Journal of World-Historical Information

JWHI

EDITORIAL

Notes Toward a World-Historical Data Resource

World-Historical Information: Facing Social Change

Human society is undergoing transformations, both positive and negative. But we are short on information about those transformations. We almost always lack information on how local shocks and transformations add up at the global level. Whether it is war, drought, massacre, economic crisis, or aging of the human population, we rarely have information on the global implications of local events. Even when we can collect information on changes in the world of today, it is difficult to know whether they are repetitions or deviations from past changes.

JWHI is devoted to studies of world-historical information. For the editorial staff, we build the publication strategy of the journal on assumptions at two levels: (1) in general, we assume the need for world-historical data, both for academic advance and to meet social needs; (2) specifically, we assume the existence of a current crisis in social inequality, a crisis that needs to be documented and analyzed at both contemporary and historical levels. Based on these assumptions, we support and publicize research in Big Data – more specifically, we encourage development of "Comprehensive, Big-Historical Data on the Human System."

In our opinion, the general and specific needs for world-historical analysis interact with each other. We need to develop an overall system for documentation of historical patterns of human society and need to focus specifically on patterns of inequality. In addition, we believe that the crisis in inequality interacts with the contemporary crisis in

Volume 2-3, No. 2 (2014-15) | ISSN 2169-0812 (online)
DOI 10.5195/jwhi.2015.32 | http://jwhi.pitt.edu

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

For world-historical data, researchers are so far moving only incrementally toward the study of large-scale social issues. Since there is a lack of social interest in the subject, the researchers are unlikely to gain much research funding. Further, even if future research develops significant findings on inequality, it is likely that there will be strong political and social opposition to implementing the policy implications, whatever they might be. Nevertheless, it is difficult to refute the need for global data relevant to the trajectory of human society. The real problem is with the difficulty of moving such research ahead.

Big Projects in Social Science and Natural Science

A look at the past shows, perhaps surprisingly, that large-scale research projects in social science and natural science have brought substantial advances in knowledge and substantial changes in the world of today. In social science, national censuses, beginning in the nineteenth century, brought immense expansion in knowledge through government efforts to compile population data. The creation of national income analysis in the early twentieth century became the basis of national economic policies. In the natural sciences, large-scale research projects created atomic science (including bombs), European collaboration created the atom-splitting CERN institute near Geneva, the U.S. and USSR built independent programs for space exploration. Worldwide collaboration brought the understanding of continental drift in the 1960s and the successful modeling of climate change over short and long time periods in the 1990s. Another international program achieved the complete documentation of the genome of humans and other species. From this perspective, the development of world-historical data on human society is the next step in such large projects to expand knowledge. But, as with the other projects, this step will not be taken automatically—it will come only if dedicated researchers can gain substantial funding to work collaboratively on a world-historical data resource over the course of years.

World-historical consciousness and world-historical interpretation have been advancing rapidly, but world-historical research is advancing much more slowly. The analysts of world history, to the degree that they work with datasets, use comparative data on nations or similar regions, with little reference to patterns at the global level. Admirable work has been completed on small scales in collecting data on economics, social structure, and human-natural interaction, but there is yet no large-scale project to document transformation of human society over the past several centuries.

What follows is a brief description of the context, the problems, and the resources for creating world-historical data. It addresses the current projects contributing to creating a worldwide data resource, the work in previous times on marshaling data on large social scales, the nature of world-historical data in its scales and its characteristics, and recommendations for steps to take to advance data preparation at the scale of worldwide human society.

Current Contributors and Preceding Efforts

There exist several social-science research groups that have conducted research on large-scale social phenomena. They focus on a wide range of topics of importance in world history. Yet they share a great deal in methodology and philosophy because of the inherent problems in gathering historical data and attempting to link local data and aggregate them to a global level. The research groups now working on large-scale social and historical analysis may be divided into two groups, especially according to the time periods under study. The differences in time period also correspond to differences in the nature of documentation, the nature of society, and the questions under study.[1]

