Big Data in History: a World-Historical Archive

Version 1.1

Patrick Manning
Director, Center for Historical Information and Analysis
University of Pittsburgh

1. Challenges of Big Data in History

The time has come for creating and analyzing a global dataset on human societal activities. Such a dataset can provide a picture of global social patterns and interactions over the past four or more centuries. Most basically, this world- historical dataset is to portray long-term, global change in human society and thereby provide a basis for planning long-term, global policies for the future.

Big Data in history will provide a new, comprehensive level of documentation on the past. Currently available historical information, while immense in its overall quantity, is scattered and dispersed. Libraries and archives in great cities hold treasure troves of data on trade, politics, and religion for national and imperial centers, but each archive is separate from the others, and the totality of their records provides far less information on people of rural areas. The idea of Big Data in history is to digitize a growing portion of existing historical documentation, to link the scattered records to each other (by place, time, and topic), and to create a comprehensive picture of changes in human society over the past four or five centuries. This overview indicates the types of historical data to be assembled, the techniques for storing and analyzing these records, and the type of patterns and connections in local histories and world history that could come from creation of this global dataset. Initial stages of the global dataset are focusing on evidence about the economy, society, politics, health, and climate. Later on, the project will address Big Data on ideas, culture, and values.

Creating this global historical data resource is now feasible, not only because of advances in information technology but because of breakthroughs in communication and collaboration among historians and social scientists. The exciting advances of Big Data in the natural sciences provide encouragement and specific techniques that will draw historical data together. In the study of climate, a huge collaborative effort at an international level has developed models and

empirical evidence on global climate in recent centuries and also in the distant past. In astronomy, there has been a parallel collection of great quantities of new data which give a steadily improving picture of the universe and its patterns of change—from the local level of our planetary system to the scale of the entire universe. In biology, a great research effort has just achieved a new level of precision in description and analysis of the human genome. The problems of creating a dataset on human history will be different from those in natural-science fields, but the general level of feasibility of the project is roughly parallel. The Center for Historical Information and Analysis formed in 2011 to define the specifics of this task, gather resources, and begin the actual work. As we have learned, each stage of our work must balance two complementary tasks: building human collaborations and building our technological capacity. At present, and focusing both on collaboration and capacity, we have organized our project into three basic missions.

Mission 1 is to gather and archive historical data. The collaborative dimension of this mission focuses especially on data collection. We have begun the work of creating our Phase 1 Archive, assembling datasets now existing or to be created out of historical resources. Data can come from the scanning of thousands of pages of tables on trade and administration of units worldwide as published in the British Parliamentary Papers from 1810 on; data can come from transcription or perhaps scanning of handwritten administrative records accumulated over hundreds of years in the imperial archives in Beijing, Istanbul, or Lisbon. Such important work raises the problem of the high cost of transcription, digitization, and verification of historical data. Yet while the cost of this work is high, the value of the data may be even higher. Another dimension of data collection is the system to facilitate incorporation of data brought by researchers who wish to submit data they have developed, documented, and perhaps analyzed. For this task we are investing substantial energy in developing a "crowd-sourcing" application that enables remote users to interact with the archive, enter the documentation for their data, and then submit the data. For all evidence being submitted to the Phase 1 Archive, it is necessary to document the data fully and also to verify the consistency and accuracy of the data. The "metadata" – the statements of data description—must fully define the source of each dataset, the creator or compiler of the data, and the precise definition of each variable and each value. Of particular importance is the attribution of source, ownership and compilation of data, as the project seeks to maintain complete and accurate attribution, not only for each dataset submitted by a contributor but also for all the derivative datasets incorporating materials from that dataset. A further distinction in the collection of data for the Phase 1 Archive is the difference between "data-rich" domains and "data-poor" domains. For some times and places, we have remarkably thorough and complete historical records: for instance, the documentation on Sweden and on Japan since the seventeenth century is remarkably complete. In contrast, the historical documentation on Thailand, Nigeria, and Bolivia is much thinner and much more uneven. Yet Thailand, Nigeria, and Bolivia each had significant populations and important roles in global society and economy (and they have also had literate strata for centuries), so we must find an appropriate way to include them. Finding ways to estimate missing data will be an essential element of this project to create consistent, worldwide series of data. Within a few years of work, we expect our Phase 1 Archive to hold terabytes of data collected and submitted by researchers worldwide.

Mission 2 is to aggregate data up to the global level. Our Phase 2 Archive is to be a unified, global-historical collection of consistent and connected data on human experience during the past several centuries. This Phase 2 Archive will result from transformation, linkage, and aggregation of resources from the Phase 1 Archive, where they have earlier been submitted by contributors. Organization of data in the unified archive requires that the documentation of space and time be at once systematic and flexible—that is, able to account for all the different ways that space and time are labeled in historical documents. In practical terms, the archive will begin with the systematic assembly of population data, and will go on to address a wide range of variables from the local to global levels. Then takes place the work of data "harmonization"—that is, the cleaning, fusion, integration, transformation, and aggregation of submitted datasets into larger and more comprehensive datasets. (The nuances among these various types of data transformation are described later on.) The archive will facilitate analysis of links among variables and among the various levels of social organization. To integrate materials of all different sorts, work will advance on locating links and contrasts in social science theory at local and global levels. In addition, the work of aggregating and linking data, since it brings further transformation of data, will also generate "incremental metadata" to record each of these modifications. This system of global historical data, once initially developed, will be able to expand with the location of additional data or inclusion of additional collaborators. In particular, to understand the quantity of historical data likely to be included in the Phase 2 Archive, it is important to become clear about the difference between *existing data* and *new data* in history. This can be explained by comparison with Big Data in the natural sciences: while one part of this project is the collection and digitization of known historical records, another part of it will result in the discovery and creation of immense amounts of *new* historical data. As in geology and astronomy, even though the facts of the past remain unchanged, today's developments of theories and techniques will result in the development of huge amounts of empirical information on the past. Finally, the CHIA project will seek ways to link its efforts with two large data-collection projects, CLIO-INFRA (economic historians based in the Netherlands) and TerraPopulus (census analysis based at the University of Minnesota). These connections will expand the global collection of data to include population in the post-1950 period and economic historical data worldwide. Through inclusion, aggregation, and documentation of the full amount of such data, the Phase 2 Archive will expand to the petabyte level.

Mission 3 is to visualize, analyze, and mine the data. While analysis is logically the last step of work with a dataset, in fact visualization and analysis of data will take place at every stage of the work. Initially, we are creating a "faceted search" through which users can explore worldwide data on population, climate, silver flows, and wars for the twentieth century, to show how specific datasets may be selected by space, time, and topic, and then displayed in an introductory interface. This and later versions of tools for visualization and analysis will provide feedback and ensure that the needs of users will remain central to the design and construction of the system in Phase 1 and Phase 2 of the archive. Visualization of the data, while it will surely include elementary summaries of variables and their descriptive data, must also give adequate representation of the multiple dimensions and multiple forms of interactions among variables. Data visualization must go far beyond spread sheets: it must draw upon the sophisticated, multidimensional representation of variables that have been developed in studies of climate and genomics. Drawing on resources from the Phase 2 Archive, it will be possible to analyze interactions throughout the global system of historical data. In some cases these will be expanded

analyses of known relationships—as between climate variation and agricultural output; in addition, techniques of data mining may permit identification of relations and patterns far beyond those now known.

The sections to follow expand on each of these points. On one hand, this essay will emphasize the range of challenges to resolve and obstacles to surmount before we can create a world-historical data resource. On the other hand the same essay is intended to demonstrate that comprehensive and energetic collaboration can solve those problems. Within just a few years, we can achieve a strong beginning for global-historical analysis in the social sciences. In the same process, we can document past links between human society and natural processes. The biggest and most important result of this collaborative project will be the creation of a worldwide, historical data resource on human society. The formulation of world-historical data will facilitate the needed research on social patterns at levels from the local to the global.

The CHIA project pledges to maintain open-source, open-access, non-proprietary standards throughout its work in constructing a world-historical archive. As we see it, the contributions of this project belong to everybody. Further, CHIA pledges to maintain high standards of attribution: the contributors of data must be recognized for their submissions and system developers must also be recognized for their contributions. Two types of problems, however, will arise at every stage in this work. First, problems of collaboration – of the social organization of this collective work. The work requires tight overall organization, but also requires a maximum of flexibility and independence for participants at every level. To succeed, it must become a very large-scale project with links to every corner of the world. Second, problems of capacity – of sufficiently high-level technology and of sufficiently large resources in space and funding to carry out the many demands of the project. Mission #1, gathering data, focuses on *collaboration* in the collection and sharing of historical data. It emphasizes building *new capacity* with its focus on peer review of datasets and by using a crowd-sourcing application that enables contributors to submit data easily from remote locations. Mission #2, aggregating data, focuses on *collaboration* in collecting data from all areas of the world. It focuses on *new capacity* through the creation of a unified global archive, relying on harmonization of heterogeneous data and the aggregation and linking of data at all levels. Mission #3, visualizing and analyzing data, focuses on *collaboration* through incorporating data from parallel groups involved in assembling global data. It focuses on *new capacity* through developing multi-dimensional systems for representing data and through application of data-mining techniques to locate unsuspected linkages within the global dataset.

Along the way to achieving the main goal of the project, a number of important, collateral benefits will arise. That is, we will have almost immediate benefits of this project even though it will take years to produce a genuinely Big Data dataset. There are immediate and intermediate steps that we can take that will be of great value to the study of social sciences.

- **Crowd-sourcing application** for collecting and archiving historical and social data will open the bottleneck that has so far prevented systematic study of human society at a large scale.
- **Peer-reviewing of datasets** through the Journal of World-Historical Information will bring recognition of the scholarly value of creating datasets, and will ensure that high standards for creating historical datasets are created and maintained.

- **Collaboration among social scientists** all over the world, through the creation and maintaining of a global system of historical data, will bring additional sharing of data and analysis.
- **Technical and analytical skills of social scientists** will advance through the process of collecting and analyzing data, and demonstrate the parallels and the links of social sciences and natural sciences.
- **Theoretical debate.** The expanded effort to link and apply social science theories, especially in order to fill in missing historical data, will strengthen theory and analysis in social science.

All of these advances in social science and information science are likely to come from systematic work on creating world-historical data. Most important, however, is the basic objective: since we have now become fully cognizant of the global nature of human society—by which we mean the intensity of global social interactions and the interactions of the human and natural world—it is a top priority for us to learn about the historical record, at regional and global levels, of our evolving human society.

2.The need to know our global past

All of us are making plans for a global future. Leaders in government, the economy, in society—along with ordinary people at every level—are trying to keep global change in mind as they make plans. The periodic global summits on environmental issues are one example: they were held at Stockholm (1972), Rio (1992), Kyoto (1997), Johannesburg (2002), Copenhagen (2009), and Rio (2012). The World Bank seeks to develop projections of global economic change. Military leaders in the U.S., NATO, China, Russia, and other countries make contingency plans for global conflict. We also see plans and projections on global health, energy use, and levels of education. Families and individuals conduct planning for a global future by identifying career choices, education plans, allocations of family wealth, and even marriage choices.

