**Some merit**

**Elmer Ray Block**

**Obama Synthesis Project**

The Obama Synthesis Project is a web-based digital database built from three major sources: (1) books about Barack Obama; (2) books authored by him, (3) books he reads. Clearly the major source for the database is the first set of materials (well over 10,000 items). Though not discussed in detail, the second source would apparently include anthologies of speeches, inauguration addresses, and transcripts of debates; it is much more extensive than simply books authored by Obama. The means of documenting the third is not entirely clear nor is the justification for including this category of material. Block does mention an "important but perhaps inadvertent role" of the president as "Bookseller in Chief" but there might not be a clear way of measuring impact. It might just as well be the case that Obama is reading books that are already popular best sellers. At any rate, the inclusion of these books has a loose connection to the first two groups of material though it sheds light on a very different issue.

Block mentions that the OSP database will enhance "ongoing conversation about Obama's deeds on the campaign trail and in the Oval Office". Not mentioned, though clearly a significant facet of Obama research, is that the nature of the books being written provides a snapshot of the motivated public's perception of the president. This reviewer found on Amazon 47 results from a search on "Barack Obama" and "birth certificate" and another similar set of material containing the president's name and "socialism". The OSP will contain material (metadata and perhaps digitized books) from both scholarly and political scholars and analysts as well as from the popular and provocative literatures on this president.

The project offers a convincing data management plan that will ensure the integrity and accessibility of project results The host institution has committed to maintaining the OSP for at least five years. It is not clear, however, whether this commitment is for the online portal to the digital collection or to the database and other text/image files that might be gathered during OSP development. It would seem that certain archive formats could be easily preserved offline (pdf/A, text and XML databases, .tiff) even if the integrated presentation could not be sustained.

My concerns with this project, concerns that motivated only a Some Merit evaluation stem from three considerations mostly involved lack of specificity regarding certain issues.

First, it is not clear what type of metadata will be developed. Block mentions "a platform for storing, coding, and analyzing the information in Obama books" but this is quite open ended. The OSP, Block notes, will allow users to compare books across authors, publishers and others "because it will be the first freely-available web resource to synthesize digital book information and, as a result, it will provide a new level of access to the texts that helped define the Obama literature". Block and his collaborators will devote six months to "develop tools for codifying the content of Obama books". It is not clear what types of tools will be developed nor what these tools take as input and what they will produce as output beyond the undefined "codified content". And again, while during the final six months a summary of the quantitative research will be made following techniques of "research synthesis," it is not clear what will be quantified.

Second, it is not clear whether the books will be accessible in digital format. The data management plan stated that "we also plan to collect digital facsimiles of Obama books to use in developing and testing the application. The collection will, as permitted by copyright law and licensing restrictions, be made available to the public for download via the OSP project website." Later in the same section: "we do not anticipate any privacy, confidentiality, security, intellectual property of other rights or requirements that will affect my ability to store and disseminate this data". I am not sure what the PI means by proposing to "collect digital facsimiles". As the proposal makes no mention of digitization nor does the budget contain a line item for this process, I can only assume that OSP plans to freely acquire digital versions of the material (over 10,000 items are listed in Amazon!) and run into no legal issues in the free, public dissemination. I am not convinced that this can be done. I have, for example, used Google Books in a pinch; but often the precise pages I need are not accessible, or within a segment of material certain pages will be hidden from view. The copyright issue needs to be directly addressed. In regard to digital material, Block mentions a "vast, unorganized world of Obama books on the World Wide Web, and specialized queries of the text in the OSP database could put the information in thse books into much-needed historical and political context". Here it seems that the OSP is a tool for extracting material that the scholar-user desires from Obama e-books. There is no attempt to quantify how much of the Obama corpus is actually already in digital format.

Finally, though I am not familiar with the software that Block plans to use I did visit the Automation Anywhere website and view the three demos on Web Data Extraction and Web Automation. I am not convinced that this software is appropriate for the task at hand. The data extraction seems best suited to database type material that has already been assembled at one URL and which the user wishes to extract and save in a database or spreadsheet format. Web Automation seems suited to automating repeated tasks at a particular URL. Block does mention the need to both customize Automation Anywhere and to "create new software". I believe there is a typo where the project mentions $5,000 to pay for consultation from Baker Diagnostics as elsewhere the figure is $500. At any rate, the amount designated for software development and testing seems woefully insufficient.

In sum, the project needs greater clarity on the type of metadata and the source and rights to dissemination of digitized book material. More specifics are also required on the type of tagging that will facilitate user-defined recombinations of the primary source material. The budget dedicates insufficient funds to both under/graduate student support and software design (both database and graphic user interface).

**Overview**

Intellectual significance: As the nature of the metadata and the extent to which digitized books will be accessible and searchable is not clear, it is difficult to judge the intellectual significance of this project.

Impact on research and technology: The same comment as in the preceding paragraph applies.

Innovation: The innovation would be in the sheer number of resources available, though again it is not clear how the online material will be presented.

Proposal development: The project is well articulated.

Feasibility: The vast amount of materials create problems for the realization of this project; there is no prioritization of materials. Nor is the copyright issue explored in the necessary detail.

Qualifications: The PI is well qualified for the task at hand.