Journal of World-Historical Information
JWHI

Vladimir Zadorozhny, Patrick Manning, Daniel J. Bain, and Ruth Mostern

Collaborative for Historical Information and Analysis:

Vision and work plan

Abstract

This article conveys the vision of a world-historical dataset, constructed in order to provide data on human social affairs at the global level over the past several centuries. The construction of this dataset will allow the routine application of tools developed for analyzing "Big Data" to global, historical analysis. The work is conducted by the Collaborative for Historical Information and Analysis (CHIA). This association of groups at universities and research institutes in the U.S. and Europe includes five groups funded by the National Science Foundation for work to construct infrastructure for collecting and archiving data on a global level. The article identifies the elements of infrastructure-building, shows how they are connected, and sets the project in the context of previous and current efforts to build large-scale historical datasets. The project is developing a crowd-sourcing application for ingesting and documenting data, a broad and flexible archive, and a "data hoover" process to locate and gather historical datasets for inclusion. In addition, the article identifies four types of data and analytical questions to be explored through this data resource, addressing development, governance, social structure, and the interaction of social and natural variables.

Volume 1, No. 1 (2013) | ISSN 2169-0812 (online)
DOI 10.5195/jwhi.2013.2 | http://jwhi.pitt.edu

Vladimir Zadorozhny, Patrick Manning, Daniel Bain, and Ruth Mostern[1]

Collaborative for Historical Information and Analysis:

Vision and work plan

> "Those who do not remember the past are condemned to repeat it."
>
> "A man's feet should be planted in his country, but his eyes should survey the world."
>
> George Santayana

Introduction

The problems in global society—in governance, socio-economic change, health, and human interaction with the environment—span regions and disciplines. The social sciences, though sophisticated in analysis of contemporary societies, continue to generally work within regional and disciplinary boundaries. Historical social science, focused at national and subnational levels, has scarcely addressed global issues. Meanwhile, contemporary globalization and concerns about future global change naturally raise questions about past patterns of global change. For example, how did population, economy, governance, and social inequality interact with each other and with climate and disease? Has social inequality fluctuated in global cycles over the centuries? Have crises in governance (at macro and micro levels) responded to shifts in inequality or in climate?

Our NSF-funded Collaborative for Historical Information and Analysis (CHIA) aims to collect, document, store, and analyze historical data to permit cross-disciplinary analysis of human society over time.[2] Our aim is not simply to archive large quantities of data but to define and link them into a single overarching set of data facilitating study of historical interactions. We require a coherent metadata framework to link data to their sources and each other. The overall topic is immense, but we believe we have found an orderly and productive way to work on it. Creating this mass of data and metadata—as we incorporate, integrate, and aggregate information —requires a strong ontological base and a crowdsourcing procedure to entice many contributors to participate. The task of large-scale data utilization can only be resolved via collaborative efforts within a large network of researchers. Advances in social theory and information technology bring a substantial opportunity to develop methods for data collection and data analysis at a global scale and in the substantially long run.

Currently, we are constructing organizational and technical infrastructure by emphasizing the interplay between the Research Collaborative and Headquarters. The Research Collaborative will link collection of data on population, climate, and other topics with a crowdsourcing tool to demonstrate a continuously growing collection of historical data and metadata. The Headquarters will assemble knowledge on repository design to deliver a repository sufficient to house the incoming data and permit global and interactive analysis. After that, CHIA will expand its collection and processing of historical data, broaden its community of social-science researchers, analyze global patterns in historical change, and share its resources with researchers, policy-makers, teachers and students.

This project moves historical analysis into the realm of Big Data by taking historical data resources into the terabyte range. The effort will stimulate efficient research collaboration that will enable systematic large-scale consolidation of diverse historical data sources. Such data, once collected and integrated, will make it possible to test hypotheses about long-term and short-term social change at the global level. Global historical data on population and climate will launch expansion of the evidence base in social sciences. The repository and analytical system, once implemented, will be able to address a wider set of questions. Disciplinary analysis will advance, for instance by linking health to demography and by incorporating climate and health factors into economic studies. Disciplinary theory will advance through interplay of fields, so that a global network of social-science researchers will emerge. For instance, such a network is emerging to link world historians of every continent (NOGWHISTO).