The period from 1500 to the present. Most of the current projects address recent centuries, going back as far as the year 1500. These projects address the world as a whole or large parts of it. The CHIA project (Collaborative for Historical Information and Analysis – chia.pitt.edu), with which this journal is associated, is perhaps the one project to take as its explicit objective the development of a global dataset. Since 2012 CHIA, a collaboration of researchers at several universities, has been working to collect, document, and archive data of relevant to a worldwide data resource, with particular focus on creating an infrastructure that can link and aggregate discrete social-science datasets. A larger group, the International Institute of Social History in Amsterdam, houses a series of worldwide historical research projects: the CLIO-INFRA project on economic history (https://www.clio-infra.eu/), the Global Collaboratory for the History of Labour Relations (https://collab.iisg.nl/web/labourrelations), and a broad new initiative in global digital humanities, CLARIAH (www.clariah.nl). The CLIO-INFRA project is closely associated with the Netherlands-based Maddison Project (www.ggdc.net/maddison/), which carries on the economic-historical data gathering of the late Angus Maddison. For census data, the largest collection of historical data is at the Minnesota Population Center (www.pop.mn.edu) whose data collections include IPUMS, for which U.S. and international versions provide users with samples of individual data drawn from full-scale censuses. Another established unit of data definition and data collection is the Electronic Cultural Atlas Initiative (www.ecai.org). In historical political analysis, the Correlates of War project (www.correlatesofwar.org/) has assembled numerous databases on topics related to politics. The large-scale dataset that has gained the widest popular usage by general audiences is the Slave Voyages website (www.slavevoyages.org/), which enables users to explore and download data on the transatlantic slave trade. Major social-science archives include the Interuniversity Consortium for Political and Social Research (www.icpsr.umich.edu/) at the University of Michigan and the Harvard Dataverse (www.dataverse.harvard.edu). For periods from 1950 forward, major international organizations have developed numerous versions of national-level statistics on population and economic activity: the organizations include the World Bank, the International Monetary Fund, UNESCO, and OECD.

The period before 1500. Several experienced groups have been collecting data and conducting analysis, focusing especially on the world before 1500 and especially in the context of major civilizations. The Institute for Research on World Systems (IROWS: www.irows.ucr.edu/) focuses especially on documenting urban life and the extent of world-systems. The Seshat project (https://evolution-institute.org/project/seshat/) focuses on an evolutionary social perspective, with a cultural approach to early civilizations. For coherent mapping resources for the world before 1500, the Pelagios project (http://pelagios-project.blogspot.com/) has substantial funding from the Mellon foundation. Other projects for collection of historical data on early times include a project, based at

University of British Columbia on religion in early times and a study at the University of Pittsburgh on the distribution and evolution of language groups worldwide for the period from 50,000 to 2000 years ago.

Large-scale Social Research Efforts before 2000. The current need for world-historical data is certainly at a larger scale than previous campaigns for creating social statistics. Yet a review of earlier campaigns for creation of social statistics is surely relevant, if only to point out potential pitfalls. For the earlier work that carved the path toward the present campaign, was it largely similar or significantly different? Gregory King, in the 1690s, prepared well-informed estimates of population, trade, and production for England over the course of the seventeenth century. From the late eighteenth century, regional and national governments in Sweden, the United States, and the United Kingdom began enumeration of their populations in censuses. An outstanding individual census leader was Carroll D. Wright, who directed detailed censuses of Massachusetts in 1865, 1875, and 1885, and then went on to become Director of the Census for the United States. In the early twentieth-century era of industrialization, war, and depression, advances in documentation of national economies were led by Wassily Leontieff and Simon Kuznets (in the U.S.) and Colin Clark (in the UK). The researchers created the concept of Gross Domestic Product, defined it for current years, launched calculation of historical GDP for North Atlantic nations—and facilitated Keynesian counter-cyclical economic policies. After World War II, GDP figures were developed for existing and newly recognized nations worldwide. In another remarkable instance of individual energy, B. R. Mitchell published numerous series of historical statistics on national population, trade, and GDP—first for Europe and then gradually worldwide. Mitchell was succeeded by Angus Maddison, who proposed historical estimates of economic and demographic data for most regions. Increasingly, after 1950, the United Nations and UNESCO published systematic statistics on population and economy for each of their national members.

Need, Objectives, and Challenges

Need. Why is it not feasible simply to calculate worldwide data by taking the sum of national data for all the countries? Most directly because national data do not now go back very far for most nations; also because records and recording units are kept differently in different nations. More generally, understanding of global society in present and past times requires explicit attention to multiple scales of aggregation (not just the national level) and also to the many exchanges and migrations across national frontiers.

