Journal of World-Historical Information
JWHI

Matt Drwenski and Christopher H. Myers

Introduction:
Historical Databases in Process

There is a growing acceptance among social scientific historians that certain types of historical questions can best be answered through the careful conception, creation, analysis and visualization of historical information as data. Our task in the following thematic grouping is to see if historical projects using databases, but which have differing scopes, scales, and foci, can be put into a common dialogue. We argue that these diverse studies cohere around a similar method that applies database analysis to historical questions. They also present opportunities for making connections and encouraging cross-fertilization. By comparing historical database research from the local to transnational to global scales, and from simple to complex analyses and visualization, we hope to gain insights on the process of using historical databases, and the commonalities between research with and without database methods. From this common analytical process, we begin to see the practical unity between micro, macro, social, and cultural histories across themes of empire, sports, and economy.

In bringing these papers together, we hope to demonstrate that database approaches create historical narratives in a similar manner as history that eschews digital tools. Many scholars have preconceived notions about the legitimacy of quantitative history, and when it is appropriate to employ database methods. This set of articles stretches the limits of what is traditionally thought of as quantitative or digital history and breaks down some of those preconceptions. In response to these debates, we argue that database-history is not unique in its process of creating a historical narrative, but rather applies different tools. This grouping of articles thus attempts to "stress-

Volume 2-3, No. 1 (2014-2015) | ISSN 2169-0812 (online)
DOI 10.5195/jwhi.2015.9 | http://jwhi.pitt.edu

test" the process of database research at different points along the conceptualization-to-analysis continuum and at different levels of geographical scales, with attention to the methods' flexibility and applicability. In doing so, we encourage readers to think about the opportunities, benefits, and limitations of this approach.

This thematic cluster came out of graduate student research at the University of Pittsburgh's Department of History. The idea of a collaborative article emerged in Fall 2014 as a growing group of graduate students using historical databases began having discussions about methodological issues as we kept running into them. It was the ongoing work of the Collaborative for Historical Information Analysis (CHIA), under the leadership of the University of Pittsburgh's World History Center that drove the collaborative article project. During the initial meetings to define this project, a basic question emerged: can all historians engage in collaborative database initiatives like CHIA? We believed that the cross-fertilization of our five projects could start a dialogue reflecting on the construction of historical narratives in focused projects using databases.

The project developed over multiple meetings from November 2014 through February 2015. We moved away from our initial idea of an article cluster about the methodology of database historical analysis, instead focusing on common data, conception, and analysis issues. Writing the individual papers turned our focus toward how historical databases can be employed in practice. This article cluster has thus become a response to current scholarship's lack of reflection on the process of practically applying databases in historical research. In doing so, we have given several practical comments and pieces of advice that may be helpful as CHIA's aggregation of world-historical information continues.

Patrick Manning's *Big Data in History* argues that the creation of large-scale historical databases covering a variety of topics and time periods "will provide a new, comprehensive level of documentation of the past."[1] While comprehensive data history is still some way off, more feasible projects done at a smaller scale are one of the building-blocks of this future level of documentation. Although Manning's vision speaks to a higher level than a historical database project, our goal here is to examine and describe this process at the project level. The articles in this thematic cluster serve as models of how historical researchers can participate in this process. We believe that Manning's levels of function for bringing together a multitude of datasets—conceptualization, aggregation, standardization, and visualization—apply as processes and methods of databases construction, analysis, and narrative creation at the project level.[2]

The database process begins simultaneously with conception and collection, which form a feedback loop that directs database creation. In our five examples, conception starts by identifying the project's main historical problem. Built into conception is our understanding of the relationship between historical information, and social structures and dynamics. Nuanced and insightful historical conclusions result from the thought, logic, and historical reasoning put into bridging the gap between the data, context, and theory. Nonetheless, there is wide variation in the organic aspect of the researchers' choice to use databases as a tool.

Collection begins with decisions about data relevance. Whether researchers create their own dataset through archival research or compiles others' data, it is largely the same process. One must determine where in the archive or published data sources to start and focus. Since historical information cannot be collected as is, the collection process also involves analysis, aggregation, and standardization. Researchers creating their own databases by bringing together disparate data from digitized or archival primary sources must apply an explicit group of standardized variables and categories. Those using already published datasets must be able to tease out the information's original source, the steps used to collect and aggregated data, and the assumptions or judgments employed. Overall, this process begins the on-going harmonization of the data.

Standardization is the process of determining how data relate to each other, which overlaps with collection and continues through analysis. Researchers transform their data in order to create some level of homogenization of variables, units, and categories. Regardless of whether data is collected from archival sources or adapted from other previous research, existing datasets and literature guide the standardization process. This begins the process of linking the different variables, units, and categories, and addressing the data's consistencies and inconsistencies based on scale, space, and time. For example, the geographical or administrative units bounding categories must be

consistently defined before the categories can be analyzed. Although the processes of conception, collection, and standardization require researchers to continually make informed analytical judgments, these first three steps must eventually stop.

Comprehensive analysis cannot take place until the data is finalized. Historical database researchers begin with descriptive statistical analysis: counts, distributions, means, etc. that demonstrate the data's contours. The contours often sufficiently or partially address many historical questions. If applicable some research projects may deploy more complex statistical tools, such as regressions or visual distributions. Before historical conclusions can be drawn or questions answered, researchers must consider the limitations and assumptions of their database projects, as well as the relevant societal dynamics beyond the data. This is no easy task and in some cases consumes the majority of the analytic space of a project. Again, by considering the institutional, structural, social, cultural, and economic factors underlying the data, historical database researchers rely on a methodology similar to histories that forgo database analysis. Competent historians master and internalize many of these facts, processes, and developments that may help clarify their analytical conclusions.

The previous steps overlap, and build toward our ultimate one: visualization. We see data and database visualization as a broad process, ranging from the presentations of tables, charts, graphs, and images that summarize or describe the data to the creation of a written narrative that draws conclusions from the historical information. Indeed, the written and the visual are almost always employed together to convey meaning and conclusions to the reader. Overall, these processes are not entirely linear, but build upon each other. Visualization often illuminates analysis in new ways.

NOTES

[1] Patrick Manning, *Big Data in History* (New York: Palgrave Macmillan, 2013), 6.

[2] *Ibid*., 6, Figure 1.1.

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.