Related Works

The Collaborative will build on earlier efforts. As UNESCO formed to coordinate worldwide scientific research, the founding Director-General, Julian 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.

Meanwhile, 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 National Center for Biotechnology, and 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.

The barriers that have restrained unification of social-science analysis lie at once within the nature of the data and in the inherited analytical frameworks. 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. 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.

Only now are advances in policy studies, data collection, and analysis bringing the unification of social science into the realm of possibility. 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 rather than focus on uniqueness and socially specific attitudes. Global and historical interests have grown among researchers in economic history, global politics, world systems, and global health (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). In retrieval of social-science data, the most obvious innovation is large-scale digitization of print, manuscript 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, Gregory and Southall 1998). 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). New techniques in ontology engineering for medical, geographic, and heritage studies may advance the classification of historical events and other data objects (Bowker and Star 2000; Madin et al. 2008).

Major historical datasets, though principally oriented toward regional and comparative rather than global applications, are now in existence and in construction. Some of these come from members of the Collaborative: the Institute for Quantitative Social Science at Harvard, the Center for Geographic Analysis at Harvard, Great Britain Historical GIS at Portsmouth, the International Institute of Social History in Amsterdam, the CLIO World Tables at Boston University, and World-Historical Dataverse at Pitt. Our approach is complementary to such prior campaigns of data collection and analysis as the Integrated Public Use Microdata Series (IPUMS, at Minnesota), Electronic Cultural Atlas Initiative (ECAI), the Alexandria Digital Library (ADL), and others (Eltis and Richardson 2010, Zanden 2009, Saito 2010). In health sciences, investment in institutions for global collection and mapping of data has expanded especially for studies of malaria (Gething et al. 2010, Hay et al. 2009). Even better established is the World Data System of the International Council for Science (ICSU), upgraded in 2009 for the natural sciences, notably astronomy.

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. Recent work has included steps that are necessary but not sufficient to the creation of global-historical data. For example, work with GIS has created georeferenced documents that give little evidence of change over time; time-series data (mostly economic) have been prepared without georeferencing; and population censuses—detailed but addressing widely separated points in time—tend not to be georeferenced. As another example, the Electronic Cultural Atlas Initiative (Berkeley) did excellent work in developing metadata for many social and cultural variables, but did not succeed in obtaining datasets from those who held them nor in developing a fully global perspective (ECAI). Meanwhile, global-historical datasets require simultaneous documentation of values and variations in space, time, content, and scale.

Building global-historical information repository: CHIA approach

Rather than wait for a gradual accretion of localized projects to bring about large-scale analysis, the Collaborative is to conduct a large-scale initiative to speed up the compilation of global-historical data through widespread and sophisticated collaboration. Collecting a large number of datasets is not sufficient to produce global data—the data need to be merged into a single, uniform data repository. Nor is it possible to create a uniform data repository through automated processing of the existing metadata—the terms are inconsistent and, too often, there turn out to be major bits of information simply missing**.** The problem is that additional metadata must be created to account for harmonization and linkage of inconsistent local datasets and for aggregation to regional and global levels. The CHIA project is to address these issues directly through creation of a global historical data resource.

A large part of the historical data logically to be included in a global historical data resource have yet to be digitized or documented. The data have certainly not been rendered compatible with one another, especially for Asia, Africa, and the Americas, and especially before the twentieth century. While 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, these 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 (ICPSR, Dataverse Network). 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. To access such expertise, our approach emphasizes "crowdsourcing," though based on expert users.

