White Paper Report

Report ID: 99169

Application Number: PR5008810

Project Director: David Bodenhamer (intu100@iupui.edu)

Institution: Indiana University, Indianapolis

Reporting Period: 5/1/2010-10/31/2012

Report Due: 1/31/2013

Date Submitted: 2/4/2013

National Endowment for the Humanities

White Paper

Submitted:	January 31, 2013
Grant number:	PR-50088-10
Project Title:	New Tools for the Humanities: Visualizing Complex Spatial Data
Project Director:	Dr. David J. Bodenhamer
Grantee Institution:	Indiana University

The Polis Center
INDIANA UNIVERSITY–PURDUE UNIVERSITY INDIANAPOLIS

Visualizing Complex Spatial Data within a Humanities-Oriented GIS:

A White Paper

Problem: Humanists are interested in complexity, yet the GIS default for regional aggregated data is the choropleth map, which simplifies and often misrepresents social phenomena. How can we enhance the technology to reveal the socio-political and cultural complexity inherent in any geographic unit, thus making it more useful for humanities scholars?

Project: The aim of the *New Tools for the Humanities* project was to develop a platform for the visualization and exploration of complex data within a humanities-focused web GIS. The project used a successful but traditionally constructed web GIS, the North American Religion Atlas (NARA), as its starting point. Its four distinct but interrelated goals were to : (1) Re-engineer an existing and well-used interactive web mapping site on twentieth-century US religious adherence, NARA (www.religionatlas.org), as an open-source product incorporating recent advances in web technologies. (2) Incorporate new ways to visualize complex spatially-referenced data, using NARA as a test bed. (3) Develop an approach to facilitate on-the-fly re-categorization and aggregation of data to permit easier, more flexible data exploration. (4) Use the work to advance a stronger theoretical and conceptual case for a spatial approach to humanities research through scholarly publication and presentations. The sequence of tasks on this project proceeded from re-engineering the spatial platform to incorporating the new visualizations and developing a way to allow other visualizations to be added more easily to creating the functional/technical requirements and prototype for the re-aggregation tool.

The agenda of work, described below, was conducted by the Virtual Center for Spatial Humanities, collaboration among Florida State University, West Virginia University, and Indiana University Purdue University Indianapolis (IUPUI), which through The Polis Center served as lead partner for technical development. The mission of the Virtual Center is to develop spatially-oriented theories, methods, and tools to stimulate significant interdisciplinary work in the humanities.

This white paper discusses the issues involved in visualizing complex spatial data and the solutions developed for the re-named site, now the Digital Atlas of American Religion. It also outlines an approach for the dynamic typology construction which constitutes the post-project agenda of work.

Significance: The project used American religion data as its immediate subject, but the case for intellectual significance rests not on the importance of religion to American history—this

argument is self-evident—but on the need to visualize complex data in order to gain a better purchase on the immense variety and dynamic character of spatially referenced data of interest to humanists. Virtually all data come to us with a spatial component. It originates from or describes some place, and we can use its location to relate it to all other data about this location either directly or within some nested hierarchy. For instance, by using location we can relate a diary from a neighborhood to an image or census data from the same place, as well as link it to similar data from a city, county, or a state, etc., and we can portray these relationships on a map or in some other spatially referenced visualization. Humanities scholars have begun to use these data, but GIS technology, especially in its web form, does not allow easy exploration or visualization of their complexity. Consider religious adherence data for any of the nation's 3,000+ counties: within current web GIS platforms, we can view the dominant denomination within the county, which produces a useful and interpretable pattern of complexity when viewed on a national, regional, or state scale. But we cannot achieve the same view for the diversity of religious adherence within a county or group of counties. We can at best map two, perhaps three denominations, using cross-hatching or other crude techniques to enhance the choropleth map presentation that is standard in GIS. What these maps reveal are similar to the red state/blue state maps common to television coverage on election nights.

