

Toward a Collaborative Model for Social Science Cyberinfrastructure Development: Lessons from the Digital Docket Project

Wayne V. McIntosh¹, Michael C. Evans²

¹ University of Maryland, Department of Government and Politics

² University of Maryland, Department of Government and Politics

Email address of corresponding author: wmcintosh@gvpt.umd.edu

Abstract. We argue – based on our experience with the “Digital Docket Project,” a collaborative NSF-funded research initiative that seeks to contribute to the development of cyberinfrastructure at the intersection of law, social science, the humanities, and information science – that essential elements of social science cyberinfrastructure can be enhanced when social scientists and information scientists engage in collaborative partnerships. We draw from our experience to (1) explain exactly why we think this is the case (and under what conditions); (2) delineate the major obstacles preventing scholars from forging collaborative relationships of this sort; and (3) make specific policy recommendations for overcoming these obstacles and realizing the benefits for social science cyberinfrastructure development to be gained by such interdisciplinary collaborative projects.

Introduction

In this paper we discuss the opportunities and challenges attendant upon an NSF-funded research initiative – the “Digital Docket Project”¹ – a primary objective of which is to contribute to development of cyberinfrastructure at the intersection of law, social science, the humanities, and information science². For the purposes of this paper, we take “cyberinfrastructure” to denote the layer of information, expertise, standards, policies, tools, and services that are shared broadly across communities of inquiry but developed for specific scholarly purposes: cyberinfrastructure is something more specific than the network itself, but it is something more general than a tool or a resource developed for a particular project, a range

1 NSF, Human and Social Dynamics, grant #BSC-0624067 / 2006-2009 / Wayne V. McIntosh, Principal Investigator. The Digital Docket is a collaborative project between the University of Maryland’s Department of Government and Politics and College of Information Studies, and the Department of Political Science at Towson University. See: <http://www.umiacs.umd.edu/~digidock/>

2 In this paper, we use the term “information science” to refer broadly to those fields devoted to primary research on digital text storage, processing, and analysis, such as computational linguistics, information retrieval, and (their intersections with) human-computer interaction.

of projects, or even more broadly, for a particular discipline. (Courant, Fraser, et al.; p. 8)

The immediate scholarly objectives of the Digital Docket Project are to analyze the dynamics of law as promulgated by the Supreme Court of the United States (SCOTUS) while advancing the state of the art in computational linguistics, information retrieval, and human-computer interaction theories and technologies. The expected broader impact of this project, however, is to generate and develop a document database (conforming to the latest standards of text encoding and archiving) and text analysis tools and that will not only enhance legal research by academics, practicing attorneys and judges, but also better enable researchers to extract, discover, visualize, interpret, and/or analyze relevant information from the growing body of digitized texts available throughout the natural sciences, social sciences, and humanities, as well as those conducting non-academic research in both the public and private sectors.

We believe our experience with the Digital Docket Project vindicates the core argument of this paper, which is that the key to cyberinfrastructure development in the social sciences is the promotion of greater dialogue and collaboration between information scientists and social scientists. Such relationships can be highly efficient and productive while at the same time fostering a virtuous creative cycle, where increased awareness by social scientists of the *capabilities* of information technologies interacts with an emergent understanding by information scientists of the particular research *needs* of social science research. As social scientists gain greater awareness of the capabilities of digital technologies, they can develop research questions that they otherwise would be unlikely to consider, and information scientists, in turn, are led to develop useful new tools that would not have been conceived absent collaboration. Interestingly, the process has the additional benefit of increasing the knowledge by all participants of both the cyberinfrastructure capabilities already available and the benefits to be gained by innovation. When participants are cognizant of the potential for their work to contribute more broadly to the development of cyberinfrastructure, and are willing and able to take the necessary steps, such activities have the potential to spawn new projects that push forward the “layer of information, expertise, ... tools, and services” that constitute the core of cyberinfrastructure.