Collecting, Integrating, and Documenting Data. For the assembly of global data, "metadata" must include several types of data description. The overriding rule is that each data value within a dataset must be fully defined in terms of its source, its dimensions, and any transformations or aggregations it has undergone between its original source and its current position in the dataset. 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. The controlled vocabulary for *where* would be a gazetteer or GIS, though it would have to account for variations in boundaries and labels of locations; an analogous and flexible vocabulary is needed for *when*. The controlled vocabulary for *what* is the most challenging, as there is no established thesaurus for statistical concepts, although classifications have been developed for occupations and diseases. The incorporation of such existing detailed classifications means that data ingest work can start before the high level framework – the overall project ontology – is finalized. We are developing the crowdsourcing data-integration infrastructure that will facilitate this task (Brodie 2010).

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.[3] 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. In sum, the volume of metadata will likely equal or exceed the volume of data in the global dataset.

The maintenance of this huge amount of metadata will be laborious and expensive. The need for these additional categories of metadata only becomes clear as we move toward aggregation to global-level data. To extrapolate further, one can imagine that an algorithm for transforming data values is found to require correction – for instance, deflation of value statistics by an improved price index – in which case corrections would have to be made throughout and additional metadata would need to be recorded. With fully upgraded metadata, based on strong standards, it will be possible to recalculate, on the fly, each value and the associated error margin, thus preserving the value of the repository and its elements over time. For contrast, the alternative is that whole datasets might have to be abandoned and recreated from the beginning. In particular, many of the global indices comparing national statistics for the past fifty years appear to contain data but no substantial metadata. Without metadata, 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. It is thus more efficient in the long term to build comprehensive metadata into the global or repository.

For collecting and documenting data we propose a collaborative architecture that will allow us to consolidate heterogeneous historical data sources in a scalable way. This work sits on the boundary between the Research Collaborative and Headquarters, but we present it as part of the Research Collaborative because it depends fundamentally on interaction with users (the archive, in contrast, belongs clearly in Headquarters). 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 to enter data into the repository. We assume that information providers will submit their data, which may initially be held in the form of Excel files, in association with wrappers provided by administrators of the repository. Such wrappers are programs 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 associated with the archive. If the data set includes information not covered by a target schema of metadata, such as new variables, 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 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). In a subsequent step, both information providers and consumers will also be able to submit their subjective data reliability assessments. Such *external* reliability assessment will be combined with *internal* reliability assessment protocols based on analysis of data inconsistencies in the integrated repositories. The data reliability assessment will occur in the process of data curation and data fusion.

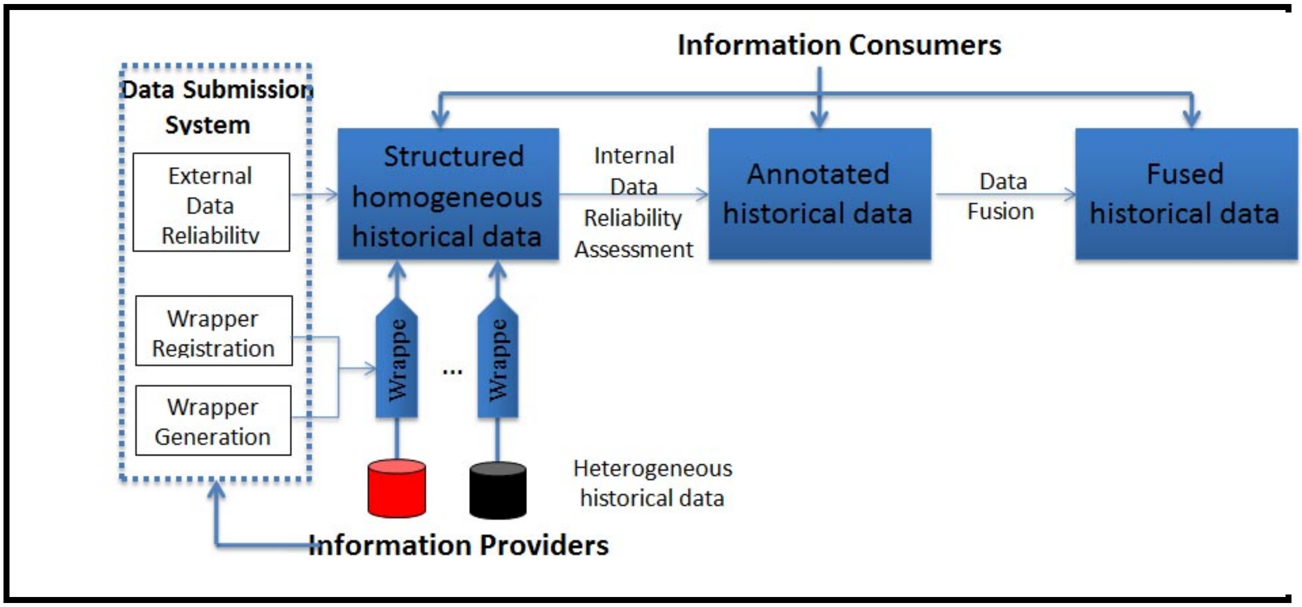
Information Consumers
Data Submission System
External Data Reliabilitv
Wrapper Registration
Wrapper Generation
Structured homogeneous historical data
Internal Data Reliability Assessment
Annotated historical data
Data Fusion
Fused historical data
Wrappe
...
Wrappe
Heterogeneous historical data
Information Providers