What is needed is a repository, supported by an extensive infrastructure and staff, with powerful systems of analysis and visualization. It must be open to all, with facilities designed for researchers at all levels and for students and the general public. Its technical infrastructure will include a distributed archive and computation, including high-speed computing. Its organizational infrastructure will include relevant types of crowd-sourcing enabling qualified users to participate in data input, documentation, and editing. More broadly, it will function through collaboration among groups building and using the resource. The resource will provide data by topic, link data from topic to topic, and test research hypotheses on historical and social change at all levels.

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

Bottlenecks in scientific work of social science and information science. Certain recurring areas of difficulty continue to inhibit the work of creating a world-historical data resource. First, conceptually, is the unfamiliarity of global frameworks: researchers have grown up thinking of the world as a collection of independent units rather than as a broad and interactive system. As a result, elementary errors in designing global research are not uncommon. Second, social-science datasets are not easily linked or combined. The ICPSR and Dataverse repositories hold many thousands of social-science datasets (in Excel, SPSS, or SAS format). But there is yet no easy way to search data across these discrete datasets, nor to combine them into a larger-scale dataset, thus inhibiting global analysis. (Text-based datasets, in contrast, are much more easily searched.) A third bottleneck is the difficulty of documenting social-science data in systematic and comprehensive terms. To be linked and aggregated effectively, each constituent dataset needs a full set of metadata (including dataset name, source and rights, variables, scope [topic, space, time, scale], and previous transformations). To enable comprehensive documentation, comprehensive ontologies must be maintained on scope and on data transformations. A fourth bottleneck is missing data: at any stage, social-science datasets commonly have large proportions of missing data. In some cases the missing data exist but are not accessible; in other cases they do not exist but can be estimated or simulated. Fifth, the practice of scientific collaboration needs to advance. For instance, how can we arrange for regular meetings – perhaps biennial – most of the groups mentioned above, to help advance their work?

Bottlenecks in social application of scientific research. In the interface between social-science researchers and the public, the most immediate bottleneck is the difficulty or arranging research funding—at local levels and global levels. For research of global significance, one must think of global organizations (such as the UN, UNESCO, or World Bank) and of international collaborations, private or public. Yet here it is worth noting that, of the large-scale research efforts noted above, whether in natural science or social science, the U.S. government has been the main supporter of really big research. The second bottleneck to large-scale historical research is social opposition. In recent years, with the destruction of priceless historical monuments, we have been reminded that some social interests wish to destroy the past, rather than preserve it. At a less obviously destructive level, many social interests are fearful of any change to the social order (or of potential loss of their privileged positions), and therefore oppose research and analysis that appears oriented toward change. The third bottleneck is human reluctance to join in large-scale social collaboration. So far, it is only in war that whole societies can be brought to work as one.

Start-up problems. World-historical research, since it is at an early stage, encounters many difficulties. Available data, in general, are fragmented, with short time frame, inconsistent dimensions. Research groups and their theories and data are fragmented too, by discipline, language, and world region. Since research groups have been limited in funds, they have been insufficient in their application of application of available technology. Past projects have exhibited weakness in articulating global concerns, absence of investment in global integration of knowledge, lack of practical concern for the dimension of time and the influence of long-term factors, and insufficient attention to ontology the systematic classification of data. Similar early-stage difficulties, however, have occurred on all big research projects. Where big projects succeed, it is because an interest in the issue drives work ahead, leading to critique and replacement of the initially inadequate techniques and analyses.

Characteristics of Big Historical Data

"Comprehensive Big-Historical Data on the Human System." This term, while perhaps too long, conveys the essential nature of a world-historical data resource and its contents. Here we take it apart, word by word.

"Data." The form of data to be included in the world-historical resource will begin with statistical data—that is, structured tabular data—but will expand by stages to include text, images, and data in other media. The initial focus on tabular data stems from the availability and relevance of tabular data. But once global analysis of tabular data has advanced beyond the initial stages, it will make sense to complement it with text and other sorts of data, to gain a comprehensive view of the human system over time.

"Historical." The term "historical data" has multiple meanings. It ranges from handwritten Medieval documents to the terabytes of digital information that will ultimately by held in linked, global repositories. Here are some of the categories:

- Data in print and manuscript data, in known and accessible repositories.
- Data in repositories that are not accessible
- Digitized data
- Digitized data that are documented and/or linked
- Data created by theoretical and analytical projections and simulations
- Conceptual data – all the data one could imagine.