In the first two sections of this paper we discuss how the Digital Docket’s collaboration between political scientists and information scientists has facilitated such a “virtuous creative cycle” with reference to our two major first-year endeavors: creating the Digital Docket’s U.S. Supreme Court Text-Centered Database and the Digital Docket Explorer tool. In the second part of the paper we reflect upon our experience to consider obstacles to forging beneficial collaborative interdisciplinary relationships and propose policies that could foster an environment more conducive to widespread collaboration

Cyberinfrastructure Development, Case I: Lessons from Building the Digital Docket’s U.S. Supreme Court Text-Centered Database

In the first part of this section we overview the U.S. Supreme Court Text-Centered Database, the purposes it serves for the Digital Docket project, and its projected broader impact on SCOTUS digital scholarship. Then, in the second part we delineate several lessons for cyberinfrastructure development illuminated by our experience thus far.

Overview of The Digital Docket's U.S. Supreme Court Text-Centered Database

Upon completion (by Summer, 2008), the Text-Centered Database will consist of a publicly accessible, online collection of all available SCOTUS opinions (~12,000), briefs (~50,000), and oral argument transcripts (~8000) from all SCOTUS cases from the beginning of the Warren Court (1953) through the most current term. The immediate purpose in creating this database is to facilitate our analytical work. Although many Supreme Court document databases exist, they all lack two features necessary for conducting the large scale automated text processing that we have proposed to do with the Digital Docket Project. First, none are organized according to the case and document characteristics of greatest interest to social scientists, such as authorship, voting blocs, ideological direction of opinions, and law and society issue categories. Instead, all extant text collections are designed to help lawyers and judges argue and decide relatively narrow legal questions. Thus, the most sophisticated legal text collections – Westlaw and Lexis Nexis – provide elaborate *legal* issue coding schemes and citation analysis algorithms to enable finding cases *most relevant* to specific legal questions. They do not allow for queries designed to quickly access *all* opinions from cases (and / or their associated briefs and / or oral argument transcripts) that have in common *extra-legal* characteristics, such as those listed above. To overcome this shortcoming of existing collections, we are in the process of linking the documents in our corpus to a variety of variables of relevance to SCOTUS scholars. Although we will eventually generate our own metadata, at this point we have begun by linking the documents to variables from four datasets already compiled, and widely used, by political scientists: (1) Harold J. Spaeth's (2007) "Original United States Supreme Court Judicial Database;" (2) James L. Gibson's (1996) "United States Supreme Court Judicial Database, Phase II;" (3) the supplemental metadata provided by Vanessa A. Baird in her (2007) "Merged Phase I and Phase II Supreme Court Database;" and (4) the Epstein, Walker, et al. (2007) "U.S. Supreme Court Justices Database."

Second, even if a scholar were to go through the laborious effort of acquiring (through existing sources) all the documents required for large scale computational analysis, the scholar would still be left with the task of pre-processing the documents so that (1) its textual *content* is formatted appropriately for the analysis and (2) its *metadata* are associated with the text in a manner compatible with the software application that will be used to conduct the analysis. Since we are using a variety of applications, some off-the-shelf and some (such as the Explorer Tool) that we are developing ourselves, and since they all have distinct formatting requirements and ways of relating metadata to text, we are designing our database to allow for (1) instant batch downloading of all documents (2) in a variety of formats (3) with full control over the exclusion or inclusion (and location) of document elements, such as headings, footnotes, quoted sections, and citations, and (4) the ability to associate available metadata to the text content in a variety of ways, such as tagged and imbedded within the document files themselves or in separate files with unique identifiers to link them. This gives us maximum flexibility for focusing our analyses on specific text sections of theoretical relevance, and for analyzing our corpus with a variety of tools and techniques available now and into the foreseeable future.

While the immediate purpose for creating the Text-Centered Database in this manner is to enable rapid document acquisition, formatting, and pre-processing for the array of analyses we will conduct in the final two years of the project, we anticipate that the Text-Centered Database will impact digital scholarship more broadly in several ways. For one thing, we expect it to enable and (thus) encourage other SCOTUS scholars – law professors, political scientists, historians, sociologists, and so on – to incorporate digital text analysis techniques

into their research. Furthermore, in building this database, we have been guided by the “principle of maximized adaptability:” we have gone to great lengths to assure that it is organized in such a manner that it can easily adapt to changes in encoding standards; evolution in the functionalities and basic architecture of computer assisted qualitative data analysis software (CAQDAS) applications; and advances in the theories and tools developed by information retrieval, data mining, and human-computer-interaction researchers. As new generations of SCOTUS scholars begin their research careers, we should expect that they will be adept at applying text analysis tools to digital collections; our database should be adaptable to the times and, thus, address their evolving needs. In consideration of likely future innovations, we have laid the groundwork for integrating a host of “Web 2.0” applications that enable the creation and analysis of dynamic user-generated content. Although not within the scope of our present project, a future Web 2.0-enabled Supreme Court Text-Centered Database could be a rich pedagogical and research tool for K-12 teachers as well as instructors and scholars at the college level.