Figure 1: Historical data integration architecture based on crowdsourcing

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, because 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. These historians are in effect a parallel to the local and amateur astronomers who have made Galaxy Zoo a crowdsourcing success (GALAXY ZOO; see also FOLDIT and Transcribing Bentham). The additional complication is that historians will be adding civerse new data rather than transcribing in coordinated campaigns Still, we expect to be able to bring the requisite expertise to crowdsourcing to make CHIA a success.

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

How are we to design the elements and the structure of an archive of this magnitude? We need to maintain a global strategy, yet be tactically flexible. CHIA is working with two different architectures for the archive, expecting that we will benefit from both of them and make better overall decisions based on this breadth of experience. The two basic designs are those of the Dataverse Network (created and hosted at Harvard's IQSS) and a design related to Data Cubes created and hosted by Great Britain Historical GIS. The Datavese Network is an open-source application to publish, share, reference, extract and analyze research data that facilitates making data available to others. It currently works across datasets in the following ways: 1) it ingests SPSS and STATA files, extracting the well-structure file and variable metadata, 2) it converts that metadata into an XML Data Documentation Initiative (DDI) format (Dataverse is fully compliant with DDI schema),[4] and 3) it provides search and subsetting of the data set based on that metadata. The main part missing for enabling it to be spatially and temporally specific is a controlled vocabulary or standardization of time and place, which will allow users to easily compare or even merge different data sets. The GBHGIS data architecture holds all names in Unicode (UTF-8), its data model supports multiple languages, and the system operates in Postgres, an open-source database. The historical geographic ontology amounts to creating a kind of data in which all individual data values exist more or less independently, not as parts of "datasets." The GBHGIS implementation of this approach currently holds 14,099,469 data values each in one row of a relational table, with one column holding the actual numbers and other holding those where/when/what/source dimensions. It is also DDI-based, but a different kind of data structure, which can directly support analytic operations; it does use controlled vocabularies including one for data semantics, though it does not yet follow any established standard (Southall 2011). CHIA works with both of these approaches, expecting each to play a role in the overall global archive. For the Dataverse Network approach, the World-Historical Dataverse project at Pitt has engaged a researcher for a two-year project of collaborating with IQSS staff to upgrade the Dataverse Network into an archive more closely approximating one appropriate to CHIA; the GBHGIS approach is being expanded in current work on georeferencing maps. In addition, discussion has now started up between these two groups and areas of collaboration have already been discovered (Grossner 2010).

Within both the Research Collaborative and Headquarters dimensions of CHIA, we expect to encounter complex choices at every stage (Kriegel et al. 2007). We will work incrementally, but at every stage the project needs to have a committee scrutinizing incremental work for its global implications. Further, the project needs to put a significant resources into overall conceptualization and specific details of all-at-once global approaches. That is, the project must begin its earliest stages accounting for the huge data quantities it will ultimately ingest. The particular focus of the first stage of the CHIA project is on clear procedures for defining and executing work and on precise documentation of all elements of this expanding data collection, to lay the groundwork for long-term dependability and renewability of the data at all levels.

