Proposal
Responding to NSF 98-63
Digital Library Initiative-Phase II Program, FY 1998

An Operational Social Science Digital Data Library

Harvard University
Cambridge, MA 02138
July 15, 1998

Principal Investigators:
Gary King, Department of Government
Sidney Verba, Department of Government and Harvard University Library

Co-Principal Investigators:
Micah Altman, Harvard-MIT Data Center
Nancy Cline, Harvard College Library
Dale Flecker, Harvard University Library

Summary

This is a proposal to develop a Virtual Data Center (VDC) for quantitative social science data. The VDC will make a vast amount of social science data available to a wide range of types of users, varying from experienced researchers seeking data for advanced research, to undergraduates writing term papers, to citizens seeking numerical answers to specific questions. The VDC will not only make data available for use, it will provide technical and organizational means of capturing new data-sets for the scholarly community and, thereby, provide a live, growing, and evolving resource.

Our current prototype, now in operation at Harvard and MIT, implements a production special-purpose digital library service for social science data. It automatically, fills orders, delivers data, enables exploration, subsetting and conversion of data, and sychnronizes our holdings with remote holdings at ICPSR.

With NSF's help, we will take this to the next level: (1) Generalizing the software infrastructure and interfaces, and adding an entire middle layer. (2) Using formative user studies to refine and extend the interfaces (3) develop a modular design which allows free alternatives for all major software components

These changes allow us to directly accomplish the following: (1) link together multiple (distributed) collections of social science data, each using our system, (2) interoperate, to an extent with other digital library services (3) freely distribute copies of the software.

This also will serve as an applications testbed, and as a way to test and adapt previous digital library research to the rigors of a production environment. Many issues such as naming, property rights, and payment are hard problems that are as yet unsolved. Indeed, a full solution to many of these problems can only come about when communities as a whole adopt standard approaches. We do not expect to solve them here, but we will create a interim solution for social science data that explores how a real production system can begin to address such problems, using insights from previous digital library research. This interim solution will be the one of the first to address a number of digital library issues in a production environment, and so might be used as a production framework for more complete solutions, as technologies for naming, metadata, payment and other services develop

This work will also put us in a position to extend the project in two major ways: combining journal articles and data within the same system, and doing large-scale user studies.

The VDC will make a vast amount of social science data available to a wide range of types of users, varying from experienced researchers seeking data for advanced research, to undergraduates writing term papers, to citizens seeking numerical answers to specific questions. The VDC will not only make data available for use, it will provide technical and organizational means of capturing new data-sets for the scholarly community and, thereby, provide a live, growing, and evolving resource

1 The Need

The need to be met by the VDC can be seen from three perspectives: that of data users, data producers, and data managers.

1.1 Data Users

This is not a project in which information will be made available in the hope that some one will want it. The demand for social science data exists, and will only grow with easier availability. The use of data to researchers is obvious, but students and citizens also need to access to data if they are to understand the world and the issues of public policy that the nation faces. They also need to understand data to manage their own lives effectively - whether that entails managing their health or their money. When we introduced our prototype system, use of social science data at Harvard & MIT increased dramatically. Our project will bring social science data closer to students in elite universities and in community colleges, and closer to citizens through public libraries.

1.1.1 Types of Users:

The VDC will serve:

Fact seekers: Undergraduates, for term papers, and citizens for general information, need answers to specific numerical questions. Graduate students and advanced researchers also often need specific facts.

Teachers and students: statistics courses and courses that use statistics in substantive areas need access to data sets for class use. Many such data sets exist, designed for pedagogical use. The rich array of data available through the VDC will allow students to use materials close to their varied substantive interests -- and will also teach them something about the valuable skill of data location.

Advanced researchers: Graduate students and faculty, of course, are prime users of quantitative data. needing easy access to a vast array of data sets for analysis.

1.1.2 Types of Use:

Finding data: A vast amount of social science data is available. Some if it, such as data deposited in national data archives, like the ICPSR, are not difficult to find, if you know it is there. For data created by individual researchers or even private concerns, nothing ensures that the data will appear in searches or even exist in any form after a few years. The VDC will enhance the availability of the leading data sets, such as Census data, the National Election Studies or the General Social Survey, that are already accessible through consortia and other well run repositories. But our goal is to develop a system that encompasses the full range of data from large and small studies, much of which is in primitive form. These small studies are at the heart of much research, but are often used only once because they cannot be easily located or used. Built into the project will be the development of a technical and organizational capacity to capture and make accessible such data.

An Initial Look. Data users -- whether novices seeking a few facts or a simple data set for a class project or advanced researchers with more complex needs -- require a quick means to “browse” data sets to see if they serve their needs. They need simple frequency distributions, cross-tabulations, or scatter plots. But data come in many formats; they are often hard to browse, access and use. A simple query about one data set may take an extraordinary effort. Another goal of the project, is to make such preliminary inquiry easy.

Acquiring the data: Users need to acquire a full data set, subset it by choosing relevant columns and rows, and convert the data to the format of their statistical, database, or spreadsheet program. This too is difficult because of the variety of formats.

1.1.3 The Need for Speed

Whatever use of a data set is made, access needs to be fast. Until very recently, the only way to get a data set that was not locally available was to wait for it to arrive in the U.S. mail on 9 track tape, a process that would normally take 4-6 weeks. And it would arrive in some strange format. This slowed the researcher and made original research in class projects virtually impossible. Recently, the ICPSR and other organizations have allowed selective electronic access to their data by a single (normally librarian) representative at each university. It appears, however, that in many places the software necessary to make this an easy transition is not available and customer service has not improved proportionately.

1.2 Data Producers

The VDC will make it possible for data producers to:

1.2.1 Make Data Available.

Many data sets are not transferred to archives because researchers are reluctant to make the effort needed to make data available. Designed as it is for heterogeneous data sets, the VDC will make such transfer easier.

1.2.2 Maintain Control Over Data

Individual researchers and others who create data in the course of their work are often willing to provide data upon request so long as they retain control over the source. Ask for the data directly and you are welcome to it. But these producers do not wish to put their data in the public domain. Since individual researchers are not professional archivists, data like these tend to disappear. The largest data collections, like the National Election Studies or General Social Survey, are routinely deposited in the ICPSR; however, only a small proportion of the data created in the course of research get deposited in archives. Some unarchived data appear on the web, they are not properly indexed or cross referenced.

1.2.3 Preserve Data

The VDC will enable data producers to share and also preserve the data used for individual research articles.

1.3 Data Managers.

The VDC is not aimed at replacing venerable national data archives like the ICPSR, but rather complementing them and helping them do their present job even better. The VDC will remove a large customer service burden from the archives by automatically handling data acquisition and distribution, sending notifications of data updates, handling data conversion and documentation, and maintaining metadata in a consistent format. When VDC systems are widespread, data archives will be able to focus their scarce resources on critical issues of preserving data and furthering the science of data archiving: storing archive copies of worldwide data, enhancing large data collections, creating new data and documentation formats, and developing tools. Data archives will be able to automatically "crawl" the Internet, searching for new data sets, copying them and preserving them against loss. The data archive staff will be able to devote more time to preparing and enhancing valuable data sets, and developing tools and standards.

2 Meeting the Need Through the Virtual Data Center.

2.1 Design Principles

The VDC is a system to be used by a wide range of institutions in relation to a wide range of data. It is not meant only for those institutions at the cutting edge of technology or only for those data sets formatted at the highest levels. Our primary goal is to produce and sustain an operational digital library that is easy to use, easy to adopt, and scalable. Universities should be able to use the system to create "main" libraries with many thousands of data-sets. Individual researchers should be able to acquire software for all core features at negligible cost, and to easily open their own "branches" of the digital library, in which they would share a few sets of data from their own research. Moreover, unsophisticated users should have no difficulty finding the data held in "main" or "branch" libraries alike. Because our focus is on operations and services, we do not offer radical new designs, standards, or algorithms, but instead propose to borrow as much as possible from previous digital library research, to use open standards and, when available and robust, free software.

Some data projects are developing general standards for linked data and codebook, work like this is exceptionally important. Data provided in these high end formats are more accessible and more interoperable than current formats, and provides the basis for meta-data standards. We will work closely with these projects, such as the document-type definition (DTD) codebook project sponsored by the ICPSR, and the Digital Library Federation's Social Sciences Databases project.

As a production system, we are strongly concerned with handling social science data in its present state, as data that are in accessible standard forms and data that are in less standard forms. One can think of the data on a continuum. At one end are data with high-quality coding. We will provide the highest level of service for these data. At the next level, data that are submitted in the format of one of several standard statistical packages, such as SPSS or SAS, will be available for automatic subsetting, format conversion, analysis, and distribution at the level of the variable or observation. Data in nonstandard statistical formats could be accommodated by continuing to expand our set of "standard" formats, by converting them to more standard formats, or by treating them as files. At the far "low" end of the continuum would be data sets in ASCII files with "README" files for documentation -- basically files with unknown contents. For this last case, we could not subset or convert (without special coding), but we would still be able to provide searching and various activities at the meta data level.

Once the VDC is in place, we believe users will more quickly perceive the differences among the various types of data documentation, and will start to demand higher level coding. Thus, the VDC should not only complement high end strategies, but it should also help encourage data producers to start following these guidelines.

2.2 What the VDC Will Do?

The VDC will provide an institutional mechanism for capturing data, it will allow users to locate the data, and allow them to access and use the data. The ultimate goal is a user friendly, large-scale network of social science data. How would this work? Consider a university with a few hundred unique data sets, perhaps from local researchers or dissertation students, in the position of setting up a new data center, perhaps as part of its library. The ICPSR or other organization might offer to take their unique data sets, but that is often not a popular proposal for many data centers who wish to continue to offer a unique product. As an alternative, the VDC project will offer the following. First, we will provide a free distribution and a list of (relatively inexpensive) hardware to purchase. The distribution will be easily self-installable. Once this university has paid the appropriate access fees to the national data archives, and chosen a system (from the options the VDC provides or others) for authentication of its users, they will have the same services available to all their students and faculty as Harvard and MIT have now: They can type in key words, search across all the data sets in several national data archives, view abstracts of data sets, look at lists of variable names, and when they are available, read codebooks on line. They might then run some descriptive statistics for a few chosen data sets (if the data are not available locally, the system will transparently retrieve the data in the background from the ICPSR or other organization).

