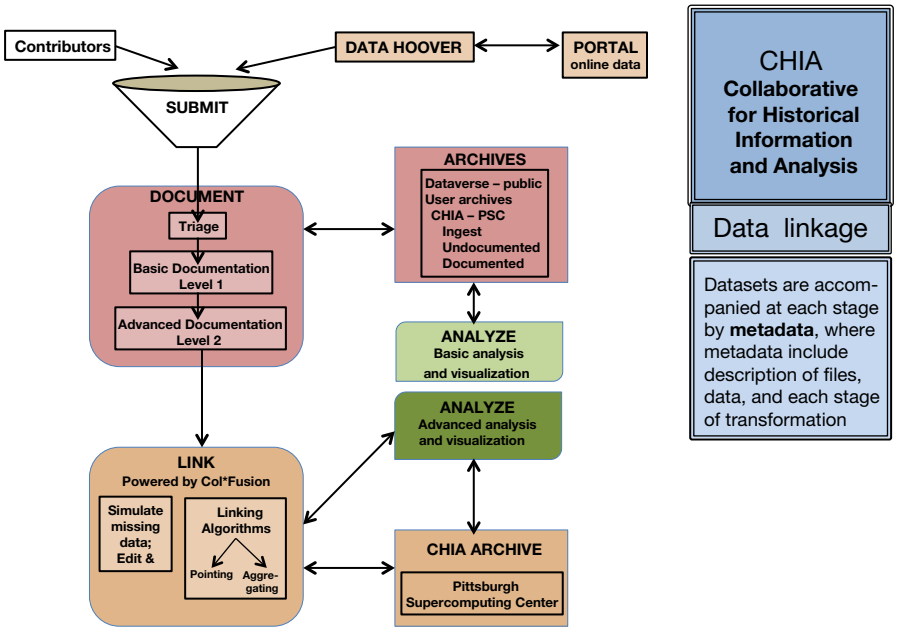

Fig. 2 CHIA system of data processing.

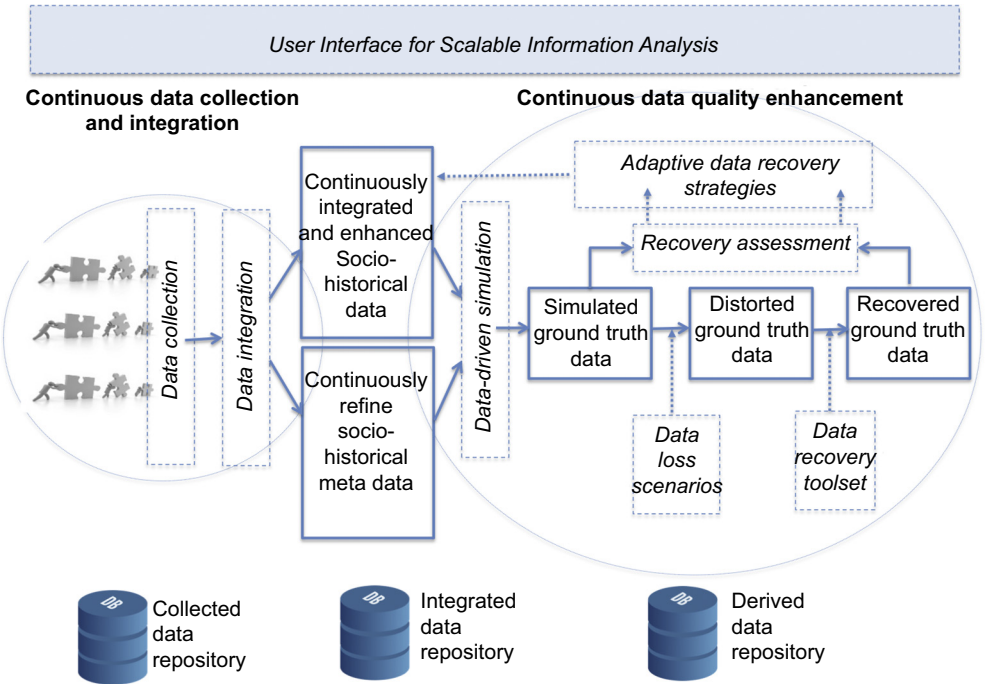

Fig. 3 CHIA system of repositories.

IDR. The challenge of data integration has been actively explored for a long time, with architectures of current data integration systems varying from data warehousing to virtual databases (Doan et al., 2012). Resolving data heterogeneities has been the focus of active research and development (Brodie, 2010; Haas, 2007). A separate body of research deals with public health and historical data integration (Tycho, Colfusion, van Panhuis et al., 2013). Since it is problematic to develop a predefined schema for large-scale data integration scenarios, especially if the data sources are added dynamically, CHIA approach is to use a bottom-up data integration approach assuming no predefined mediated schema (Doan et al., 2012, ColFusion). Our IDR will support an extendable target schema automatically generated from submitted datasets. In order to integrate datasets, we need to apply schema matching and schema mapping algorithms for all pairs of the datasets in our CDR. We call this process *relationship*

discovery. A relationship between datasets can be established based on various syntactic and semantic similarities between variables' metadata (such as variable name or description).

Target schema. Maintaining a predefined target schema where data from heterogeneous data sources are supposed to be loaded to as in traditional data warehousing approach is too restrictive. Instead, we will maintain a global *Relationship Graph*—a undirected graph where nodes represent datasets and edges represent relationships. Currently, we experiment with Neo4j graph database, the leading graph database, to store and traverse the Relationship Graph (Neo4j). We develop an extendable architecture that enables easy way to utilize different schema matching and schema mapping tools. We explore novel approaches combining automatic relationship discovery with crowdsourcing techniques. In addition to submitting datasets, users can provide feedback on automatically discovered relationships and/or create relationship manually.

DDR. The DDR will support the integrated data quality enhancement module. The DDR repository will maintain both structured and semistructured information reflecting various information fusion and reliability assessment options and application constraints to perform large-scale data analysis. Therefore, we will implement DDR as an integrated system utilizing both relational and NoSql data stores (McCreary and Kelly, 2013). In design and development of the DDR, we will use our considerable prior experience in data fusion (Zadorozhny et al., 2015; Zadorozhny and Lewis, 2014), information reliability assessment (Zadorozhny and Grant, 2015; Pelechrinis et al., 2015; Ren et al., 2014), and scalable dynamic data analysis (Kang and Zadorozhny, 2015; Cherniak et al., 2013; Cherniak and Zadorozhny, 2013). Since historical data integration results to be sparse and aggregated at various level representations, the DDR will support various missing data imputation and disaggregation methods.

Implementing Components of the Information Processing Architecture

We plan to explore various options to fully implement the information processing architecture explained in the previous section. In particular, we plan to utilize the Dataverse Archive and other related approaches as explained below. We also collaborate with Pittsburgh Supercomputing Center to provide an efficient HPC support for scalable information processing.

Harmonizing metadata. Incorporated datasets, initially held and documented in the Dataverse Archive, move next to higher levels of documentation and integration, so that they can be aggregated progressively to levels eventually approaching the global. The process of integration itself required documentation, so that the quantity and depth of metadata grow at each stage. Ontologies will be developed allowing worldwide comparisons and data linkage; one must also link the broad ontology to the locally specific ontologies that already exist. A world-historical gazetteer is in development; similar ontologies will follow for time, topic, sources, topical connections, and project procedures.

Choice in aggregating datasets. How are we to design the elements and the structure of an archive of this magnitude? We need to keep an eye on our global strategy, yet be tactically flexible. That is, we need to be certain that we will end up with a fully functional global historical dataset, but we must also be able to navigate the many steps up to that result. CHIA is working with two different architectures for the archive, expecting that we will benefit from both of them and make better overall decisions based on this breadth of experience. The two basic designs are those of the Dataverse Network and a design related to Data Cubes created and hosted by Great Britain Historical GIS (Dataverse, GBHGIS).

The Dataverse Network is an open-source application to publish, share, reference, extract, and analyze research data that facilitates making data available to others. It currently works across datasets by ingesting SPSS and STATA files, extracting the file and variable metadata, converting that metadata into an XML DDI format (Dataverse is fully compliant with DDI schema), and providing search and subsetting of the dataset based on that metadata (Dataverse). The main part missing for enabling it to be spatially and temporally specific is a controlled vocabulary or standardization of time and place, which will allow users to easily compare or even merge different datasets.