What humanists need are ways to examine the diversity of human experience at varying scales and to consider these measures with other dimensions of activity and belief. Instead of viewing only the dominant denomination within a county, we should be able to visualize all the denominations by their relative size as well as by other measures (e.g., theological orientation), both singly and in combination. In doing so we will achieve a much more complex but more realistic understanding of human experience as captured in administrative records. In this instance, we will be able to see, for example, that Marion County, Indiana, has a dominant Catholic presence when mapped by standard techniques, but in fact Catholics represent only 20 percent of the religious adherents, with Methodists, Presbyterians, and Disciples of Christ measuring 18, 15, and 12 percent respectively. Or that a contiguous county, Morgan, has a dominant Baptist presence but almost no Catholics, a circumstance that provides an analyst of Midwestern religious experience—both counties are in the Indianapolis MSA—a quite different understanding than is available from a choropleth map. Significantly, visualizing complex religion data does not pose a unique problem, and the tools developed for this proposal can be used to visualize other complex data, such as demographic and economic datasets, that have multiple facets within any given geography. The various US censuses are foremost among the spatially referenced datasets that contain complex data, but such data are rife throughout the humanities, and they extend across all data types and formats.

What we developed is a way to overcome the current limitations of GIS by making maps more complex and more visually dynamic, thus fitting what we know about human experience, but

also by making them more easily interpreted. Consider, for example, election maps. The convenient division between red states and blue states is profoundly misleading. That convention has implications that reach all the way down from the state to the county to the individual voter, with a powerful set of assumptions built in: people vote in ways that tend to be generally static, unreflective, and bundled. Thus, the common nomenclature of red and blue assumes that people who live near one another tend to share political ideas because of their common history, ethnicity, and economic experience. New spatial visualizations, on the other hand, will give us a glimpse of what deep contingency might look like over time. By allowing us to see space and time at a distance, in relatively abstract ways, the maps can show us dissolving and crystallizing patterns otherwise invisible in rows of numbers or static maps based on the standard visualization schema.

Visualizing spatial complexity is essential but not sufficient for the needs of humanities scholars. We also need ways to re-categorize data quickly and then to visualize the new typologies we have created for our research. So the project also developed requirements for a tool to facilitate on-the-fly re-aggregation of data for use in the new visual framework. For the study of American religion, this tool will allow scholars to categorize denominational families by criteria of their own choosing (and aggregate data by the new categories) rather than adopt family trees and aggregations developed by other scholars for other purposes. One scholar, for example, may place a denomination within a family ordered by theology, whereas another scholar may want to use ecclesiology as the criteria for inclusion. Currently, the process for making this change might include a laborious restructuring of the database; our goal was to make this process more dynamic, thus allowing scholars to experiment with the data easily. Its application, when completed, extends far beyond religion data. For use with US censuses, voting statistics, or any other spatially referenced data, for example, the tool should permit researchers to group and map categories easily and in new ways or to aggregate by existing scholar-defined classifications.

Lessons. Even when simple in design, digital projects often encounter unexpected problems for which no easy solution exists. This circumstance certainly was the case with DAAR. The issues we encountered were not unknown to us—we understood the potential for problems before beginning the work—but on occasion we underestimated the time required for a solution. One result was that the solution was not as elegant or as simple as we had hoped, at least not within the limits of time and budgets. The following discussion notes some of these problems and identifies the solution we reached; it then offers guidance for researchers/developers engaged in similar projects.

1. Handling change over time in religion adherence and membership data is much more complicated than we anticipated. There are numerous instances in which

denominations split or merged, sometimes with the same name and sometimes with a new name. The technical challenge is how to present data consistently across time and especially how to ensure comparability of the data. Without it, the user will have an inaccurate portrait of the shifts in religion adherence for the denomination in question and also for the larger context in which these changes occur. This issue is well-known among scholars of American religion, so the problem was not conceptual; rather, the question was how to implement a solution technically within a GIS.

We quickly recognized that a hierarchical database could not accommodate a solution without extensive programming, which easily could introduce errors that are not easily discernible. A graph database/neural network structure holds much more promise because it can maintain the existence of relationships among the splitting or merging denominations, most of whom after all still were part of the larger denominational family (e.g., Baptists). But GIS does not accommodate graph databases easily, at least for this purpose, so we had to enlist appropriate experts to reach a conceptual solution, which we intend to implement after the end of the NEH-funded portion of the project. Even so, the path was not as straightforward as it seems because computer scientists who are expert in graph databases often have little experience with GIS and spatial data. The result was less than what we had hoped: we now understand more completely the complex set of issues involved in developing the desired tool.