An undergraduate might stop here, but for a graduate student or senior researcher, the VDC will also subset the data set (by the chosen rows or columns), convert it to the appropriate format, and deposit it on the user's hard disk. This would all be available instantly and automatically. Once this basic service has been created, this new data center may also take its unique data sets and deposit them in a special subdirectory on the system. Without any additional preparation, the VDC system will automatically recognize these data sets. They will become searchable, subsettable, analyzable, and convertible as are the other data in the system.

The VDC project will also be far more than porting a product now available at Harvard and MIT. For example, it can handle unique data sets that might be available at other universities. We will expand support for the federation of digital libraries through a publish-subscribe interface, along with publisher proxies for other services. High performance will be possible through caching and optional mirroring of meta-data and digital object repositories, allowing both simple sharing and hierarchical distribution of data. This will allow administrators of the VDC to create collections that encompass remote sites, and for universities to unify multiple collections within their departments. Once the VDC is installed at a site, it will be able to connect to the network of existing VDC sites to share metadata. This means that any user at any site doing a simple search can choose to automatically explore all the data in the ICPSR and other national archives, all local unique data, and all the unique data at other VDC sites. In return, the local site can choose to make their metadata and data available to others through the same system. Our extensive conversations with those generally reticent to make data available by depositing it in the ICPSR tell us that they would willing and in most cases eager to provide data in this way.

To facilitate this process, those with unique data sets will be able to choose one of several methods of access, for example – unrestricted, signing an authenticated “electronic guest book”, writing an explanation of the desired use and asking for permission, or entering in a credit card number. This gives the provider control over the master copies of the data (which would still reside locally), visibility in providing the data to the scholarly community, and access to the vast array of unique data at other sites. Each additional unit that connects to the VDC will make the whole network more valuable. We will even provide a personal version of the VDC that will enable a scholar to hook their individual web site into the system, thus capturing for the scholarly community the data being made available by the fast growing and relatively undisciplined practice of putting data up at isolated web sites but not in any unified catalog.

We have found great interest in data centers around the country in our system, and we believe that it will be quickly adopted. Many will use our software directly. But we plan to design the software so that it is highly modular, and fully open and extendable. We imagine many “snap-in modules” that could be written to improve the system, and we hope to encourage the user community to contribute them. For example, we have experimented with a quick way to get a sense of those data sets that are organized geographically by automatically generating maps colored in by chosen variables. (The Harvard-MIT Data Center has much experience in this field, and currently in conjunction with the Harvard Map Collection, has begun a two year "Geospatial Liboratory [sic]" project to create a separate digital library of geospatial data that can be linked to the present project (http://data.fas.harvard.edu/hdc/hmdcproj/liboratory.shtml). Other modules might include specialized statistical software or systems to handle unusual types of data organizations or formats. These modules can be written to extend our system or even to replace parts of it. In fact, a few of the most sophisticated university data centers, with their own software in place already, might wish to avoid our software altogether. If this happens, these few sites will still be able to contribute since we will also provide protocol gateways. Our specific software is a proof of possibility, and will be of use to the vast majority of sites, but it is not necessary to be part of the VDC system.

Once the scholarly community is using the system, we hope to open access to commercial data providers, under the condition that they write a module that snaps their data into our system. Once that is done, users could purchase commercial data on our system and would no longer need to worry about unique data formats and specialized programs; they would get it in the same automatically convertable, subsettable, analyzable formats as the rest of VDC data. (Conceivably, there could be a small tax on data providers to support the continuing development of the system.)

We have very close connections with the ICPSR, its director, and the Council. Gary King gave a complete presentation of the Virtual Data Center project to a meeting of the ICPSR Council, on which he serves. This was followed up by much feedback and encouragement. As the largest ICPSR data user, the Harvard-MIT Data Center is well acquainted with the staff of the ICPSR, and we have received many helpful suggestions along the way. We plan to continue and reinforce these contacts, so the development of the VDC serves to strengthen the ICPSR.

Eventually the ICPSR could perform this activity for the scholarly community, as they are one of the only organizations with professional practices such as cataloging, verifying, and archiving with off-site backups. Permissions would need to be secured from local contributors, with agreements written regarding future use of backed up data (for example, the ICPSR might agree to only distribute data if the local site vanishes).

3 What Has Been Done?

We do not begin these studies from scratch.

3.1 No Digitization Needed

The main previous work on which we draw is the vast amount of digital quantitative data that exists. The point is so obvious that it might be missed: the quantitative data in our system require no conversion to digital form. One of the advantages of this project is that it adds a large amount of value to material already in digital form.

3.2 Drawing on the Work of Others.

Our goal is to create a system that integrates many tools and applications; we do not wish to develop new tools and applications when we can apply existing one. The VDC will, as much as possible, be developed using robust pre-existing tools.

3.3 Our Previous Work:

Most significantly, we can build on work we have done already. We have developed a prototype at the Harvard-MIT Data Center which is being used extensively to automatically collect data from remote archives, and to deliver a large and varied amount of data to a scattered and heterogeneous set of users. Our system will search across all available data sets, at Harvard, MIT, and several national archives. It will automatically subset and convert data to chosen formats, and it scales up quickly so that new data sets in a large variety of standard formats added to the system will also be instantly subsettable and convertable. Our prototype (http://data.fas.harvard.edu/), and has greatly accelerated data-based research and teaching within Harvard University and MIT. This data center is now one of the most heavily used academic data centers in the world. In 1997, it served over 10,000 data sets, and automatically answered over 100,000 queries from all over Harvard and MIT.

We now want to build on our experience to make these resources more generally available. In fact, this process has already started. We have an agreement with the Henry A. Murray Research Center (http://www.radcliffe.edu/murray/) to make their unique holdings available to qualified investigators through this system, on an experimental basis.

4 VDC Design and Development

No tool duplicates the core functions of the VDC[1], but the VDC will be, as much as possible, created from robust pre-existing tools. This section discusses the current system, the design goals and principles behind the VDC plan, the primary features of the VDC, the functional components of the VDC, and development methods and schedule.

Our current system has been successful in enabling the Harvard and MIT communities to search for, obtain data through world wide web. Some of this data is produced on-site, but most is retrieved from the ICPSR; our system manages automatically the retrieval and caching of ICPSR data, and the the process of keeping our metadata and data holdings consistent with the holdings of remote archives, as studies are updated, added and delete. It supports simple authentication through the use of Harvard and MIT i.d. numbers. It also enables summarize data, subset it, and convert it to their preferred format – all while on-line.

Our current system was produced in a service-oriented environment, and has a very simple architecture.

[1] One product sharing some elements of the VDC is the commercial "Data Warehouse." Although superficially similar to the VDC, and providing some of the same functions, there are many crucial differences between the data warehouse and the VDC. Although a data warehouse can, in principle, provide a combination of data management, metadata-management and user interface features (Hackathorn and Inmon 1994; Inmon 1996). In practice, data warehouse products are designed to meet very different goals that the VDC.

In essence, data warehouses are designed to capture, clean, and store huge quantities of business data, in a highly centralized organization, administered by a dedicated staff. The data that data warehouses deal with typically are "side effects" of business transactions (e.g., receipts, logs) and so must be captured, scrubbed and cleaned extensively. Huge quantities of data are processed and updated on a daily basis. Data warehouses are typically cumbersome, expensive to purchase or license (typically costing one to several hundreds of thousands of dollars(Hurwicz 1997).), are complicated to setup and maintain, do not provide an easily integrated set of tools, and are based on proprietary software and communications protocols (Darling 1996).

Virtual Data Centers will be designed to service the research community, which is decentralized. The database component of the VDC will be much simpler, and the VDC as a whole will be lightweight, easy to set up and to administer, with minimal hardware requirement. A social scientist will easily be able to set up a VDC even to share a single dataset from a single publication. Rather than developing a single centralized server capable of storing terabytes of data, millions of information objects, and being queried tens of thousands of times per second (Kimball 1996), VDC's will typically comprise a network of servers which are each one-hundredth of that size.

It comprises a simple two-tier client-server architecture, with ad-hoc extensions to synchronize automatically with other remote data collections. Description and normalization meta-data, naming, and communication with other data archives are all ad-hoc. In addition, some of the components of the system are based upon proprietary commercial software, such as SPSS.

The current architecture is too simple to provide a general framework for either the treatment of digital objects, or the management of federated collections. The VDC will be a significant redesign and re-implementation of this system. First, we will re-implement this prototype on a foundation of a general digital object infrastructure: this will involve re-conceptualizing the design in terms of objects, generalizing the data-structures and interfaces we use to handle social science data so that they are flexible enough to be used for other types of digital objects, and building a middle layer into the system to enable interoperability. Second, we will incorporate free, open, tools to provide alternatives to the commercial products we use now. Third, we will extend our current digital object preparation tools, to aid researchers to migrate their data into the digital library. Development will be guided by user studies of the current services.

4.1 VDC Features

The initial VDC features comprise four categories: data preparation, data access, user interface, and interoperability. Approximately half of these features are provided, in some form, by the current prototype and are marked with a "*". We list the primary features of the VDC in Table 1, and then we outline the design below.

These features challenges with which any large-scale production system that operates in an open environment has to come to terms. Yet many features, such as naming, property rights, and payment, raise hard research problems that are as yet unsolved. Indeed, a full solution to many of these problems can only come about when communities, as a whole, adopt standard approaches. We do not expect to solve these here, but we intend to create an interim solution for social science data that incorporates insights from previous digital library research to explore how these problems can be approached in a real production system. This interim solution will be one of the first to address a number of digital library issues in a production environment, and so might be used as a production framework for more complete solutions -- as technologies for naming, metadata, payment and other services develop. We also expect to produce a framework in which we can develop services for other types of digital objects, such as journal articles, and which will allow us to launch major user-studies.

