

palgrave▶pivot

BIG DATA IN HISTORY

Patrick Manning



palgrave►pivot

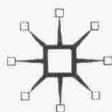
Big Data in History

Patrick Manning

*Andrew Mellon Professor of World History &
Director, Collaborative for Historical Information and
Analysis, University of Pittsburgh*

常州大学图书馆
藏书章

palgrave
macmillan



© Patrick Manning 2013

All rights reserved. No reproduction, copy or transmission of this publication may be made without written permission.

No portion of this publication may be reproduced, copied or transmitted save with written permission or in accordance with the provisions of the Copyright, Designs and Patents Act 1988, or under the terms of any licence permitting limited copying issued by the Copyright Licensing Agency, Saffron House, 6–10 Kirby Street, London EC1N 8TS.

Any person who does any unauthorized act in relation to this publication may be liable to criminal prosecution and civil claims for damages.

The author has asserted his right to be identified as the author of this work in accordance with the Copyright, Designs and Patents Act 1988.

First published 2013 by
PALGRAVE MACMILLAN

Palgrave Macmillan in the UK is an imprint of Macmillan Publishers Limited, registered in England, company number 785998, of Houndmills, Basingstoke, Hampshire RG21 6XS.

Palgrave Macmillan in the US is a division of St Martin's Press LLC, 175 Fifth Avenue, New York, NY 10010.

Palgrave Macmillan is the global academic imprint of the above companies and has companies and representatives throughout the world.

Palgrave* and Macmillan* are registered trademarks in the United States, the United Kingdom, Europe and other countries.

ISBN: 978-1-137-37898-9 EPUB

ISBN: 978-1-137-37897-2 PDF

ISBN: 978-1-137-37896-5 Hardback

A catalogue record for this book is available from the British Library.

A catalog record for this book is available from the Library of Congress.

www.palgrave.com/pivot

DOI: 10.1057/9781137378972



Big Data in History

Other Palgrave Pivot titles

- Mitchell Congram, Peter Bell and Mark Lauchs: **Policing Transnational Organised Crime and Corruption: Exploring Communication Interception Technology**
- János Kelemen: **The Rationalism of Georg Lukács**
- Susan D. Rose: **Challenging Global Gender Violence: The Global Clothesline Project**
- Thomas Janoski: **Dominant Divisions of Labor: Models of Production That Have Transformed the World of Work**
- Gray Read: **Modern Architecture in Theater: The Experiments of Art et Action**
- Robert Frodeman: **Sustainable Knowledge: A Theory of Interdisciplinarity**
- Antonio V. Menéndez Alarcón: **French and US Approaches to Foreign Policy**
- Stephen Turner: **American Sociology: From Pre-Disciplinary to Post-Normal**
- Ekaterina Dorodnykh: **Stock Market Integration: An International Perspective**
- Bill Lucarelli: **Endgame for the Euro: A Critical History**
- Mercedes Bunz: **The Silent Revolution: How Digitalization Transforms Knowledge, Work, Journalism and Politics without Making Too Much Noise**
- Kishan S. Rana: **The Contemporary Embassy: Paths to Diplomatic Excellence**
- Mark Bracher: **Educating for Cosmopolitanism: Lessons from Cognitive Science and Literature**
- Carroll P. Kakel, III: **The Holocaust as Colonial Genocide: Hitler's 'Indian Wars' in the 'Wild East'**
- Laura Linker: **Lucretian Thought in Late Stuart England: Debates about the Nature of the Soul**
- Nicholas Birns: **Barbarian Memory: The Legacy of Early Medieval History in Early Modern Literature**
- Adam Graycar and Tim Prenzler: **Understanding and Preventing Corruption**
- Michael J. Pisani: **Consumption, Informal Markets, and the Underground Economy: Hispanic Consumption in South Texas**
- Joan Marques: **Courage in the Twenty-First Century**
- Samuel Tobin: **Portable Play in Everyday Life: The Nintendo DS**
- George P. Smith: **Palliative Care and End-of-Life Decisions**
- Majia Holmer Nadesan: **Fukushima and the Privatization of Risk**
- Ian I. Mitroff, Lindan B. Hill, and Can M. Alpaslan: **Rethinking the Education Mess: A Systems Approach to Education Reform**
- G. Douglas Atkins: **T.S. Eliot, Lancelot Andrewes, and the Word: Intersections of Literature and Christianity**
- Emmeline Taylor: **Surveillance Schools: Security, Discipline and Control in Contemporary Education**
- Daniel J. Hill and Daniel Whistler: **The Right to Wear Religious Symbols**
- Donald Kirk: **Okinawa and Jeju: Bases of Discontent**
- Sara Hsu: **Lessons in Sustainable Development from China & Taiwan**
- Paola Coletti: **Evidence for Public Policy Design: How to Learn from Best Practices**
- Thomas Paul Bonfiglio: **Why Is English Literature? Language and Letters for the Twenty-First Century**