Comment: The issue here was not a lack of familiarity with the data but rather a failure to appreciate its complexity for the feature we sought to add. It is a straightforward task to map the split or merged denomination within a GIS; it is quite another thing to manage the continuing relationship of groups to each other. We might have solved the problem more quickly had we surfaced it sooner and tested our assumptions about the data more thoroughly and earlier in the project. The irony is that we used the data from the American Religion Data Archive, the best documented source of the data, in part because it recorded the splits and mergers. But we learned to our dismay that the ARDA sources are not coded consistently from census to census. Here, we relied too much on its documentation and on our discussions with its expert staff.

Using graph databases within a GIS represents a conceptual leap; it will enhance GIS for humanities research but the difficulties of implementing it within this project exceeded the project scope and budget. If we had the opportunity to re-engineer the project, an early in-house workshop between computer scientists, domain experts, and project staff may have yielded an implementable solution, but even then it is doubtful the project budget would have supported the long-term engagement of computer scientists or their

advanced students, who have their own agendas and whose schedules did not easily correspond to our needs.

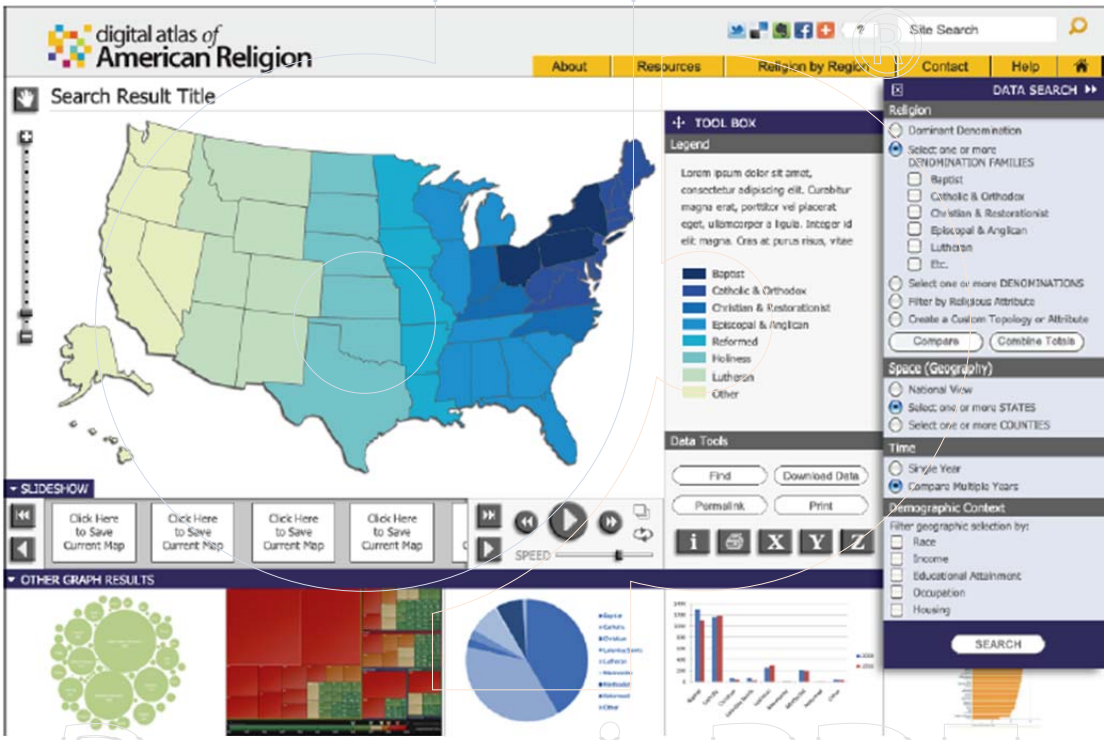
The larger issue is how much investment of time and resources is reasonable as preliminary work prior to submitting a proposal. We did our due diligence in planning the project: we knew the data and had worked with it within a traditional GIS framework for over a decade, and we understood the nature of the problem we were trying to solve. But we did not know, nor could we have anticipated, the vagaries of the data when applied to the problem, which sought to add functionality not accommodated easily within GIS. And we likely could not have discovered the problem without doing the detailed analysis funded by the project.

2. State and county boundaries change over time during the history of the US, and the project benefitted enormously from the NEH-funded Atlas of Historical County Boundaries (Newberry Library), which furnished a complete set of shape files for the 20th century. Even so we had to do extensive programming to reference the appropriate boundary file closest matching the year of the data, to handle cases when no data are available for a particular county or state, and to distinguish territorial data appropriately.