But in proposing these plans for the global future, what information do the planners have about the global past? What hints do the planners have about the actual specifics of global conditions, even today? What errors might result from projecting patterns of a certain city or country as proxies for global realities? Are the current changes in cultural fashion or in unfolding social movements really "for the first time," as the participants commonly claim? It is risky to make decisions at the global level while blind to past patterns—especially since it is now possible for us to learn about past global patterns in human society. Similarly, we now have the capability to document comprehensively the global patterns of today and see to what degree they are similar to or different from patterns of the past. This is the need for world-historical data and analysis.

The problems in global society—in governance, health, social inequality, population change, and human interaction with the environment—stretch across regions and disciplines. The crucial point is that one cannot conduct global analysis without global data. Social sciences, though sophisticated in analysis of contemporary societies, continue to work within regional and disciplinary boundaries and with historical data that are scanty, especially for times before 1950. Advances in social theory and information technology, however, bring a substantial opportunity

to develop data and theory at global scale and over several centuries. Global data will document the overlaps of such human dynamics as the regular pace of generational replacement, sharp business cycles, erratic political changes and the influence of climate and disease.

We cannot effectively address the global-historical analysis of social-science issues without developing large quantities of dependable global data. But in order to develop dependable global-historical data we must overcome a range of obstacles—organizational, technical, and conceptual. Organizational obstacles include the silos of various disciplines (in which researchers speak only to others of the same discipline) and the differences in data collection separating wealthy and poor regions of the world. Technical obstacles include the heterogeneity of the data, the complexities of data description, and the amount of skilled labor required to digitize documents in print, manuscript, or image form. Conceptual obstacles include the reality that social science theories remain within discrete domains that restrict global analysis. Thus, micro and macro theories within given disciplines are not clearly connected; disciplines are connected only marginally to each other, and some visions of "the world" leave out whole continents, notably Africa and South America. Further, in the frustrating feedback loop with which we began, global patterns in social science have not been explored in much detail for lack of global data.

As an example of how large-scale analysis can fall short of being global, empires are mostly investigated one at a time: they are studied occasionally as they interact with a competitor but almost not at all as interactive parts of a world-wide political system. Thus the British Empire, even at its most dominant, held only a portion of the world's imperial possessions: actions of other empires constrained the British. Fascinating and groundbreaking analyses await us once we assemble historical data systematically at regional and global levels. We will be able to trace the rise and fluctuation of global systems of money and credit.[Reinhart and Rogoff 2009] We will have detail on shifting global patterns of population, mortality, and migration. We will trace the unfolding of governance at local, national, and imperial levels, and changing systems of family structure.

Examples of Global historical interaction and change. Here is a compressed narrative of the last five centuries of human history, drawing especially on a few variables, as if they were drawn from a dataset. The central factors in this concise global story are human experiences with *lifespan*, *migration*, *textiles*, *silver*, *empires*, and *social inequality*. Even this simplified view of world history shows the differences of local and global patterns. It shows surprising variations and interactions and raises questions about global change. Let us begin with population and lifespan: five centuries ago, world population was slightly more than half a billion, not even one tenth of what it is today. World regions in rank order of population were China, India, Africa, Europe, the Americas, and other parts of Asia (Southeast, West, and Central), as they are today. At that time, lifespans averaged some 30 years, slightly lower in tropical areas. All regions experienced high levels of infant mortality, accompanied by high levels of mortality of mothers in childbirth.

Expanding oceanic voyages brought migration of sailors, soldiers, and merchants. Global trade required an effective currency: candidates were African gold, Chinese copper, and even pearls or diamonds. But the opening in 1550 of great mines in Bolivia and Mexico ensured that

silver would become the world's main currency. Each year, Spanish fleets carried silver to Seville and single galleons carried silver across the Pacific to Manila. Buyers in Amsterdam, Moscow, Constantinople, Bombay and Canton competed for a share of Mexican silver. In exchange, Chinese silks, Indian cottons, European woolens, and elegant ceramics traveled in opposite directions. But growing interregional contact brought serious problems as well. Sailors on long voyages suffered from scurvy – a vitamin C deficiency from lack of vegetables. Human contact also brought spread of infectuous disease. In the Americas, higher death rates brought disastrous population decline during the sixteenth century as Old World diseases reached communities that had never before experienced them. Similarly, smallpox spread in Africa and syphilis reached Japan.

Expanded social interactions of several types accompanied global travel and trade. Great empires arose suddenly in about 1500 in India and Iran, and also for the Ottomans, Spanish, Portuguese, and Russians—then the Songhai empire in West Africa fell. In about 1650 another set of empires arose, partly displacing the previous powers: the Dutch, French, English, and the Qing state in China. Settlers from Europe, Africa, and Asia moved to crowded lands and to empty spaces. European migrants to Asia and the Americas, mostly in military service, were outnumbered by African migrants to the Americas and the Mediterranean, mostly in slavery. In the eighteenth century now-under-populated Americas began population grew in the Americas with the arrival of European and African migrants; in western Africa, population began to decline because of enslavement. Indian cottons went to Africa in exchange for many of these captives. Still, silver flowed from Bolivia and Mexico.

The Atlantic, Indian, and Pacific oceans became increasingly linked in commerce and migration, and eighteenth-century health conditions gradually improved in Eurasia and the Americas. Atlantic empires came to depend on forced labor to produce sugar, tobacco, coffee; Asian empires produced sugar and opium. The English and French fought a century of wars on every continent. The English replaced the Mughals in India, but lost the United States. The empires of England, France, Spain, and Portugal declined in size from 1780 to 1820—only China and Russia maintained their size. Silver mining declined in independent Bolivia but expanded in Mexico. Now it was merchants from the independent United States who shipped silver to Europe, Asia, and Africa. England, relying on slave-grown American cotton fiber, began to displace India as the main global source of cotton textiles. Emancipation of slaves in the Americas and of serfs in Eastern Europe and Russia brought migration of newly freed people. The rise of steamships overlapped with emancipation.

From 1850 steamships changed commercial transport and revolutionized migration. Migrants first left Europe – Irish in the lead. Overall, this great wave of migration from 1850 to 1940 evened out world population. Migrants went mostly from densely populated regions—Europe, India, China, Russia—to take up work in regions of sparser population. Over 30 million each went to the Americas, Southeast Asia, and Northeast Asia. Africans too continued to migrate, now from one part of the continent to another. Emancipation and migration worldwide brought new social mixtures that generated new ideas of racial hierarchy on every continent. At the same time, new industrial economies became rich by comparison to the previously leading economies of China and India. Shortages of money in an expanding global economy pushed up the value of silver and gold. Gold rushes broke out from 1848 to the end of the century, when

South African deep mines were established. But still the Mexican silver continued to flow from the mines. Britain led in creating a gold standard for international money from 1850 through 1930, though silver remained the main monetary metal.

New social problems and huge gaps of inequality developed during the nineteenth century. As the steamships carried migrants and commodities rapidly and on schedule among all ports of the world, they also carried microbes. Cholera, which had previously been restricted to India, now showed up in ports throughout the world. Typhoid, tuberculosis and other bacterial diseases spread rapidly around the world. Levels of health, education, and income improved dramatically in Europe, North America, and Japan, but also with growing gaps of rich and poor. China and India became economically poor regions during the century. Remarkably, lifespans grew in many areas of the world during the nineteenth century, though in Africa, where expanded enslavement reduced lifespans. Still, terrible droughts and horrible famines broke out in climate shifts that are now understood as the El Niño Southern Oscillation. The results brought great loss of life in the 1870s and 1880s in India, China, Brazil, and Africa.

Empires expanded again, if briefly. Late in the nineteenth century, European powers absorbed all of Africa and the Pacific, and most of Asia. Yet they collided in two world wars, from 1914 to 1945, bringing revolution and ultimate collapse of all but two of the imperial powers. Imperial decline from 1945—"decolonization"—brought recognition of a hundred additional nations on all the continents and in the islands. The United Nations arose as a global forum and a Cold War pitted the U.S. and the Soviet Union against each other until 1992. Money changed, as its supply depended primarily on the checking account balances available in the U.S. Even so, the ups and downs in silver trade – still coming mostly from Mexico, remained an important factor in the global economy. Textiles manufacture moved back to Asia—especially Japan, India, and China. Yet synthetic fabrics – nylon, acrylic, polyester, and more—supplemented and displaced cotton and woolens. Another long wave of migration began in the 1960s, but international migration was smaller than before in volume because of national and racial restrictions, and migrants were now treated as a threat rather than tolerated. More significantly, migration built huge cities on every continent, and made human population more uneven.

The world of the twenty-first century brought new inequalities and new equality. Labeling and hierarchical ranking of people by "race," which crystallized in the seventeenth century and expanded in the nineteenth century, gradually declined. Higher education served wealthy countries and wealthy families, but the great global disparities in literacy and lifespan separating rich lands from poor lands declined. By 2010, the average lifespan worldwide exceeded 70 years. African countries remained behind in lifespan (especially those hit hard by the AIDS pandemic), yet African lifespans nearly doubled between 1950 and 2010. More than ever, however, economic inequality continued to expand within nations and between nations, separating poor from wealthy countries and poor families from rich families.

In this hurried but complicated narrative of global interaction, only a few key factors were used to recount human history. Adding in such factors as family and environmental change would add to the complexity and the connections. The story of each factor depends on change in the others. Documenting the global interactions in human society is a substantial task, and it may

lead to results that are surprising. A world-historical data resource—Big Data on the human past—will give us a better version of large and small contours in the overall story.

The realistic potential for documenting the global past. Is it realistic to plan on creating a world-historical data resource? Recent technological advance, especially in electronic communication, certainly makes it easier. Electronic scanning can translate texts and images from analog to digital media: Google Books has pressed further ahead on digitizing existing print works. Creating a scan of an existing document preserves it in a new format, but we still need improved technology in scanning print and manuscript files so that they can automatically be translated into searchable, digital text.[1] The internet, which is gradually converging with telephone communication, makes it possible to communicate by voice, image, text, or data files anywhere in the world, if the relevant devices are available. Geographic Information Systems, which had become a successful commercial system by the 1980s, allowed for a steady expansion in spatial definition of electronic files. Also from the 1980s, the emergence of supercomputing systems brought a steady emphasis on large-scale storage and analysis, working notably on climate. Computational systems involving the interaction of large numbers of variables and big datasets are an important part of this development. In a more recent development, the notion of collective intelligence has led to focus on interfaces that enable large numbers of individuals to participate in collection and analysis of data on a given problem. In addition to these advances in technology, every discipline has expanded greatly the quantity and the detail of its scientific knowledge. For instance, the great expansion in knowledge of the earth's climate history and the growing detail of history of disease are now ready to be combined with data on population and on agricultural production to give us clearer explanations of past changes and interactions of these variables.