List of Illustrations

Figures

1.1	World-historical data resource, showing functions and activities	6
2.1	Types of global-historical data	19
3.1	Structure of CHIA: interactions of participating groups	34
4.1	Mission 1: assembling and documenting data	55
5.1	Mission 2: harmonizing and aggregating data	66
6.1	Mission 3: visualization and analysis	73

Table

2.1	Some global changes and linkages, 1500–2000	24
-----	---	----

Preface

This little book arises out of the intensive yet widely distributed effort to create a world-historical data resource. Collaborative effort has centered particularly at the University of Pittsburgh, where the World History Center has provided an institutional base and a source of funding, and where the School of Information Science has provided administrative support, faculty research and commentary, and research by graduate students. The Dietrich School of Arts and Sciences, led by Dean N. John Cooper, created the World History Center and supported it with small grants and appointment of a post-doctoral fellow. The Office of the Provost, with particular thanks to George Klinzing, provided essential financial and moral support at a difficult juncture.

Several other universities and research institutes have been essential to confirming the international and collaborative nature of this project. The Center for Geographic Analysis and the Institute for Quantitative Social Science at Harvard University offered leadership at the start, soon followed by the International Institute of Social History in Amsterdam and the University of Portsmouth. Colleagues at Boston University, University of California – Merced, and Michigan State University then joined in, and others are to follow. The most important institutional support arrived with the 2012 award from the National Science Foundation for three years of work to build the infrastructure of the Collaborative for Historical Information and Analysis.

At an individual level, I express my appreciation to those who contributed to development of the vision that is articulated not only in this book but especially in the

ongoing work of research and development. Adam McKeown joined me in the initial sketches of this work at Northeastern University, 1998–2000. In the years 2007–2008, the project gestated and established its connections with parallel groups. Siddharth Chandra became my first partner in this work at Pitt; Geoffrey Bowker provided early inspiration, while Hassan Karimi and Ron Larsen added insights and assistance. From other institutions, Peter K. Bol, Gary King, Ruth Mostern, John Gerring, Bob Woodberry, Lex Berman, Humphrey Southall, Marcel van der Linden, and Ulbe Bosma contributed their links to one after another of the parallel groups pursuing this work. From 2009 to 2011 this growing group wrote up collaborative grant applications which cemented their relationships and clarified the tasks and their explication. At Pitt, Johan Mohd Sani built the World-Historical Dataverse website and its content; Daniel Bain, Wilbert van Panhuis, and Donald Burke led in imaginative work to link health, climate, and population records; and Vladimir Zadorozhny persisted in demonstrating the key role of crowdsourcing in large datasets. From 2011, CHIA took its formal form, and within a year its first major funding had been obtained, in the form of National Science Foundation Award 1244672: the award greatly assisted the completion of this book. Kai Cao joined the research staff at Pitt and Emily Palmer performed essential administrative tasks. Then new expansion in collaboration brought ties with Jan Luiten van Zanden of CLIO-INFRA (Amsterdam), Ebrima Sall of CODESRIA (Dakar), and Pablo Gentili of CLACSO (Buenos Aires). Inevitably, other important figures are neglected in this recounting of an expanding network of scholarship.

The current stage of the CHIA project is to build infrastructure – the archive and the system for data ingest and documentation. Yet the most important element of the infrastructure, as we are learning, is the human system of collaboration – the willingness to collaborate, share, provide frank commentary, and to meet deadlines. We have made some progress and look forward to further advances. It has been my great pleasure to attempt to coordinate so many of the pieces of this big project. The errors in this overview are my own, but I expect my colleagues to speak up quickly and make the necessary corrections.


This work is dedicated to the historians at all levels who are collecting information on the human past. It is dedicated as well to the audience of researchers, teachers, and students who will – if all goes well – be able to explore world history through the resource of the public archive that will result from this project.