The GBHGIS data architecture holds all names in Unicode (UTF-8), its data model supports multiple languages, and the system operates in Postgres. The historical geographic ontology amounts to creating a kind of data in which all individual data values exist more or less independently, not as parts of "datasets." The GBHGIS implementation of this approach holds millions of data values each in one row of a relational table, with one column holding the actual numbers and the other holding those where/when/what/source dimensions. It is also DDI based, but a different kind of data structure, which can directly support analytic operations; it does use controlled vocabularies including one for data semantics, though it does not yet follow any established standard (Southall, 2011).

CHIA is committed at a significant level to working with each of these approaches. We expect that such combinations of approaches will recur at every stage of this large project. The historical data domain is characterized by various querying capabilities and a multitude of data sources. For example, most of the sources provide two types of search capabilities: targeted search based on an identifier and sophisticated search engines based on Boolean keyword queries.

Collaboration with the Pittsburgh Supercomputing Center. CHIA relies on PSC's unique computational and data-handling resources for the advanced information processing required for large-scale data integration, data quality enhancement, and data analysis. Specifically, our infrastructure will leverage PSC's *Bridges* HPC (high-performance computing) resource for data-intensive research (Nystrom et al., 2015, Bridges). *Bridges* is a uniquely capable HPC system designed to enable new kinds of research by providing an extremely flexible software environment, interactivity, and large-memory nodes to allow seamless scaling of in-memory database operations and data analytics. We will leverage those capabilities of *Bridges* to drive the Social Weather Service (SWS). We will exploit the computational power of *Bridges* to explore the use of advanced, hardware-accelerated analytics to enable complex data exploration and to improve interactivity for the research collaborative workflow. To gain access to a substantial allocation, we will develop a research proposal to XSEDE (XSEDE). We have already applied for and been granted a start-up allocation for preliminary experiments and benchmarking.

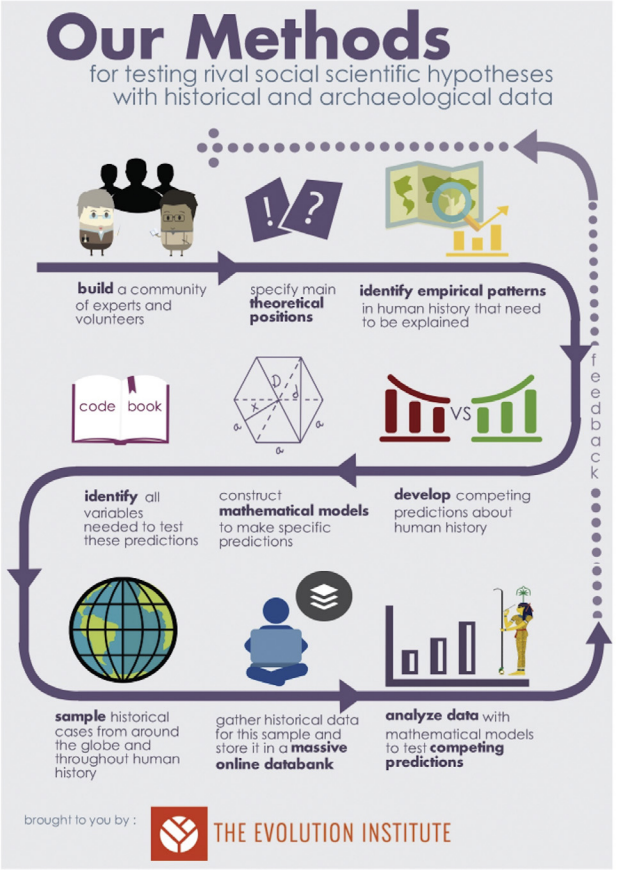

Fig. 4 Seshat methodology for testing theories with historical and archaeological evidence (http://seshatdatabank.info/methods/).

3.09.5.2 Information Infrastructure at Seshat

Seshat's research method (Turchin et al., 2015, François et al., 2016) involves the following steps:

Explicate theory. Specify the main theoretical positions and identify empirical patterns in aspects of human history that need to be explained, e.g., the evolution of cooperation among large groups of people.

Alternate hypotheses. Enumerate a set of competing hypotheses that make different empirical predictions about human history, including the construction of mathematical models that incorporate historical data.

Code book. Develop a "code book" that identifies all of the information we need to gather in order to test these competing hypotheses as individual variables.

Global sample. Draw a representative sample of historic polities for the relevant spatial and temporal frame.

Data collection. Gather historical information from each polity following the variables outlined in the code book, and store this in a massive online Databank.

Analysis. Employ various modeling and statistical analyses to test the competing hypotheses about a given theory. These analyses will determine which hypotheses most parsimoniously explain the empirical patterns exhibited by the collected historical and archaeological material.

Seshat data. To perform analyses on historical data, we define and operationalize such concepts as *social scale*, *inequality*, and *intensity of military competition*. We then systematically collect data on these concepts, remaining aware of the complications and complexities involved in such an endeavor. Seshat data ultimately consists of a complex interrelated set of information about the past. For instance, the numeric data point "50,000,000–60,000,000" refers to the range of best estimates of the population of the Roman Empire in the year CE 100; it is, thus, a quantified number that is georeferenced (to the borders of the Roman Empire, defined separately in the Databank) and temporally bounded (CE 100). Critically, Seshat data contain both quantitative and qualitative information, as here there is disagreement between scholars about the precise size of Rome's population. Including the range of estimates, citing relevant sources and attaching descriptive, explanatory paragraphs to each quantitative data point, allows us to avoid the pitfalls of trying to smooth over the messy, inconsistent nature of historical data (**Fig. 4**).

Data types. Seshat data comprise three separate types of information (François et al., 2016; Brennan et al., 2016). Critically, data can express both scholarly disagreement and uncertainty by giving ranges or multiple values (i.e., when there is disagreement by domain experts whether a particular variable was absent or present in a given polity, we code both). This avoids issues of overspecifying and reductivism when translating the complex historical record into analyzable data.

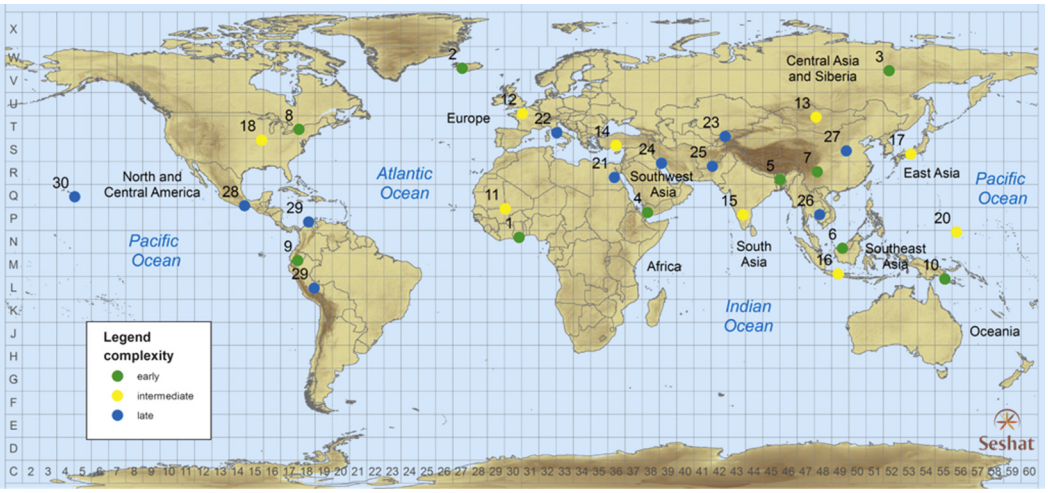

Fig. 5 Seshat's global sampling scheme. Turchin P, Brennan R, Currie TE, Feeney K, Francois P, Hoyer D, Manning J, et al. (2015) Seshat: The global history databank. *Cliodynamics: The Journal of Quantitative History and Cultural Evolution* 6(1): 77–107.

Machine-readable code. The first is a machine-readable code providing quantitative data ready for analysis.

Qualitative Explanation. Every coded variable for every polity in our Databank includes a descriptive passage providing contextual information and explanation for the codes chosen.

Citation. Lastly, each data point contains citations to the sources—primary and secondary source material along with domain expert input—that supplied the information.

Data life history. Metadata captures all relevant information about each Seshat data point, including who uploaded what information, when they did it, what sources they used, and how data has changed over time. Keeping track of such user metrics is essential to assess the quality of the data stored in the databank, to evaluate the efficiency and reliability of users on an individual basis, and to regulate against misuse of data or data creation process.