Category	Features
Digital Object Preparation/Intake	• Naming: Uniquely identify digital objects
	• Aids for preparing and converting data in common formats.*
	• Aids for preparing and converting metadata in common formats.
Digital Object Management	• Repository management (addition, deletion, modification of objects).*
	• Metadata queries.*
	• Cacheing and Mirroring: performance enhancements for repository and metadata management, location of digital objects
User Interface	• Views/browsers: including features to display*, summarize*, subset*, aggregate, convert* and merge data and documentation.
	• Conceptual maps: provide maps of the content of the library to users for searching and browsing*
	• Administrative interfaces: for establishing sessions, authentication* and payment, and administering collections.
Middleware and Interoperability	• Direct support for Corba IIOP.
	• Gateway support for common protocols.*
	• Facilities for the integration of multiple collections.
	• Publisher-proxies for the federation of heterogenous collections.

Table 1: Primary Features ("*"'s indicate features that are at least partially supported in the current

version).

4.2 Structure of the VDC

Rather than designing a specialized digital library for social science data, we will build a general digital library that provides specialized services for this data. At the same time, we will maximize the use of existing, freely-available and openly architected software.

In order to handle digital objects in a general way, we rely on an infrastructure services for naming objects, for communication among clients and servers, and for metadata. Free statistical products and databases will be used for the actual data manipulation.

Figure 1 shows a sketch of the design of the system, illustrating some of the protocols that are likely candidates for incorporation into the system. This design will, of course, change during the course of the project – often, an initial design serves simply to make clear what does not work. A significant of the project will be to refine the design through multiple rapid prototypes.

User Interfaces. Initially, users will probably interact with the system using web browsers. Through an interface that combines HTML, XML, and Java applets, users will connect to web-servers that will act as proxies to the middle layers. Together, these will obtain objects and metadata through the middleware layer from repositories, and enable the user to display, summarize, subset, aggregate, convert and merge data and documentation.

Naming services. Naming services are fundamental to the digital library. Digital objects must be uniquely identified, and this identification should not change if the object is moved to a different location or if the repository in which the object resides changes location. A number of experimental schemes for uniform location-independent naming exist: including "PURLS" (Shafer, Weibel et al. 1997) and "URN"'s (Daniel 1997), the CORBA naming service (Object Management Group 1997), and the Handle System ® (Orth 1998) (also see http://www.handle.net).

Handles and PURLS are the most well developed of these naming systems, and may be the best suited towards the needs of digital libraries. However, naming systems have a number of limitations that we will have to address. First, naming schemes have either been implemented in a limited fashion (like handles) or not yet implemented at all (like URNS); it is unlikely that in the near future a universal scheme will be adopted across data archive. Second, most naming schemes do not distinguish among multiple copies of the same digital object, so a proxy system will have to be used with the handles resolution system to direct user's to close instances of each object. Third, outside data collections, such as ICPSR, will probably not be adopting handles or any other true naming service in the near future – proxies will have to be devised for each external service (we plan to create such a proxy for ICPSR holdings, based upon the study #'s that ICPSR assigns to data collections). Fourth, because the system is likely rely on CORBA for its middle-tier, a service will need to be provided for mapping handles into CORBA names. In our project we will explore the use of handles and other naming schemes for objects.

Middleware. Middleware provides a foundation for the interoperation of diverse clients and servers. The two largest competitors in this arena are DCOM and CORBA. DCOM is a closed standard developed and controlled by Microsoft . CORBA is an open standard, with a number of open implementations, that may be better suited for the development of a free, extensible system. CORBA provides particularly flexible object-oriented middleware services, and has been used successfully in a number of other digital library projects, but has not been used in large-scale production systems. Stanford's Digital Library Interoperability Protocol serves as another protocol layer on top of CORBA that further defines how digital library elements can interoperate.

There are three significant limitation of CORBA that we will have to avoid. The first is that many of the secondary CORBA services are still in a state of flux, so we will only be relying upon those core CORBA services that are relatively stable. The second limitation is that although external connections to CORBA services are well defined, the programmer API's for CORBA are highly dependent on the CORBA implementation used (Orfali, Harkey et al. 1997); so, we will use "bridges" to insulate the collection management services from these API's[2]. The third limitation, for the immediate future, is that other digital libraries with which we will want to interoperate do not use CORBA (e.g., ICPSR's digital library uses its own protocols); so we will construct or adopt proxies for ICPSR's home-grown protocol as well as for other common protocols.[3]

Many simple automated clients may not understand CORBA, or other potential middleware choices. So, we will also explore the use of limited gateway interfaces for simple automated clients. Many libraries use Z39.50 to share metadata information, and we will explore the use of Z39.50 as a gateway into our services. HTTP is also attractive, because of its wide use, but the structure of HTML limits the services that can be provided by automated client. The emerging XML standard offers better prospects for distributing structured data using HTTP.

Metadata. Most existing schemas for social science metadata data are relatively simple, and can be efficiently stored as relations, and efficiently queried using boolean operators in an SQL89-type syntax. There is no universal metadata schema for social science data, but much of the Dublin core is applicable, although not comprehensive, and would provide a reasonable basis for users wishing to find sets of data of interest. The initial implementation, will map metadata added to the system into the Dublin core, will support queries using SQL89. We will also explore using the Stanford STARTS protocol in this system.

There are a number of limitations to this approach. First, significantly more detailed metadata are required for data services more complex than location and delivery of sets of data. Second, other digital objects will require different, and more complex metadata. Third, STARTS (although it is extensible) and SQL89 queries are inadequate for some knowledge domains.

The Dublin core provides general metadata that could be used to find datasets of interest, but does not attempt to provide detailed descriptions of the components of digital objects. In order to provide data-exploration services, data viewers, subsetting and conversion of data formats much more metadata is needed – typically at the variable level. Special challenges are raised by services that attempt to automatically aggregate and merge different data collections into coherent extracts (e.g., merging census data collected at the block-group level with public opinion poll data collected in congressional districts and economic indicators for metropolitan areas). These services require extensive metadata at the variable level to ensure correctness (see (Greene 1997) for a study of some of these issues). There are a number of avenues we will explore to address this situation. First, ICPSR and DLF are developing an SGML DTD for datasets that would provide much of the needed variable level information (http://www.icpsr.umich.edu/DDI/codebook.html), and we will develop tools to read these DTD's. Second, many data formats for statistical programs (such as SAS and SPSS) contain embedded variable level metadata, or variable level data implicit in the study-level data, which we will capture, normalize and convert.[4] Third, we will experiment with ontologies for variable level metadata to support the correct merging of heterogenous sets of data on common fields such as time and geographic location.

Social science data categorizations will change, and other digital objects will have other metadata schemas that are better suited to their domains. A framework is needed that can be used to support metadata for a variety of objects as the system develops. In order to avoid being wedded to a particular schema, we will explore the use of Warwick Framework, which acts as a container for different "packages" of metadata, such

[2] For descriptions of bridge, proxy, and publisher techniques see Gamma, et al. (1995), and Buschman, et al. (Buschmann, Meunier et al. 1996).

[3] A note on protocols and standards: SQL is a query language for databases, described in a number of standards documents (Date and Darwen 1992). The STARTS (Gravanbo, Chang et al. 1997), InterPay (Cousins and 1995), and DLIOP (Hassan and Pepcke 1997) are digital library protocols developed by Stanford. MARC is a federal metadata standard (Library of Congress Cataloging Distribution Service 1993). CORBA (Vinoski 1997) and HTTP (Fielding, Gettys et al. 1997) are commercial protocols. Z39.50 is a protocol for queries (ANSI/NISO 1992).

[4] We expect that the metadata profile for many sets of data will have gaps, even after embedded and implicit metadata has been captured. In some cases this will limit the services that can be provided for individual sets of data, but we expect to provide a range of services for a wide variety of data by exploiting this information.

as the Dublin core (Weibel and Lagoze 1997). No one has implemented this framework, as yet, but it may provide a means to encapsulate different metadata schemes.

In addition to requiring different metadata, other domains may require different query semantics. For example, spatial relations are often used to query geospatial databases, but can only be incompletely expressed in simple SQL89 dialects. We do not plan to support other query languages in the first round, but we will provide a mechanism for specifying the language associated with each section of a query, so that the system can be extended to handle queries crossing multiple languages.[5]

Data manipulation. Most social science data schemas can be modeled easily as relations techniques, once the peculiarities of proprietary data formatting are transcended. Data that is inducted into the system will be converted into a common base format, and freely available database and statistical packages will be used to provide conversion, summarization, subsetting, aggregation and exploration services.

Distributed Data: It is often inefficient or inconvenient for digital objects to have a single location. But allowing objects to exist in multiple locations raises a number of issues: How to locate the “closest” objects? How to maintain consistency and completeness? How to perform updates? One approach will be to explore a simple caching and mirroring of metadata and/or data, using a publish and subscribe model for updates. We will investigate the use of both simple single-level mirrors and more complex hierarchic collections of data.

Access Services: In a production system, property rights must be honored, yet there is no uniform way of describing these rights, nor a universally accepted payment method. Still, payment and rights management are happening on the Internet now. Our goal will be to develop or borrow a simple set of authentication and payment mechanisms[6], informed by theoretical research, and maintain simple property rights metatdata, so that the system can be used in the real world.

