This paper was submitted to the National Science Foundation as part of its SBE 2020 planning activity (www.nsf.gov/sbe/sbe_2020/). Its inclusion does not constitute approval of the content by NSF or the US Government. The opinions and views expressed herein are those of the author(s) and do not necessarily reflect those of the NSF or the US Government.

Understanding Global Change: How best to organize information?

Humphrey Southall (Department of Geography, University of Portsmouth)
Patrick Manning (World History Center, University of Pittsburgh)
Merrick Lex Berman (Center for Geographic Analysis, Harvard University)
John Gerring (Department of Political Science, Boston University)
Peter Bol (Center for Geographic Analysis, Harvard University)

Abstract:

Understanding the larger socio-economic challenges facing our society requires a long-term global perspective, but in practice such perspectives are almost impossible to achieve because the necessary datasets are fragmentary or non-existent. All too often, historical research is based on a single country or a small group of advanced economies; or on just the last thirty or forty years.

We need to assemble not just historical statistics but closely integrated metadata, including locations and reporting unit boundaries, so that researchers can explore alternative approaches to achieving consistency over space and time without requiring an army of assistants for each new project.

We explore a range of possible approaches, concluding that existing social science data repositories are insufficiently integrated; that we cannot leave it all to Wikipedia, although an open collaborative approach is essential; that Geographical Information Science technologies are necessary; but they are not sufficient, and concepts from other areas of Information Science are also needed, notably including ontologies and linked data.

A set of more specific research challenges are identified, including the need to link vector- and ontology-based data structures for social science history with raster and grid-based resources in environmental history.

(1) **Many of the *grand challenges* facing society require long-term global perspectives**: limits on resources, environmental changes, economic conflict and change, potential political conflicts, and questions of familial and social change. Our only data for addressing these issues come from the past, and while natural-science data are commonly organized at the global level, most social-science data are organized at the national level. **The development of policy solutions to these societal challenges creates a *grand challenge for social science researchers*: how to assemble and analyse inevitably fragmentary historical information**. One approach is to limit research to a small sub-set of countries with high quality historical data, such as those of the North Atlantic Population Project, but this can only be a partial solution and we also need a truly global body of information, documenting variations in quality.

(2) **Specific *grand challenges*** to be addressed by a global historical data assembly include:

This paper was submitted to the National Science Foundation as part of its SBE 2020 planning activity (www.nsf.gov/sbe/sbe_2020/). Its inclusion does not constitute approval of the content by NSF or the US Government. The opinions and views expressed herein are those of the author(s) and do not necessarily reflect those of the NSF or the US Government.

- **Trends in human inequality**, measured by income, wealth, literacy, health. Indications are that these fluctuate over time and link to social protest.
- **Economic cycles**. Fluctuations in output and commerce, from short-term business cycles to apparent cycles of several centuries, should be related to the shifting balance of population and output in world regions.
- **Money and finance in world history**. Global stocks and flows of silver and gold can be estimated for the past several centuries, enabling long-run monetary history. Research into the history of global finance is needed to contextualise recent crises.
- **Trends in health and disease**, infectious and chronic, as these have varied over time and space, and as new diseases spread.
- **Changes in social structure** linked to demographic change, for instance because of varying age-group proportions resulting from changing rates of birth and death.

(3) **Analysis at the global level is fundamentally different from analysis of individual localities, nations, or even empires**: the globe is a closed system encompassing the totality of what are seen locally as external effects. In all the above examples, global patterns will differ from those seen in a few well-documented regions.

(4) **Existing global datasets cover too short a time period**. "Global perspectives" are necessarily about geographical variation: consistent global data on physical geography, broadly defined, are now relatively easily available via remote sensing; and globally consistent socio-economic data are available, more problematically, through the work of the World Bank, the United Nations Statistics Division (UNSD), etc., either gathering data globally or harmonizing the work of national agencies. The first Landsat satellite was launched in 1972; the UNSD *Millennium Development Goals* database goes back only to 1990; even the UNSD *Demographic Statistics* begin only in 1948, with far more limited coverage than today. The clearest exception is climate change, where the assembly of consistent long-run data has genuinely been a grand challenge over the last two decades; but society also faces socio-economic challenges.

What form should a "global historical data assembly" take?

(5) **Existing data archives**, such as the Inter-University Consortium for Political and Social Research, are of course assemblies of data, and their holdings are now generally available on-line via automated repositories. However, their contents are divided into *datasets* with highly variable internal structures, and the high-level documentation created for the repository provides clearly insufficient detail to support actual analysis. Significant research has been done to try to automate data mining across these collections, but no magic bullet has been found: a more integrated structure with deeper consistent documentation, down to the data item level, is required.