Data creation. Seshat data are created and verified through overlapping procedures (François et al., 2016). Data are generally first created by Seshat's team of researchers, research assistants, and volunteers. Experts help structuring this data gathering by identifying key data sources. Scouring the historical and archaeological record, data are manually inputted into the Databank for each historical polity in our sample following the code book. This involves both primary and secondary materials and consulting with domain experts to be approved, augmented, qualified, or rejected. Data are also uploaded directly by domain experts like historians, archaeologists, or religious study scholars. Ultimately, each data point will receive input from more than one domain expert and involve numerous rounds of collection and refinement. This iterative process ensures that the most accurate and up-to-date information is being stored and assessed for analysis. Further, an impressive range of digital tools are being developed to help both experts and research assistants to populate data fields more quickly through the EU-funded ALIGNED project (ALIGNED). These tools include notably Web scrapers, which query very large data collections like JSTOR or Google Books and offer "likely candidates" of relevant material or texts to our research assistants for data creation. Such tools will vastly improve the speed and efficiency of data collection while also allowing us to track effectively the progress of data collection by individual researchers, crucial to record the full life history of each data point held in the Databank, a key step identified above to create large, useful, well-integrated datasets for global historical research. Indeed, a major goal of this world historical research repository is to facilitate the creation and sharing of such data by other research projects worldwide; reaching the "critical mass" of accessible data benefits all research equally.

Storage and open access. Once created, Seshat data are stored in our massive online databank as structured, well-defined graph data (see "Data and Metadata" below). Once data in a particular section are sufficient for analyses, they are reviewed, verified by a host of domain experts, and analyzed by members of the Seshat team along with our collaborators. Once the dataset for each cluster of variables reaches maturity, the data are made fully and freely available through Seshat's public website, along with user-friendly interactive tools that allow researchers to explore, combine, and visualize the data (**Fig. 5**).

Global sample. For efficiency, initial data gathering focuses on a subset of historical polities, though eventually we plan to cover all historic polities (Turchin et al., 2015). For the initial sampling scheme, the world is divided into ten major regions, and three natural geographic areas (NGAs) are selected for each region. An NGA is defined spatially by the area enclosed within a boundary drawn on the world map. It does not change with time. Its rough spatial scale is 100 km × 100 km (but can vary severalfold). These 30 NGAs were selected with two goals in mind: (1) to include as much variation among sampled polities as possible, at least along the social complexity dimension, and (2) to ensure that representation of different parts of the world was maximized. The three NGAs in each world region correspond to the relative antiquity of complex polities within it. Accordingly, one NGA was selected in an area that developed complex societies very early, e.g., northern China or Mesopotamia, and another that developed centralized polities (chiefdoms and states) relatively recently, often not until the colonial period. Finally, the third NGA was intermediate in the rise of complex society, such as the Orkhon Valley or the Deccan Plain in India.

Polities. We define polity as any independent political unit. Kinds of polities range from villages (local communities) through simple and complex chiefdoms to states and empires. What distinguishes a polity from other human groupings and organizations is that it is politically independent of any overarching authority; it possesses sovereignty. For each NGA, then, we gather data concerning *every* polity or quasi-polity that ever held some or all of the territory comprised by that NGA.

3.09.5.3 Information Infrastructure at IISH

The overall organization of IISH is into discrete research projects, each with its own director. Recent reorganization within IISH, however, is leading to upgrading systems of archiving and analysis. Archival collections, including those of CLIO-Infra and the Global Collaboratory on Labor Relations, are being moved to the archive of Harvard Dataverse, which is becoming increasingly widely used.

3.09.5.4 Information Infrastructure at IROWS

IROWS works through the preparation of individual working papers (supplemented by follow-up publications); research on many of the working papers is externally funded. Archives of the datasets and statistical analyses are housed independently.

3.09.6 Data and Metadata

Collection and digitization of historical data is the essential launching point for Big Data in Human History, but effective data collection relies equally on construction of detailed metadata.

3.09.6.1 Data

Here we identify data not by their topic but by their qualitative and quantitative characteristics.

Distributed data. Sociohistorical data may be spread over various data sources and organizations.
Heterogeneous data. Sociohistorical data may be represented in different data formats and in files with different types and structures.
Sparse data. Sociohistorical data may be fragmented—missing values are common.
Aggregated data. Sociohistorical data may be aggregated in different ways over various time intervals and space regions.
Inconsistent data. Sociohistorical data items reported by different data sources may contradict each other, resulting in data inconsistencies.
Unreliable data. Sociohistorical data sources may have different degrees of reliability.
Data intensivity. Historical data collections do not generally come from sensors. The Tycho database on US disease surveillance, created by the Pitt Global Health Center, is currently on the order of ten gigabytes and growing (Tycho). It is arguably the largest data resource, historical or other, in the field of public health. However, the approach of duplicate manual entry, while successful in this case, is too costly to be generalized to collection of enough data to give a valuable and relevant record of the human past over 10,000 years. The associated project linking US disease surveillance (from Tycho), climate (from NOAA), and population since 1900 (from ICPSR), is also of the order of ten gigabytes in size. They range from simulation data on African population history to reports on historical labor systems, religious censuses by missionary organizations, and data digitized from population and health records of colonial India. These indicate the continuing heterogeneity and steady expansion in the CHIA datasets. Work on world-historical population under this RCN project will collect roughly a terabyte of data. Further, as we calculate it, the volume of climatic data since 1600 to be incorporated—including high and low temperature and average humidity—when developed systematically and historically, will also total over one terabyte in volume. In particular, in addition to direct observations of temperature and humidity (some of these will be collected by TerraPop), indirect measures of the same variables are widely available through the expanding specificity of climatological analysis, documenting climate in specific places and rather precise times, for centuries and millennia. The crowdsourcing data-ingest procedure will result in continuously growing multi-terabyte volumes of data.

3.09.6.2 Metadata: Supporting a Fully Global Data Resource

Assembling a large number of datasets is not sufficient to produce global data—the data need to be merged into a single, uniform data repository. Nor is it possible to create a uniform data repository through automated processing of the existing metadata—the terms are inconsistent and, too often, there turn out to be major bits of information simply missing. The problem is that additional metadata must be created to account for harmonization and linkage of inconsistent local datasets and for aggregation to regional and global levels. The CHIA project is to address these issues directly through creation of a global historical data resource.

A large part of the historical data that need to be included in a global historical data resource have yet to be digitized or documented. The data have certainly not been rendered compatible with one another, especially for Asia, Africa, and the Americas, and especially before the 20th century. It is true that a large body of historical data already exists, generally on the Internet and more specifically in such repositories as the ICPSR and the Dataverse Network. Even these, however, are disaggregated sets of data with two very distinct levels of documentation—the high-level documentation of the repository system (SAS or SPSS) and the documentation provided for constituent datasets by their creators. Most statistical data assembled by historical researchers,

further, are held in Excel and other spreadsheet software, with no systematic documentation facilities. Thus, no magic bullet will turn existing repositories into globally analyzable bodies of data: existing data need essentially to be redocumented, and newly entered historical data need to be documented comprehensively. Such documentation requires both a consistent framework and the expertise of academic researchers—those who constructed or transcribed the data or others with similar expertise. To access such expertise, our approach emphasizes "crowdsourcing," though based on expert users.

Collecting and Documenting Data: the Research Collaborative. For the assembly of global data, "metadata" must include several types of data description. The overriding rule is that each data value within a dataset must be fully defined in terms of its source, its dimensions, and any transformations or aggregations it has undergone between its original source and its current position in the dataset. Dimensions here include both the quantity being measured and its temporal-spatial delimitation. Consider the simplest case, the addition of a single number. At least four pieces of information need to be added beyond the number itself: *what* is being measured; *where* the information or reporting unit is in reference to; *when* (date or period); and the *source* of information (including the contributor). To hold this information in a consistent structure, answers to these questions need to be selected from controlled vocabularies (though these can be extended by users). The controlled vocabulary for *where* would be a gazetteer or GIS, though it would have to account for variations in boundaries and labels of locations; an analogous and flexible vocabulary is needed for *when*. The controlled vocabulary for *what* is the most challenging, as there is no established thesaurus for statistical concepts, although classifications have been developed for occupations and diseases. The incorporation of such existing detailed classifications means that data-ingest work can start before the high-level framework—the overall project ontology—is finalized. In section 5d, we consider the crowdsourcing data integration infrastructure that will facilitate this task.