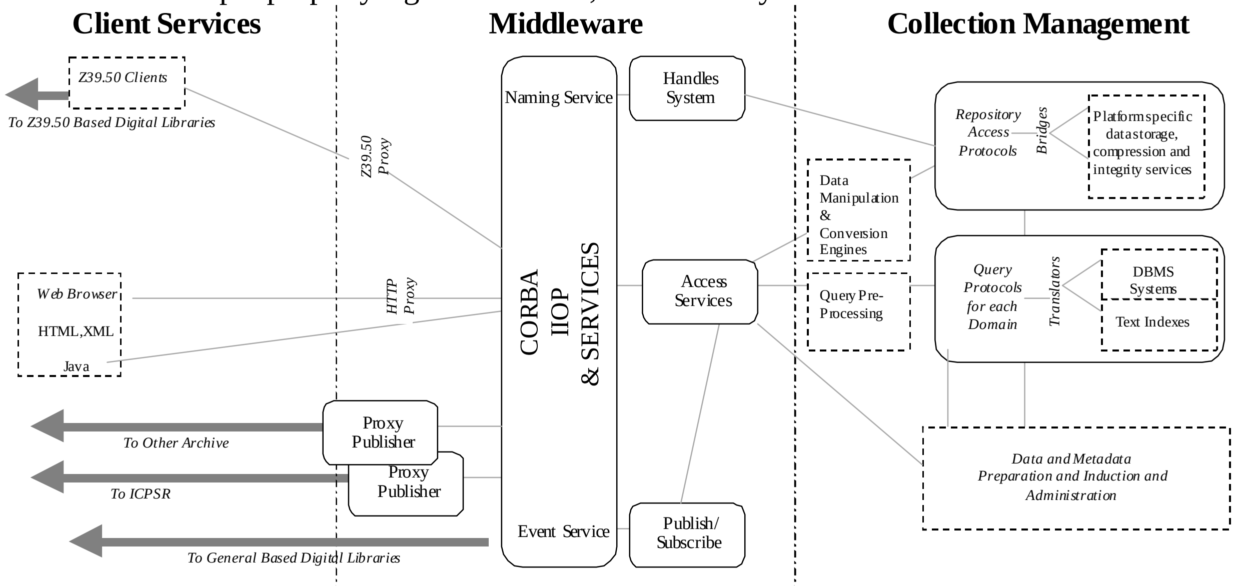


Figure 1: One Potential Architecture for the Digital Library

4.3 Integrating Free Components into A Flexible Portable System

When Edison invented the electric light, he also had to design a system that would deliver electricity to his customers, and found companies to manufacture it (Cowan 1997, 119); even ten years ago, digital library development faced similar challenges. In the last five years however, there has been unprecedented growth in

[5] Different metadata packages will be expected to share only a few core elements, at most. We will also support hooks for query pre and post processing, in order to allow queries from one domain to be converted, as much as possible, into another (for example, converting geographic place names into latitude-longitude coordiantes for geospatial queries). This might be accomplished through the use of subsuming queries and filtering, see Baldanado, et al. (1997).

[6] For a general approach see (Cousins and 1995; Ketchpel, Garcia-Molina et al. 1996).

the scope and technology of the national information infrastructure (Kahin 1995). Fortunately, there now exists a much larger base of existing open software, and a significantly greater convergence on common protocols, than existed ten years ago.

It is now possible to develop upon completely free and open systems, using open operating systems such as Linux and FreeBSD, multi-platform and platform-independent development languages such as Perl, Python, and Java. The GNU project (http://www.gnu.org) has established a large base of applications and utilities, many of which run on both Windows NT and a variety Unix of platforms; these include file management utilities, databases and statistical software. Users will be able to access the system through free web browsers (Netscape has now made the source code for the next generation of its browser freely available.) and Java applets. The VDC project will take advantage of this base of open software components to create a flexible digital library system that is free and portable.

We will use a combination of multi-system tools, and multi-platform languages to develop code designed to be platform independent. We realize, however, that even supposedly platform-independent tools do not actually behave identically on all platforms. Two operating systems are likely candidates for testing and devlopment, each of which has distinct advantages. The primary advantage of Windows-NT is that the various flavors of Windows are in common use, and Windows-NT will continue to gain popularity as Microsoft phases out Windows95 and Windows98. The primary advantage of Linux is that it is a completely free, open, and relatively robust and flexible operating system. Producing a Linux tested version of the VDC will mean that all features will be open, even those implemented by the operating system.

Although much current data carries licensing restrictions, significant sets of data could be provided without restriction to the general public. In addition to “replication” data for individual articles, which can usually be distributed to the public, much government data could be shared more widely. Such data includes studies, now held at ICPSR, by the National Center for Health Statistics and the Department of Justice. Furthermore, much government data is available digitally at government documents libraries, but is not available on-line. We will work with Harvard’s government documents library to make a selection of its digital holdings, now available only on CD-ROM, available via the VDC system.

Moreover, the availability of the VDC system may encourage researchers to share their data more widely. For example, the VDC prototype at Harvard, has enabled us to make the massive, new, NSF-supported, *Record of American Democracy* (King, Palmquist et al. 1997) data[7] publicly available (available from http://data.fas.harvard.edu/ROAD/).

One of our CO-P.I.’s, Gary King, has a long experience developing highly complex public domain software (available from http://GKing.Harvard.Edu/stats.shtml). He has developed what is now widely used software for evaluating redistricting and forecasting election results that have been used by academics, public officials, judges, and partisans in many legislative redistricting cases. He has also produced software for the statistical analysis of event counts that has been widely used in the academic community. His new software on ecological inference is being used across a range of academic disciplines, as well as in government and private industry.

4.4 Rapid Prototyping

A guiding principle of designing complex systems is “plan to throw one away” (Brooks 1982, 116). Following this principle, we will employ a rapid prototyping model for development – building and distributing internal versions of the software and external beta versions early, and releasing new versions often. Source code will be open, and comments, feedback, and contributions to the project will be solicited early.

Development will proceed in several phases: design, alpha implementation, testing and quality assurance, beta testing, and major release. During the design phase, we will develop a complete formal

[7] The Record Of American Democracy (ROAD) data includes election returns, socioeconomic summaries, geographic maps, and demographic measures of the American public at unusually low levels of geographic aggregation, spanning 1984-1990. It is a large set of data (approximately 2GB), and has been downloaded thousands of times since it was made available at the beginning of the year.

description of each module and its methods, and of all interfaces. The alpha implementation will implement the objects and interfaces, and test them internally at Harvard and MIT. In each phase of the process, we will seek comments from the general academic community via Internet RFC's, conferences, and beta testing programs. The beta release will be an open release, soliciting comments from many sites at other universities.

Harvard and MIT are ideal test-beds for this development because they are microcosms of the larger research community: They have, together, a huge community of users of quantitative data from all fields and disciplines; and data is distributed (physically and politically) throughout both institutions, as it is in academia. In addition, a number of other university data centers will be able to immediately use the new VDC, and will also serve as alpha-test sites

We are working closely with librarians from the Harvard University Library, the Harvard College Library and the MIT libraries to identify centers of quantitative data at Harvard and MIT that would serve as alpha-test sites for the new VDC. We expect that the new VDC would be used immediately by the Harvard Map Center, one of the oldest collections in the University, which has collected extensive electronic geographic data collections, and by the Government Documents library, which has extensive holdings of data on CD-ROM.

4.5 Formative research with users.

Since quantitative data already has thousands of users we plan to conduct studies at the start and throughout the duration of the project, to determine how users understand and work with the system: its interface, data organization, and other features. In addition, we will explore how faculty might plan to use the system in their teaching, and how faculty, students, and citizens might work with it for research purposes. Thus, we plan that these studies will explore the following issues: (a) interface design; (b) organization of information; (c) analysis features; (d) how the data might be used in higher education courses, and how it might affect course design (e.g. assignments, student work products and the like); (e) key concepts in the teaching of the social sciences that might be especially supported by the system; (f) how use of the system might affect social features of course work, such as team projects and the like.

We will work with a variety of users in these studies in order to thoroughly understand the needs of a range of users who would benefit from the system. These variations will include levels of disciplinary expertise: undergraduate students, graduate students, faculty, citizens.

The studies will be designed to examine how people use the system, and how they understand and think about it. Methods will include individual task sessions interviews in which people will be asked to carry out various tasks with the system, thinking aloud as they are doing so. Individual in-depth and focus group interviews will explore how people might use the system for different purposes and contexts.

5 Enhancing the VDC.

The project for which we are seeking support is part of a longer range set of projects. There are two enhancements to our current work that we anticipate: systematic user studies and linkages of the data in the VDC to text. During the current project, we will move in the direction of these enhancements, though their full development and implementation will wait for subsequent phases.

5.1 Additional User Studies.

Computer applications, especially those in fields which cannot be complete automated, often fail because user's are left out of the design process, or brought in only at the end (Landauer 1995). Furthermore, we have few systematic studies of the use of technology libraries and in learning. In this project, user studies will be integrated throughout, and will be used both to shape the design of services, and to evaluate the results. Indeed, we plan to develop an extensive series of studies to determine how users understand and work the system: its interface, data organization and other features. We intend for these studies to go well beyond the usual informal analyses or classroom questionnaires. Professor Jan Hawkins of the Harvard Graduate School of Education is a specialist in the development of technology for classroom work and will design with us a series of studies of the use of the VDC in different types of classes and in different institutions. The research will feed back into the various stages of our design for the VDC, including the enhancements we hope to add, and the research should provide basic information of a more general sort as to the use of technology in higher education. The first phase of this research will place in the framework of the project for which we are applying. The next two phases will take place with support we hope to add in the future.

A major component of our long-term plan is a series of studies of how these data are used in the classroom (or in the dorm room). We wish to move well beyond the usual methods of learning about use through intuition or anecdote or though simple questionnaires given out in class. We want an in-depth study of use across a wide range of types of users in varied institutions. The research will feedback into the various stages of our design for the VDC, including the enhancements we hope to add. And the research should provide basic information of a more general sort as to the use of technology in higher education.

In the first phase, under this grant, formative research studies will be conducted to determine how users understand and work with the system: its interface, data organization, and other features. (See section 4.5.)

In subsequent phases, we will work with a variety of users in these studies in order to thoroughly understand the needs of a range of users who would benefit from the system. These variations will include levels of disciplinary expertise: undergraduate students, graduate students, faculty, citizens. The studies will also include exploring variation by institutional context, including: universities (Harvard, MIT); state and/or community colleges; public libraries. These data are important to design refinements to ensure simple and robust use in intended contexts. We will thus conduct formative field studies beginning in the second project year. We will collaborate with a small group of institutions to conduct these studies, including Harvard/MIT, state and or community colleges in the region, and one or more public libraries. At the higher education institutions, we will recruit/select a small group of faculty who teach a range of courses, and who are willing to integrate the system into their teaching during this project year. It is likely that these faculty will be identified in the institutions in which the formative user studies were done. We will work with faculty to help them to understand the features of the system, as support for incorporation into their teaching. (At Harvard, we will be able to take advantage of the fact that Harvard College will be introducing a new set of courses on Quantitative Reasoning into its Core curriculum. Thus we will have a set of new courses, aimed mostly at non-quantitative students in the humanities and in some of the social sciences, for which data sets will be an important part of the basis of instruction.)