Contents

List of Illustrations	vi
Preface	vii
1 Challenges of Big Data in History	1
2 The Need to Know Our Global Past	14
3 CHIA: Its Collaborative Mission, Structure, and Innovation	29
4 Mission 1: Assembling and Documenting the Data	44
5 Mission 2: Creating a Comprehensive Historical Archive	61
6 Mission 3: Analyzing and Visualizing Data Worldwide	71
7 Comparisons: Big Data across Time and Disciplines	84
8 Priorities for CHIA; Benefits of CHIA	101
References	108
Index	116

1

Challenges of Big Data in History



Abstract: *This book makes the case for expanding the worldwide historical archive now under development through the Collaborative for Historical Information and Analysis (CHIA). The opening chapter emphasizes that the time has come for using available technology to create a coherent record of human social change in recent centuries, so as to link patterns of past change to the great policy decisions our society faces. The chapter underscores the fundamentally collaborative nature of CHIA's expanding worldwide project, comparing it to existing projects in climate modeling and genetic databases. It describes the five major levels of the archive – tied together by the three missions for developing them – and the global analysis that will emerge from the fifth level. It introduces the succeeding chapters on the construction of the global historical archive and its intended use by researchers and the general public.*

Keywords: collaborative; social change

Manning, Patrick. *Big Data in History*. Basingstoke: Palgrave Macmillan, 2013. DOI: 10.1057/9781137378972.

The time has come: meeting the challenge

The time has come for creating and analyzing a global dataset on human societal activities. Such a dataset can provide a picture of worldwide social patterns and interactions over the past four or more centuries. 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. Often, plans have been made for the future with little idea of the dynamics inherited from the past and little sense of the directions in which those dynamics are restructuring the present. And while the past is both known and forgotten at local and national levels, at the global level we have almost no knowledge of the historical forces and experiences that have unfolded within human society.

The organization of Big Data in history will provide a new, comprehensive level of documentation on the past. Currently available historical information, while enormous in its overall quantity, lies scattered and dispersed among many repositories. 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 other, and the totality of their records provides rather scanty information on the numerous people in the rural areas.¹ 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 the various changes in human society over the past four or five centuries. This volume provides an overview indicating 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 focus on evidence about the economy, society, politics, health and climate. Later on, the project will address Big Data on ideas, culture, and values.

The challenge is huge: there are great quantities of data to be collected and processed, and the work of processing will be complex. Big Data in history is not like the data harvested from mobile phones or commercial records. Today's phone and other data are consistent in format because they are born digital, including straightforward metadata to describe the data. Instead, historical data mostly exist in small files that need to be digitized, documented, and transformed to become parallel to other

datasets before they can be analyzed. The cost of collecting and archiving historical data is therefore much higher than for contemporary data.

But if the cost of historical data is great, the value is even greater. Most basically, data from the past can give us insight into change over time, a factor addressed only minimally in studies limited to the last few years. Historical views of key variables may enable us to learn about processes of growth, cycles, and interactions that are now unknown. In addition, when properly analyzed, data from the past can be aggregated to yield reconstruction of global patterns in the past. At present, we have some idea of global patterns for today, but the past to which we compare it consists of records only from a collection of localities, and probably unrepresentative localities at that. Being able to compare the global patterns of today with global patterns in the past may provide us with a different idea of global social change.

Creating a global historical data resource, while a complex task, is also one that has become feasible. The organization of Big Data in history can now be accomplished, 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 that 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.

This volume introduces the Collaborative for Historical Information and Analysis (CHIA, www.chia.pitt.edu) and its project for a world-historical archive. With a growing number of colleagues, I have been working since 2007 to build the project for a world-historical archive addressing variables in social sciences, health, and climate to document the past 400 years. The CHIA collaborative, based at the University of Pittsburgh, includes participants at universities in the U.S. and Europe – and at research centers in Africa, Latin America, and Asia. CHIA has now gained substantial initial

funding from the U.S. National Science Foundation (NSF), and smaller amounts of funding from several other sources. Concisely put, the purpose of CHIA is to create a single, comprehensive archive linking an immense range of historical data across space, time, topic, and scale. Fortunately, modern information technology will make it possible to achieve this objective through a virtual archive distributed across many sites (though it is necessary to have shared protocols and standards throughout). Historical data are huge in quantity and are deeply diverse in quality – but are the only source of information on change over time.

This book is to show how the task is being taken on. First, the book is to articulate, for scholars and policymakers, the need for a world-historical archive. Second, it is to attract collaborators to the project – developers of archive structure, contributors of data, and users and evaluators of the data resource. The third purpose of the book is to launch a broad and critical discussion on the ends and the means of a world-historical archive, especially by potential users of the archive, to clarify which paths of development will be most broadly useful. Fourth, the book is to compare the CHIA project with other large-scale data collection projects in social sciences and natural sciences, to obtain insights into the pitfalls and achievements of such projects and thereby help set priorities for CHIA.