In addition to the "what, where, when, source" of the originally entered data, additional transformations and aggregations will be required. Original submissions of data need to be cleaned of errors and integrated to resolve duplications and inconsistencies across datasets. Thereafter—along with the transformation of submitted data by language, geography, time, weights, measures, and other criteria to make them compatible with other contributed datasets—comes the creation of "incremental metadata" to document further transformations. That is, along with aggregation *of data* by scale (both geographic and temporal) in order to have consistent regional and global datasets created out of the smaller datasets comes the creation *of incremental metadata* to document the aggregation. In sum, the volume of metadata will likely equal or exceed the volume of data in the global dataset.

The maintenance of this huge amount of metadata will be laborious and expensive, but the effort will be worth the cost. The need for these additional categories of metadata only becomes clear as we move toward aggregation to global-level data. To go perhaps one level further, one can imagine that an algorithm for transforming data values is found to require correction—for instance, deflation of value statistics by an improved price index—in which case corrections would have to be made throughout and additional metadata would need to be recorded. With fully upgraded metadata, based on strong standards, it will be possible to recalculate each value precisely, on the fly, thus preserving the value of the repository and its elements over time. Further, the effort will diminish as a "critical mass" of needed metadata is reached, lessening the burden and effort needed to expand as new issues arise. The alternative is that whole datasets might have to be abandoned and recreated from the beginning. In particular, many of the global indices created and widely circulated to describe national statistics for the past 50 years appear to contain data but no substantial metadata, so that if price indices or commercial volumes were to be recalculated, there would be no available basis for recalculation: the choice would be to use outdated figures or simply junk the dataset.

3.09.7 Ontology

To be able to link data across all the scales of this work, a relatively comprehensive ontology is necessary, with dimensions of space, time, topic, and level of aggregation. The latter, perhaps under-conceptualized is to identify boundaries separating local levels of space–time–topic from broader levels, and deciding where to put (for instance) global space/local time/specific topic in the overall scale of aggregation. Ontological work typically builds upward from the lowest level or downward from the most general level, and the two approaches rarely match up. Groups need to seek out good ontological work that has been done at any level and should seek to link the various areas of strength. Spatial ontology is the best developed. The work of Pelagios and the current proposal for a world-historical gazetteer point in the direction of a resource that can encompass the past 3000 years—a good start.

3.09.7.1 Spatial Ontology

Gazetteers—controlled vocabularies of places—are the main form of spatial ontology. The expansion of digital datasets, many of them requiring representation of places, has put many researchers in the position of independently developing gazetteers for their research, but the independent gazetteers are inconsistent. In the most successful single enterprise, a combination of the Pleiades gazetteer for the ancient world and the Pelagios Commons for using open data methods has achieved a coherent system for labeling places in the world before 1500 (Pleiades, PelagiosCommons, Mostern et al., 2016). Parallel work under way at CHIA will create a world-historical gazetteer for the period after 1500. In the CHIA project, a "Spine" of some 10,000 world-historical places is documented, working from a set of world-historical atlases. This Spine is then linked to an Ecosystem of independent research projects, each developing its locally specific gazetteer but linking it explicitly to the places also listed on the Spine. The linking

mechanism is the Pelagios Interconnection Format. The combination of Spine and Ecosystem, sharing a set of standards developed as part of the world-historical gazetteer, will allow for the development of a steadily growing network of linked descriptions of places.

3.09.7.2 Temporal Ontology

Time, because it is one-dimensional in some ways, is commonly thought to require only a very simple representation. As a result, temporal ontology is generally given a low priority, so that the temporal dimension of datasets is commonly labeled ineffectively and in confusion style. But even if time extends only forward and backward in a single dimension, it nonetheless has many characteristics. Time is both quantitative and qualitative; it is both absolute and relative. Quantitative time is measured in multiple units—commonly in days, months, years, and centuries. Yet the scales of time, from microseconds to millions of years, can be difficult to represent. Definitions of qualitative time, represented especially in named periods, have recently been formalized for named historical, art-historical, and archaeological periods in the work of PeriodO (PeriodO).The issues requiring clarification in a temporal ontology also include the distinction between closed-ended and open-ended periods (e.g., 20th century vs. postwar), cycles in history (e.g., El Niño cycles), time measured according to various calendars, relative time (e.g., generational time or years of a reign), and issue of simultaneity (how long it takes news to move from one place to another). All of these issues must be defined appropriately and consistently in order to permit the combination of multiple historical datasets into aggregated dataset that can represent large parts of the human experience.

3.09.7.3 Topical Ontology

Topical ontology is necessarily the most complex dimension in description of the elements of the historical documentation from which we seek to reconstruct analytical models of the past. The topics of the past range infinitely in their qualitative differences, and the various elements of historical topics overlap so that they can rarely be set in hierarchical or fully distinctive relationships. Nevertheless, to compare and link local datasets in order to aggregate and assess global patterns, some approach to overall classification of historical topics is necessary. In fact, since humans are always defining and implementing categories, historical researchers have plenty of models on which to draw. For instance, the disciplines of the academic world were set up to discuss the elements of society and the natural world, so they provide a ready-made categorization of elements of the world. Similarly, library classification systems, notably the US Library of Congress system, provide relatively comprehensive classifications of the issues in which researchers are interested. More specific topical ontologies have also developed in areas of intensive social practice: in medicine, commerce, and museums. From these models, and from the experience of developing spatial and temporal ontologies for human history, it will be possible to develop appropriate topical ontologies.

3.09.7.4 Aggregative (Scalar) Ontology

Spatial, temporal, and topical ontologies all range from the smallest scales to progressively larger scales. In considering interactions among our collaborating projects, we will need to make explicit statements of the scale of various topics and issues, to decide how to fit our results together. We will surely find that various research groups will work in different regions of the overall ontological map. In addition, we will locate and give attention to voids—regions of ontological space that have been neglected. We will seek out handy overall measures of the scale of our work, such as the number of bytes of data in a given area of work, but we will need to develop more sophisticated measures of scale. For the overall project of the ontology of the historical human system, we will have to work from both the specific and the general standpoints. Practical work on ontology consists of defining and applying metadata at the immediately practical level, as Theoretical work includes considering the larger configurations of dimensions and definitions. Necessarily, research will proceed at the two levels separately, with the hope that they will connect from time to time.

3.09.7.5 Ontology Development in the Seshat: Global History Databank

Seshat Ontology. Seshat has developed a unique ontology for social-historical data classes, designed to be reconcilable with existing spatial, temporal, topical, and scalar ontologies (Seshat Ontology). The basic entity classes of Seshat data are Organizations and Territories. Territories have fixed geographical bounds that do not change with time, whereas Organizations are defined by temporal bounds and may be associated with specific Territories at specific intervals, such as a religious group occupying a particular set of polities during a defined time frame. Importantly, we define relationships between classes. Most important is the *controls* relationship, in which a social organization controls a given territory at a given time. For example, the Roman Principate, a type of social organization we define as a polity (see above), controlled the Latium NGA from BCE 31 to CE 284. Further, several additional organizations (e.g., interest groups such as religious groups, ethnic groups, economic firms) may have an *exists within* relationship to a particular territory which is controlled by a polity; namely, the early Christian Church *existed within* the Roman Principate polity which *controlled* the Latium NGA from BCE 31 to CE 284 (**Fig. 6**).

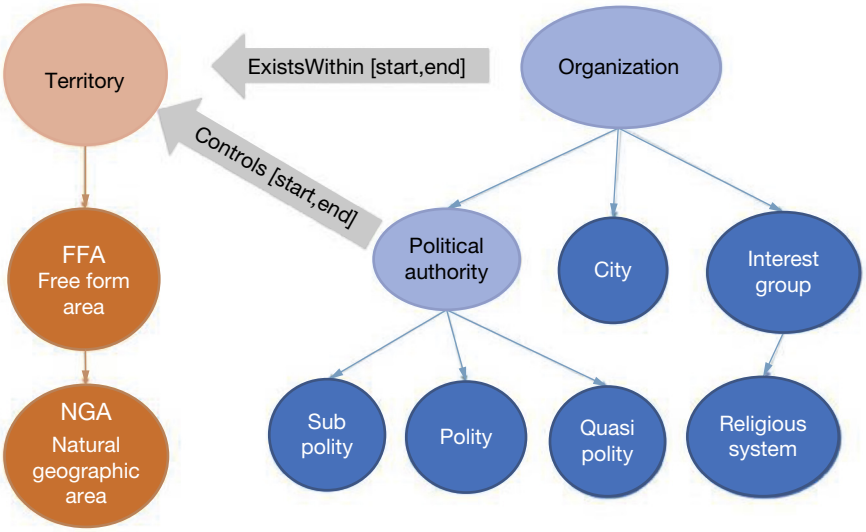

Fig. 6 Seshat metamodel showing entity classes and subclasses and the two key relationships: *controls* and *exists within*. Turchin P, Brennan R, Currie TE, Feeney K, Francois P, Hoyer D, Manning J, et al. (2015) Seshat: The global history databank. *Cliodynamics: The Journal of Quantitative History and Cultural Evolution* 6(1): 77–107.