We will seek a range of social science courses with respect to discipline and teaching style for the field studies. How is the system used in these various contexts? What kinds of assignments and student work are associated with its use? Our overall goals in this work will be to understand how the system can be used in higher education and for public information, the problems that may arise, and to identify specific potential impacts of its use on course design and student learning that will guide the design of the final outcome studies.

Faculty will be interviewed prior to the use of the system in their course(s), including documentation of prior course design. A sample of students in each class will also be interviewed prior to participation. We will regularly observe classes, as well as sessions in which students use the system as part of course assignments. Course documentation will be collected, as will samples of student assignments that include work with the system. Interviews will again be conducted at the conclusion of the course. These data will be analyzed with particular attention to suggested refinements in design, and to accumulating understanding of the best use contexts in social science courses, and effective curriculum design for use of these resources. We will interview faculty research users about how the system functions for their research needs. We will also identify a public library where citizens will be given ready access to the system. We will collaborate with the library staff encourage a variety of people to use the system for their inquiries. Interviews will be conducted before and after these, in general, one-time use sessions. This will enable us to determine the features of the design that are required for potentially one-time use for citizens' questions.

The final research phase will focus on outcome studies: to understand the consequences of use of the system for the design and organization of courses, and for learning outcomes. The specific design will grow out of the outcomes of the formative field studies in which we will identify specific concepts, course designs and activities, and social organizations of course work that are most likely to be affected by the incorporation of the system. The research will be focused on pre/post studies of system use, and where possible, comparative studies of system use in similar courses taught by the same faculty member with and without the system. We plan to examine student learning of carefully selected concepts/material that are the focus of the system resources and course assignments, and the quality of student products. We will also collect data that characterize the course context and process of use through interviews, documentation, and observational methods.

These studies will also be conducted in different types of institutions, including universities and state or community colleges in several locations around the country.

5.2 Linking Data to Text:

Quantitative social science data are tools for research. The analysis of such data appears in professional journals, in scholarly books, and more and more often in more popular media. For the scholar, the connection between text and data is natural, but these connections are sometimes cumbersome or difficult to make. Data that back up an article are often difficult to find and even more difficult to analyze. Thus, our ability to replicate the work of others and to build on it diminished. A similar problem exists for scholars who move from data to published work based on the data. It may not be easy to trace the publications that emerge from a data set -- so that we can build on rather than duplicate that which has come before. Scholarship would be greatly enhanced if one could move easily from data to text and from text to data.

The connections are, in some sense, even more vital for less sophisticated users of data resources. Researchers know that the reports of research results in a data table are the result of a complex error prone process. Students and others not aware of the nature of social science research may believe the analyses that appear in a publication come from some unquestionable scientific process and are to be copied down and believed. There is no lesson more important for a novice student in the social sciences to learn than the complexity and uncertainty of the research process that moves from data to published text.

The major enhancement we plan for the VDC is the development of links to social science text. Imagine one is reading an article in a social science journal. Reference is made to a data set that is the basis of the article or a footnote appears to the works of a scholar whose data are relevant. The reader immediately locates the data and proceeds to check the findings in the article or go beyond what has been done. “The author should really have included this or that variable.” “The author should have applied a different procedure.” The reader can do it! Conversely, a research or student decides to write a paper on subject X. She finds a good data set. But has anyone else published on that topic based on these data? The student can call up the published works on the subject.

The general goal is clear. Defining it more precisely is more difficult. And accomplishing it more difficult yet. One needs to locate a body of text and a body of data into which links can be placed, design a place in the data structures of body of text and set of data where links can be placed, and provide user interfaces to move from one to the other. In many cases, one would need permission of copyright holders of journals (though one might hope they would be permissive in this since such links only enhance the value of their intellectual property). Naming schemes are needed to locate materials at both ends of the connection. Issues of user authentication and rights of access will be more complex, since users in varied locations would have to be linked to journals and data that also might have varied locations.

These are major and complex issues, but they are problems faced in relation to many other digital library applications. We would, for instance, take advantage of such developments as those in relation to naming that are being worked on through the Handle system being developed by CNRI. The data base for the quantitative data will be the data under our control in the VDC -- or some selection of it to begin with. The data base for the text may be the JSTOR collection of journals -- though, again, we might limit ourselves to the journals in one social science field. We have discussed these possibilities with both CNRI and JSTOR, and both are interested in pursuing a collaboration. This part of our long-term plan is not part of the current proposal. We cite it here because we think it is a very exciting next step. And we will design the VDC in consultation with these other entities so that transition to this next stage would be possible. It is the intent of our project to explore how digital libraries can be used to deliver complete intellectual works in social science -- unifying data and research articles. We plan to create a digital library of such intellectual works.

6 Coordinating with Other Institutions

This project will foster the established links we have made among social science data providers, data users, and the library community. This will be facilitated by the fact that this is a joint project between the

Harvard-MIT Data Center (a center managed and used by active social scientists) and the Harvard University Library. This system will not be built in a vacuum – but will connect with Harvard's own digital library development.

The Principle Investigators come from both the social sciences and the library. Sidney Verba, is both Director of the Harvard University Library and a Professor of Government. Gary King, is Director of the Harvard-MIT Data Center, a Professor of Government, and now also serves on the ICPSR Council.

We are working closely with professionals at the ICPSR, as well as others working in the field of digital information. Harvard is a founding member of the Digital Library Federation, we have its endorsement for this project, and will be exploring with them ways in which the VDC can be used to complement their initiatives to make high-profile data more widely available. We are working with the Corporation for National Research Initiatives to integrate its system for naming digital objects into our VDC. And we will, in the second phase of the project, work with JSTOR in an effort to join journal articles with sets of data.

The Harvard-MIT Data Center serves both Harvard and MIT is having discussions with the university of Michigan to provide the same services to their faculty, students, and staff. Even without advertising our work, many other data centers have asked for early version of our system when it becomes available, and have offered to be beta-testers.

These contacts will be important throughout the development stage, in order to remain consistent with developing standards, and so that one or more of these institutions will eventually take over and institutionalize key parts of the project.

7 In the Long-Run:

The Harvard-MIT Data Center, and its VDC prototype, is a part of Harvard University's and MIT's library resources. We expect that, as well as extending the features of the current data center, and extending those features to other data centers, the new VDC will be the key to providing a much sought after "virtual union catalog" that indexes quantitative data at multiple institutions. Although Harvard will, for a time, keep such a catalog of the range of data holdings, the data itself will remain distributed and controlled by the owning institutions. We will work closely with the Harvard and MIT libraries to integrate it into the library system. When we are successful, we will not continue to run the VDC indefinitely, and expect that its future will be ensured by its adoption at many sites inside and outside of Harvard. When the system becomes routine, we plan to transfer operational management of it to another institution, such as the ICPSR and/or consortium of participating universities.

Since the VDC will be capable of running on a most modern personal computers and workstations, will be free and extensible, and will allow unprecedented access to the world's research data, we expect that this system will eventually be adopted by a large variety of research centers and universities. Since it will be based upon open, general protocols, such as DLIOP and Z39.50, the VDC will be a useful component of and gateway to federated digital libraries of many types. In addition, we will take a number of steps to encourage the widespread use of the system: During the beta-testing phase, we will distribute the VDC to selected sites at other University data centers. We will also make the VDC freely available, and prepare installation programs that will make it simple for researchers to install and run.

Widespread data sharing will have a broad and significant impact on research, and teaching. Information access technology has the potential to speed the diffusion of research within and across disciplines (Crane 1988, 128). Easy access to the data that are the root of published research will be an enormous boon for verification, extension of methodology, and secondary analysis (Clubb, Austin et al. 1985). Easy access to real data will improve teaching (Sieber and Trumbo 1991).

We hope the VDC will also affect the norms of data sharing in academia. The sciences and social sciences have come to accept data sharing, at least in principle. There is also a movement in political science to provide data as a condition of publication, and a large number of journals have adopted such policies (King 1995). In many other fields the premier journals, such as *Science*, *Nature*, the *American Economic Review*, the *Journal of the American Statistical Association*, the *Journal of the American Medical Association*, *Social Science Quarterly, Lancet* (and other journals conforming to the "Uniform Requirements for Manuscripts

Submitted to Biomedical Journals"), and the various journals published by the American Psychological Association all, in principle, make the provision of important data on which the article is based a condition of publication[8] (Clubb, Austin et al. 1985). In addition, several government organizations, such as the National Science Foundation, require that data resulting from funded projects be made available to other users (Hildreth, Aborn et al. 1985; Sieber 1991). Despite these requirements and benefits, much data is not shared - and the discipline has been held back consequently. The norms of sharing data are weak, requirements are fuzzy, and verification of data sharing is lacking. By providing a system for quickly, easily, and verifiably sharing data, the VDC will enable more journals to adopt and enforce data sharing and replication policies. The VDC may even result in better public policy by reducing the opportunity of manipulating the impact of a study by withholding the data on which it is based (Hildreth, Aborn et al. 1985).

[8] We gathered information from the “Contributors” section of the 1997 edition of each journal. The exact terms of the requirement to share data vary both in principle and in practice. The "Uniform Requirements for Manuscripts Submitted to Biomedical Journals" has, arguably, the weakest policy, stating simply that the editor may recommend that important data be made available.

References Cited

ANSI/NISO (1992). ANSI/NISO Z39.50-1992, American National Standard Information Retrieval Application Service Definition and Protocol Specification for Open Systems Interconnection. Bethesda, MD, NISO Press.

Baldanado, M., C.-C. K. Chang, et al. (1997). “The Stanford Digital Library Metadata Architecture.” International Journal on Digital Libraries **1**: 108-121.