But more than technology has changed. We are beginning to develop an adequate conceptualization of the notion of global patterns in human society. The idea of humanity as a whole has been understood for a long time, but primarily in the minds of a few visionaries. Religious leaders envisioned the fate of all humanity in relation to the Creator; emperors considered the possibility of conquering the whole world; observers of the heavens compared the earth and its peoples to the heavenly bodies. Humans spent most of their time, however, focusing on their families, communities, and the states within which they lived—and they still do. But in recent decades almost everyone, for one reason or another, has come to spend time thinking about the world. This greater consciousness of the world and of human society is the first reason why it is now more possible than before to begin organizing information for the world as a whole.

Thinking globally means more than thinking about large areas. It means long periods of time and it means a wide range of topics. In addition, thinking globally means considering not just analysis at the largest scale, but the interactions of life all the way from the smallest scale to the largest and back again. An analogy with the field of biology is relevant. At one level, biology is the study of plants and animals – whole organisms. But plants and animals range from the tiny to the huge. In addition, biologists study the elements of each organism: they study cells and the constituents of cells, down to the molecular level. At the other extreme, biologists study whole herds of one species, and they study great groups and evolutionary orders of plants and animals. Yet on every level it's all biology, and each part of the earth's great biological system interacts

with many others. Similarly for human history: the overall view we seek to document ranges from individual behavior through various groups to all of humanity, and it includes many types of activity over short and long periods of times. This sort of global understanding is now expanding in the social sciences.

Advances in conceptualization have not simply expanded in scale. Within the past half century the notion of *systems* has developed productively in many areas of intellectual work. Systems are conceived as collections of interacting components which combine to sustain a larger whole. Systems are described in mechanical, thermodynamic, and organic terms, but also in social and environmental terms. Systems have structure, interconnectivity, and behavior. Some systems have purposes and exhibit adaptive behavior. Systems can be modeled in ways that distinguish closed from open systems; study of systems draws attention to information systems within them. The application of systems-thinking is valuable at multiple levels in this project: we can treat an archive or an application as a system and we can treat the whole of historical society as a system. Linking these extremes, we can think of our participating colleagues—who build and maintain a world-historical data resource—as comprising a system in themselves.

Overall, I argue, the implementation of a global-historical data resource requires the unification of social-science analysis. That is, the various social sciences, while they will continue to have their particular arenas of application and specific purposes, must become more explicitly linked to each other. There have been important efforts to trace links and commonalities in human social and historical behavior, but typically they have reached limits because they could only explain so much.(Marx 1975, Compte 1967, Wallerstein 2001) More commonly, social scientists have been content to remain within their domains, explaining more and more about less and less. Social-science analysts have sought out data of homogeneous quality and in finding it have tended to stay within national units, short time frames, and standardized data such as censuses. Crossing boundaries in time and space requires facing heterogeneity: it involves linking terms with changing meanings, linking maps with inconsistent scales, and addressing multiple languages, varying weights and measures. Policy-makers are learning that long-term processes, previously ignored or undetected, have significant implications for the decisions they seek to implement: early events may have generated structures with lasting impact.(Nunn 2009) In analysis, recent decades have seen dramatic change in the outlooks and scholarly practice of social scientists and the techniques available to them. After the decline of colonialism and racialism, it has become easier for social scientists to seek out common experiences and motivations for our species as a whole rather than focus on uniqueness and socially specific attitudes defined by race or nation. Global and historical interests have grown among researchers in economic history, global politics, world systems, and global health.[2]

In retrieval of social-science data, the most obvious new technology is large-scale digitization of print, manuscript and image data. For the organization of data, new techniques in GIS make it possible to define and analyze units that are modifiable in area and time.(Southall 2011) Other advances enable an attack on problems in missing data: on one hand with ways of getting useful information out of incomplete datasets; on the other hand by using advanced techniques of simulation and estimation to fill in the blanks.(Honaker and King 2010, Manning

2010) Collaboration, meanwhile, has advanced more slowly in social sciences than in natural sciences. The inherited notion of the individual investigator is still valuable in every field: in astronomy, for instance, individual amateur astronomers still make important contributions to the field. But for historical studies, learning how to conduct collaborative research, write co-authored papers, and jointly seek support from major institutions is essential for large-scale analysis. While patient individual work has chipped into this barrier, the main hope for advance lies in a large-scale campaign of data retrieval, transformation, and flexible integration that will make historical data accessible for global analysis.

3.CHIA: Collaborative mission, structure, innovation

Philosopher George Santayana understood the need for a large-scale historical resource long before it was feasible to create one. He warned (in 1901) that, "Those who cannot remember the past are condemned to repeat it." Less known, but likely as important, is his challenge that "a man's feet should be planted in his country, but his eyes should survey the world." The Center for Historical Information and Analysis (CHIA) exists to accelerate and empower research surveying the global human record. As a public system, it will be used synergistically by policy-makers adopting decisions, scholars identifying global processes, and educators developing student skills in global analysis. It will ingest comprehensive, multidisciplinary data and provide tools to uncover the patterns of social interaction and the processes driving these interactions. Its research and analysis will integrate the approaches of social, health, and environmental sciences with those of information sciences. The result, an improved understanding of past patterns in society at all levels, is fundamental to assessing future challenges and predicting the success of proposed solutions.

The long-term purpose of CHIA—in the time frame of perhaps a decade—is to facilitate the creation and maintenance of historical data sets from local to global levels, from short term to long term, linking variables on many areas of human experience. The resultant summation of human experience can reveal the varying patterns and dynamics of social change. While past social, economic, and cultural dynamics may not carry automatically into the future, they should not be neglected in our attempts to make plans and form policy. The Center intends to link social sciences to each other and to the principal problems in human society, at scales from the local to the global over the past four centuries and into the future. It seeks to encourage a culture of data sharing among social scientists. And it expects to develop a global, integrative repository and analytical framework supporting specific research projects on four domains of social life: *human-natural interaction, population change, development of socio-economic inequality, and governance (local and global)*. New knowledge of these past patterns will surely shape policy formulation.

In the intermediate term, roughly five years, CHIA intends to develop a strong and expanding research team which will unleash a rapid inflow of historical data to be documented and archived. CHIA will develop an overall ontology for world-historical documentation and analysis, including an expanding system of metadata to describe data and assist in their integration and aggregation. CHIA will conduct interactive analysis at regional and global levels of variables in social sciences, health, and climate; and develop systems of visualization that will assist in analysis and provide feedback for collection and definition of data.

CHIA is a collaboration of scholars in social sciences, information sciences, and natural sciences from institutions in the United States and several other parts of the world. It has formed to respond to the need for global, historical data in the social sciences and in recognition of the current advances in the potential for assembling such data. CHIA took form in 2011 as a new initiative launched by veterans of several earlier campaigns to collect, preserve, archive, and analyze social science data on a large scale. Typically of the new type of academic work, the founding meeting took place by internet discussion rather than in a single room. The group adopted a comprehensive approach, seeking to emphasize new conceptions, not just accumulation of data. Aspects of the work as identified by CHIA include the data, a place to put data, keeping track of all the different types of data, documenting individual data (by time, place, topic, and scale), gaining a sense of the links among data as seen through theory, conducting advanced analysis to discover significant patterns within the data, and making the entire resource available to researchers, teachers, and students everywhere.

The organizational structure of CHIA was designed to fit its task. Administrative headquarters are at the University of Pittsburgh, where the project's objective first gained recognition with university financial support. CHIA is housed in the World History Center, founded in 2008 as an institution for "research, teaching, and international collaboration on the global past, with attention to policies for the global future." CHIA is governed by an Executive Committee consisting of its director and four leading scholars in social science and information science, according to a set of statutes adopted by its founders. The Executive Committee admits Affiliates who apply for membership by proposing and submitting deliverables—that is, data, procedures, or consulting work which contribute to constructing a world-historical data resource. Membership is renewed every six months. The Executive Committee appoints an Archivist who supervises the CHIA archive, the submission of data to it, and the procedures of data documentation and analysis. The Executive Committee governs and appoints the editors of the *Journal of World-Historical Information*, its official journal. Further, the Executive Committee may set policy and approve applications for funding of research activities. Affiliates are groups of three or more researchers based at a research institution who propose and submit deliverables; they are otherwise self-governing or governed by their local institutions. In addition, individual researchers may become members of CHIA by agreeing to submit historical data. The Executive Committee meets nine months a year in three-day online meetings; the full membership of CHIA meets once a year in an online meeting. Within this framework, individuals and groups within CHIA define and carry out research projects, using their own resources, resources of their institutions, or through successful application for external funding.

CHIA bring significant innovations to this effort to unify researchers in large-scale collaboration: several such innovations are marked in bold within this paragraph. The global strategy of CHIA requires a consistently **global vision** combined with intensive interactions at local and intermediate levels. This global vision ensures that, while details of the project draw researchers into work on many specifics, overall project strategy remains fixed on developing world-historical data on human society To implement this global strategy, the project emphasizes building a research collaborative with **a balance of affiliates and headquarters** in communication and decision-making. Headquarters is the center of gravity, maintaining the archive and facilitating communication. The affiliates are institutionally-based groups of

researchers who carry out the tasks of data collection, documentation, and analysis, commonly in collaboration with each other. In addition, the CHIA collaborative is to facilitate the sharing of data by individual researchers through crowdsourcing and analysis of the archival holdings by individuals and groups. CHIA will construct **a global archive** which performs multiple functions and is developed through appropriate stages. Phase 1 of the archive incorporates a wide range of relevant datasets, linking them to create a web of global-historical data (reaching across time, space, scales from local to global, and data from various disciplines). In the juncture between Phase 1 and Phase 2, CHIA will serve as 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. In Phase 2, the archive is to integrate and aggregate the originally submitted datasets of Phase 1 into a consistent, world-wide set of data. CHIA emphasizes **cross-disciplinary alliances in academic fields**: this means close, research-focused relationships among researchers of distinct social-science fields and equally close work with researchers in natural-science fields and in information science. All of these cross-disciplinary links support the development of a project to **link and unify social-science theory**, both to expand the realm of theory and to advance the collection, estimation, and analysis of social-science data. To achieve wide participation and collaboration in the collection and documentation of historical data, CHIA has committed itself to development of **crowd-sourcing in data incorporation** along with a "data-hoover" approach to solicitation of data. Further, within the developing world-historical archive, CHIA emphasizes **"harmonization" of data,** a systematic set of procedures to render data mutually consistent so that they can be aggregated up to a global level. In addition, CHIA is committed to open access to its materials, and emphasizes the **participation of researchers, teachers, and students,** who will be able to consult data, use the project's visualization tools, and conduct data analyses at every stage of the project.