Only small proportions of existing world-historical data are now accessible for analysis by researchers, and only small portions of the available data have been digitized. Only when the heterogeneous data from wide-ranging sources have been documented (as to their source, dimensions, and transformations) can they be linked into larger units. After exhaustive collection of existing data, there will still be numerous missing data. Therefore, the work of data collection is conceptual as well as empirical: simulation and estimation of missing data, based on known or theorized relationships, can fill in some of the gaps.

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

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

"Comprehensive." A world-historical data resource must be comprehensive. By "comprehensive world-historical data," I mean well-documented and linked data, conveying information on interconnected topics of human existence during the last several centuries. The topics include information on the size and characteristics of population; evidence on the activities and structures of economy, society, and politics; information on technology and other sorts of knowledge; documentation on health and disease; and data on the environment and human-natural interactions, especially in climate. But "comprehensiveness" also means "selectivity." Because the range of a world-historical data resource must be so multi-dimensional, only so much data can be selected for each segment. For these reasons of scale, it would be best to begin with a prototype—a relatively small-scale version of the data resource that includes all the main functions and dimensions that the full-scale version will have, in order to confirm that the full data resource will produce valuable results.

Social crises: Inequality and Climate Change

The current crisis in climate change is the most obvious instance of a challenge that is global. In fact it is two great challenges—finding whether policy can affect ongoing climate change and learning whether policy will or can be implemented. On the technical side, the first, unmistakable evidence of contemporary anthropogenic warming came in the 1970s with the work of Syukoro Manabe and Kirk Bryan, measuring and modeling rising levels of atmospheric carbon dioxide. From that time, we can trace the scientific work of modeling and documenting climate change—at levels from local to global—so that we now have a remarkably detailed outline of climate change in our own time, stretching back to the distant past. This work is a great achievement in knowledge about the earth's environment and its interaction with human society.

On the social side, human transformations are affected on one hand by policy makers—and on the other by unregulated human activity. Knowledge of the nature and pace of these transformations could improve the quality of life and increase the chance of avoiding major disasters. Socio-economic inequality, for instance, appears to be on the rise, threatening the stability of society and contradicting values of social solidarity. What has been the global past of inequality and what is its trajectory? The common focus on national units inhibits understanding of global interactions and aggregate patterns. Threats of war and other violence, concern about financial and economic crisis, the desire for improved social welfare—these and other factors require data and analysis of world-historical data.

Proposed Steps to Developing World-Historical Data

To conclude, here are some recommended priorities for research in developing world-historical data. Because it ranges so widely, each group should avoid becoming overly routinized within an established system. A regular search for new possibilities, in technology, in theory (both in social science and in information science) and in collaborative techniques, will turn up valuable innovations. Of particular importance is the development of technology for documenting and aggregating social data—breakthroughs in this area would be highly beneficial. A second specific focus should be on developing improved techniques for estimating missing data, and then for verifying their value. A third specific focus should be on crowd sourcing—developing techniques for locating qualified individuals who can not only submit data but conduct work in documentation, editing, and analysis. Fourth, consistent attention should go to locating specific topics, within this framework, which can be advanced with relative speed and efficiently, and which have relatively significant social implications. The CHIA project has chosen to emphasize social inequality in this regard; other topics will surely prove to be of great value. Setting these priorities correctly is important, as world-historical data development requires some clear successes in order to draw attention to its potential and to show its feasibility.

More broadly, an academic discussion needs to be developed on the significance of investment in world-historical information: perhaps competitions could be set up to encourage advance in world-historical knowledge. While the research will clearly be distributed rather than centralized, it would be best for the world-historical data resource to have clear institutional support: UNESCO, in principle, is best placed to play this role. Finally, work on world-historical data is oriented to the public interest, so it must emphasize consistent display of results and procedures to the public, and must facilitate public commentary on the results and the priorities of the analysis.

Patrick Manning

NOTES

[1] In addition to the two time periods identified here, research is also progressing on the human system for a longer and earlier period, from earliest human times to about 5000 years ago, though it generally relies on different disciplines and methods.

This journal is published by the University Library System of the University of Pittsburgh as part of its D-Scribe Digital Publishing Program and is cosponsored by the University of Pittsburgh Press.