Brooks, F. P., Jr. (1982). The Mythical Man-Month. Reading, MA, Addison-Wesley.

Buschmann, F., R. Meunier, et al. (1996). A System of Patterns: Pattern Oriented Software Architechture. New York, John Wiley & Sons.

Clubb, J. M., E. W. Austin, et al. (1985). Sharing Research Data in the Social Sciences. Sharing Research Data. S. E. Fienberg, M. E. Martin and M. L. Straf. Washington, D.C., National Academy of Sciences.

Cousins, S. B. and e. a. (1995). Interpay: Managing multiple payment mechanisms in digital libraries. Proceedings of Digital Library, Austin, TX.

Cowan, R. S. (1997). A Social History of American Technology. New York, Oxford University Press.

Crane, D. (1988). Invisible Colleges: Diffusuion of Knowledge in Scientific Communities. Chicago, University of Chicago Press.

Daniel, R. (1997). Request for Comments: 2169 A Trivial Convention for using HTTP in URN Resolution, Network Working Group.

Darling, C. (1996). “How To Integrate Your Data Warehouse.” Datamation(May 15.).

Date, C. J. and H. Darwen (1992). A Guide to the SQL Standard. Reading, MA, Addison-Wesley Publishing Company.

Fielding, R., J. Gettys, et al. (1997). Request for Comments: 2068 , Hypertext Transfer Protocol -- HTTP/1.1, Network Working Group.

Gamma, E., R. Helm, et al. (1995). Design Patterns: Elements of Reusable Object-Oriented Software. Reading, MA, Addison-Wesley.

Gravanbo, L., C.-C. K. Chang, et al. (1997). STARTS: Stanford Proposal For Internet Meta Searching. Proceedings of the 1997 ACM SigMOD.

Greene, S. (1997). “Metadata for Units of Measure in Social Science Databases.” International Journal on Digital Libraries **1**: 161-75.

Hackathorn, R. D. and W. H. Inmon (1994). Using the Data Warehouse. New York, John Wiley & Sons.

Hassan, S. W. and A. Pepcke (1997). Stanford Digital Library Interoperability Protocol. Technical report. SIDL-WP-1997=0054, Stanford University.

Hildreth, C., M. Aborn, et al. (1985). Report of the Committee on National Statistics. Sharing Research Data. S. E. Fienberg, M. E. Martin and M. L. Straf. Washington, D.C., National Academy of Sciences.

Hurwicz, M. (1997). “Take Your Data to the Cleaners.” Byte **22**(1).

Inmon, W. H. (1996). Building the Data Warehouse. New York, John Wiley & Sons.

Kahin, B. (1995). The U.S. National Information Infrastructure Initiative: The Market, the Net, and the Virtual Project. National Information Infrastructure Initiatives: Vision and Policy Design. B. Kahin and E. J. I. Wilson. Cambridge, MA, MIT Press.

Ketchpel, S., H. Garcia-Molina, et al. (1996). UPAI: A Universal Payment Application Interface. USENIX 2nd e-commerce workshop.

Kimball, R. (1996). The Data Warehouse Toolkit. New York, John Wiley and Sons.

King, G. (1995). “Replication, Replication.” Political Science & Politics **28**(3): 444-52.

King, G., B. Palmquist, et al., Eds. (1997). The Record of American Democracy, 1984-1990, Harvard University, Cambridge, MA [producer], Ann Arbor, MI: ICPSR [distributor].

Landauer, T. K. (1995). The Trouble with Computers. Cambridge, MA, MIT Press.

Library of Congress Cataloging Distribution Service (1993). USMARC Format For Authority Data. Washington, DC, Library of Congress Cataloging Distribution Service.

Object Management Group (1997). CORBAservices: Common Object Services Specification. URL:, http://www.omg.org/corba/sectrans.htm.

Orfali, R., D. Harkey, et al. (1997). Instant Corba. New York, John Wiley & Sons.

Orth, C. W. (1998). Handle Administration Protocol Specification. http://www.handle.net/handle_admin.html, Corporation for National Research Initiatives.

Shafer, K., S. Weibel, et al. (1997). Introduction to Persistent Uniform Resource Locators. Dublin, OH, OCLC Online Computer Library Center, Inc.

Sieber, J. E. (1991). Sharing Social Science Data. Sharing Social Science Data. J. E. Sieber. Newbury Park, CA, Sage Publications, Inc.**:** 1-19.

Sieber, J. E. and B. E. Trumbo (1991). Use of Shared Data Sets in Teaching Statistics and

Methodology. Sharing Social Science Data. J. E. Sieber. Newbury Park, CA, Sage Publications, Inc.**:** 1-19.

Vinoski, S. (1997). “CORBA: Integrating Diverse Applications Within Distributed Heterogeneous Environments.” IEEE Communications Magazine **14**(12).

Weibel, S. L. and C. Lagoze (1997). “An Element set to Support Resource Discovery.” International Journal on Digital Libraries **1**: 176-186.

Letters of Support

Four letters of support are attached from Kevin Guthrie, President of JSTOR, Don Waters, Director of the Digital Library Federation, and William Arms, Vice-President of the Corporation for National Research Initiatives and Richard Rockwell, Director of the Inter-university Consortium for Political and Social Research.

[Via Electronic Mail]

Date: Mon, 13 Jul 1998 17:21:35 -0400
From: Kevin Guthrie <KG@jstor.org>
To: "'sverba@harvard.edu'" <sverba@harvard.edu>
Subject: DLI II Proposal

Dear Dr. Verba,

This is a letter in support of Harvard's proposal for Phase II of the Digital Library Initiative. At JSTOR we are learning a great deal about what it takes to build a useful digital collection. Although we have made enormous strides toward reaching our objectives, we recognize that we are just one of many thousands of components in what will become the "digital library" of the future.

Your proposal to create a Virtual Data Center for quantitative social science data is an important project and represents another component of that digital library. For this library to function effectively, links will have to be made between the many collections in ways that are both convenient and meaningful to scholars. Your proposal to make links between the source data and articles that depend upon it, if implemented, would be of enormous value to social science researchers. We look forward to working with you to investigate the possibilities for making these links between articles in JSTOR and relevant datasets. It is our understanding that the actual task of creating such links will await a later phase of your project.

We provide this endorsement with one caveat. Copyright ownership of the material in JSTOR remains with the original publishers. We will have to get the permission of the print publisher before commencing work on making linkages.

I wish you the best of luck with your proposal.

Sincerely,

Kevin M. Guthrie
President
JSTOR
188 Madison Avenue
New York, NY 10016
Phone: 212.592.7345
Fax: 212.592.7355

COUNCIL ON LIBRARY AND INFORMATION RESOURCES

Commission on Preservation and Access • Leadership • Economics of Information • Digital Libraries

DIGITAL LIBRARY FEDERATION

July 9, 1998

Gary King
Professor
Department of Government
Harvard University
Cambridge, MA 02138

Dear Gary,

I very much appreciated the opportunity to read your proposal for a grant in the federal Digital Library Initiative —Phase II.

As you know, the Digital Library Federation, in which Harvard, is a founding partner, is in the process of implementing its initial program agenda (see enclosure). The motivating goals for the Digital Library Federation (DLF) include:

- Organizing, providing access to, and preserving knowledge that is born digitally and thus is available to scholars, researchers, students, and the general citizenry in no other form.
- Providing an accessible and durable knowledge base that helps improve the quality and lower the costs of education.
- Leveraging digital library facilities for managing intellectual property in support of efforts to redesign the scholarly communication process.
- Extending the reach of higher education to new segments of the citizenry.

The emerging program of DLF to advance these goals includes activities that serve to stimulate the development of:

- The network and system requirements and means for various and diverse classes of users to authenticate themselves as individuals authorized to gain access to diverse and distributed bodies of knowledge.
- The requirements and best practices for repositories to store and provide access to collections of digital works of knowledge.
- Interoperating systems of descriptive, structural, and administrative information that users need to discover and retrieve objects from digital collections.
- Archival mechanisms that preserve the integrity and usability of culturally significant digital objects over a long term.
- The organizational options, including economic models and intellectual property strategies, for effectively managing usable digital libraries.

The Harvard proposal identifies a set of research goals and activities that intersects the interests of the DLF in many significant respects. Your underlying assumptions match ours about the need to link together multiple and distributed collections with sets of interoperable services. By focusing on social science datasets, you are opening lines of inquiry that are especially important to the Federation. Like the DLF, you are seeking better to understand the means of providing a rich set of library services for information that is born digitally. Such information exists in no other form and so has unique requirements, compared to materials that are digitized from other formats and simply mimic those forms in the digital environment. Moreover, the lines of development you have proposed for the Virtual Data Center complements and will likely advance a related project — called the Social Sciences Database Project — that Federation partners, including, Harvard, are pursuing.

The Social Science Databases Project of the DLF is specifically focused on improving the usability of codebooks in relation to the data they describe, particularly in datasets that are heavily used in the undergraduate curriculum. Of special interest to the DLF is the work you are proposing to develop the means for the rapid intake of data and the tools needed to manage these data in a persistent repository and to enable rapid exploratory and experimental analysis of them. DLF is also keenly interested in your proposed investigation of ways to link data to texts that describe and analyze them, and in the research you propose with users to define requirements for interface design, data organization, and data analysis in higher education contexts.

Harvard seems well positioned to undertake the lines of research described in the proposal and to produce results that are both innovative and useful. The early version of the VDC, which I saw in March, provides a solid foundation for further development. As you proceed in your research, I would look forward to bringing the resources of the DLF to bear in providing forums for you to test your working assumptions, and in helping to find the means to move your research findings and the resulting Virtual Data Center rapidly into production operations.

With all good wishes,

Sincerely,

Don

Donald J. Waters
Director

Encl.