To summarize, the Collaborative is to create an infrastructure for retrieving, holding, and analyzing world-historical data. The Collaborative will be an institution of sufficient scale and authority to address the analytical and organizational challenges of documenting human society in recent centuries. In the long run, 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. 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, the Collaborative 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 Collaborative will lead good practice in the inevitable debates about the ownership of data and citation and recognition of the contributors of data.

Using Global-Historical Information Repository for Specific Research Directions

- Our Global-Historical Information Repository will be utilized in a wide range of interdisciplinary research projects. The specific research directions will address some of the most fundamental issues in the social sciences with a methodology that explores scales from the individual to the social, spaces from the local to the global, and times from the immediate to the long term, including possible future change (Manning 2003).

- ***Social and natural variables.*** The couplings of social and natural systems are challenging to understand because of the complicated feedbacks governing these couplings. 1) What are the relative contributions of climate variability and demographic changes in determining patterns of infectious disease? Recent studies at the University of Pittsburgh Graduate School of Public Health have developed the hypothesis that population density and composition are significant determinants of incidence of certain diseases—while previous research has focused on environmental factors. Testing this hypothesis requires the compilation and organization of previously unorganized disease surveillance records and linking them to population data through Collaborative's data infrastructure. While this analysis will begin with twentieth-century United States data, the insight into process and analytical tools developed will be applied to accruing Collaborative data holdings and be extended to longer-term analysis over wider areas, including data-poor low- and middle-income regions (Vora et al. 2008). 2) As another example, simple hydrological variables such as rainfall can be compared with commercial, population, and migration data to reveal spatially explicit patterns in the couplings between these systems for the nineteenth and twentieth centuries. One may test the hypothesis that natural factors influenced the decisions made during early colonization and land division in North America (Bain and Brush 2008; Pastore et al. 2010; Pierce et al. 2009). Population and migration figures for recent times are readily available; for earlier times, improved estimates for Europe, Asia, and Africa are developing. (McKeown 2004; Lucassen and Lucassen 2009; Manning and Nickleach in process). Knowledge of the coupling among human systems is fundamental to our understanding of how societies develop and sustain themselves (Chandra, Kaljanin, and Wray in press).

- ***Development.*** "What has been the trajectory of development in economy and health?" The issue of development—as measured through social indices—has been a principal topic in social-science analysis. In this study, indices of development are traced for political and social units worldwide. Hypotheses on the relations among production, commerce, health, and population can be tested for regions and for the world as a whole (Maddison 2003). Inclusion of data on most areas of the world over time will permit the identification of global and regional interactions in processes of development. Thus the Collaborative can compare the impact of colonization and decolonization worldwide (Gerring, et al. 2011). Further, while the study of women in politics may begin at national and jurisdictional levels, to cover longer times the analyst must include the local levels to which women's political activity was long constrained, plus the transnational effects of feminist movements (Smith 2008).

- ***Governance.*** "How have systems of governance evolved?" This comprehensive study will trace the history of government by tracing it in parallel at local, provincial, national and imperial levels, addressing institutions, systems of law, processes of representation and redress, and violence. It will evaluate the hypothesis that innovations in governance arise out of local governance as much as from central government. Analysis of multi-level data may show when local government systems influence those at higher levels, and when imperial structures determine local structures (King, Honaker, Joseph, Scheve 2001). Attention to local government reveals more information on the role of social movements in governance more broadly. Long-term analysis should reveal the emergence of the global state system, but also the ways that more localized state systems linked to emerging great-power rivalries (Gerring et al 2011).