(6) **At the opposite extreme, Wikipedia** is a vast assembly of global information which, in its more structured form as DBpedia, forms the hub of the semantic web (see http://linkeddata.org) and is being systematically mined by many computer science projects. Three limitations of Wikipedia make it an unsatisfactory solution: statistical data are scattered almost randomly through a vast body of text; far too little information is available on the provenance of data; and while it is a comprehensive resource on

This paper was submitted to the National Science Foundation as part of its SBE 2020 planning activity (www.nsf.gov/sbe/sbe_2020/). Its inclusion does not constitute approval of the content by NSF or the US Government. The opinions and views expressed herein are those of the author(s) and do not necessarily reflect those of the NSF or the US Government.

computing topics, coverage of *local knowledge*, especially historical, reflects a clear shortage of contributors, the majority of place-specific pages being "stub entries" rarely updated. The disambiguation of geographical names is unreliable, creating special dangers for automated analysis. Even so, Wikipedia demonstrates the immense power of a more open approach to data assembly, and is better suited to automated *analysis* than data archive repositories.

(7) **Neither data archive repositories nor Wikipedia provide systematic geographical frameworks, so should we be building a global historical GIS?** A large limitation of existing social science repositories is that dataset-level documentation rarely reveals the spatial structure of data; even when *coverage* is recorded, *granularity* is not, so we do not know whether a dataset covering the US does so at national, state, county or tract level. Wikipedia have bolted on a reasonably effective framework for recording point locations globally, but fail to distinguish systematically between these "places" and the administrative areas, i.e. polygons, to which most statistics relate. These become major limitations when the analysis is historical, as states and empires expand then disintegrate, and as the statistical reporting areas change even though the places they are named after stay more or less the same. It is essential for long-run global analysis that the data assembly record as precisely as possible the areas to which individual statistical data values relate, and include facilities for converting between reporting geographies to obtain long-run comparability.

(8) **However, existing GIS technology is not enough**. Recent years have seen much promotion of *historical GIS* as a nascent sub-discipline, but the field is not as new as sometimes claimed: the first major historical GIS, The Great American History Machine, was developed at Carnegie Mellon in the mid-1980s; the main commercial GIS software products, and companies, have their roots in the quantitative geography of the late 1960s and early 1970s. Those commercial tools are often poorly suited to historical contexts, so while GIS technology has a substantial part to play in building the necessary global data structure, that structure cannot be a conventional GIS:

- GIS data models treat locations as the framework to which all other information must be linked, but historically locations are often uncertain. This is an ever-larger problem as we go further back in time, or study less developed nations. We need to be able to hold data for entities with unknown locations, adding locations later to permit analysis and sometimes treating locations as interim hypotheses.
- In historical research, we deal with uncertainty primarily through carefully recording the provenance of information, but standard GIS file formats provide inadequate facilities for such documentation, or variant names.
- The *global historical data assembly* needs to hold data on diverse topics reflecting the needs of many disciplines. Even if we focus on specifically statistical information, conventional GIS technology simply does not address how to manage millions of data values measuring thousands of variables.
- The data structure must obviously be based on open data standards, and should certainly follow the standards of the Open Geospatial Consortium (OGC) where appropriate, so linking to modern federal geospatial data infrastructure. However,

This paper was submitted to the National Science Foundation as part of its SBE 2020 planning activity (www.nsf.gov/sbe/sbe_2020/). Its inclusion does not constitute approval of the content by NSF or the US Government. The opinions and views expressed herein are those of the author(s) and do not necessarily reflect those of the NSF or the US Government.

the structure must also link to the wider non-spatial web of knowledge. Specifically, while it must be more curated than Wikipedia or Geonames, it must formally link to them, as **linked data** exposing Uniform Resource Identifiers**.**

The way forward

(9) Global historical research must borrow from the leading edge of information science, geographical and otherwise:

- We must implement geo-spatial ontologies using object-relational database software, paralleling the USGS strategy for the National Spatial Data Infrastructure, using a geographic ontology to mediate between multiple axes. Such architectures avoid the limitations of traditional GIS outlined above.
- Documenting statistical semantics is as important as recording geography. The US National Historical GIS extended the Data Documentation Initiative standard, enabling data item level documentation, but used it with a relatively conventional collection of datasets supporting only download. More recent UK research removed this limitation, more tightly integrating DDI metadata with a large collection of individual data items, enabling automated visualisation and, arguably, analysis.
- We need a resource designed for very broad sharing, created by and designed to support a sophisticated division of labor among researchers. The core resource should be cloud-hosted and support a range of application programming interfaces enabling at least three access modes: a web-based reference interface for audiences numbering at least in the millions; simple web sites through which individual research projects digitally publish their findings, drawing their geographical context via Web Map and Web Feature servers; and more specialised analytical tools drawing their subject matter from the same cloud servers.