For CHIA to accomplish its missions of data-gathering, aggregation, and analysis—and thereby to provide answers to big questions on the global dynamics of social change—it must create a suitable cyberinfrastructure and a stream of data. In the terms of a recent initiative of the National Science Foundation of the U.S., completing each mission requires "building capacity and community." The criteria for the overall infrastructure are *analytical*—this is the capacity side of the CHIA mission: i) the data repository must be able to store spatial and temporal attributes flexibly to capture changes in borders and a variety of time scales; ii) organizing the data requires estimating large amounts of missing data and transforming data for consistency; iii) processing of data will benefit greatly from existing and new data-mining methods; iv) analysis requires linking of disparate theories; and v) visualization requires conveying analytical results and displaying elements of large data sets at various temporal and spatial scales. The infrastructure criteria are also *organizational*—that is, the collaboration side of the CHIA mission: i) the Center must achieve a readiness to contribute data among social, health, and environmental scientists; ii) researchers must cooperate in maintaining high and consistent standards in documenting data; and iii) theorists in various fields must collaborate to link their models. The need to work simultaneously and collaboratively on these issues helps explain why the Center has taken form.[3]

In fact, while the three missions are in some sense distinct, the task of creating global historical data requires constant interaction and feedback among the various parts of the project.

For instance, in one sense the visualization and analysis of data come, logically, after the collection of data. On the other hand, the results of visualization and analysis will provide new ideas on what data are most important and how they should be defined—so that the needs of analysis help determine what data should be collected and estimated. Similarly, the assembly of data that have already been archived into regional and global aggregations of evidence comes logically after the collection of data. On the other hand, one needs a sense of which types of global-historical data will be most valuable in analysis in order to set priorities for collecting the localized datasets which are to be the building blocks.

To summarize, CHIA will create an infrastructure for retrieving, holding, and analyzing world-historical data. CHIA will be an institution of sufficient scale and authority to address the analytical and organizational challenges of documenting human society in recent centuries. It will develop new data standards that account for heterogeneity, procedures for documenting and integrating heterogeneous data, and permanent housing for both raw and transformed data.[4] It will facilitate cross-disciplinary analysis and visualization, sustaining synergies among researchers in social, health, environmental, and information sciences. It will lead to elaboration of theory to connect existing theories. In organizational terms, CHIA will facilitate a campaign encouraging social scientists to collect and submit historical data for shared access and analysis. Out of this campaign there may arise an improved system of reward and recognition for sharing data. The Center will lead in articulating good practice in the inevitable debates about the ownership of data and citation and recognition of the contributors of data.

This is by no means the first effort to assemble historical social-science data on a large scale. The CHIA group is aware of the challenges, achievements, and failures of earlier groups. For instance, systems of national income accounting—established for most major national economies in postwar years—stand as an immense achievement in research, accounting, and analysis. These national accounts were created not only for current years but for past years, going back generations. Such historical data collection and analysis is in many ways parallel to the project we are now taking on – except that ours is at least an order of magnitude larger in scale. The CHIA project devotes significant attention to reviewing earlier work and to benefiting from lessons learned in previous projects. The penultimate section of this essay reviews previous projects on social science data in further detail.

The CHIA plan is to focus on collecting data for the era prior to 1950. We recognize that, for years since 1950, much progress has been achieved in developing global data through the energies of modern national governments, the United Nations, the World Bank and other groups.[5] The work of CHIA will focus initially on earlier times, for which data are less well developed yet equally important. The project is to focus especially on the data which have yet to be digitized, have yet to be published, and on the regions where even handwritten documents are scarce. As the collection of pre-1950 historical data becomes larger and better organized, CHIA's attention will turn to linking global data from times before 1950 with data on the world since 1950.

4. Mission #1: Assembling the Data

What sort of data belong in the world-historical dataset? The objective is to combine many types of data and trace their interactions. The initial work of data collection focuses especially on population data. Population records come from formal censuses, from local and religious censuses, and from military records. Trade data are available from port records, from commercial tax registers, and from the records of individual business firms. Money supplies and money flows can be documented from commercial records. Climate data, directly measured by instruments for recent times, can also be indirectly measured by recent geological research. Valuable and accessible health data include accounts of epidemic disease and studies of death rates. Other relevant data include food production, trade, and consumption; social data on births, marriages, deaths, and communities; religious records on individuals and groups; and reports of travelers. Most of the data mentioned here are quantitative, but qualitative data—in text and images—also have value. These qualitative data can be described by time, place, author and other descriptive information; procedures of data mining can then extract patterns from the data.

One massive category of historical data consists of information that has been published but not digitized in searchable form. From the early nineteenth century, many governments published annual records on trade, taxation, and government expenditure. Newspapers published clearings and arrivals of ships. Such documents are available not only for western Europe but for Russia, Japan, Latin American countries, the Ottoman Empire, and for European colonies in all parts of the world. In some cases electronic records have been made of these documents in the form of PDF files, but PDF files do not enable searching on individual characters on each page. Optical Character Recognition (OCR) scanning can be used to digitize such printed files, though the accuracy of OCR is not yet high enough to provide dependable results on quantitative data. It is possible that a procedure involving multiple OCR files of each page, with comparison of the results, will enable automatic digitization of print data files.

In addition to published data, and especially for earlier times, the majority of the historical data we seek is in manuscript form—handwritten documents located in archives. The most convenient such data are located in well-run archives of national governments or great institutions: in Britain, France, the Netherlands, Russia, Japan, China, Turkey,and in the archives of the Catholic Church, especially in Rome. Private archival data exists in manuscript form in the records of businesses or individuals or social institutions. Even more data are held by families: in an extraordinary instance, families in and around the fabled West African city of Timbuktu, in recent years, have donated great quantities of previously hidden historical documents, mostly in Arabic language, which have since been placed in regional archives for preservation, digitization, and historical analysis. Digitization may require manual or mechanical entry to turn manuscript into digital files, and special work to digitize tabular data. To demonstrate that this work is feasible, one may turn to the case of the work done by David Eltis, who read through the Foreign Office archives in London and found the detailed reports of surveillance of slave trade throughout the Atlantic. He transcribed and then digitized the data, then added them to a general database. His results, now digitized and published as part of the Voyages database, provided important new results – that the Atlantic slave trade of the nineteenth century, though now "illegal," continued at the same rate as in the eighteenth century until it halted just after 1850.[6]

Commonly, however, social-science scholars remain reluctant to share or publish the historical datasets that they have constructed with such care. What is needed is a change in the

values and professional practices of scholars. That is, they should agree to submit systematically the data they collect to publicly available repositories where others can check their work and can use the same resources for additional research. Efforts to develop such data-sharing practices have been carried out for generations, some of them with remarkable success, as in the Human Relations Area Files created by anthropologists.[7] For the most part, however, such efforts have not succeeded.

One device that should facilitate the sharing and publication of datasets is peer-reviewing. That is, one begins by recognizing the integrity of each dataset, as prepared by its compiler. Published datasets represent immense efforts in compiling, editing, verifying, and documenting historical data. These data collections need to be recognized as contributions in themselves and should be granted such recognition in published statements by authorities in the field. The *Journal of World-Historical Information*, a newly established academic journal associated with CHIA, is working to ensure that historical datasets are reviewed by qualified scholars who confirm or question the assembly of the data, their value for historical analysis.

How is the CHIA project to acquire datasets? CHIA seeks at once to survey social scientists about the datasets the hold and to identify techniques for encouraging them to submit copies of their datasets to archives. This is the "data-hoover" project, referring to a human analog of the Hoover vacuum cleaner what is intended to draw in all of the available historical data. At the junior level, a "data-hoover" researcher is to survey faculty members at one or more universities, to identify the amount of historical data that are held by various researchers—and, hopefully, to learn of techniques for encouraging researchers to submit copies of their data to a general repository. At senior level, the "data-hoover" is an academic diplomat who meets with editors of journals, to encourage them to require that authors of published articles submit the data to back up the arguments of their articles. Such a requirement has already been established by the *American Economic* Review, and the CLIO-INFRA research group has been circulating a proposed Policy on Data Availability with the same objective.

Standards for Data Documentation: Basic Metadata. One can expect that most data submitted to the CHIA project will come in the form of electronic spread sheets. But in addition to the values in the cells there is need for specific definition of each of the variables and a considerably fuller list of documentation of each dataset and its data. In sum this corresponds to the metadata or data documentation. Researchers in quantitative social science have developed a succession of standards providing the specific organization and extent of data documentation: the Dublin Core and the Data Documentation Initiative (versions 1, 2, and 3) are key examples. These data standards provide information on the sources and compilers of data. In addition, to achieve interconnection of historical datasets around the world, the metadata must provide consistent descriptions of the places and times to which data refer.

Here is some additional detail on the underlying nature of data documentation. One important point to start with is that the metadata must be linked to or even part of the dataset itself. 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. 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 located; *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 (sets of predefined terms—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.

This and other stages of documentation are contributions to the overall **ontology**—the overarching classification system—of the global archive. Various aspects of the ontology are established at different stages of the project. Initially it includes what we here call metadata—the description of values and variables in each data set and the recording of the sources and compilers of data. 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. Later stages of ontology include more comprehensive categorization of types of data, definitions and classification for the linkage and aggregation of datasets, and definitions for the analysis and visualization of data.

Phase 1 Archive and "faceted search." A collaborative team within CHIA has begun working on an architecture that will allow us to consolidate heterogeneous historical data sources in a scalable way. Phase 1 relies on three connected components. First is a set of dataverses—archives within the Dataverse Network system maintained by the Institute for Quantitative Social Science at Harvard University.[8] Second is the development by staff at Harvard and the University of Pittsburgh of a "faceted search" of selected global historical search. The facets of the search are a map, a time line, and a text box in an interface which contains the spatial, temporal, and topical references of the archived datasets. The user then sets search criteria by specifying the spatial, temporal, and topical limits desired, and the faceted search program returns a list of studies or datasets that meet those criteria. The third step is that, once the user has retrieved the selected study data, the data can be explored either spatially through WorldMap or statistically through the applications of Dataverse Network. To complete this work, we require an archive, data, metadata, a search system, and tools for analysis. For the archive, the Dataverse Network system already exists, and can archive a wide variety of studies. For the data, the studies to be prepared for Phase 1 include a limited number of relevant world-historical data types, limited initially to the twentieth century. These include populations at national levels and (for very populous nations) at provincial levels; periodic climate conditions for identified places and times within the same spatial units; silver flows of production and trade; basic statistics on wars of the twentieth century; wheat and rice production and trade figures. For the metadata, a system of data documentation must be developed that enables all the variables to be aggregated and compared on a consistent data, with special attention to their description in space and time. For the search system, the team's developers will conduct programming in Java. For the analytical tools, users may employ the Dataverse Network (for statistical analysis) and the WorldMap application of the Harvard Center for Geographic Analysis (for visualization).