- ***Social structure.*** "How have social structures changed and interacted?" A worldwide study on this issue has been launched at an initial level by the International Institute for Social History, with funding from the Netherlands Research Foundation (NOW) and the Gerda Henkel Foundation (Germany). This group, by analyzing the types and sub-types of labor conducted in populations from 1600 forward (distinguishing types of labor defined as reciprocal, tributary, and commodified, plus the non-working) is revealing parallel changes in the organization of work in many parts of the world. Out this work comes the hypothesis that the range of types of labor grew more complex up to the early nineteenth century, followed by a simplification in labor structures in later times as wage labor became more prominent. This project's organization of research on historical labor shows the feasibility of large-scale, collaborative research, and provides a fine model for procedures in data collection and data organization (Van der Linden 2008). A decentralized collaborative of researchers—the Global Collaboratory for the History of Labour Relations—collects and documents data, interacting with IISH in processes of cleaning and verification. IISH Research Director Marcel van der Linden will expand this role, serving as Director of Affiliated Researchers for the collaborative.

The outcome of these studies will bring global and historical analysis to bear on these important questions, and will be able to confirm and advance the approach of the Collaborative.

NOTES

[1] The authors acknowledge the contributions of other members of the CHIA group who have made substantial contributions to the preparation of this article. At the University of Pittsburgh: Wilbert van Panhuis, Donald Burke, Kai Cao, Hassan Karimi, and Carlos Sanchez. At Harvard University: Gary King, Peter K. Bol, Mercè Crosas, Benjamin Lewis. At Portsmouth University: Humphrey Southall. At Michigan State University: Siddharth Chandra. At the International Institute of Social History: Ulbe Bosma. At the University of California – Irvine: Geoffrey Bowker.

[2] The institutions participating in the NSF-funded project include the World-Historical Dataverse Project at the University of Pittsburgh (PI Patrick Manning, co-PI Vladimir Zadorozhny), the Institute for Quantitative Social Science and the Center for Geographic Analysis at Harvard University (PI Gary King), the CLIO project at Boston University (PI John Gerring), the Asian Studies Center at Michigan State University (PI Siddharth Chandra), and the University of California Group for Historical Information and Analysis at the University of California – Merced (PI Ruth Mostern). Duration of the NSF project is from 1 January 2013 to 31 December 2015: it is a Research Coordination Network project in response to the Building Community and Capacity initiative of the NSF Directory for Social, Behavioral and Economic Sciences. Other groups participating in the activities of CHIA are the International Institute of Social History (Amsterdam), Great Britain Historical GIS at the University of Portsmouth, Information Science at the University of California – Irvine, and the Council for Economic and Social Development in Africa (CODESRIA, Dakar).

[3] This work of integration must also grapple with the challenges of fuzzy data.

[4] "The Data Documentation Initiative (DDI) is an effort to create an international standard for describing data from the social, behavioral, and economic sciences. Expressed in XML, the DDI metadata specification now supports the entire research data life cycle." http://www.ddialliance.org.

REFERENCES

Institutions:

ADL. Alexandria Digital Library, a distributed digital library with collections of georeferenced materials. http://www.alexandria.ucsb.edu/.

Dataverse Network. Open-source, digital library system for the management , dissemination, exchange, and citation of virtual collections of quantitative data. http://thedata.org.

ECAI. Electronic Cultural Atlas Initiative, a consortium to create a distributed virtual library of cultural information with a time and place interface: http://www.ecai.org.

GALAXY ZOO. A large-scale project of galaxy research: http://www.galaxyzoo.org/.

ICPSR. Interuniversity Consortium for Political and Social Research. http://www.icpsr.umich.edu/.

FOLDIT. Solve puzzles for science: http://fold.it.

NOGWHISTO. Network of Global and World History Organizations: a federation of affiliates based in North America, Europe, Asia, Africa, and Latin America, and affiliated with the International Committee of Historical Sciences: http://www.uni-leipzig.de/~gwhisto/.

Transcribing Bentham. Online transcription of manuscripts by Jeremy Bentham. http://www.ucl.ac.uk/transcribe-bentham/.

Citations:

Bain D.J., Arrigo J.A.S., Green M.B., Pellerin B.A., Vörösmarty C.J. 2008. “Historical Legacies, Information and Contemporary Water Science and Management.” *Water*. 2011; 3(2):566-575. http://www.mdpi.com/2073-4441/3/2/566/.