An additional way that we expect our database to contribute to cyberinfrastructure development is by providing an example to social scientists of the virtues of (1) storing data in relational databases (as opposed to the two-dimensional statistical tables or spreadsheets normally used) and (2) linking interpretive codes of digital text directly to their original source. From the standpoint of maximized adaptability, there is really no reason that social scientists should not generally follow both examples. A relational database frees data-compilers from making path-dependent decisions about the unit of analysis or data format that a dataset will follow. If compilers were to take this approach, end-users would be able to decide what unit of analysis and format is most appropriate to their individual project needs. Perhaps more import, by using a relational database such as MySQL, compilers can link primary text data to the variables (or metadata) in the dataset. This can be beneficial for at least two reasons. First, as with our database, researchers can conduct analysis on the text data directly. Second, it promotes greater transparency, and, thus, should reduce bias and other types of error in variable coding by giving all users of the dataset convenient access to the primary source data. Such practices should become standard in social science research, but, regrettably, they seem to be rarely if ever followed, probably in no small part because acquiring knowledge about relational database design and usage is not yet an established professional expectation. We hope our efforts will encourage more social scientists to use relational databases when compiling new datasets and to link to primary source data when possible.

Lessons for Cyberinfrastructure Development

We think our experience with this database – from its conception three years ago to now (after one year of implementation)—points to three important and interrelated lessons for how to promote the creation of cyberinfrastructure resources such as this for digital scholarship in the social sciences. The first is admittedly quite obvious: the database would not come into being – or, at least, not in a form conducive to cyberinfrastructure needs – absent external financial support. Without funding, we would have been forced to do many things differently, and our project would have significantly lower input potential for infrastructure development. For example, we most certainly would not have adhered to the “the principle of maximized adaptability” if not for the funding we received, as doing so does not serve our immediate purpose in producing analytical results. Just to create a functional database conducive to our immediate research needs has been a daunting task. Over the past year we have had three political science graduate students and ten volunteer undergraduate research assistants devoted to such seemingly simple tasks as collecting the documents and

relating them to extant metadata. These students have poured hundreds of hours of work into this aspect of the project, and it is highly unlikely that we would have achieved even these relatively modest objectives absent funding. But the funding from NSF for our project—coupled with the expectation that our project should (in part) contribute to cyberinfrastructure—enabled and led us to consider the broader potential for our project to contribute to digital scholarship on SCOTUS. Our commitment to a database design that will assure its long-run adaptability emanated from our frustrations with attempting to conduct large-scale computational linguistic studies with currently available databases. Although some of these are state of the art text collections, they are all rigidly designed for narrow purposes. Charged with the responsibility of committing some of our resources to cyberinfrastructure and desirous of creating a flexible state of the art database, we decided to solicit assistance from the Maryland Institute for Technology in the Humanities—a research center on the University of Maryland campus staffed with humanities scholars and computer scientists—since they have substantial digital curation experience and are leaders in the promotion of cyberinfrastructure in the humanities. As we discuss below, this has proven to be a beneficial partnership. But the present point is that given the incentive structure faced by political scientists, we most certainly would not have taken these steps if not for the support and leadership provided by the NSF.