In this way, users will be able to experience the initial level of world-historical data exploration. Phase 1, including the final stage of visualization, is to convey the workings of the

project and encourage more holders of data to contribute copies of their datasets to the growing global archive. This work is expected to have two advantages. On one hand, it is for getting the initial kinks out of the programming for the expanded archive, assembly of data, construction of metadata, programming the search, and conveying selected data for visualization. On the other hand, it is to display a simplified version of world-historical data analysis, so that potential contributors of data and potential users of the resource will understand its potential more clearly.

Crowd-sourcing for data incorporation. In previous efforts to gather large quantities of data, the bottleneck has been the limits on the willingness and ability of researchers to submit their data to a common repository. Important and valuable initiatives such as the Electronic Cultural Atlas Initiative (ECAI, http://www.ecai.org) and ChronoZoom (www.chronozoomproject.org/) have fallen short of their targets in collecting data for lack of a means to open this bottleneck. When explored in detail, the bottleneck in data submission turns out to result from overlapping problems stemming both from the outlook of researchers and the inherent difficulties of completing and conveying historical datasets. The researchers have concerns about the impermanence of online resources, about recognition and citation of their work—they find that the academic world gives little recognition to either the cost or the value of historical datasets. Those who do seek to submit data find that the submission process is complex and that the data, once submitted, are difficult to find.

Crowd-sourcing has developed recently both as a technology and as a philosophy of collective intelligence. As a technology, crowd-sourcing uses online interfaces with public access to gather and exchange information. One major success in use of this technique is the Galaxy Zoo, in which amateur astronomers, working with online images, completed typological descriptions of thousands of galaxies with unexpected rapidity.(http://www.galaxyzoo.org) As a philosophy of collective intelligence, crowd-sourcing works by decentering research and relying on widely dispersed knowledge. The shift to an approach based on collective intelligence involves a major reorganization of the work style of historians and other scholars—more time listening to others and more time explaining things to others. The result, however, may have the benefit of engaging the expert knowledge of historians, now dispersed among individuals, and focusing it on building a collaborative resource, so that the world-historical data resource reflects in some sense the collaboration in human society that was necessary to create and sustain the societies that historians analyze.

The CHIA group has begun a substantial investment of effort in developing a crowd-sourcing application for the collection of world-historical data. Figure 1 shows a general architecture that utilizes collective intelligence to form a global repository of historical data. This architecture efficiently combines methods of crowdsourcing with wrapper/mediator technology. We assume that information providers will submit wrappers that utilize an application programming interface (API) to extract information from their corresponding data sources and to map the information to a standard homogeneous representation. If the data set includes information not covered by a target schema, we extend the schema correspondingly. The data submission system allows providers to register their wrappers as a part of the data-access layer of the global repository. The system will also support a wrapper-generation functionality to facilitate the wrapper development process. The wrappers can be used either to access data remotely or to load/replicate parts of the data at different nodes of the distributed repository (i.e.,

to optimize data analysis, or to consolidate a repository profile to deal with a specific application domain). Reliance on crowd-sourcing brings benefits for data ingest, data documentation, and data-reliability assessment. With regard to the latter, both information providers and consumers will also be able to submit their subjective data-reliability assessments through an online interface. These are *external* reliability assessments which will be combined with *internal* reliability assessment protocols based on analysis of data inconsistencies in the integrated repositories. The data reliability assessment will be a part of in the process of data curation and data fusion.

Continuing effort is required to ensure that the crowdsourcing device conveys an attractive interface to users: it must provide contributors with considerable practical benefits in order to attract them. We are hopeful, however, that we can develop a successful and user-friendly interface. The users we expect to attract are drawn from the many experienced historians, both professional and amateur, who are skilled in the domain knowledge of the many subfields of history and are devoted to collection and study of data.

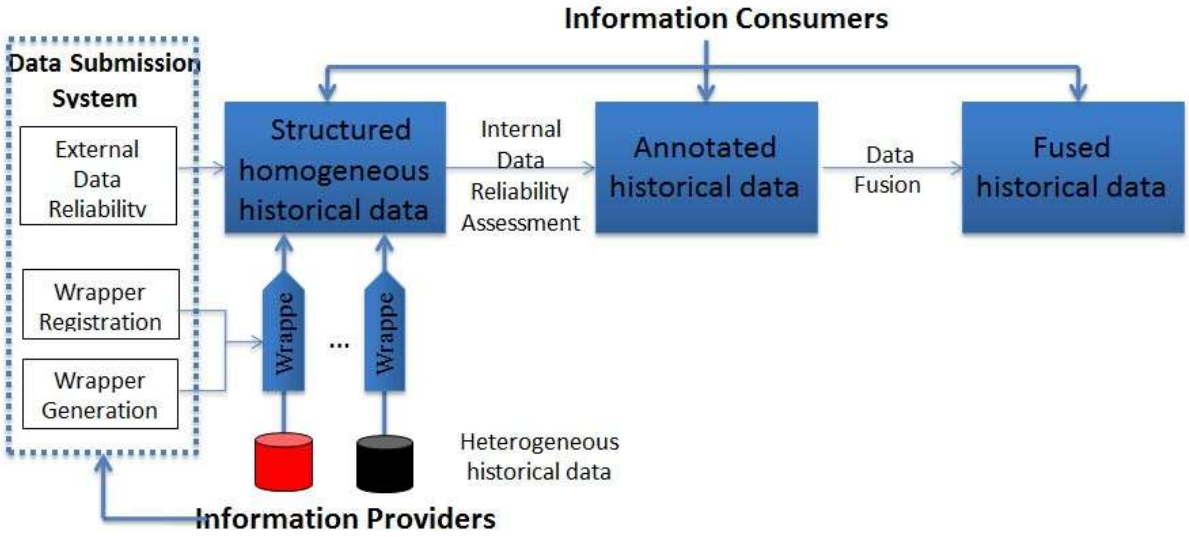

Figure 1: Historical data integration architecture based on crowdsourcing

Preliminary analysis and visualization. To succeed in building the CHIA system, starting with the Phase 1 archive, we must have constant feedback on the many aspects and interconnections of the system as it is constructed. We need to display the data immediately in order for staff of the project to monitor the input, links, and calculations. And we need to provide for output to general users at the earliest possible stage, in order to accelerate the dissemination of global-level data, to draw more users and contributors into working with the system, and to begin contributing to public information and education as soon as possible. As the initial analysis and visualization will indicate, the work of assembling data in Mission #1 points in many directions and requires parallel work by a number of groups. As we turn to address Mission #2, many of these varying tasks will be drawn together to focus on reformulating the data collection into consistent materials that provide a global picture.

5. Mission #2: Creating a Unified Historical Archive

Step-by-step assembly of data, as above, is necessary as a start along the road to creating a world-historical data resource. Only through these successive steps will we encounter the numerous problems that await us and develop the numerous innovations required for work at this increasingly large scale. But if we move only at this incremental rate it will take forever for us to develop a seriously world-historical data resource. This is shown in the slow process of consolidating local and national data into global data.

Historical data have certainly not been rendered compatible with one another, especially for Asia, Africa, and the Americas, and especially before the twentieth 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 re-documented, 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. One of the benefits of using a crowd-sourcing approach to gathering data is that it will cause all contributors of data to develop and work from a common schema of data definition.

Rather than wait for a gradual accretion of localized projects to bring about large-scale analysis, CHIA seeks to conduct a large-scale initiative to speed the transition into global social analysis. The data, technology, and readiness of researchers to collaborate are within reach. This collaborative has the experience and the organizational skills, and now seeks large-scale institutional support. CHIA can decisively address the remaining gap: it can advance global analysis in social sciences by leading in creation of a consolidated system of information and also by resolving many attendant technical and organizational challenges. Whether CHIA remains the sole center for assembling global historical data and analysis or whether it attracts other major groupings and becomes part of a larger collaborative effort, the launching of this center will speed and strengthen the long-overdue process of systematically documenting the human record.

Phase 2 Archive: A Global Data Resource. To repeat, gathering 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.

"Harmonizing" is a term adopted here to refer to several different types of modification of raw data necessary to create a coherent, global dataset. 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 the smaller datasets, comes the creation of incremental metadata to document the aggregation.[9] Once the Phase 2 archive is fully developed, its volume of metadata will likely equal or exceed the volume of data.

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, once 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. 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 fifty 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.

Rich data, poor data: the whole world. The typical approach to collecting historical data has been to find the best existing collections of data and work with them. Sometimes it is the case that these are the most important data as well as being the most available. But otherwise this may not be the case. For instance, the most readily available migration data—focusing on Europeans crossing the North Atlantic— long made it appear that these Europeans were the most migratory of humans. Just over a decade ago, however, a systematic look at Asian data showed that migration from China and India each roughly equaled that from all of Europe, for the period 1850-1940.

In order to develop comprehensive world-historical data, it is necessary to gather data on all the regions, all the populations, and all the time periods. As a result, researchers will need to devote extra effort to regions and time periods for which data are in short supply. That certainly means that CHIA researchers will need to concentrate on regions such as Africa, Southeast Asia, and Central Asia, inviting scholars in those regions to affiliate with CHIA. Data collection in data-poor regions will require intensive application of established techniques and development of new techniques. That is, archival and family-held data, in numerous languages, will need to be located and digitized. Scattered publications will need to be located, relevant data identified, and then digitized. When direct data are not available—as on population, trade, or politics—researchers will have to work to develop indirect estimates. So work with data-poor domains will require advanced techniques for estimation and simulation of missing data. For instance, current work is relying on techniques of simulation to prepare decennial estimates of African population

from 1650 to 1950, including numerous regions within Africa.[10] In ways such as this, the study of data-poor regions can advance the CHIA project overall: development of techniques for estimating missing data will clarify theoretical relationships among social-science variables, and the resulting advances in estimation and cross-disciplinary theory can then by applied to data-rich domains as well.

In expanding work on historical domains where data are in relatively short supply and of relatively poor quality, social scientists can learn from the work of natural scientists, whose search for data has led them to work closely with research institutions around the world. Especially in the fields of astronomy, climatology, and the various fields of biology, researchers work increasingly through collaborations with universities, research institutes, and individual scholars from all over the world. In another way that social scientists building the Phase 2 Archive can learn from the advances of natural sciences, we can expect that the estimation and simulation of historical evidence for data-poor regions will sharpen the distinction between *existing data* and *new data* in history. This can be explained by comparison with Big Data in the natural sciences: while one part of the CHIA project is the collection and digitization of known historical records, another part of it will result in the discovery and creation of immense amounts of *new* historical data. As in geology and astronomy, even though the facts of the past remain unchanged, today's developments of theories and techniques will result in the development of huge amounts of empirical information on the past.