Bain, D. J. and G. S. Brush. 2008. “Gradients, Property Templates, and Land Use Change.” *Professional Geographer* 60 (2): 224-237. http://www.informaworld.com/smpp/ content~db=all~content=a 791162319.

Bowker, Geoffrey. 2008. *Memory Practices in the Sciences (Inside Technology).* Cambridge: MIT Press.

Bowker, Geoffrey, and Susan Leigh Star. 2000. *Sorting Things Out: Classification and Its Consequences (Inside Technology).* Cambridge: MIT Press.

Brodie, M. 2010. “Data Integration at Scale: From Relational Data Integration to Information Ecosystems.” *Proc. Of 24th AINA*. Abstract: 10.1109/AINA.2010.184.

Chandra, Siddharth, Goran Kaljanin, and Jennifer Wray. 2011. "Mortality from the Influenza Pandemic of 1918-19: The Case of India." *Demography*, in press.

Chase-Dunn, Christopher, and Salvatore Babones, eds. 2006. *Global Social Change: Historical and Comparative Perspectives.* Baltimore: Johns Hopkins University Press.

Eltis, David, and David Richardson. 2010. *Atlas of the Transatlantic Slave Trade.* New Haven: Yale University Press.

Gerring, John, et al. 2011. "CLIO World Tables: A Global Historical Database." Unpublished paper, Boston University. http://people.bu.edu/jgerring/documents/ColonialismLegacies.pdf.

Gerring, John, Philip Bond, William Barndt, Carola Moreno. 2005. "Democracy and Growth: A Historical Perspective." *World Politics* 57: 323-64.

Gething, Peter W., David L. Smith, Anand P. Patil, Andrew J. Tatem, Robert W. Snow & Simon I. Hay. 2010. "Climate change and the global malaria recession." *Nature* Vol 465|20 May 2010| doi:10.1038.

Giddens, Anthony. 2003. *Runaway World: How Globalization in Reshaping our Lives*, 2nd ed. New York: Routledge.

Gregory, I. N., and Humphrey Southall. 1998. "Putting the Past in Its Place: the Great Britain Historical GIS." S. Carver (ed.) *Innovations in GIS 5* (Taylor & Francis, London), 210-221.

Grossner, K. 2010. "Event Objects for Spatial History." In R. Purves, R. Weibel (eds.) *Extended Abstracts Volume, GIScience 2010, Zurich*. http://www.giscience2010.org/index.php?page=Representing-space...

Hay, Simon I., Carlos A. Guerra, Peter W. Gething, Anand P. Patil, Andrew J. Tatem, Abdisalan M. Noor,Caroline W. Kabaria, Bui H. Manh, Iqbal R. F. Elyazar, Simon Brooker, David L. Smith, Rana A. Moyeed, Robert W. Snow. 2009. "A World Malaria Map: Plasmodium falciparum Endemicity in 2007." *PLoS Medicine* 6, 3: 286-301. PMID. 1000048

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. http://gking.harvard.edu/files/pr.pdf

Huxley, Julian. 1946. *UNESCO, its purpose and its philosophy*. London, UNESCO.

King, Gary, James Honaker, Anne Joseph, and Kenneth Scheve. 2001. "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation," *American Political Science Review*, Vol. 95, No. 1 (March, 2001): Pp. 49-69. http://gking.harvard.edu/files/evil.pdf.

Kriegel, Hans-Peter, Karsten M. Borgwardt, Peer Kröger, Alexey Pryakhin, Matthias Schubert, Arthur Zimek. 2007. "Future trends in data mining." *Data Min Knowl Disc* 15:87–97. DOI 10.1007/s10618-007-0067-9

Lucassen, Jan, and Leo Lucassen. 2009. "The mobility transition revisited, 1500–1900: what the case of Europe can offer to global history." *Journal of Global History* 4: 347-377. http://journals.cambridge.org/action/displayAbstract?fromPage=online&aid=6486060

Maddison, Angus. 2003. *The World Economy. Historical Statistics.* Paris: OECD.