3.09.8 Analysis

Historical analysis of global society requires distinctive large-scale projects but also that the projects collaborate to identify next steps in articulating and testing global hypotheses.

3.09.8.1 CHIA

Analysis of inequality. The theory of inequality, focusing on feedback networks, is expected to identify distinct dynamics at varying time frames, social arenas, and aggregative scales.

Time frames in social inequality. In one clear distinction, analysis must account for varying time frames in the propagation of social inequality. Here are examples of analysis in a short-term or crisis-centered time frame and in the time frame of a century or more.

Crisis. A key case of crisis is the El Niño climate shifts of the late 1870s and the immense droughts and famines that resulted from them—especially in India, China, Brazil, and East Africa. Institutionalist, social historical, and human-natural interaction paradigms will each be explored through study of these regions over a period of some 60 years, centered on the late 1870s (Davis, 2001). As a second example of crisis, the years 1915–1922 brought a parallel shock of deep social devastation, including wars, pandemic, and famine. These theoretical perspectives are to be applied to data on crises of up to 5 years in length.

Centennial analysis. Nunn has proposed long-term impacts of slave trade, tariffs, adoption of food crops, and gender relations (Nunn, 2008; Nunn and Qian, 2010). This corresponds to the commonly 100-year time frame of the rise and fall of communities, cities, and states. Resource endowments may be exhausted after a century of exploitation. For each time frame, numerous categories of social and natural relations need to be considered in interaction. Categories include *labor systems* (self-employment, slavery, wages, etc.); *occupational categories* (by racial, ethnic, or familial grouping); *governance* (degrees of democracy and representation); *social institutions* (state, religion, education); *prices and wages* (set by value but also by processes of bargaining); *migration* (cross-community migration rates for urban, rural, seasonal, or long-term migration); and *commodity chains* (for minerals, agricultural commodities, manufactures). Other categories include individual agency and social movements; economic cycles; epidemics; resource endowments and exhaustion; and climate shocks such as drought and flood (Haldon et al., 2014; Ludlow and Manning, 2016). Analysis across these categories should identify feedback links that correlate with significant shifts in inequality, at various time frames.

Crisis analysis: The SWS

Consider a potential inequality pattern in social development as a society moves through periods of notable transitions. The inequality may have both positive and negative impacts on a society. From one side, it may result in healthy competition, producing new ideas and increasing human well-being (Van Zanden et al., 2014). However, as it approaches a certain threshold, the inequality may cause severe disruptions and instabilities. The social instability may cause significant degradation of human well-being, involving civil unrest and slowing down social progress. Such instability patterns can occur at different scales and may vary in duration and severity. The importance of timely discovery of the factors that may cause social instability is hard to underestimate. The idea here is to address the challenge of design and development of a scalable information processing infrastructure to support social early warning and forecasting system, which we refer to as the SWS.

Integrated table for inequality analysis

Country	Year	LE	IS	TC	PW	GL	GE	TO	AN	AF	AW	BL	BS	BI	BD	BR	GC	GCP	GI	WS	AC	AH	II	LD	GY	GCM
.....																										

Code	FullName
LE	Life Expectancy At Birth
IS	Inverse Sum Freedom House
TC	Theil Coeff, UTIP
PW	Per capita private wealth
GL	Group Largest %
GE	Gender Equality Index
TO	Top 1% income share
AN	Adjusted net national income per capita (constant 2005 US$)
AF	Adolescent fertility rate (births per 1,000 women ages 15-19)
AW	Average working hours of children, working only, female, ages 7-14 (hours per week)
BL	Benefits incidence in poorest quintile (%) -All Labor Market
BS	Benefits incidence in poorest quintile (%) - All Social Assistance
BI	Benefits incidence in poorest quintile (%) - All Social Insurance
BD	Battle-related deaths (number of people)
BR	Birth rate, crude (per 1,000 people)
GC	GDP per capita (constant 2005 US$)
GCP	GDP per capita, PPP (constant 2011 international $)
GI	GINI index (World Bank estimate)
WS	Wage and salaried workers, female (% of females employed)
AC	Armed Conflicts (Internal)
AH	Avg Height
II	Income Inequality (GINI)
LD	Latent Democracy Variable
GY	Gini (WYD)
GCM	GDP per Capita, Maddison

Fig. 7 Schema for conceptual integrated table for inequality analysis.

In order to use the inequality-related variables for analysis of social stability, they should be measured comprehensively for different countries (or analogous units) over a sequence of time intervals (e.g., years). The analysis can be performed at various scales of aggregation. One key information science breakthrough we seek is to link the currently discrete social science datasets into larger databases that can ultimately reach a global scale, through data-driven linkage. Conceptually, this task can be thought of as a challenge of creating an integrated table for inequality analysis (ITIA), in which available multidisciplinary data can be laid out to facilitate the data linkage. Fig. 7 shows a part of the schema of the ITIA based on the inequality categories in Fig. 1 with explanation of each schema variable.

A complete collection of data streams in the ITIA would allow for comprehensive exploration of human inequality. If we had complete historical information on inequality, historians could begin to address some of the most central questions of modern history. More specifically, the people of Latin America and the Caribbean have witnessed a dramatic increase in life expectancy and literacy over the past century but the nations have remained economically unequal (Bértola and Ocampo, 2012). Recent studies have also suggested that nations have returned to a high point of economic inequality in very wealthy countries. However, the available data are commonly too sparse and too inconsistent in many categories, in other time periods, and in less wealthy countries for social scientists to construct such narratives for inequality at the global level.

The ITIA table should integrate available historical datasets, which are commonly small, complex, and, until recently, compiled by human agency. So far, we have identified over 210 separate published datasets related to the inequality variables from over 20 different sources. We integrated those datasets and, looking only at the national-level scale of analysis, we created an ITIA table of over eight million data points. Linking this information is problematic, even after the major data integration challenges are resolved (cleaning data, unifying data units and formats, accounting for territorial changes over time, etc.). We will need to combine incomplete data of dissimilar quality and scale, and of differing granularity. Drilling down to levels of greater detail and examining a single country, we can see the scattered nature of the data, which only grow sparser as we move further back in time.

The challenge of sparse data. For this purpose, we need to develop and apply advanced data imputation methods. Such methods can range from relatively simple techniques, such as using the sample mean, sliding window mean, or last observation for missing values, to more sophisticated approaches utilizing related variables, observations, and semantic constraint-based historical evidences. We propose to use large-scale simulation for approximation of the ITIA table. The simulation infrastructure will be used for imputation of missing data under most likely assumptions and application constraints. For an example of simulation work to impute missing population data for Africa, see Manning et al. (2015a,b).

The challenge of aggregated data. Consider a simple example in Fig. 8, with average life expectancy statistics (LE) and top 1% income share statistics (TIS) reported at different time intervals. We would like to estimate what value of LE better corresponds to given

Life expectancy

From	To	LE
1981	1985	59.9
1986	1990	61.5

Total 1% income share

From	To	TIS
1985	1986	9.1

Fig. 8 Merging two variables at different time aggregation.

TIS. In database terms, we would need to join two tables together on their corresponding time intervals. Performing an equi-join on the reports' start and end time would yield an empty table.

We will need to develop advanced approximate join techniques that would intelligently provide the best effort to join tables on different aggregation levels. In general, to join aggregated data streams, we can either join aggregated reports directly (Aggregated Join) or first disaggregate reports to a common time unit and then use equi-join (Disaggregated Join). The Aggregated Join approach should utilize a proper similarity measure between the aggregates (e.g., relative overlaps of aggregation intervals). The Disaggregated Join approach is based on temporal disaggregation methods, representing aggregated reports as low-frequency time series and then disaggregating them into high-frequency series D1 and D2, using the most suitable temporal disaggregation method. The result of the disaggregation would be two time series with estimated values for matching time units. The process of merging aggregated data streams is resource consuming, and it involves trade-offs between accuracy of the produced results, execution time, and consumed computational resources. We will explore those trade-offs for large-scale inequality analysis scenarios using both real and simulated data.