   Comment: **T**he extraordinarily careful and professional job done by Newberry Library staff was invaluable, but we discovered that our other data sources, especially the various religious censuses maintained by the American Religion Data Archive at Penn State University did not always have geographic identifiers consistent with the Newberry data. Part of this was a timing issue. Newberry was still finalizing its product when our project started. The final historic boundary files released in late 2012 contained all of the identifiers we needed. But the larger issue occurred when linking the boundary files with the data, which revealed a number of data problems that were not apparent until we visualized them. Because this step occurs naturally as a late-stage process, it is likely that even a careful analysis of the data without its geographic visualization would not have revealed the problems we encountered.

   We carefully documented the problems and will inform ARDA of their existence, a step which should resolve the problem for these datasets when used outside the DAAR environment. Still, it raises the likelihood that other datasets (population census, voting records, etc.) may have the same problems, which suggests that users of those datasets with a GIS will need to develop a strategy to identify and correct the problems quickly to avoid distortions in the data representation and analysis—and stay within project scope and budget.

3. This project had the explicit aim of enhancing GIS-derived visualizations to make them more robust and more user-friendly for humanists. We did not want to send users to other software to render the desired visualizations, so we chose to integrate solutions into the GIS and to link them dynamically so that a change in one visual format would be rendered as well in other visualizations. We also sought initially to display all visualizations within a single frame to allow for easy comparison (see image below), a goal we shelved for this phase because it was more complicated to get the visualizations to work synchronously and would take more staff time than was available without sacrificing more important tasks.



An unanticipated challenge was in the selection of tools for the development environment, primarily because we sought open-source solutions whenever possible. At the Polis center, we work regularly with a large array of tools, languages, and platforms, but most of them are proprietary, which we use because they usually are more robust and better supported/documented than open-source software. This project confirmed our standard practice. We used open source PostgreSQL, but we experienced many performance issues with complicated queries. The open-source tools that met our visualization requirements also had performance and integration issues. We also ended up using IBM ILOG Elixir software for the controls, but even this comprehensive toolkit required modifications to accommodate our needs. For instance,

Elixir controls did not consume all the dimensions of our visualizations: cartographic maps use one dimension for rendering the colors, whereas tree maps use two dimensions (one for the color and one for the size). So we had to devise a solution to make all controls reflect the same data across all visualizations.

But this solution, which solved one problem, also created another one. The database and the number of possible visualizations (as maps or other graphics) is very large, and IBM delivered vector files as intermediate products, thus making the task of rendering maps unreasonably slow and cumbersome, especially for county geographies. We developed a two-fold solution. We used intermediate XML between the databases and visualization application, which enabled us to minimize data queries to the database by caching the XML files beforehand. Because XML is good for hierarchical data formats, it allowed us to transfer data and render the visualizations efficiently, although it also means that the addition of new data requires re-caching to maintain performance. We also use a Scalable Vector Graphic (SVG) maps for some geographies, even though in doing so we created an inconsistency in map creation stage, with some maps rendered by IBM Elixir and some by SVG.

Fortunately for the immediate delivery, DAAR consumes all the religious census data used regularly by scholars. The addition of data from the other sources will require re-caching, as will any re-aggregations, so we continue to consider more flexible solutions for this issue. We also will seek a way to resolve the inconsistency in the map creation process. The user will experience nothing different, but the development path forward will become easier when we can use a single process for map creation.

Comment: Within the digital humanities, the use of open-source software has become something of an article of faith that supposes it as good as proprietary software for most problems. We have never subscribed to this idea because our experience in system development suggests otherwise—and, to be fair, we have been able to pass the cost of licenses, etc., on to our clients. But here we sought to do as we had promised in our proposal, namely, use free and open-source software if at all possible. We quickly learned again what we already knew: many open-source products are not well-documented or well-supported. As a result, using them may become much more expensive than project directors assume.

An example may be helpful here: IBM Elixir did not provide us with all the functionality we needed for our controls, so we called the development team, which saw our problem as one from which it could learn. The upshot was that we received both expert advice and significant help in thinking through solutions that worked for us. We also

sought to include an ESDA module (an exploratory spatial data analysis tool allows experimentation with and correspondence of data in both graphed and map forms). It was an impressive open-source product in its demonstration—and freely available to us—but it came without sufficient documentation and the university-based development team, although friendly, was too consumed in their own issues to provide meaningful consultation, offering instead to install the module on our platform for a sizeable fee. It was an offer we could—and did—refuse for several reasons, not the least of which was the instability of the team (a professor and several graduate students).