This brings us to the second and third lessons we have drawn from our experience building the database: collaboration between social scientists and information scientists to create digital text collections promotes cyberinfrastructure development by more efficiently allocating energy and expertise (second lesson) and raising awareness among social scientists of the capabilities, standards, and possibilities of digital curation (third lesson). Both lessons, therefore, point to the importance of promoting collaboration between social scientists and information scientists, albeit for slightly different reasons. The former emphasizes the economic dictum that efficiency is promoted when trading partners focus on that to which they have a comparative advantage. By collaborating, information scientists are able to focus on what they do best and social scientists can focus their time and energy on social science. Little is gained when social scientists devote their energy and attention to the nitty-gritty technical aspects of digital curation. In general, they lack the intimate knowledge of the standards and techniques that go into such endeavors and, more importantly, time spent acquiring expertise in digital curation is time not spent developing expertise in a field of substantive political interest. When social scientists attempt to “play information scientist,” they run the risk of re-inventing the wheel or, worse, making one that is not round. Our collaborative experience with the College of Information Studies and with MITH illustrates this well. The political scientists on our team had no experience with digital curation when the project began, but did have extensive knowledge of the datasets (essentially consisting of metadata) commonly used to conduct statistical analysis on judicial behavior. Furthermore, we had previously worked with existing text collections to acquire about 1500 documents in order to conduct computational linguistic and information retrieval analyses on opinions and briefs from a subset of cases from 1978 to 2005. We therefore had a good idea of the features an ideal Text-Centered Database would have, while the information scientists understood well the technical ins and outs in constructing a useful database. In our case, it was clearly more efficient (at least in the short run) to maintain specialization and a robust division of labor. Due to this organizational structure, our partnership with MITH has produced a database of a quality unthinkable when we conceived the project. Every element of the database construction process – including document parsing and tagging; relating text to metadata; importing data and metadata into a MySQL database; and building a user-friendly web-based interface—has been enhanced by this efficient division of labor.

Paradoxically, however, creating an efficient inter-disciplinary division of labor requires that both sides have basic literacy of the other's field; otherwise, effective communication is impossible. We could not have proceeded if political scientists did not learn the basic concepts behind, and possibilities of, digital curation. Likewise, the information scientists could not have progressed effectively if they did not come to understand the concepts behind the metadata well enough to create an appropriate database structure, or if they had not learned the Supreme Court decision making process well enough to know how to parse documents and link them properly to the metadata. So, while it is inefficient for either partner to attempt to *acquire expertise* in the field for which the other is already expert, it is essential that both partners *acquire basic literacy* of each other's field. Basic literacy is a very different thing from expertise. Indeed, the former is much less costly to acquire. Furthermore, basic awareness is necessary for social scientists to collaborate effectively with information scientists and to be good "end-users" of digital text collections. Beyond a certain point, however, the value of greater knowledge about digital curation reaches a threshold. In sum: *basic awareness* of information science on the part of social scientists is necessary but insufficient for the development of cyberinfrastructure in the social sciences, whereas the development of expertise among social scientists is unnecessary and insufficient. Efficient productivity is optimized, therefore, when this proper balance between specialization and cooperation is met.

This relates to the third lesson, which emphasizes the importance (for cyberinfrastructure development) of basic literacy, or awareness, by social scientists of the capabilities and standards of digital curation. Whereas the efficiency of a collaborative project is enhanced by the basic information scientific literacy on the part of social scientists, it is also true that collaborative projects will tend to promote such awareness and that such awareness in turn is a prerequisite for social scientists to become active contributors to the development of cyberinfrastructure. Indeed, by our working definition, the existence of such awareness can itself be considered an important part of the "the layer of information" of which "cyberinfrastructure" partly consists. The collaborative structure of the Digital Docket project led our political scientists to enter into a new "epistemic community." This experience has gradually broadened our perspective on the possibilities of digital curation, as well as other ways information technology can enhance social science research more broadly. This experience has instilled in us an appreciation for the benefits of using relational databases to create datasets and, when possible, linking primary source text data to them; the importance of creating a flexible database structure; and the promise of eventually establishing a Web 2.0 interface that enables dynamic user-generated content. We think similar awareness-raising experiences will be attendant upon any serious collaborative effort between social scientists and information scientists.