(10) The following methodological areas need further research, partly to support the future construction of a global historical GIS:

- **Integrating and translating between raster and vector content**: Historical GIS research, so labelled, has been dominated by human geographers and historical demographers focused on vector geographies, especially administrative boundaries and transport networks. However, there is also a strong but largely separate tradition of research in environmental history which emphasises raster data, while map librarians have scanned many historic maps but rarely geo-referenced them or extracted features. These strands of work need to be more closely linked; historical GIS researchers need more training in raster GIS techniques; and, in particular, research is needed on the use of image processing technologies to extract vector features from historic maps.
- **Synthesising boundary lines**: The construction of detailed national historical GIS systems has been very expensive because of the need to research boundary lines. Much of this expense can be postponed by an ontology-based approach, but statistical analysis requires boundary polygons. A low-cost way to spatially enable an administrative ontology is to add point coordinates for the administrative centres of the lowest-level units. The construction of Thiessen polygons around those

This paper was submitted to the National Science Foundation as part of its SBE 2020 planning activity (www.nsf.gov/sbe/sbe_2020/). Its inclusion does not constitute approval of the content by NSF or the US Government. The opinions and views expressed herein are those of the author(s) and do not necessarily reflect those of the NSF or the US Government.

points is a long-established methodology, but enhanced algorithms are needed which can incorporate additional information we often have available: we sometimes know the total area of each unit; we know that boundaries tend to follow physical features such as waterways; we know that historical boundaries were often the same as modern ones.

- **Ontologies of place**: Constructing large ontologies of administrative units is relatively simple because they are legal entities, already defined quite formally. To understand past geographies of social life we must also work with information about "places", existing in texts and discourse. Constructing simple gazetteers, i.e. non-hierarchic word-lists, is well understood, but can we build thesauri of places, i.e. with hierarchy; polyhierarchic place thesauri; or ontologies of place, with multiple hierarchies and differentiated relationships between places? Much work has been done in recent years on digital gazetteers, but it has been led by the Alexandria Digital Library, focusing on "geographical features", with a physical existence in the landscape, and paying limited attention to linguistic issues. We need to link to linguistic place-names researchers; to research algorithms for extracting placenames from text; and to work with multiple languages and non-Roman alphabets.

(11) The necessary investment cannot be justified solely by the needs of social science historians, but such a data structure would serve many audiences:

- Health researchers need to track historical epidemics, measure the impact of past environments on individual health, and monitor long-run trends.
- Although environmental historians' core data requirements differ, much relevant evidence comes from documentary sources and statistical surveys, needing gazetteers and base maps for interpretation.
- There is a very large educational audience in schools as well as colleges.
- Such an assembly of *intelligence*, broadly defined, has military applications; indeed, past military surveys would be a major source, while recent conflicts have shown the need for broader historical knowledge.
- Locality information in the system would be of broad public interest, especially to people tracing ancestral origins. This matters partly because there are various ways of generating income from such large audiences to sustain the overall resource.

(12) One necessary precursor is ***training***, e.g. on open data standards, and other programs to develop links between social science historians and information scientists. Historians also need ***input*** into data standards.

References:

Hill, L. (2006) *Georeferencing: the geographic associations of information*. Cambridge: MIT Press.

Manning, P. (2003) *Navigating World History: Historians Create a Global Past*. New York: Palgrave Macmillan.

Southall, H., (2008), "Visualization, Data Sharing and Metadata", pp. 259-275 in Dodge, M., McDerby, M., and Turner, M., *Geographic Visualization.* Chichester: Wiley.

This paper was submitted to the National Science Foundation as part of its SBE 2020 planning activity (www.nsf.gov/sbe/sbe_2020/). Its inclusion does not constitute approval of the content by NSF or the US Government. The opinions and views expressed herein are those of the author(s) and do not necessarily reflect those of the NSF or the US Government.

This work is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-sa/3.0/ or send a letter to Creative Commons, 171 Second Street, Suite 300, San Francisco, California, 94105, USA.