The other lesson is obvious but deserves attention nonetheless. Even though an impressive array of tools exists to allow us to do various things with our data, it is no easy trick to make them all work together. Often the solution is inelegant and appears to be jury rigged, even when it delivers the intended product. The problem, of course, is that digital humanists often seek to repurpose tools for the data and questions that interest them; because the issues that interest us are complex, we want complex (integrated) toolkits. Developing these tools from scratch is costly and time-consuming, and it runs the risk that the commercial market will introduce tools that quickly surpass the functionality we imagined for any single feature or group of features. But stitching together tools is not easy; it too is costly, even with the use of APIs and other bridging techniques. Much as we might hope otherwise, it is unlikely that this situation will disappear, so digital humanists will need to calculate costs carefully (including direct, marginal, and opportunity costs) to ensure that they take the most efficient and flexible development path.

4. We developed DAAR as a web product for use with the largest number of browsers, but we did not anticipate the rapid shift to tablets—and especially the iPad—as a preferred medium for interacting with the data. Web platform was based on Microsoft IIS, with WordPress and the CMS. We chose Adobe Flash Player as our main visualization delivery technique, which eliminated the need for customizing applications for different browsers, but Flash is not supported by Apple technologies. We propose an HTML 5, JavaScript, and SVG-based solution for the next cycle of development, even though this solution will introduce its own problems because many advanced visualization controls are not built on these technologies.

   Comment: Technology changes rapidly, and we must adapt. The issue is not that DAAR is less valuable—the web expression will have continuing (and greatly enhanced) value—but rather that we must acknowledge and accommodate the shift in the

preferred media (especially for students) for consuming this information. When we make this transition—and how much it will cost—has yet to be determined. Even had we been prescient enough to gauge this seismic shift accurately, we still may have chosen to follow the development path we chose because the technologies required for the longer-term solution (e.g., HTML 5) have only now become robust enough for us to use them.

General Recommendations: Most of the lessons learned in this project can be generalized for use by other digital humanities initiatives; indeed, most of them are already well known to developers and project managers. Rather than repeat what is discussed in the section above, it may be more beneficial to address ways NEH might help to mitigate problems, even if we acknowledge that projects of this sort will always encounter them no matter how well planned or well managed the effort is.

NEH would be well advised to do two things to help its digital humanities projects achieve the most success possible in an uncertain and rapidly changing technical environment:

1. Require that each project begin with a technical review by independent consultants. To be sure, the standard review and award process makes a judgment about the technical merits of the proposal and the ability of the identified team to undertake the project. But this process, including development of the proposal, antedates the beginning of the project by a year or so, and the proposal itself rarely includes a sufficiently detailed technical plan. More important, the reviewers and panelist comments are rarely enough to warrant a full-scale post-award assessment of the technology plan. Such a review at the beginning of the project would accomplish two purposes: it would provide as much assurance as possible that the project team had a sound, practical plan for development, and it would allow consultants to share lessons learned from similar projects with all team members, many of whom may not be as aware as they need to be about these other projects.

2. Require a mid-point technical review by the same set of consultants, with the aim of helping the team to think carefully about the problems it has encountered and to assure NEH that the team is developing appropriate solutions.

This process of continuing review and consultation should be constructive, with NEH supporting the cost and sending a program officer to the review. Consultants should be retained by NEH but chosen in consultation with the principal investigator (and project manager, if these are separate positions). Although the composition of a technical consulting team will vary depending on the aims of the project, it should include individuals with sufficient development

experience and wide knowledge of the available and emerging technologies. It may or may not include a domain expert, although one would be helpful for projects in which domain and technical expertise cannot be easily separated.

One other NEH action would be helpful upon the end of the project. Periodically, NEH should sponsor a technology showcase at which all completed digital humanities projects would be required to exhibit, with attendance by the PI and the lead technical developer. The aim would be two-fold: (1) provide a highly visible public launch of the project; and (2) use the occasion for shared learning among project teams, with these lessons to be distilled and disseminated to the digital humanities community and other interested parties.

David J. Bodenhamer
Executive Director and Professor/Project Director
The Polis Center
Indiana University Purdue University, Indianapolis
intu100@iupui.edu