Corporation
for
National
Research
Initiatives

July 10, 1998

Research
and
Development
for
the
National
Information
Infrastructure

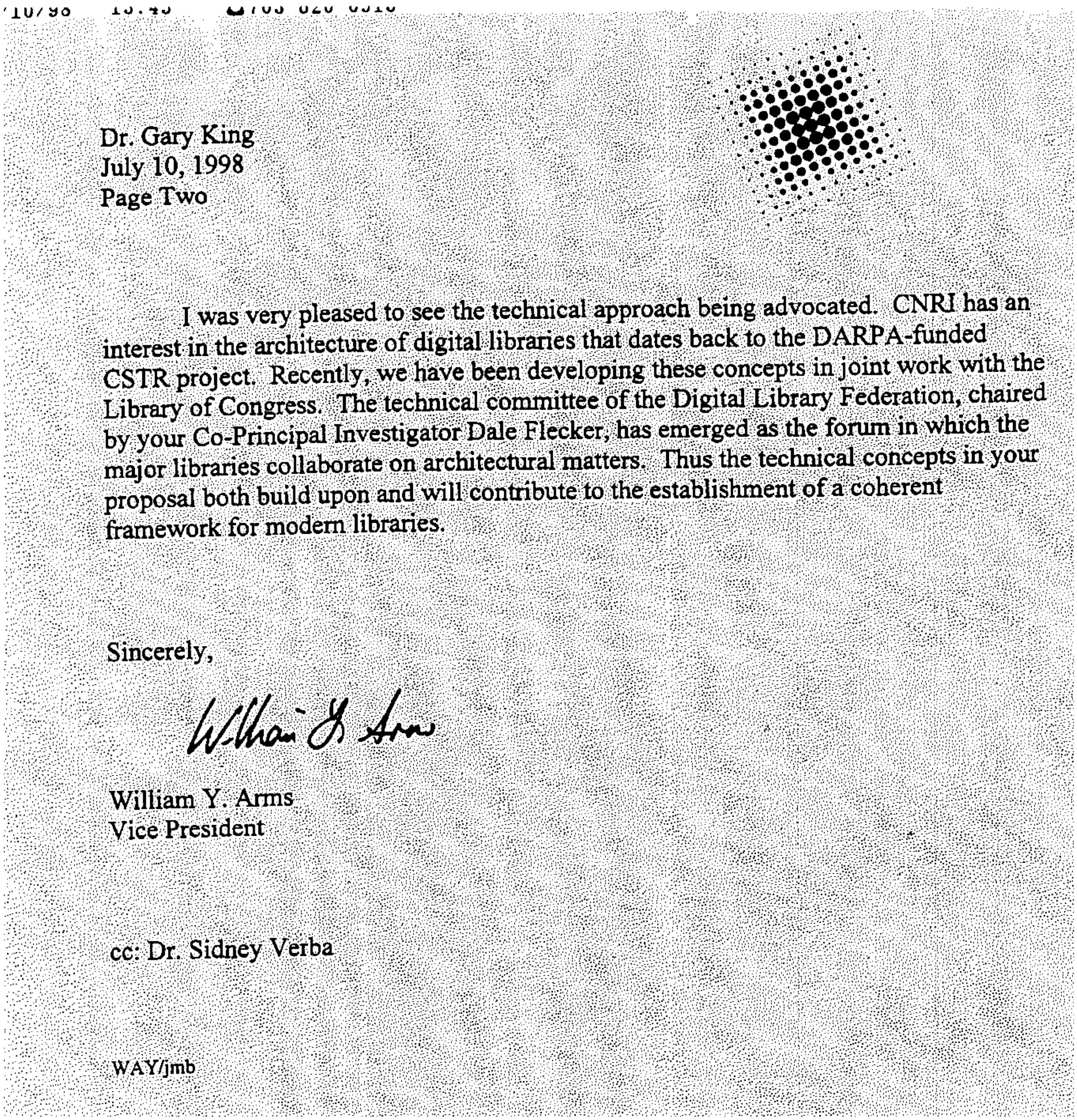
Dr. Gary King
Department of Government
Littauer Center
Harvard University
Cambridge, MA 02138

Dear Dr. King:

This letter is in support of your proposal "An Operational Social Science Digital Data Library."

The area in which you plan to do research is both interesting and important. The management of quantitative social science data has been well served by the ICPSR, but its design and services date from many years ago. In particular, the ICPSR is based on a view of libraries that is centralized rather than distributed. The Digital Libraries Initiative provides a great opportunity to study the underlying user needs and develop a modern distributed architecture for this area. Earlier DLI projects have greatly advanced the state of the art in libraries for video segments and geospatial data. This project could do even more for social science data and its attendant literature.

The team that you have assembled is particularly well qualified to carry out this research. The Harvard-MIT Data Center brings considerable expertise; you have several highly experienced faculty who use this data in their research and teaching; and Harvard's libraries are outstanding.



'10/98 15:45 ☎703 620 0910

Dr. Gary King
July 10, 1998
Page Two

I was very pleased to see the technical approach being advocated. CNRI has an interest in the architecture of digital libraries that dates back to the DARPA-funded CSTR project. Recently, we have been developing these concepts in joint work with the Library of Congress. The technical committee of the Digital Library Federation, chaired by your Co-Principal Investigator Dale Flecker, has emerged as the forum in which the major libraries collaborate on architectural matters. Thus the technical concepts in your proposal both build upon and will contribute to the establishment of a coherent framework for modern libraries.

Sincerely,

William Y. Arms

William Y. Arms
Vice President

cc: Dr. Sidney Verba

WAY/jmb

ICPSR

Richard C. Rockwell
Executive Director

June 22, 1998

Prof. Gary King
Department of Government
Harvard University
Cambridge, MA 02138

Dear Professor King:

The Inter-university Consortium for Political and Social Research, a membership organization of more than 325 colleges and universities in North America, would be pleased to cooperate with Harvard University to assist you in meeting the goals of your current proposal to the Digital Libraries Initiative - Phase II competition. We see our skills and knowledge as being strongly complementary with yours.

Your proposal for a Virtual Data Center promises to bring a new level of accuracy, convenience, and speed to the tasks of identification, retrieval, and subsetting of data sets. The prototype of this service now in operation at Harvard and MIT shows not only that these aims can be accomplished but also that this service will be used. The design of the VDC by a leading researcher, yourself, with your colleagues, ensures that it will be responsive to actual user needs.

Harvard University was one of the early members of ICPSR and has long been a supporter of our work. I am very pleased at the prospect of a stronger relationship with Harvard in this project and wish you the best in your application in the Digital Libraries Initiative competition.

Sincerely,

Richard C. Rockwell

Richard C. Rockwell

cc. Gregory Haley, Columbia
Ann Gordon, USC

Inter-university
Consortium for
Political and
Social Research

At the
Institute for
Social Research,
University of
Michigan

P.O. Box 1248
Ann Arbor, MI
USA
48106-1248

426 Thompson St.
Ann Arbor, MI
USA
48109-1248

1-734-764-2570
(voice)

1-734-764-8041
(fax)

http://www.icpsr.umich.edu
richard@icpsr.umich.edu

Micah Altman

Harvard-MIT Data Center, Department of Government,
Littauer Center M-34, Harvard University, Cambridge , MA 02138
Phone: (617) 496-3784, Fax: (617) 496-5149, E-mail: maltman@data.fas.harvard.edu

Education

Ph.D.	California Institute of Technology,	Social Sciences,	1998
A.B.	Brown University,	*Magna Cum Laude;* A.B., Ethics and Political Philosophy A.B., Computer Science	1989

Honors & Awards:

Weaver Award (best paper in representation), *American Political Science Association*	1998
Pre-doctoral Fellowship, Harvard-MIT Research Training Group in Political Economy	1996-7
John Randolf Haynes and Dora Haynes Fellowship	1995
Anna and James McDonnell Memorial Fellowship	1994
Graduate Research Assistantship	1993
Sigma Xi, Phi Beta Kappa, Magna Cum Laude	1989

Positions

1998	Associate Director	Harvard-MIT Data Center
	Postdoctoral Fellow	Dept. of Government, Harvard U.
1996–1997	Research Fellow	Harvard-MIT Data Center
1990–1992	Member of Technical Staff	Silicon Graphics, Inc.
1989	Software Developer	Sun Microsystems

Publications

"Predicting the Electoral Effects of Mandatory District Compactness," *Political Geography*, Forthcoming Spring 1999

"Traditional Districting Principles, Judicial Myths vs. Reality" *Social Science History* 22(2), (Forthcoming) 1998

"Is Automation the Answer: The Computational Complexity of Automated Redistricting," *Rutgers Computer and Law Technology Journal* 23 (1), 81-142, 1997

[*Published Data and Documentation*]
Gary King; Bradley Palmquist; Greg Adams; Micah Altman; Kenneth Benoit; Claudine Gay; Jeffrey B. Lewis; Russ Mayer; and Eric Reinhardt. 1997. "The Record of American Democracy, 1984-1990," Harvard University, Cambridge, MA [producer], Ann Arbor, MI: ICPSR [distributor]. 1997

Recent Collaborators not Listed Above

None

Graduate Advisors

R. Michael Alvarez, Associate Professor of Political Science, California Institute of Technology
D. Roderick Kiewiet, Professor of Political Science, California Institute of Technology
J. Morgan Kousser, Professor of History and Social Sciences, California Institute of Technology
Scott E. Page, Associate Professor of Economics, University of Iowa

Nancy M. Cline

Harvard College Library
Widener Library, Harvard University
Cambridge, MA 02138
Phone: (617) 495-2401; Fax: (617) 496-4750; E-mail: ncline@fas.harvard.edu

Education

B.A.	(English)	University of California, Berkeley	1968
M.L.S.	(Librarianship)	University of California, Berkeley	1970

Positions

1996-present	Roy E. Larsen Librarian of Harvard College, Harvard University
1988-96	Dean of University Libraries, The Pennsylvania State University
1984-88	Assistant Dean and Head, Bibliographic Resources and Services Division (Acquisitions, Collection Development, Cataloging, Serials), The Pennsylvania State University Libraries
1980-84	Chief, Bibliographic Resources Department (Acquisitions, Serials), The Pennsylvania State University Libraries
1971-80	Head, Government Documents Section, The Pennsylvania State University Libraries
1970-71	Pennsylvania Documents Librarian, The Pennsylvania State University Libraries

Related Publications & Presentations

"Government Information Futures, the importance of government information to the public," American Library Association's program meeting of the Government Documents Round Table, June 26, 1989; published in DttP, vol. 18, no. 2 (June 1990), pp. 117-119.