Theory for estimating missing data. Population data will be developed as the basic core of data for the global data resource. Population is included, in one way or another, in all social-science theory and data collections. The establishment of a relatively universal dataset on human population for the past several centuries – with attention to regional breakdown, composition by age and sex, and other available demographic data – will provide the empirical grounding for the global dataset, to which other data will be gradually added. In addition, the work of building the global population dataset will help clarify the handling of population data in various aspects of social science theory. Feedback within these processes of analysis will help to improve the quality of regional and global population data, including changes over time.

Analysis in the social sciences has developed impressively in the last fifty years, with many advances at micro-, macro-, and (increasingly) meso-levels of theory and research. 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 sub-theories in each discipline to each other. For instance, behavioral approaches have become influential in microeconomics, but it is not yet clear whether the behavioral approach has implications for macroeconomic analysis.[Etzioni 2011] Sociological studies at micro and macro studies show considerable divergence; studies in comparative politics focus on national government almost to the exclusion of trends and traditions in local governance.[Calhoun and Duster 2005] Further, general reviews of social science tend to address social sciences by comparison of their parallel silos rather than focus 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, and cultural affairs.

The social sciences have thus responded to globalization more with intensive development of sub-theories 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.[11] All in all, the current state of social science analysis accords low priority to studies that are long-term in their time frame, multi-scale or 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.

At least conceptually and perhaps in practice, the Phase 1 and Phase 2 versions of the CHIA archive will be held within a single resource, hopefully housed by a Supercomputing Center. At the most basic level, this comprehensive archive will hold the datasets submitted by individual and institutional contributors, so that users may consult and cite the data at that level. At the next level of integration, the archive will hold revised datasets that have undergone cleaning, various types of harmonization, and are described in terms of the uniform CHIA system of documentation—including both "basic metadata" for each dataset and "incremental metadata" to account for transformation and harmonization of datasets to make them mutually consistent. At still another level of integration, the archive will hold aggregated datasets, in which the harmonized but localized datasets are assembled into regional and global datasets over short or long periods of time—plus the additional metadata to describe the aggregation process. The volume of data in this comprehensive archive will reach the petabyte level.

6. Mission #3: Analyzing Data at the Worldwide Level

The combined tasks of visualizing and analyzing data at the worldwide level require overall clarity, access to detail, and the identification of unexpected patterns. One of the great successes in global visualization has been the "Gapminder" framework as developed by Hans Rosling. In his lively presentations at annual TED meetings, Rosling was able to display surprising changes in global social and economic development.(http://www.gapminder.org/) In addition to this step forward, however, many more advances are required in the visualization and analysis of global data. For instance, Gapminder is limited to a comparison of national units for not much over one century in which those units have existed. At base, it includes just two dimensions, although clever handling of colors and bubbles enables the inclusion of additional variables. Gapminder does not display multiple levels of aggregation, from local to global, and its display of time is limited to one-year cross-sections. In sum, the display of world-historical data must go beyond cross-sectional national comparisons to include multiple levels, varying spatial aggregations, and exploration of change over varying temporal sequences.

Relevant spatial units include not only modern nations. For proper comparison, the huge national units of Russia, China, the United States, India, and Brazil need to be analyzed in terms of units comparable in size to European or African countries. Yet any attempt to trace such units over four centuries encounters the shifting political and imperial boundaries as well as great changes in population density. The exploration of global patterns over time must take place not only through year-by-year chronology, but also through longer periods (to address cycles in

economy and climate) and shorter periods (to assess seasonal variation). Time must be considered not only in absolute, chronological terms, but also in relative terms, to account for the life cycle of individuals and the creation and maturation of economic and social institutions. The system of visualization must address both aggregated and disaggregated variables: we want to know not only about changes in the level of wheat production and trade over time, but also about changes in the total caloric intake of humans over time.

To illustrate how the CHIA project has begun to address these big issues, here is an example of research, analysis, and visualization that addresses the interconnection of three quite different types of data. The project began with data on health: a project of the Global Health Center at the University of Pittsburgh led to development of the Tycho Project, which digitized disease-surveillance records for the United States from 1892 forward.(http://www.tycho.pitt.edu) In collaborative work with CHIA, these data on disease have been supplemented with data on climate and population, again for the U.S. beginning in 1892. Figure 2 displays disease data for measles, diphtheria, and polio; Figure 3 displays average annual climate data for eight eastern U.S. cities; and Figure 4 combines disease, climate, and population data in a graph which shows the average annual variation in measles incidence, per capita, in cities with varying seasonal temperatures. These results indicate an impact of environmental factors on measles incidence.[12]

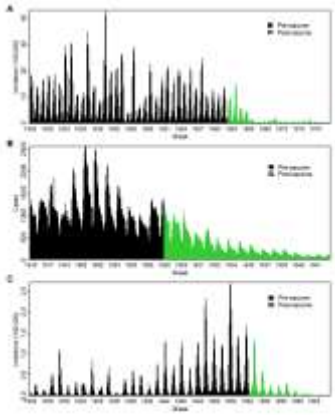

Fig. 2. Disease

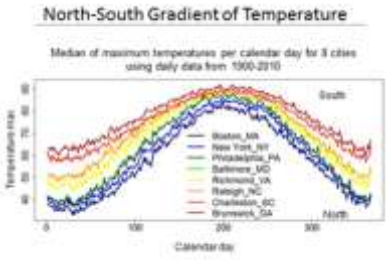

Fig. 3. Climate

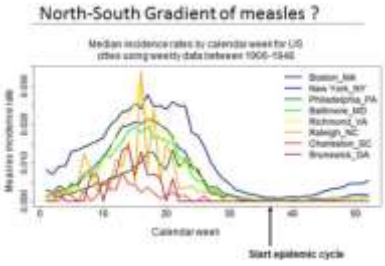

Fig. 4. Disease, Climate, Population.

This example may help to convey both the potential and the complexity of visualization and analysis at the global level. Climatological research has clarified global patterns—the primacy of the power of solar glare on lands of the northern hemisphere in generating the patterns of global climate—but it has also documented the specificity of local climates within the overall global pattern. But visual representations balancing the global and local patterns of climate have yet to become really convincing.

The effort to represent global patterns is made more difficult by the long-established philosophy of scientific analysis. Positivistic philosophy, gradually perfected in the scientific work of the nineteenth century, focused on breaking big problems into small problems, developing solutions to the small problems, and extending those results to the understanding of larger issues. (Darwin's mechanism of natural selection in biological evolution and Alfred Marshall's price theory in economics stand out as examples.) To this day, the tendency to

analyze by breaking big problems into small problems remains preeminent. Thus a search on Google goes from the general to the specific and a search on Wikipedia leads progressively on a tree out to the leaves of the most specific aspects of each topic. The problem of global analysis does not reject this focus on specificity, but it requires that it be balanced with a focus on the global and encompassing nature of systems. If a Google search progressively narrows the investigation, can there be a search system that progressively broadens the scope in an orderly way. Will we have access to search systems that explore issues at multiple scales? As we explore the dynamics of historical change, will we have tools to facilitate shifting the temporal frame and the spatial frame of those dynamics? As we await responses to these questions, we can also look in other directions for possible devices to assist in analysis and visualization.Among the possibilities are the techniques in data-mining that are being developed for large data collections. These computational techniques conduct eclectic explorations of all possible relationships among data within a resource, and report instances where correlations arise. Once the CHIA resource becomes sufficiently large in the volume of its data, it will be time to try out data-mining techniques.

CHIA for researchers and teachers. The approach of CHIA requires that its resources be available to general users—especially researchers and teachers—as soon as possible. As a result, it is clear that CHIA plans for analysis and visualization prepare for three main types of users: CHIA project staff seeking feedback on how to improve aspects of the program; researchers seeking sophisticated results on world-historical patterns; teachers seeking basic results on global patterns that can inform and inspire their students. Since earlier sections have emphasized project development and advanced research, this section emphasizes the ways that CHIA can serve the interests of teachers and students.

Since conceptualizing the world first presents itself as a geographical issue, teachers may well choose to begin by collecting data to compare and link regions of the world. The organization of materials in the CHIA resource will make it possible to alternate among regional, national, and continental views of the world. Next, since the focus of human society is on people, teachers and students may wish to emphasize population: total populations by region, but also breakdowns by age and sex, by birth and death rates, and by migration. Students could then look at weather patterns, past and present, for various areas of the world, and link weather to other factors. If the contents of the CHIA resource can be documented in sufficiently clear terms, it will be possible for teachers and students to ask how government has worked and changed around the world, at levels from the locality to the empire. Students will also be able to trace the occasional breakdown of government in war and conflict. They will be able to compare the changing types of work over time and space and get a sense of how families have changed in their size, their activities, and how they bring up their children. Naturally there will be plenty of evidence on major products in agriculture, handicrafts, and industry—on their quantities, trade, production, and changing design. With time, CHIA will be able to add to its archive such cultural information as details on languages, cultural practices, education, and literacy. Students will have an opportunity to explore as far as they want, and will also get practice in dealing with an overload of information.

Users will be able to comment on data and to correct it as appropriate. Teachers will be able to download units from the archive and enable to select and analyze within these limits. Yet

the problem of having teachers download units from the archive is that of intellectual property: how will it be possible to maintain recognition of the source archives, the historians and the developers whose work underlies the data? How will it be possible to convey to students a sense that the data are not just abstract, true facts but statements resulting from the work and collaboration of a stream of workers?

In parallel with steady improvement and expansion of its system of data collection, documentation, archiving, and analysis, the CHIA group expects also to reach out to other groups of researchers, in various disciplines, who have been developing large-scale data resources. One outstanding such group is CLIO-INFRA, composed of leading scholars in economic and social history, based at the International Institute for Social History in Amsterdam. Like CHIA, CLIO-INFRA is a collaboration of institutions: the University of Tübingen, Groningen Growth and Development Centre,. International Institute of Social History, Utrecht University, and the Data Archiving and Networked Services (The Hague). This group has set up "data hubs" to work on collection of specific sorts of economic-historical data, focusing for instance on human capital, population, and standard of living; the group works closely with other economic historians who have developed historical data on wages and prices.

A second collaborative group works primarily on population data, especially censuses by national governments, and links of population with environmental data. This group—TerraPopulus—based at the Minnesota Population Center, has recently gained support for creating a consolidated digital data resource out of government censuses of nations throughout the world, focusing on the period since 1960.(www.terrapop.org) This project represents a major step forward in global data resources, notably in population data. It relies significantly on the work of the United Nations Office of Population, which has worked diligently to clarify and improve the quality of census returns and population estimates for countries throughout the world, but especially for recently decolonized nations. Most of the data are already in digital form, though the TerraPopulus group still faces the serious task of making the various datasets sufficiently consistent to be able to aggregate them across national lines. Similarly, there is the question of whether to make subnational population data available, especially for large and populous countries. For CHIA, meanwhile, work with groups such as CLIO-INFRA and TerraPopulus data holds forth the promise of adding substantial economic and demographic data to a collaborative archive, which would form the core data of a universal historical data resource.