Cyberinfrastructure Development, Case II: Lessons from the Creation of the Digital Docket "Explorer Tool"

Overview of the Digital Docket's "Explorer Tool"

The other major first-year project component that contributes to cyberinfrastructure is the development of an exploratory search tool called the Digital Docket "Explorer Tool." At a broad level, this part of the project aims to construct innovative interfaces to help scholars in the humanities, social sciences, and law explore large text collections. We think of it as developing "creativity support systems" for scholarship that will augment research activities—by utilizing the tools to visualize relationships between different texts, to drill down into a text, to draw connections otherwise not apparent, and most important of all, to

form hypotheses that provide the basis for further exploration. Since computers by themselves do not generate knowledge, we believe a system's primary role should be to facilitate the processes of hypothesis formulation, evidence gathering, and so on. Information retrieval and related text processing techniques (data mining, linguistic analysis, text classification and clustering, and the like) can discover characteristics of texts and connections among them that may be of interest to scholars. The Explorer Tool is one such creativity support system. It incorporates two novel features. First, it attempts to combine data visualization technology with standard search functionality to enable researchers to track "influence" over time by visualizing lexical and semantic patterns and trends over time in an intuitive manner. Second, it is designed to place equal emphasis upon *content*, or what text segments are *about*, and *metadata*, or the *context* that defines and situates the text. The immediate purpose of the Explorer Tool is to assist in gaining insight into the roles and influences of various actors in the judicial process, through analysis of written records, such as briefs written by litigants and other interested third parties, and opinions written by justices. However, through the collaborative process of developing and applying the tool, we have come to see its more general applicability for tracking concepts and/or ideas in various types of corpora over time. Indeed, we are now deliberately constructing it to be flexible enough to assist with the creative interrogation of medical records, late 18th century American historical documents, and archived blog pages, to name only a few possible applications.

Lessons for Cyberinfrastructure Development

We think our experience in creating the Explorer Tool points to a good model for promoting the development of analytical tools to support social science digital scholarship. The benefits of collaboration between social scientists and information scientists are especially apparent for the digital text analysis tool aspect of cyberinfrastructure. For information scientists to apply their technical expertise to building such innovative tools, at least three conditions must be met. First, they need the incentive to do it. This, again, points to the importance of funding opportunities, as discussed above. Second, they require specific research needs by practicing scholars on which to focus their attention. Third, they require a stock of relevant theoretical knowledge from which to draw to address the problem. Our experience with the Explorer Tool demonstrates that when information scientists work with social scientists to develop tools designed to address the latter's research needs, the second condition is met and the third can be promoted. Let's consider each in turn.

Our collaborative effort to develop the Explorer Tool began with a specific substantive research question: What influence do *amicus curiae* briefs have on the SCOTUS' written opinions? After experimenting with purely automated (clustering) methods, we found that the best approach would be one that combined fuzzy search with visualization. This would allow us to view "legal memes" – i.e. lexical similarities – in different types of documents (briefs or opinions) over time. Coming to this realization was itself a product of collaboration as it required both the information scientists' sense of the technical possibilities and the political scientists' feedback on whether different approaches actually were useful to addressing the research question. We discovered the optimal tool for the task by going back and forth in this manner. This also demonstrates how close collaboration can broaden the scope of social scientists' research questions. In this case, the political scientists had initially assumed that the focus would be on the influence of *amici* on the opinions from a single case. The information scientists instead demonstrated how well the incidence of different memes in a variety of documents over time can be visualized. This has led to entirely new (and better) research questions regarding trends and origins of change in legal language over the course of many decades. Once fully developed, the tool will be sufficiently adaptable to address any

collection of texts of interest to researchers across academic disciplines. We think the initial focus on a specific and narrow research question helped (paradoxically) to produce a quality tool of broader application by encouraging the information scientists—whose primary concern is theoretical—to pay close attention to practical details that are necessary for useful research-support tools.

In the long run, collaborative projects such as this can contribute to cyberinfrastructure not only by producing specific tools, but also by contributing to the stock of theoretical knowledge from which future information scientists can draw to develop other tools. This is not a necessary result of collaboration, but steps can and should be taken to make this outcome more likely. Our experience suggests that social scientists can play an important role in bringing this about. One necessary condition is a willingness by social scientists to engage in the give-and-take necessary to make sure the collaborative project serves the long run research interests of information science, in addition to their own immediate needs. For our project this was something of a negotiation, since there was also danger from the opposite direction: that the collaboration will be one-sided in serving the interests of information scientists without producing a tangible benefit for political science research. Ultimately, our partnership was marked by enough interdependence and mutual respect that we were able to settle on mutually beneficial terms: the information scientists would produce something useful to us in return for our patience as they proceeded in such a manner as to maximize the theoretical knowledge on information retrieval and human-computer interaction gained by the process. A computer science doctoral student has actually made the development of this tool, and observation of its interaction with end-users, the focus of his dissertation.