Challenges of inconsistent and unreliable data. These are tightly related. Data inconsistency is commonly caused by inaccurate historical reports, which may indicate a nonreliable source of data. In many cases, data inconsistency can be revealed through analysis of relationships between existing reports in the historical database. Historical data sets may have different levels of reliability due to the quality of components such as the primary source of information and data collection methodology. In order to assess the reliability of a report, we need to account for the data inconsistencies it causes. Assuming that the system continuously receives new historical reports, we can compute a reliability value for the source of these streams, which evolves with respect to new evidence.

Scalable information-processing infrastructure for the SWS. We seek to devise an advanced information processing architecture combining two complementary processes: (1) sociohistorical data collection and integration and (2) integrated data quality enhancement. Both processes continuously populate and improve the quality of integrated sociohistorical data and metadata repositories, which form the information core of the SWS system. As shown in Fig. 3, work is conducted in each of the three repositories: the CDR (where historical data are initially stored), the IDR (where historical data are integrated and documented), and the DDR (where simulated data are developed and refined, drawing on the input of historical data).

At the level of sociohistorical data collection and integration (in the first two repositories), the data collection component accumulates heterogeneous datasets from various data sources in the CDR, providing a user-friendly data submission interface. We have considerable experience with developing such interface in our Col*Fusion system. The data integration component then resolves data heterogeneities and produces the IDR, which stores both homogeneous data and metadata reflecting application-specific data integrity constraints.

The principal contribution of the project takes place at the level of integrated data quality enhancement, where various elements of simulation combine to produce a comprehensive system-wide set of data of enhanced quality. This component aims to improve integrated data by imputing missing values, performing proper data disaggregation/aggregation for multi-resolution data analysis, discovering data inconsistencies, and assessing data reliability. A large-scale data-driven simulation environment supports this functionality. As shown in Fig. 3, the data-driven simulation creates Simulated Data, starting from the IDR as a source of documented historical data. Additionally, the simulator utilizes metadata from the IDR, reflecting theoretical and typological relationships in social sciences and natural science to constrain data variables in a meaningful way. In addition, relationships among variables as hypothesized in social science theory will be included in the process of simulating missing data. After that, as shown at the right of Fig. 3, the simulator explores various data-loss scenarios and applies different information recovery methods to yield estimates of missing data at different time and location scales. We will assess efficiency of recovery methods with respect to various data-loss scenarios. The evaluation will be used to devise adaptive strategies for data recovery, enhancing quality of the integrated sociohistorical data in a most efficient and reliable way. The results of this continuous data exploration and data quality assessment under different data-loss scenarios will be stored in the DDR. Therefore, the DDR will be accumulating valuable information for further research and development in related areas. Next we consider in more details the three major repositories that the SWS infrastructure will populate and maintain to support its scalable information processing.

The SWS users will perform advanced sociohistorical data analysis at different scales. The user interface will allow researchers to embed their data in a large context of the SWS repository and explore various "what-if" scenarios under different data-loss assumptions. Researchers will also be able to experiment with their own recovery strategies, adding them to our extendable Data Recovery Toolset. We will validate our approach by performing inequality analysis of the enhanced integrated data to detect well-known historical events in social developments characterized with notable transformations and instability (e.g., change of political system). Success in such tests will indicate that the underlying methods and algorithms are robust.

3.09.8.2 Seshat

Seshat Data are designed to be flexible and inclusive, amenable to numerous types of analysis depending on the specific goals of current and future research projects. In general, the goal of Seshat analysis is to let the empirical data we collect adjudicate between different theories on a given topic by comparing the strength of support for different statistical models.

Model comparison. Key to our approach is not simply searching for correlations that match a particular hypothesis, e.g., that more equal societies have higher measures of well-being like public goods provision, but systematically evaluating a range of competing

hypotheses. This involves fitting multiple models, with varying numbers of parameters (Turchin et al., 2012). As an example, growth in economic productivity may produce stark wealth or income inequality, though it improves overall well-being through increasing the scale of wealth and goods available throughout the society. A competing hypothesis to test is that the cultural systems that promote prosocial norms and produce public goods decrease inequality, leading to more overall well-being.

Model simulation. Producing simulations based on complex mathematical models and assessing the fit of the simulations against the true historical record is another powerful tool for assessing the validity of particular sets of theoretically informed causal relationships between variables (Turchin et al., 2013; Bennett, 2016).

Complexities of historical data. While this basic approach is fairly straightforward, in practice, there are a number of complexities that need to be taken care of. First, because historical data have spatiotemporal structure, our statistical approach deals explicitly with spatial and temporal autocorrelations. If these issues are not adequately controlled, there is a risk of artificially inflating the strength of relationships between variables. Second, we need to allow for the possibility of nonlinear effects. For example, the effect of a predictor variable could be ∩-shaped: initially rising and later falling. A simple way to test for such eventualities is by adding quadratic terms. Third, historical datasets are likely to contain "holes," as mentioned above. That is, there are likely to be many gaps in our knowledge about the values of key variables in any particular time and location. This missing data problem can be dealt with by the statistical method of multiple imputation (Rubin, 1987; White et al., 2011).

3.09.8.3 IISH

The taxonomy of labor relations aims to map different kinds of labor relations in various world regions in the period 1500–2000. The taxonomy, created at the Research Department of the International Institute of Social History in 2007, distinguishes among four types of labor, nonwork, reciprocal labor, tributary labor, and commodified labor, connected with the household, the community, or the market, and extends these categories with 19 different labor relations at the individual level. The resulting datasets can also capture combinations of labor relations, including seasonal migration and work cycles, "economies of make shift," proto-industry, and "petty capitalists." By mapping labor relations in various parts of the world, this dataset allows us to identify important shifts from one type of dominant labor relations to another. For example, shifts from tributary to wage labor, increases and decreases in slave labor, intensification of labor efforts within households due to the increased labor market participation of women, and flexibilization of labor contracts. The collaboratory concentrates on five cross-sectional years (1500, 1650, 1800, 1900 [adding 1950 for Africa], and 2000). The systematic collection of standardized data worldwide in the period 1500–2000 has led to increasing collaboration with colleagues from all world regions, especially in the "Global South."

3.09.8.4 IROWS

Recent IROWS working papers focus on such topics as social movements in premodern polities, urban and polity size swings from the Bronze Age, networks of systemic interaction, and Chinese involvement in sub-Saharan Africa (IROWS, Chase-Dunn and Lerro, 2014; Chase-Dunn and Hall, 1997).

3.09.9 Prospects for This Research and Analysis

Each of the participating groups working in Big Data in Human History has heavy responsibility in sustaining its local work: meeting project deadlines, submitting proposals for funding, coordinating staff activities, and more. All of the groups are collaborative in principle, so they seek to devote time to exchanges of knowledge both within and beyond the group. The idea of Big Data in Human History is that there should be some additional benefit to collaboration reaching across as many groups as possible that are doing large-scale social science research.

3.09.9.1 Meetings

It was agreed that members of Big Data in Human History should meet each year, with meetings alternately at the Social Science History Association (in the United States, in even-numbered years) and at the European Social Science History Conference (in Europe, in odd-numbered years).

3.09.9.2 Academic Journals

Several academic journals are published by participants in Big Data in Human History. The Big Data in Human History community will benefit if participants publish in various of these journals, so that they are able to sustain a common discourse on large-scale social science analysis. They are *Cliodynamics: The Journal of Theoretical and Mathematical History* (edited by Peter Turchin of Seshat and published by IROWS at the University of California, Riverside); the *International Review of Social History* (published in Amsterdam by the International Institute of Social History); the *Journal of World-Historical Information* (published by the World History Center at the University of Pittsburgh); and the *Journal of World-Systems Research* (published at the University of Pittsburgh, Jackie Smith, editor).

3.09.9.3 Common Standards

The development and application of common standards throughout the community of scholars working on large-scale social-scientific research will increase communication among research results. Among the possible areas for standards discussed in this article are standards for metadata in the documentation of data files; ontologies in space, time, topic, and level of aggregation; conventional time periods in which research is conducted; and standards for rates of migration as well as economic output for social units.

References

Acemoglu, D., Robinson, J.A., 2012. Why nations fail: The origins of power, prosperity and poverty. In: Crown Publishers, New York.

Bain, D.J., Brush, G., 2008. Gradients, property templates, and land use change. The Professional Geographer 60 (2), 224–237.

Bennett, J., 2016. Repeated demographic-structural crises propel the spread of large-scale agrarian states throughout the old world. Cliodynamics: The Journal of Quantitative History and Cultural Evolution 7 (1), 1–36. http://dx.doi.org/10.21237/C7clio7128530.