7. Reconsidering Previous Historical Data Collections

The era since World War II has brought accelerating efforts to collect social science data on national levels and growing efforts in comparison of national data. Looking back from the present, we can conceptualize these earlier programs as campaigns in global studies. The catastrophes of world-wide war, ending with the explosion of atomic devices in two major cities, provoked widespread reflection on global patterns and on the direction of human society overall. As the United Nations Organization formed in San Francisco in late 1945, it adopted a charter including UNESCO, the United Nations Educational, Social, and Cultural Organization—an organization to facilitate international scientific collaboration in all fields, in addition to its missions in education and culture. The first Director-General was appointed in 1946 and, in British-born biologist Julian Huxley, UNESCO found an energetic and visionary leader who laid

out plans for collaboration on a scale that seemed headed toward compensating for the relentless national competition and enmity that had brought such devastating warfare. Huxley articulated the scientific philosophy of UNESCO as "a scientific world humanism, global in extent and evolutionary in background," but within it expected social sciences to work by comparison of cultural or national groups.(Huxley 1946)

The social sciences, working within those cultural boundaries for the following half century, achieved remarkable advances in scope and method. The area-studies movement brought substantial expansion of study on Asia, Africa, and Latin America, encouraging comparative analysis of national and local subunits. Macroeconomics arose as an important new field; social and economic history developed productive quantitative methods; and spreadsheets brought a quantum leap in applications of demography.(Preston et al. 2000) Yet that same postwar era brought massive globalization—in which global literacy and health advanced impressively, while political constellations changed repeatedly. In that context it is remarkable how little the social sciences have done to adopt a new mission of developing coordinated study of human society. One still awaits the big advances in linking information and analysis across disciplines, time, and space. The Center can facilitate the next step, linking the disciplines and linking historical and contemporary studies within them.

Since the formation of UNESCO, the natural sciences have developed institutionalized, well-funded systems to support research that unified analysis from micro-level to universal scales: examples include CERN in physics, the U.S. National Center for Biotechnology, the Long Term Ecological Research Network. These institutions facilitate advances in research, at once responding to and creating the current explosion in scientific information.(Bowker 2008) Where is the equivalent higher-order study in the social sciences? The explosion in social-science information is arguably just as rapid, not only through creation of new data but also through growing access to historical data brought by new techniques. The historical data include advances in health, earth science, and genomics with social-scientific implications.

An overview of the period since World War II reveals three stages or generations of global analysis in social-science. 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. At the initial stage, the formation of UNESCO was part of a postwar boom in enthusiasm that included what may be called the first generation of global studies—a policy-oriented, eclectic, short-term set of concerns led by attention to "modernization" and focusing on political, economic, and sometimes ecological issues. These early days of global studies centered on analysis of international relations in the early Cold War era, when atomic war was a daily threat. This era of disciplinary splintering was complemented by the rise of encompassing trends in social-science analysis. Concerns about population growth and the emergence of ecological movements brought expansion of global study to further disciplines.(Manning 2003) The first generation of research in global studies created a number of major institutions and important databases.[ICPSR, HRAF, OECD, NBER, WB] For economic historical data over the longer run, B. R. Mitchell began his monumental individual statistical compilation with European historical

statistics, and gradually expanded both his geographic and temporal scope.[Mitchell, 2003] These studies were overwhelmingly national though sometimes imperial in scope.

Technology in the first generation of global studies relied initially on manual compilations but changed sharply with the expansion in social statistics and early computers. The postwar quantitative studies in politics, economics, sociology, and history, however, focused on small datasets and on community-level or at most national-level scope. An important if marginal exception was the international literature on the volume and direction of the Atlantic slave trade. In conceptual terms, the first generation of global studies, the era of worldwide decolonization, brought great advances in global scholarly thinking at the practical level, in that the dramatic expansion of area studies scholarship brought a more inclusive approach to social-science analysis. In most cases, however, analysis was nationally and regionally specific and the level of global conceptualization was limited to "the West and the rest." In the era of rapidly expanding development studies, the overwhelming focus of development analysis was on the current postwar era—an implicit assumption that historical trends were of little interest for much of the world.

Global studies gathered steam and entered a second generation in the 1980s, as the term "globalization" came to represent the expansion in global economic and cultural interconnections. Datasets created in this second generation of global studies were more explicitly transnational and transdisciplinary, and they were structured in the more flexible and relational technology developed in those years. This was the era of spreadsheets and relational databases. In some cases, large institutions took on research and data display in this updated framework. Among these institutions were Hitotsubashi University (Tokyo), the United Nations Population Division, and a few others.[HITOTSUBASHI, UNPOP] Much of the original research in global atudies, however, was carried out by individuals and small groups of scholars, whose datasets and analyses therefore risked being neglected rather than integrated into the larger task of global analysis.[GEHN] Angus Maddison developed global estimates of population and gross domestic product that were less specific but more extensive, addressing much of the world for the last millennium.[Maddison website]

Academic programs in global studies began to form in U.S. universities: most continued to operate within the eclectic and short-term approach of the first generation, while a few undertook the research and analysis of the second generation. Most of the U.S. National Resource Centers in International Studies, in contrast, focus primarily on international studies and international business in the short term, without significant emphasis on research or long-term analysis. An outstanding program of the second generation is the Earth Institute at Columbia University, which has undertaken major research in natural and social sciences as well as its policy-oriented support of the United Nations Millennium Program on global inequality, although its historical studies are limited to climate.[EARTH]

With the era of globalization, attention to data on global issues expanded, facilitated by the advances in computer technology, the development of personal computers, and the emergence of practical Global Information Systems (GIS). As a result, the expanded social-science datasets of this era gave great attention to spatial designation but not much attention to temporal analysis. The two pioneering projects were the Great America History Machine

(GAHM), years ahead of its time in mapping U.S. census and electoral data via a windowing interface, and the joint project between the Newberry Library and the University of Wisconsin Automated Cartography Laboratory which mapped changing U.S. county boundaries.[Miller and Modell 1988, Langran 1992] A series of European projects followed shortly after, the largest being the Swedish National Topographic Database.[Goerke 1994] The Alexandria Digital Library emerged as an impressive gazetteer system.[Hill et al. 1999]

Two national historical GIS systems stood out, in Britain and the U.S. The Great Britain Historical GIS achieved high performance by holding all statistical metadata in a set of extensively de-normalized relational tables, tightly linked to statistical data values held in one column of a single very large table rather than in external text files.[Southall et al. 2009] The U.S. National Historical GIS (NHGIS) relied substantially on a large-scale project for defining metadata, the Data Documentation Initiative.[Blank and Rasmussen 2004] Both addressed the issue of setting statistics into these spatially focused datasets. In later years the same issue was pursued in the Colonial Legacies Project (which has since become CLIO World Tables), compiling information from global sources into core statistical datasets, creating pools of data coded to administrative reporting units.[CLIO; see also Eltis 2009] Despite these advances, there still remain problems in spatial representation of historical data.

Other important projects which developed during this second generation of global studies included the CIDOC Conceptual Reference Model for cultural heritage materials[CIDOC]; the system for sampling national census results developed by the Integrated Public Use Microdata Series (IPUMS) for the the U.S. and for several other countries; the Electronic Cultural Atlas Initiative (ECAI); and, for the natural-science community, the World Data System of the International Council for Science (ICSU).[WDS] For these as for the geographically focused datasets, the lack of sufficient links of data values to each other through metadata means that they remain principally as repository projects—they raise interest in the possibility of interactive analysis, but they do not themselves permit such analysis. The second generation of global studies, the era of contemporary globalization, advanced to giving explicit attention to global patterns and led to serious efforts at documenting contemporary global patterns. Nevertheless the approach to documenting global patterns focused primarily on reification of contemporary national units—a poor framework for evaluating long-term change.

Are we now opening a third generation of global studies? Still, researchers are nowhere near to having a set of global historical data against which to test emerging large-scale theory. Global theory, in turn, remains vague for lack of comprehensive data to explore. Rather than wait for a gradual accretion of localized projects to bring about large-scale analysis, the Center is to conduct a large-scale initiative to speed the transition into global social analysis. The data, technology, and readiness of researchers to collaborate are within reach; this collaborative has the experience and the organizational skills, and now seeks large-scale institutional support. Creation of the proposed CHIA can decisively address the remaining gap: it can advance global analysis in social sciences by leading in creation of a consolidated system of information and also by resolving many attendant technical and organizational challenges. Whether CHIA remains the sole center for assembling global historical data and analysis or whether it attracts other major groupings and becomes part of a larger collaborative effort, the launching of this

center will speed and strengthen the long-overdue process of systematically documenting the human record.

8.Priorities for CHIA; Benefits of CHIA

The list of tasks identified above is daunting. The project could easily run out of resources or lose track of the key issues and end up with little to show for an immense expenditure of energy. So it is very important to identify primary tasks in this work which are central and productive. Here is an initial suggestion of top priorities for work in the next few years. Each of these priority steps is at once *necessary*, in that the project must complete it in order to continue, and *sufficient*, in that it will bring benefits not only to CHIA but to social science analysis and cross-disciplinary work in general.

1. **Global repository, global analysis**. CHIA must maintain its objective of designing a big but extendable data repository in order to avoid distractions and make the best choices in development of this global resource. The benefit for social sciences will be a systematic effort to clarify global perspectives.

2. **Crowd-sourcing for data collection and documentation**. This application create a feasible and attractive option for those ready to submit and document data for a comprehensive repository. It will benefit social science in developing collaborative practice.

3. **Peer reviewing of historical datasets.** The peer-reviewing of datasets, notably in the *Journal of World-Historical Information*, will bring scholarly critique of datasets and maintain high standards and best practices. The benefit for social science is to give recognition to the value of creating and publishing historical data.

4. **Cross-disciplinary theory.** Theoretical study will address interconnection at micro and macro levels to show interactions across disciplines and will facilitate estimation of missing data and linking of variables. The benefits for social sciences is the further emphasis on macro-theorization and disciplinary interplay.

5. **Collection of historical data.** Collection of numerous datasets will expand the ability of the CHIA archive to identify global historical patterns. For social science, the development of global historical data will lead to new analytical insights.

These five top priorities apply to the initial several years of the CHIA project. For each of these objectives, we already have insights and resources that point us in the direction of achieving the goal. Yet there remain major problems that we have only begun to address. One of these is the high cost of data collection and digitization of historical data. The initial CHIA repository can hope to provide a valuable sample of world-historical data—enough to launch major new initiatives of research and interpretation. But to confirm the global hypotheses that will arise from initial investigation, it will be necessary to continue at length in research, digitization, harmonization, and aggregation of data on the human past. The available funds for historical research and indeed for training of skilled historians are not now available. One may

hope that the initial results of world-historical analysis will show the value of such work, and will lead to additional support for its continuation. A second unsolved problem is that of intellectual property. The complex transformation, aggregation, and distribution of data through the CHIA archive will increase the difficulty of retaining and conveying the identity of the sources, compilers, and developers of the data at every level; yet the maintenance of a clear record of each step in data manipulation is necessary to preserve data integrity and to make necessary updates. This dilemma will require concentrated attention at every stage of work.