The other way that social scientists can help to assure that the collaborative relationship results in a contribution to the theoretical stock of information science is to give feedback and and/or participate in controlled experiments that enable information scientists to glean insight about the strengths and weaknesses of various functions and interface designs. One of our political science graduate students is working on a dissertation project that requires careful study of the works of Alexander Hamilton, Thomas Jefferson, and James Madison. He is using the Explorer Tool to analyze that corpus while at the same time allowing our computer science graduate student to conduct a “multidimensional in-depth long term case study” (MILCs) of his usage of the tool. Through this process, both are contributing to the body of theoretical knowledge in the computer science (human-computer interaction) subfield of “Document Collection Visualization.”

Obstacles to Beneficial Collaboration and Policy Recommendations for Promoting this Collaborative Model for Social Science Cyberinfrastructure Development

So far we have attempted to delineate what we see as (1) the potential benefits for the development of cyberinfrastructure in the social sciences to be gained by forging collaborative relationships between information scientists and social scientists, and (2) specific conditions that are necessary to increase the likelihood that such collaboration will indeed contribute to cyberinfrastructure development. To sum-up, so far we have argued that:

1. Any project, no matter how it is organized, is more likely to contribute to cyberinfrastructure if (a) it is funded (b) with the requirement that it produce useful

digital text collections (conforming to latest standards) and/or digital text analysis tools;

2. Collaborative relationships between social scientists and information scientists can contribute to cyberinfrastructure development by creating an efficient division of labor, provided that both parties are sufficiently literate in the other field to be able to communicate effectively *and* each focuses on that to which they have a comparative advantage (i.e. the substantive area of inquiry by social scientists and the technical information technology aspects by information scientists);
3. Although it is not necessary (and potentially counterproductive) for social scientists to acquire expertise in the technical aspects of digital curation and analytical technology development, an important aspect of cyberinfrastructure is the widespread basic literacy and functional competencies of social scientists so that they are able to (a) *use* the digital collections and analytical tools that are produced and (b) *contribute* to the development of those collections and tools as participants in collaborative partnerships with information scientists (consistent with the conditions described in (2) above);
4. Collaboration has the added benefit of contributing to the development of that basic awareness and functional competency of social scientists by putting them into the information scientific “epistemic community”;
5. Information scientists are more likely to create useful analytical tools if they work closely with social scientists in addressing the specific IT-related research needs of the latter. In many cases, tools initially created for specific purposes can be adapted so that they are more generally applicable;
6. These collaborative partnerships can contribute to cyberinfrastructure development in the long-run by adding to the body of theoretical knowledge (of database and analytical tool design) from which future information scientists can draw;
7. Social scientists can increase the likelihood that the process of making analytical tools (for specific research purposes) will make that contribution to the stock of information scientific theoretical knowledge by being willing to help information scientists take steps necessary to glean theoretically important insights from the development process; and
8. An especially beneficial step information scientists can take is to observe and study how social scientists use the tool(s) to conduct their research so that they can discover the strengths and weakness of different functions and interfaces.

While our experience has led us to believe these propositions about the benefits of collaboration for cyberinfrastructure development, it has also made us cognizant of the challenges that need to be met if these relationships are to become more widespread. The remainder of this final section is divided into two parts. First, we will overview what we see as the primary obstacles to the creation of collaborative relationships that contribute to cyberinfrastructure development. Second, we will make specific policy recommendations for promoting beneficial collaborative relationships in light of these obstacles and our discussion above.

Obstacles to Collaboration

The bounded nature of academic disciplines in most universities creates artificial distance, particularly between the social sciences and the “hard” sciences, which are housed in distinct

colleges and often located in physically dispersed quadrants on the campus. This reduces the possibilities for local communications, except for distribution of mass mailings, thus diminishing the probability that researchers from different quadrants will be exposed to theories and approaches outside their own disciplines that might be transferable. This also means that in our highly balkanized universities, there is currently a significant “social distance” for social scientists and information scientists to traverse in order to forge mutually-beneficial collaborative relationships that, as an indirect result, promote cyberinfrastructure in the manner described in the sections above. We thus argue that this social distance, coupled with an incentive structure that discourages attempts to overcome it, constitutes a significant constraint to cyberinfrastructure development. Here, we briefly unpack what we see as the underlying causes of, and incentives that perpetuate, this "social distance" between social scientist and information scientists.