This book is a comprehensive introduction to CHIA and its tasks. It is relatively technical, in that it attempts to describe most of the main tasks of CHIA and its archive, and to show how they connect to each other. Thus the book must address many fields of social sciences, natural sciences, humanities, and information science, and address their specialized vocabularies in doing so. Quite a different sort of introduction to CHIA focuses on the historical and social lessons that will result from the CHIA archive. It gives more details on why it is necessary to gather data about the history of the world as a whole, recent global changes, and how ordinary people will benefit from supporting and using the CHIA program. That introduction, available on the CHIA website, explains why every society should become involved in supporting the collection of historical data and study the results of this new historical resource (Manning 2013).

The first two chapters of this book convey the character of a global historical data resource and the social needs that will be met by such a resource. This initial chapter provides a tour through the facets of the CHIA project – its purpose, main areas of activity, the groups of people it

seeks to engage, and the challenges that it is likely to encounter. Chapter 2 begins with a fuller discussion of the need for big, global data in history, and follows up by describing the character and patterns of global historical data along with the nature and feasibility of the project.

Chapter 3 helps demonstrate the ability to collaborate, even in social sciences and history, on such a large project. It describes the specific objectives of CHIA and the way in which the project's collaborative structure is designed to bring broad geographic scope, great flexibility, local autonomy, and still benefit from strong organization and clear direction. The specific techniques for facilitating collaborative work are presented as innovations that have served CHIA well in its early stages and may prove more generally valuable. The chapter concludes with an explicit invitation to those who might become fellow workers on the archive, contributors of data, or users and evaluators of the data resource.

Three missions in creating a world-historical data resource

Chapters 4, 5, and 6 address the specific missions of the overall project. The chapters introduce the interacting elements of 'the Archive', as illustrated in Figure 1.1. The Archive, also identified as the 'world-historical data resource', refers to the whole system of repository, documentation, and analysis. At the same time, each of the five levels within it is also seen as an archive. The Archive is a comprehensive, distributed, and linked repository and analytical system containing relational datasets. Datasets are maintained at distinct but overlapping levels of the Archive, from datasets newly received and beginning the process of incorporation to global, interdisciplinary collections of data along with the results of analysis and visualization. If we think of the Archive as a repository, we can say that it refers to all the files of data and applications that are accessible to the CHIA project, which may be held in a wide range of repositories. In terms of particular servers, they include a server in the School of Information Science at the University of Pittsburgh, the resources of the Pittsburgh Supercomputer Center, housed on the Cloud, and the CHIA files held within the Dataverse Network of Harvard University, which in turn are stored on the Cloud.

Figure 1.1 displays the Archive as a whole and many of the functions and processes within it. It shows the five principal levels of the overall

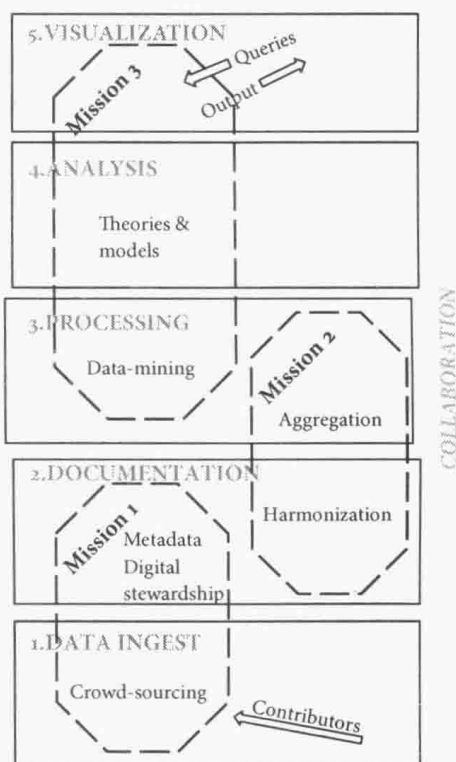


FIGURE 1.1 *World-historical data resource, showing functions and activities*

data resource and the three missions that link the levels by moving and transforming datasets from one level to the next. Each level consists of a distributed archive holding datasets that undergo a specific sort of housing and treatment – and each level is complicated by interactions and feedback loops. Here is a concise summary of the levels, followed by a fuller description of each. At Level 1, files contributed to CHIA are reviewed and archived in their original form for at least three distinctive purposes.² Files entered through the CHIA crowdsourcing ingest process and documented through the Col*Fusion application move into the Level 2 archive and by stages to the higher levels of CHIA. Level 3 holds files fully documented, harmonized and linkable to other CHIA files. Level 4 holds aggregated files created by the merging of existing CHIA files, up to the global level. Level 5 holds files including those created by analysis and visualization of the data. To keep track of the full range of files in the Archive, a system of administration and protocols is expanding progressively. Visitors and users of CHIA will be able to explore the

data at all five levels using the CHIA visualization tools, with simpler tools for Levels 1 and 2 and more elaborate ones for Levels 3, 4, and 5.