While the task of developing world-historical data remains daunting, we can already see glimmers of the patterns that may be documented and explained in the years to come. As usual, one turns to studies of climate to begin this list of examples. The phenomenon known as El Niño Southern Oscillation (ENSO), which brings alternating periods of wet and dry weather to opposite shores of the southern Pacific Ocean, began to be studied seriously only in the 1970s. By now, however, El Niño episodes have been documented in timing and intensity for the past several thousand years. In studies of human population, we have detailed worldwide documentation only for the last half century; documentation for earlier times is restricted to a few highly literate regions. Yet with the techniques now available, it will be possible to estimate rates of birth, death, and migration around the world to give new and improved estimates of past population and population change, so that we can hope for valuable estimates of world population during the past four centuries with just a few years of additional research.[13] And, to conclude with an example that is at once more specific and of wide interest, we have the hope of preparing detailed estimates of the flows of silver from mines (especially in Mexico and Bolivia) to purchasers and users worldwide for the past four centuries. Silver became the principal form of liquid wealth around the world from about 1600: by tracing the quantities of its production and its flows around world regions, we will gain new insights into money supplies, commercial transactions, financial fluctuations, and the economic linkages among regional economies.

BIBLIOGRAPHY

Organizations

CHIA. Center for Historical Information and Analysis (http://chia.pitt.com).
ChronoZoom (http://www.chronozoom.org/).
CIDOC. International Committee for Documentation (http://cidoc.mediahost.org/).
CLIO. CLIO World Tables (http://peanut.bu.edu).
CLIO-INFRA. Clio Infrastructure, supported by the European Commission initiative Digital ResearchInfrastructure for the Arts and Humanities (http://www.clio-infra.eu).
DVN. The Dataverse Network (http://thedata.org).
EARTH. Earth Institute at Columbia University (http://www.earth.columbia.edu/).
ECAI. Electronic Cultural Atlas Initiative (http://www.ecai.org).
FOLDIT. Fold It: Solve Problems for Science (http://fold.it/portal/).
GALAXY ZOO. Galaxy Zoo (http://www.galaxyzoo.org/).
GEHN. Global Economic History Network(http://www.lse.ac.uk/collections/economicHistory/GEHN.htm).
GBHGIS. Great Britain Historical GIS (http://www.port.ac.uk/research/gbhgis/). See also *A Vision ofBritain through Time* (http:// www.visionofbritain.org.uk).
HITOTSUBASHI. Hitotsubashi University (Tokyo) (http://www.hit-u.ac.jp/laboratories/index-e.html).
HRAF. Human Relations Area Files (http://www.yale.edu/hraf/).

ICPSR. Interuniversity Consortium on Political and Social Research (http://www.icpsr.umich.edu/); LabourRelations. Global Collaboratory on the History of Labour Relations (https://collab.iisg.nl/web/labourrelations)
MPC. Minnesota Population Center (http://www.pop.umn.edu/).
NBER. U.S. National Bureau of Economic Research (http://www.nber.org/data/).
OECD. Organization of Economic Cooperation and Development (http://www.oecd.org/statsportal/0,2639,en_2825_293564_1_1_1_1_1,00.html).
Science of Team Science (http://www.scienceofteamscience.org/).
Taxacom Listserv (http://taxacom.markmail.org/).
Terra Populus. Terra Populus : Integrated Data on Population and Environment (http://www.terrapop.org/)
Transcribe Bentham (http://*www.ucl.ac.uk/transcribe-bentham/*
TYCHO. Project supported by the Bill and Melinda Gates Foundation (http://www.tycho.pitt.edu/).
UNPOP. United Nations Population Division (http://www.un.org/esa/population/unpop.htm).
WB. World Bank (http://data.worldbank.org/).
WDS. World Data System of the International Commission of Science (http://www.icsu-wds.org/).
WorldMap. WorldMap (http://worldmap.harvard.edu).

Books and Articles

Bain, D. J. and G. S. Brush. 2008. "Gradients, Property Templates, and Land Use Change." *Professional Geographer* 60 (2): 224-237. Bowker, Geoffrey. 2008. *Memory Practices in the Sciences (Inside Technology)*. Cambridge: MIT Press.
Calhoun, Craig, and Troy Duster. 2005. "Sociology's Visions and Divisions." *Chronicle of Higher Education* 51 (49): B7.
Chase-Dunn, Christopher, and Salvatore Babones, eds. 2006. *Global Social Change: Historical and Comparative Perspectives.* Baltimore: Johns Hopkins University Press.
Comte, Auguste. 1975. *Cours de philosophie positive.* 2 vols. Paris: Hermann.
Eltis, David, Halbert, M., et al. 2009. "Voyages: The Trans-Atlantic Slave Trade Database" (http://www.slavevoyages.org/).
Etzioni, A. 2011. "Behavioural economics: Next steps," *Journal of Consumer Policy*, 34(3), 277-287.
Gerring, John, Philip Bond, William Barndt, Carola Moreno. 2005. "Democracy and Growth: A Historical Perspective." *World Politics* 57: 323-64.
Giddens, Anthony. 2003. *Runaway World: How Globalization in Reshaping our Lives*, 2nd ed. New York: Routledge.
Goerke M, ed. 1994. *Coordinates for Historical Maps*. Gottingen: Max-Planck-Institut fur Geschichte.
Hill, Linda L., James Frew, and Qi Zheng. 1999. "Geographic Names: The Implementation of a Gazetteer in a Georeferenced Digital Library," *D-Lib Magazine*, 5(1).
Langran, G. 1992. *Time in Geographic Information Systems*. London:
Honaker, James, and Gary King. 2010. "What to do About Missing Values in Time Series Cross-Section Data." *American Journal of Political Science* Vol. 54, No. 2 (April, 2010): Pp. 561-581.
Huxley, Julian. 1946. *UNESCO, its purpose and its philosophy*. London, UNESCO.
Maddison, Angus, website (http://www.ggdc.net/maddison/).
Manning, Patrick. 2003. *Navigating World History: Historians Create a Global Past*. New York: Palgrave.
Manning, Patrick. 2010. "African Population: Projections, 1851-1961." In *The Demographics of Empire: The Colonial Order and the Creation of Knowledge*, eds. Karl Ittmann, Dennis

D. Cordell, and Gregory Maddox, (Athens, OH: Ohio University Press, 2010), pp. 245-275.

Marx, Karl. 1967. *Capital: A Critique of Political Economy*. 3 vols. Edited by Frederick Engels. New York: International Publishers.

Miller, D. and Modell J. 1988. "Teaching United States history with the Great American History Machine," in *Historical Methods*, 21:121-134.

Mitchell, B. R. 2003. *International Historical Statistics: Africa, Asia and Oceania, 1750-2000* (Basingstoke and New York: Palgrave Macmillan, 4th edn.

O'Brien, Patrick K. 2006. "Historiographical traditions and modern imperatives for the restoration of global history." *Journal of Global History* 1: 3-39. Pomeranz, Kenneth. 2000. *The Great Divergence: China, Europe, and the making of the modern world economy*. Princeton: Princeton University Press.

Preston, Samuel , Patrick Heuveline, and Michel Guillot. 2000. *Demography: Measuring and Modeling Population Processes*. Hoboken, NJ: Wiley-Blackwell.

Reinhart, Carmen M., and Kenneth Rogoff. 2009. *This Time Is Different: Eight Centuries of Financial Folly* (Princeton: Princeton University Press).

Santayana, George. 1905-1906. *The Life of Reason, or, The phases of human progress*. New York: Charles Scribner's Sons.

Southall, H. 2011. 'Rebuilding the Great Britain Historical GIS, Part 1: Building an Indefinitely Scalable Statistical Database' in *Historical Methods: A Journal of Quantitative and Interdisciplinary History, 44:3,* pp149-159.

Southall, Humphrey, Von Luenen, Alexander and Aucott, Paula. 2009. "On the organisation of geographical knowledge: data models for gazetteers and historical GIS." In: *E-Science Workshops*, 2009 5th IEEE International Conference on. IEEE, Oxford, pp. 162-166.

Wallerstein, Immanuel. 2001. *Unthinking social science : the limits of nineteenth-century paradigms*. Philadelphia: Temple University Press.

Zimmer S.M., Burke D.S. 2009. "Historical perspective—Emergence of influenza A(H1N1) viruses." *N Engl J Med*. July 16; 361(3): 279-85.

NOTES

[1] Even for years in which PDF files existed for data on U.S. disease surveillance, project Tycho had to to re-enter all the data by hand in order to create a searchable file.

[2] O'Brien 2006; Reinhart and Rogoff 2009; Gerring et al. 2005; Giddens 2003, Pomeranz 2000, Chase-Dunn and Babones 2006, Zimmer and Burke 2009; Bain et al. 2008.

[3] U.S. National Science Foundation awards to CHIA for Building Capacity and Community(BCC, September 2012): awards 1244282, 1244667, 1244672, 1244693, 1244796.

[4] Raw data files incorporated into the Phase 1 archive will remain there, regardless of whether they are incorporated into Phase 2, and will remain attributed to their contributor.

[5] For United Nations data, see http://data.un.org/Explorer.aspx; for World Bank data, see http://databank.worldbank.org/data/home.aspx

[6] Eltis's research, first published in 1980, is now incorporated into Eltis, et al., "Voyages."

[7] Human Relations Area Files (http://www.yale.edu/hraf/), established at Yale University in 1949.

[8] Institute for Quantitative Social Science (http://iq.harvard.edu); Dataverse Network (http://thedata.org).

[9] Within the general category of "harmonization," "cleaning" refers to correcting errors in individual data values, "fusion" refers to consolidating overlapping or duplicate data, "transformation" refers to standardization of weights and measures as well as language translation, and "aggregation" refers to consolidating files to develop data files

into files of larger scale in space and time. Work of constructing the archive will further clarify the details and overlaps of these processes.

[10] Patrick Manning and Scott Nickleach, *African Population, 1650-1950: The Eras of Enslavement and Colonial Rule* (in process).

[11] Sociology and anthropology have little close contact, development economics has little to do with economic history, and the links of demography and health are only now coming under serious study.

[12] Other research from the same group indicates that demographic patterns influence measles incidence. These cross-disciplinary results open new frontiers in medical-historical research.

[13] CHIA and CLIO-INFRA are planning a joint conference in Amsterdam in 2013 to take a next step in this research.