Madin, J. S., Bowers, S., Schildhauer, M. P., and Jones, M. B. 2008. "Advancing ecological research with ontologies." *Trends in Ecology & Evolution*, 23(3), 159-168. http://www.sciencedirect.com/science/article/pii/S0169534708000384.

Manning, Patrick. 2003. *Navigating World History: Historians Create a Global Past.* New York: Palgrave.

__________. 2010. "African Population: Projections, 1850-1960." Karl Ittmann, Dennis D. Cordell, and Gregory Maddox, eds., *The Demographics of Empire: The Colonial Order and the Creation of Knowledge* (Athens, OH: Ohio University Press), 245-275. http://hdl.handle.net/1902.1/15281.

Manning, Patrick, and Scott Nickleach. In process. *African Population, 1650 – 1950: The Eras of Enslavement and Colonial Rule.*

McKeown, Adam. 2004. "Global Migration 1846-1940." *Journal of World History* 15: 155-189. http://muse.jhu.edu/login?uri=/journals/journal_of_world_history/v015/15.2mckeown.html.

Nunn, Nathan. 2009. "The Importance of History for Economic Development," *Annual Review of Economics* 1: 65-92. http://www.economics.harvard.edu/faculty/nunn/files/Nunn_ARE_2009.pdf.

O'Brien, Patrick K. 2006. "Historiographical traditions and modern imperatives for the restoration of global history." *Journal of Global History* 1: 3-39. http://journals.cambridge.org/action/displayAbstract?fromPage=online&aid=413154.

Pastore C., M. B. Green, D. J. Bain, A. Munoz-Hernandez, C. Vörösmarty, J. Arrigo, S. Brandt, J. Duncan, F. Greco, H. Kim, S. Kumar, M. Lally, A. Parolari, B. Pellerin, N. Salant, A. Schlosser, K. Zalzal. 2010.. "Tapping Environmental History to Recreate America's Colonial Hydrology." *Environmental Science and Technology* 44 (23): 8798–8803. http://pubs.acs.org/toc/esthag/44/23.

Pierce, D. W., Barnett, T. P., Santer, B. D., and Gleckler, P. J. 2009. "Selecting global climate models for regional climate change studies." Proceedings of the National Academy of Sciences, 106(21), 8441. http://www.pnas.org/content/106/21/8441.abstract.

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 S. Rogoff. 2009. *This Time is Different: Eight Centuries of Financial Folly.* Princeton: Princeton University Press.

Saito, Osamu. 2010. "An Industrious Revolution in an East Asian Market Economy? Tokugawa Japan and Implications for the Great Divergence." *Australian Economic History Review* 50: 240-261. http://onlinelibrary.wiley.com/doi/10.1111/j.1467-8446.2010.00304.x/abstract.

Smith, Jackie. 2008. *Social Movements for Global Democracy.* Baltimore: Johns Hopkins Univ. Press.

Southall, H. B. 2011. "Rebuilding the Great Britain Historical GIS, Part 1: Building an indefinitely scalable statistical database," *Historical Methods* 44: 149-159. doi:10.1080/01615440.2011.589774

Van der Linden, Marcel. 2008. *Workers of the World: Essays toward a Global Labor History.* Leiden: Brill.

Vora A, Burke D.S.**,** Cummings D.A. 2008. “The impact of a physical geographic barrier on the dynamics of measles.” *Epidemiol Infect* 136: 713-20 (2008). PMID: 17662170

Zanden, Jan Luiten van. 2009. *The Long Road to the Industrial Revolution: The European Economy in a Global Perspective, 1000 – 1800.* Leiden: Brill.

Zadorozhny, V., Raschid, L., Gal, A. 2008. ”Scalable Catalog Infrastructure for Managing Access Costs and Source Selection in Wide Area Networks.” *International Journal of Cooperative Information Systems*,17, 1. http://www.worldscinet.com/ijcis/17/1701/S0218843008001786.html.

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. Epub June 29 (2009) PMID: 19564632

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.