The Level 1 archive includes all the activities of CHIA crowdsourcing ingest and pre-processing. Its first element is the Data Hoover, a program designed to locate and ‘hoover’ (or ‘vacuum’) valuable historical datasets into the Archive. Second, contributors submit data through a crowdsourcing process, in interaction with CHIA staff. Third, each incoming dataset undergoes a review to determine its character and its place or direction in the system is recommended. Fourth, each incoming dataset is archived, registered, and made available to CHIA users online, according to the specifications of Digital Stewardship. Fifth, those files that are to be incorporated into larger analysis may undergo Pre-processing. Sixth, there are two directions for additional analysis. First of these is the ‘Spatio-Temporal Bridge’, which focuses on spatial and temporal metadata. Its analysis reveals the spatio-temporal emphasis of numerous files in a given topical area. The advantage of this application is that, with minimal processing, it provides detailed information on the density of existing studies. Finally, datasets may be incorporated into the Col*Fusion process for merging with the full set of files.

The Col*Fusion process is the step that leads most clearly to the creation of a world-historical data resource because it is able to merge datasets through analysis and revision of their metadata. Col*Fusion works through interaction of the contributor, the application itself, and with consultation by CHIA staff to upgrade and coordinate the description of data so that it is consistent with that of files already in the Level 2 archive. In this way a ‘target schema’, an overall system of metadata, develops steadily and maintains consistency as more data are incorporated into the Level 2 archive. The Col*Fusion application has the additional advantage that it is able to merge the incoming file with existing files: through repetition, this process can lead to merger of multiple files.

Once in Level 2, each dataset undergoes ‘harmonization’, a set of processes that enables datasets to be linked more fully to other datasets. Details of harmonization include cleaning data of remaining errors, identifying overlaps and conflicts of datasets, establishment of consistent weights and measures, and confirming consistency among variables documented in different languages. As in Level 1, the datasets in Level 2 and higher levels of the Archive benefit from Digital Stewardship, the system of dataset preservation and citation. For these higher levels, citations of datasets refer not only to the original datasets submitted by contributors,

but also to the transformed datasets based upon submitted files. One key result of the ingest and harmonization of numerous datasets is the expansion of the overall 'target schema, a consistent set of metadata on sources, time, space, topic, and transformations. The expanding system of metadata, facilitated by the work of CHIA staff, will contribute to the broader goal of approximating a world-historical ontology. Completion of this process of harmonization prepares datasets for transfer to Level 3 of the archive.

The advances of the CHIA project have implications going well beyond the project itself, especially in facilitating the merging of datasets. Since all of the work of CHIA is to be open-source and open-access, the combined processes of Col*Fusion and harmonization can be used independently by researchers – outside of CHIA, if they wish. Using these techniques, researchers will be able to merge datasets selected for their own purposes. While the Col*Fusion process – verifying the documentation of each variable and resolving inconsistencies in data definition – involves some labor, it addresses a problem in data analysis that has long been resistant to resolution. That is, the thousands of social science datasets archived in major repositories have tended to remain isolated from one another. That is, while datasets can be connected through analysis of the file-level metadata on their sources and overall characteristics, there has not previously been an application that connects the variable-level metadata within datasets. The combined process of Col*Fusion and harmonization, though it is still in development, seems destined to permit the full merger of independent datasets and, through recurring application of the process, widespread aggregation of datasets.

The processes in Level 3 of the Archive focus first on aggregation and then on data-mining. Aggregation, conducted by CHIA staff, is the continued merger of datasets to the point where they provide observations on macro-regional and global dynamics of the variables. Aggregation of datasets moves along at least three axes: expanding the geographic scope of datasets, expanding temporal scope, and expanding topical scope – the range of topical variables included in each dataset. The aggregated datasets include extended metadata, to document the steps in aggregation along with all previous metadata. All of the datasets, from local to global and over various time frames, are then ready to be advanced to Level 4 of the archive. Before that transfer, however, CHIA conducts an exercise in data-mining within the Level 3 archive. That is, using numeric techniques of analysis and high-speed computation, the datasets in the Level