Bértola, L., Ocampo, J.A., 2012. The economic development of Latin America since independence. In: Oxford University Press, New York.

Bizer, C., Heath, T., Berners-Lee, T., 2009. Linked data-the story so far. International Journal on Semantic Web and Information Systems 5 (3), 1–22. http://dx.doi.org/10.4018/jswis.2009081901.

Boyd, R., Richerson, P.J., 2005. The origin and evolution of cultures. In: Oxford University Press, New York.

Brennan, R., Feeney, K.C., Gavin, O., 2013. Publishing social sciences datasets as linked data: A political violence case study. In: ENRICH 2013 at SIGIR 2013, Dublin, 1 August 2013, Seamus Lawless, Maristella Agosti, Paul Clough, Owen Conlon, 2013, pp. 20–31.

Brennan, R., Feeney, K., Mendel-Gleason, G., Bozic, B., Turchin, P., Whitehouse, H., Francois, P., Currie, T.E., Grohmann, S., 2016. Building the Seshat ontology for a global history databank. In: Proceedings of 13th International Conference, ESWC 2016, Heraklion, Crete. Greece, May 29–Jun. 2, 2016, pp. 693–708.

Brodie, M., 2010. Data integration at scale: From relational data integration to information ecosystems. In: 24th IEEE International Conference on Advanced Information Networking and Applications (AINA-10). Perth, Australia, Apr. 20–23, 2010.

Camerer, C., Loewenstein, G., Prelec, D., 2005. Neuroeconomics: How neuroscience can inform economics. Journal of Economic Literature 43 (1), 9–64. http://dx.doi.org/10.1257/0022051053737843.

Campbell, B., 2016. The great transition: Climate, disease and society in the late-medieval world. In: Cambridge University Press, Cambridge.

Chandra, S., Kuljanin, G., Wray, J., 2012. Mortality from the influenza pandemic of 1918–1919: The case of India. Demography 49 (3), 857–865.

Chase-Dunn, C., Hall, T.D., 1997. Rise and demise: Comparing world-systems. In: Westview Press, Boulder, CO.

Chase-Dunn, C., Lerro, B., 2014. Social change: Globalization from the stone age to the present. In: Paradigm Publishers, London.

Cherniak, A., Zadorozhny, V., 2013. Signature-based detection of notable transitions in numeric data stream. IEEE Transactions on Knowledge and Data Engineering 25 (12), 2867–2879.

Cherniak, A., Zaidi, H., Zadorozhny, V., 2013. Optimization strategies for A/B testing on HADOOP. In: Proceedings of International Conference on Very Large Data Bases (VLDB'13), *Aug. 26–30*. Riva del Garda, Trento, Italy.

Davis, M., 2001. Late Victorian holocausts: El Niño famines and the making of the third world. In: Verso, New York.

Doan, A., Halevy, A., Ives, Z., 2012. Principles of data integration. In: Morgan Kaufmann, San Francisco, CA.

Drwenski, M. 2015. Scales of inequality: Strategies for researching global disparities from 1750 to the present, (MA thesis). University of Pittsburgh.

Flannery, K., Marcus, J., 2012. The creation of inequality: How our prehistoric ancestors set the stage for Monarchy, Slavery, and Empire. In: Harvard University Press, Cambridge.

François, P., Manning, J.G., Whitehouse, H., Brennan, R., Currie, T.E., Feeney, K., Turchin, P., 2016. A macroscope for global history. Seshat global history databank: A methodological overview. Digital Humanities Quarterly 10 (4).

Gerring, J., 2012. Social science methodology: A unified framework. In: 2nd edn. Cambridge University Press, Cambridge.

Glimcher, P.W., 2009. Neuroscience, psychology, and economic behavior: The emerging field of neuroeconomics. In: Tommasi, L., Peterson, M.A., Nadel, L. (Eds.), Cognitive biology: Evolutionary and developmental perspectives on mind, brain, and behavior. The MIT Press, Cambridge Massachusetts, pp. 261–278.

Green, M.H., 2015. Pandemic disease in the medieval world: Rethinking the Black Death. In: ARC Medieval Press, Kalamazoo, MI.

Haas, L., 2007. Beauty and the beast: The theory and practice of information integration. In: Proceedings of 11th International Conference, *Barcelona, Spain, Jan. 10–12, 2007*.

Haldon, J., Roberts, N., Izdebski, A., Fleitmann, D., McCormick, M., Cassis, M., Doonan, O., et al., 2014. The climate and environment of Byzantine Anatolia: Integrating science, history, and archaeology. Journal of Interdisciplinary History 45 (2), 113–161.

Hoyer D and Manning JG (forthcoming) Empirical regularities across time, space, and culture: A critical review of comparative methods in ancient historical research. *Historia*.

Jensen, F.V., Nielsen, T.D., 2007. Bayesian networks and decision graphs. In: 2nd edn. Springer, Secaucus, NJ.

Kang, Y., Zadorozhny, V., 2015. Process monitoring using maximum sequence divergence. Knowledge and Information Systems. http://dx.doi.org/10.1007/s10115-015-0858-z.

Korzeniewicz, R.P., Moran, T.P., 2009. Unveiling inequality: A world historical perspective. Russell Sage Foundation, New York.

Lange, M., 2012. Comparative-historical methods. In: SAGE, London.

Lucassen, L., 2016. Working together: New directions in global labor history. Journal of Global History 11 (1), 66–87.

Lucassen, L. J. (Ed.), 2014. Globalising migration history. The Eurasian experience (16th–21st centuries). Brill, Leiden.

Ludlow, F., Manning, J.G., 2016. Revolts under the ptolemies: A paleoclimatological perspective. In: Manning, J.G., Collins, J.J. (Eds.), Revolt and resistance in the ancient classical world and the near East: In the Crucible of Empire. Brill, Leiden, pp. 154–171.

Maddison, A., 2007. Contours of the world economy, 1-2030 AD: Essays in macro-economic history. In: Oxford University Press, Oxford.

Manning, P., 2013. Big data in history. In: Palgrave, London.

Manning, P., 2014–2015. A world-historical data resource: The need is now. Journal of World-Historical Information 2–3 (2), 1–6.

Manning, P., 2017. Inequality: Historical and Disciplinary Approaches. American Historical Review 122 (1), 1–22.

Manning, P., Ravi, S., 2013. Cross-disciplinary theory in construction of a world-historical archive. Journal of World-Historical Information 1 (1), 2169–2812. http://dx.doi.org/10.5195/jwhi.2013.3.

Manning, P., Zhang, Y., Bowen, Y., 2015. Volume and direction of the Atlantic Slave trade, 1650–1870: Estimates by Markov Chain Monte Carlo analysis. Journal of World-Historical Information 2-3 (2), 127–149. http://dx.doi.org/10.5195/jwhi.2015.31.

Manning, P., Nickleach, S., Yi, B., McGill, B., 2015. Demographic models for projecting population and migration: Methods for African historical analysis. Journal of World-Historical Information 2–3 (1), 23–39. http://dx.doi.org/10.5195/jwhi.2015.19.

McCreary, D., Kelly, A., 2013. Making sense of NoSQL: A guide for managers and the rest of us. Manning Publications, Shelter Island.

McNeill, J., 2000. Something New under the Sun: An environmental history of the twentieth-century world. W. W. Norton, New York.

Milanovic, B., 2016. Global inequality: A new approach for the age of globalization. In: Harvard University Press, Cambridge, MA.

Mostern, R., 2011. Dividing the realm in order to govern: The spatial organization of the Song state (960–1275 CE). In: Harvard University Asia Center, Cambridge.