We think a primary cause of the social distance between social scientists and information scientists is deeply rooted in the culture of the modern university. Social scientists simply are not socialized to think of their occupation as involving collaboration with information scientists. This remains so despite the fact that we are well over a decade into the information revolution and that information technology is rapidly and radically transforming both the possibilities for research and the nature of society itself. Due to cultural embeddedness, most senior faculty are simply training the next generation of researchers to carry-on as they did, using what are often archaic data collection and storage techniques, treating computers essentially as glorified typewriters and adding machines, holding-on to the outmoded distinction between research oriented around the written word and that which utilizes computational and statistical methods, ignoring the next wave of change (Web 2.0) that is rapidly approaching, and, most fatefully, re-inculcating the great deceit that disciplinary autarky is beneficial and that building collaborative relationships outside the discipline is unthinkable if not shameful. We think this is a tragic state of affairs, both for cyberinfrastructure development and the quality and relevance of social science research.

Another cause of this social distance becomes manifest in the rare instances that social scientists overcome engrained habits and self-limiting identities and actually consider forging collaborative relationships with information scientists. Now the problem becomes primarily one of communication: In order to effectively communicate, both the social scientists and information scientists must bear the cost of acquiring basic awareness of core concepts of each other's field of expertise, as well as the technical terminology peculiar to each. For many researchers, this is prohibitively costly, especially among untenured faculty who are under such immense pressure to publish that they cannot afford to take time to learn the language of another discipline.

This brings us to the role of the academic incentive structure in reinforcing the social distance between social scientists and information scientists to the long-run detriment of both cyberinfrastructure development and quality social science research. Two disincentives to collaboration affect senior and junior faculty alike. One is the paucity of publication opportunities for work that draws from information science. Similarly, and even more pronounced, is the complete lack of rewards (tenure, promotion, rankings, status) for creating digital text collections and/or datasets that conform to high encoding and database standards. Perhaps the most perverse condition of all, however, is the fact that the persons who have the highest stake in, and (typically) the most IT competence to contribute to, the development of social science cyberinfrastructure – young faculty and graduate students – are the ones most ensnared by the pressure to conform to institutional conventions and face the most intense publication requirements, and, therefore, are the least likely to attempt to make a significant contribution to social science cyberinfrastructure. One would think that much progress could

be made in cyberinfrastructure development if these persons could count on rewards, instead of nearly certain punishment, for attempting to branch-out and work with information scientists. At present, it is simply too risky a strategy.

Policy Recommendations

What, then, is to be done? There are, of course, limits to what can be done to overcome deeply ingrained institutional and psychological resistance to change. However, we nevertheless think NSF can do many things to mitigate the effects of these obstacles, to narrow the social distance between social scientists and information scientists, and encourage high quality collaborative relationships that will simultaneously promote cyberinfrastructure development and enhanced social science research. Here we offer four suggestions:

1. Target grants to project proposals that involve collaboration between social scientists and information scientists.

The thrust of our argument above has been that such collaborative relationships (1) have the potential to result in beneficial contributions to social science cyberinfrastructure and (2) are most likely to reach fruition with external funding. All things held equal, an NSF-funded project that involves a collaborative component of this kind is more likely to contribute to cyberinfrastructure than one that does not, for all the reasons stated above. However, our propositions above also suggest that further steps – such as requiring adherence to high encoding and other state of the art curation standards – will increase the likelihood that such partnerships will so contribute.

2. If a project proposal does not include a collaborative element, but does involve the creation of a dataset and/or digital text collection, the product should be made publicly accessible and be developed as a relational database (preferably a free application like MySQL) encoded according to the highest standards.

Among other things, this could have the desirable effect of encouraging researchers to branch-out and seek assistance from information scientists and/or digital humanities colleagues. The project would thus become a collaborative endeavor and have the desirable impact described above. NSF could also offer extra money to such projects as a reward for choosing to collaborate with information scientists.