Mostern, R., Arksey, M., 2016. Don't just built it, they probably won't come: Data sharing and the social life of data in the historical quantitative social sciences. International Journal of Humanities and Arts Computing 10 (2), 205–224.
Mostern, R., Southall, H., Berman, M.L., 2016. Placing names: Enriching and integrating gazetteers. In: Indiana University Press, Bloomington, IN.
North, D.C., Wallis, J.J., Weingast, B.R., 2009. Violence and social orders: A conceptual framework for interpreting recorded human history. In: Cambridge University Press, Cambridge.
Nunn, N., 2008. The long term effects of Africa's slave trades. Quarterly Journal of Economics 123 (1), 139–176.
Nunn, N., Qian, N., 2010. The Columbian exchange: A history of disease, food, and ideas. Journal of Economic Perspectives 24 (2), 163–188.
Nystrom, N.A., Levine, M.J., Roskies, R.Z., Scott, J.R., 2015. Bridges: A uniquely flexible HPC RESOURCE for new communities and data analytics. In: Proceedings of the 2015 Annual Conference on Extreme Science and Engineering Discovery Environment, XSEDE15. ACM, New York, NY, USA. http://dx.doi.org/10.1145/2792745.2792775.
O'Rourke, K.H., Williamson, J., 1999. Globalization and history: The evolution of a nineteenth-century Atlantic economy. In: MIT Press, Cambridge, MA.
Pelechrinis, K., Zadorozhny, V., Kounev, V., Oleshchuk, V., Anvar, M., Lin, Y., 2015. Automatic evaluation of information provider reliability and expertise. World Wide Web 18 (1), 33–72.
Piketty, T., Goldhammer, A., 2014. Capital in the twenty-first century. In: Harvard University Press, Cambridge, MA.
Ren, Y., Zadorozhny, V., Oleshchuk, V., Li, F., 2014. A novel approach to trust management in unattended wireless sensor networks. IEEE Transactions on Mobile Computing 13 (7), 1409–1423.
Richerson, P.J., Christiansen, M.H., 2013. Cultural evolution: Society, technology, language, and religions. In: MIT Press, Cambridge, MA.
Rubin, D.B., 1987. Multiple imputation for nonresponse in surveys. In: Wiley, New York.
Skocpol, T., Somers, M., 1980. The uses of comparative history in macrosocial inquiry. Comparative Studies in Society and History 22 (2), 174–197.
Southall, H.B., 2011. Rebuilding the Great Britain Historical GIS, Part 1: Building an indefinitely scalable statistical database. Historical Methods 44, 149–159.
Tilly, C., 1998. Durable inequality. In: University of California Press, Berkeley.
Turchin, P., 2008. Arise 'Cliodynamics'. Nature 454 (7200), 34–35. http://dx.doi.org/10.1038/454034a.
Turchin, P., Nefedev, S.A., 2009. Secular cycles. In: Princeton University Press, Princeton.
Turchin, P., Whitehouse, H., Francois, P., Slingerland, E., Collard, M., 2012. A historical database of sociocultural evolution. Cliodynamics: The Journal of Theoretical and Mathematical History 3 (2), 271–293.
Turchin, P., Currie, T.E., Turner, E.A.L., Gavrilets, S., 2013. War, space, and the evolution of old world complex societies. Proceedings of the National Academy of Sciences of the United States of America 110 (41), 16384–16389. http://dx.doi.org/10.1073/pnas.1308825110.
Turchin, P., Brennan, R., Currie, T.E., Feeney, K., Francois, P., Hoyer, D., Manning, J., et al., 2015. Seshat: The global history databank. Cliodynamics: The Journal of Quantitative History and Cultural Evolution 6 (1), 77–107.
van Panhuis, W.G., Grefenstette, J., Jung, S.Y., Chok, N.S., Cross, A., Eng, H., Lee, B., Zadorozhny, V., Brown, S., Cummings, D., Burke, D., 2013. Contagious diseases in the United States from 1888 to the present. New England Journal of Medicine 369, 2152–2158.
Van Zanden, J.L., 2014. How was Life?: Global wellbeing since 1820, height and standards of living in the global past. In: OECD Publishing, Paris.
White, D.R., Feng, R., Gosti, G., Oztan, T., 2011. Easy R scripts for two-stage least squares, instruments, inferential statistics and latent variables. Sociological Methodology. https://pdfs.semanticscholar.org/7183/f378339210fc60a68c20b2632da38c3f6419.pdf.
Zadorozhny, V., Grant, J., 2015. A systematic approach to reliability assessment in integrated databases. Journal of Intelligent Information Systems 46 (3), 409–424. http://dx.doi.org/10.1007/s10844-015-0359-2.
Zadorozhny, V., Lewis, M., 2014. Fusing information, crowdsourcing and mobility (tutorial). In: Proceedings of the 15th International Conference on Mobile Data Management (MDM'14).
Zadorozhny, V., Manning, P., Bain, D., Mostern, R., 2013. Collaborative for historical information and analysis: Vision and work plan. Journal of World-Historical Information 1 (1), 1–14. http://dx.doi.org/10.5195/jwhi.2013.2.
Zadorozhny, V., Lee, P.-J., Lewis, M., 2015. Collaborative information sensemaking for search and rescue missions. In: Proceedings of the 12 th International Conference on Information Systems for Crisis Response and Management (ISCRAM'15), 2015.

Relevant Websites

https://www.aaas.org – American Association for the Advancement of Science (AAAS).
http://http://aligned-project.eu – ALIGNED. ALIGNED Software and Data Engineering; European Union Horizon 2020 Project no. 644055.
http://www.psc.edu – Bridges. Bridges: A Pittsburgh Supercomputing Center Resource.
http://www.hecc.ubc.ca – Centre for Human Evolution, Cognition, and Culture, University of British Columbia (CHECC).
http://chia.pitt.edu – Collaborative for Historical Information and Analysis, World History Center, University of Pittsburgh (CHIA).
http://www.clariah.nl – Common Lab Infrastructure for the Arts and Humanities (CLARIAH).
https://www.clio-infra.eu – Clio-Infra. CLIO-INFRA, International Institute of Social History, Amsterdam; Reconstructing Global Inequality.
http://colfusion.exp.sis.pitt.edu – Col*Fusion. Collaborative Data Fusion.
http://www.correlatesofwar.org – Correlates Of War. Correlates of War project.
https://dataverse.harvard.edu – Dataverse. Harvard Dataverse.
http://datadryad.org – Dryad. Dryad Digital Repository.
https://ec.europa.eu – EC_Research. European Commission Research and Innovation.
http://www.ecai.org – Electronic Cultural Atlas Initiative (ECAI).
http://hraf.yale.edu – eHRAF-archaeology. Human Relations Area Files, Archaeology, Yale University.
http://ehrafarchaeology.yale.edu – eHRAF-culture. Human Relations Area Files, World Cultures, Yale University.
www.port.ac.uk – GBHGIS. Great Britain Historical GIS.
https://collab.iisg.nl – GCHLR. Global Collaboratory on the History of Labor Relations, International Institute of Social History, Amsterdam.
http://www.icpsr.umich.edu – Interuniversity Consortium on Political and Social Research (ICPSR).
https://socialhistory.org – International Institute of Social History (IISH).
http://www.irows.ucr.edu – Institute for Research on World Systems (IROWS).
https://pitt.hosted.panopto.com – Language&History. Patrick Manning, "Language in History".
http://www.ggdc.net – Maddison. Maddison Project.
http://dx.doi.org/10.7910/DVN/ZN1WLF – Mafrica. Mafrica, Chelsea, 2016, "Place, Population, Precipitation, War since 1800", doi:10.7910/DVN/ZN1WLF, Harvard Dataverse, V1 [UNF:6:sS4CKXAhPciatrart61R7Q==].
http://www.pop.mn.edu – Minnesota Population Center (MPC).
http://neo4j.com/ – Neo4j Graph Technology.
http://pelagios-project.blogspot.com – Pelagios. Pelagios project.

http://perio.do – PeriodO. Gazetteer of Period Definitions.
http://religiondatabase.org – ReligionDatabase. Database of Religion, associated with CHECC.
http://seshatdatabank.info – Seshat. Seshat Databank.
https://evolution-institute.org – Seshat-Evolution. Seshat – Evolution Institute.
http://aligned-project.eu – SeshatOntology. Detail on the Seshat ontology can be found in ALIGNED.
http://songgis.ucmercedlibrary.info – SONGGIS. The Digital Gazetteer of Song Dynasty China, Version 1.1. Ruth Mostern with Elijah Meeks, 2010.
https://www.terrapop.org – TerraPopulus. Terra Populus : Integrated Data on Population and Environment.
https://www.tycho.pitt.edu – Tycho. Project Tycho: Data for health.
https://www.wider.unu.edu – United Nations University: World Income Inequality Database (UNU-WIDER).
http://utip.gov.utexas.edu/data.html – University of Texas Inequality Project (UTIP).
http://www.slavevoyages.org – Voyages. Voyages, the Trans-Atlantic Slave Trade Database.
http://topincomes.gmond.parisschoolofeconomics.eu – WTID. Alvaredo, F., Atkinson, A. B., Piketty, T., and Saez, E. The World Top Incomes Database.
https://www.xsede.org – Extreme Science and Engineering Discovery Environment (XSEDE).