3. Reward (with special funding opportunities) departments, colleges, and/or universities that take steps to address the counter-productive incentive structure described above. Perhaps special status might be afforded to units where contributions to cyberinfrastructure (e.g. improved data management, creation of digital text collections, adherence to encoding standards, creation of IT-enabled research tools, efforts at spreading IT knowledge and competency through student and faculty outreach and/or through integrating it into the curriculum) count favorably and significantly for tenure and promotion.

This could have a dramatic impact on cyberinfrastructure development. Graduate students and junior faculty would find that reaching-out to information scientists could assist, rather than derail, their long-run career goals. Given that each generation inevitably has greater IT competency than their predecessors, this will mean that those most capable of contributing to cyberinfrastructure development will be given the incentive to do so.

4. Offer grants to encourage campuses to create model "eSocial Science Centers" that (1) promote collaborative research between social scientists and information scientists and (2) provide training and education of faculty, graduate students, and undergraduates on the skills, knowledge, and issues related to the intersection of information science and social science research.

We think something along these lines has perhaps the greatest potential to promote social science cyberinfrastructure development. Such centers would constitute a permanent linkage between social scientists and information scientists, thus creating a structural remedy to the social distance problem discussed above. The proliferation of such centers could in itself be considered an integral layer of cyberinfrastructure, as they would serve not only as incubators for innovation and research, but also as educational and socialization institutions on campuses, thus assuring that next generation researchers have access to state of the art technical developments and are able to acquire the competencies for integrating current tools and standards into their work. Furthermore, it assures information scientists ready access to practicing researchers, thus significantly increasing their opportunities to focus on specific research needs and to have access to scores of willing test subjects. As discussed above, we experienced first-hand the beneficial synergistic effect created by the close collaboration between scholars and information scientists in our digital *humanities* center (MITH) at Maryland. Such humanities centers could be a good beginning model for the social sciences, although the latter would have many additional digital scholarship needs beyond the curation and interpretive tools emphasized by the former. While they would certainly be dedicated to those objectives, one could envision eSocial Science Centers casting a wider net, bringing together social scientists and information scientists to address the future of statistical dataset creation and management; the possibilities for next-generation approaches to simulation and modeling; disseminating information about, and exploring new directions for, Computer Assisted Qualitative Data Analysis Software (CAQDAS), Geographic Information Systems (GIS), and other such research technologies; and addressing issues of transparency and reliability in the application of automated content analysis and other machine learning applications to the array of large social scientific corpora that will continue to be created at an increasing rate well into the foreseeable future.

Conclusion

Although there are limits to what can be accomplished by public policy alone, we see great potential for promoting social science cyberinfrastructure development through NSF policy. At the heart of our policy proposals is the conviction, derived from our experience, that with the right mix of specialization and interdisciplinary cooperation, social science cyberinfrastructure in the broadest sense can be optimally promoted. With policies designed to overcome the social distance between disciplines, and the counter-productive incentives reinforcing that distance, social science can move more quickly toward the promise of the information revolution by making better use of, and becoming a creative force in the development of, social science cyberinfrastructure.

References

- Baird, Vanessa A. (2007). *Merged Phase I and Phase II of the United States Judicial Database*. <http://sobek.colorado.edu/~bairdv/research.html>
- Courant, Paul. N; Fraser, Sarah E.; et al. (2006): *Our Cultural Commonwealth*. Report by the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences. <http://www.acls.org/cyberinfrastructure/cyber.htm>.
- Epstein, Lee; Walker, Thomas; et al. (2007). *The U.S. Supreme Court Justices Database*. <http://epstein.law.northwestern.edu/research/justicesdata.html>.
- Gibson, James L. (1996) *United States Supreme Court Judicial Database, Phase II: 1953-1993*. ICPSR version. Houston, TX: University of Houston [producer], 1996. Inter-University Consortium for Political and Social Research, Ann Arbor, MI [distributor], 1997. <http://webapp.icpsr.umich.edu/cocoon/ICPSR-STUDY/06987.xml>.
- Spaeth, Harold J. (2007). *The Original United States Supreme Court Judicial Database*. Distributed by the S. Sidney Ulmer Project for Research in Law and Judicial Politics. University of Kentucky. <http://www.as.uky.edu/polisci/ulmerproject/sctdata.htm>