"Information Resources and the National Network," paper presented at EDUCOM's Net '90 Conference, Washington, DC, March 15, 1990. Published in EDUCOM Review, Summer 1990, pp. 30-34.

"Local and Remote Access: Choices and Issues," paper presented at the Research Libraries Group Symposium on Electronic Access to Information: A New Service Paradigm, San Francisco, CA, July 23, 1993.

"Academic Libraries at Risk," paper presented at the Association for Library Collections and Technical Services President's Program at the American Library Association Annual Meeting, June 26, 1993.

"GIS and Research Libraries: One Perspective," co-authored with Prudence S. Adler, Information Technology and Libraries, vol. 14, no. 2, pp. 111-115, June 1995.

Other Selected Publications & Presentations

Government Micropublishing Programs," Institute on Federal Information: Policies and Access, American University, Washington, DC, May 21, 1979.

Cline, N. LIAS - Library Automation with an Integrated Design. Library Hi Tech 1(2): 33-48 (1983).

"A Strategic Planning Imperative: The Penn State Experience," co-authored with S. Meringolo, Journal of Library Administration, vol. 13, nos. 3/4, pp. 201-22, 1991.

"Government Documents - Assets or Liabilities? A Management Perspective," in Management of Government Information Resources in Libraries, ed. by Diane H. Smith, Libraries Unlimited, Inc., 1993, pp. 222-227.

"Staffing: The Art of Managing Change," in Collection Management and Development: Issues in an Electronic Era, ed. by Peggy Johnson and Bonnie MacEwan, 1994, pp. 13-28.

Recent Collaborators not Listed Above

none

Advisors

none

Dale Flecker

Office of Information Services, Harvard University Library
1280 Mass. Ave., Ste. 404, Harvard University, Cambridge, MA 02138
Phone: (617) 495-3724, Fax: (617) 495-0491, E-mail: dale_flecker@harvard.edu

Education

M.A.	University of Michigan,	Library Science,	1978
B.A.	Wayne State University,	History,	1965

Numerous courses in management (1966-71) and technical topics (1971-86)

Positions

1985–	Associate Director for Planning and Systems	Harvard University Library
1979–1985	Head, Office for Systems Planning and Research	Harvard University Library
1978–1979	Systems Librarian	Harvard University Library
1976–1978	Associate Library Systems Analyst	Yale University Library
1973-1976	Programmer Analyst and Systems Analyst	University of Michigan, Data Systems Center
1972-1973	Programmer	University of Michigan, Purchasing Department

Professional Memberships

American Library Association
Association of College and Research Libraries
Library and Information Technology Association

Related Presentations

“Digital Libraries and Licensing,” American Library Association Annual Conference, 1998.

“The DOI and Digital Libraries,” American Library Association Annual Conference, 1998.

“Technical considerations in digitizing projects,”
Research Libraries Group Symposium on Digital Reformatting, 1996.

Chairman, session on "Data bases and on-line information,"
Association of College and Research Libraries, West European Specialists Section Conference, 1988.

Chairman, session on "Library/computer center relationships," Library and Information Technology Association National Conference 1988.

Other Presentations

"The HOLLIS system," NELINET Program on Integrated Systems, 1984.

Program on Alternative Catalogs, Program Moderator, Association of College and Research Libraries, New England chapter 1982.

"The Harvard Union Catalog," International On-Line Conference, London, 1982.

"Using OCLC archive tapes," NELINET Program on Retrospective Conversion 1981.

Recent Collaborators not Listed Above

None.

Jan Hawkins

Center for Children and Technology, Education Development Center, Inc.
96 Morton Street, 7th Floor, New York, NY 10014,
Phone: (212) 807-4208, E-mail: jhawkins@tristram.edc.org

Education

Ph.D. in Developmental Psychology, Graduate Center, City University of New York, NY.	1987
M.Ph., Psychology, Graduate Center, City University of New York, NY.	1975
M.A., Human Development, University of Connecticut, Storrs, CT.	
B.A. Psychology and English, Tufts University Medford, MA., Summa Cum Laude, Phi Beta Kappa (University of Durham, Durham, England 1971-72)	1973

Positions

1998 (Fall)-	Professor of Practice, Harvard Graduate School of Education
1997	Vice President, Education Development Center
1993	Director, Center for Children and Technology, Director, Center for Technology in Education (OERI, to 1994), Executive Committee, EDC
1990	Director and Senior Research Scientist, Center for Children and Technology Director, National Center for Technology in Education (OERI) Dean, Research Division, Bank Street College of Education, New York, NY.
1987	Associate Director, Center for Children and Technology, Center for Technology in Education, Bank Street College of Education.
1981	Research Scientist, Center for Children and Technology, Bank Street College of Education.
1981	Instructor, Child Development, Brooklyn College.
1978-1982	Research Director, Brooklyn College Day Care Staff Development Project, Brooklyn, NY.
1976-1983	Research Associate and Analyst, Kennan Research and Consulting, Inc., NY, NY.

Related Publications & Software

Hawkins, J. Embedded informative assessment in classroom practice. National Academy of Sciences, February, 1997

Hawkins, J., C. Grimaldi, Terry Baker, Pat Dyer, B. Moeller & J. Thompson. (1996) Distance Learning Evaluation: Dutchess County, NY; Distance Learning Evaluation: New York City. New York: Center for Children and Technology Technical Reports Series.

Hawkins, Jan, Margaret Honey & Cornelia Brunner. Reflections in the mirror: Roles for technologies in education innovation. (working title for book), in preparation for Jossey-Bass

Hawkins, Jan & Allan Collins (eds.) Design Experiments. In preparation, Cambridge University Press.

Hawkins, J. (1996) Technology in education: Transitions. National Education Summit, published by IBM International Foundation.

Other Selected Publications & Presentations

Honey, M, J. Hawkins & F. Carrigg. (1998).Union City On-line: An architecture for networking and reform. C. Dede & D. Palumbo (Eds.). Yearbook of the ASCD, Washington DC: Association for Supervision and Curriculum Development.

Hawkins, J. (1996) Dilemmas. In C. Fisher, D. Dwyer & K. Yocam, Education and Technology: Reflections on computing in classrooms. San Francisco: Jossey-Bass.

Hawkins, J. (1996) Key factors affecting system wide innovation with technology. Report to the National Foundation for the Improvement of Education, National Education Association.

Hawkins, J. Future directions for technology in education. Invited talk, the Ford Foundation, New York, NY, September, 1996.

Hawkins, J. What do we know about technology and schooling?: Research issues. Presentation, President's Council of Science and Technology Advisors, Washington, DC, June, 1996.

Recent Collaborators not Listed Above

S. Malcom, Allan Collins, Margaret Honey, Cornelia Brunner, C. Grimaldi, Terry Baker, Pat Dyer, B. Moeller, J. Thompson, F. Carrigg

Advisors

Gary King

Department of Government, Littauer Center North Yard,
Harvard University, Cambridge, MA 02138
Phone: (617) 625-2027, Fax: (617) 496-5149, E-mail: king@harvard.edu

Education

Ph.D.	Political Science, University of Wisconsin, Madison,	1984.
M.A.	Political Science, University of Wisconsin, Madison,	1981.
B.A.	*Summa Cum Laude*; Highest Honors in Political Science; State University of New York, College at New Paltz,	1980.

Positions

1990–	Professor of Government, Department of Government, Harvard University,
1987–	Director, Harvard-MIT Data Center
1989–90	John L. Loeb Associate Professor of the Social Sciences, Department of Government, Harvard University
1987–-89	Associate Professor, Department of Government, Harvard University
Summer 1985	Visiting Assistant Professor, Department of Political Science, University of Wisconsin, Madison
1984–87	Assistant Professor, Department of Politics, New York University,

Related Publications

A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data, Princeton: Princeton University Press, 1997, (replication dataset: ICPSR s1132).

Designing Social Inquiry: Scientific Inference in Qualitative Research. Princeton: Princeton University Press, 1994 (with Robert O. Keohane and Sidney Verba).

Unifying Political Methodology: The Likelihood Theory of Statistical Inference. Cambridge, England and New York: Cambridge University Press, 1989.

The Elusive Executive: Discovering Statistical Patterns in the Presidency. Washington, D.C.: Congressional Quarterly Press, 1988 (with Lyn Ragsdale).

``Replication, Replication,'' *PS: Political Science and Politics*, with comments from nineteen authors and a response, Vol. XXVIII, No. 3 (September, 1995): 443-499.

Other Selected Publications

``On Political Methodology,'' *Political Analysis*, Vol. 2 (1991): Pp.1-30. (replication dataset: ICPSR s1053).

``How Not to Lie with Statistics: Avoiding Common Mistakes in Quantitative Political Science,'' *American Journal of Political Science*, Vol. 30, No. 3 (August, 1986): Pp. 666-687.

``Not Asked and Not Answered: Multiple Imputation for Multiple Surveys,'' *Journal of the American Statistical Association*, forthcoming (with Andrew Gelman and Chuanhai Liu).

``Estimating the Probability of Events That have Never Occurred: When Is Your Vote Decisive?'' *Journal of the American Statistical Association*, forthcoming, March 1998 (with Andrew Gelman and John Boscardin).

``Why Context Should Not Count," *Political Geography* Vol. 15, No. 2 (February, 1996).: 159-164.

Recent Collaborators not Listed Above

Martin Tanner, Jonathan Katz, Micheal Tomz, Jason Wittenberg, Curtis Signorino, James Honaker, Anne Joseph, Kenneth Scheve, Kenneth Benoit

Advisors

Leon Epstein, University Professor Emeritus, University of Wisconsin

Sidney Verba

Department of Government
M33, Littauer Center, Harvard University
Cambridge, MA 02138
Phone: (617) 495-5740, Fax: (617) 496-5149, E-mail: ywei@harvard.edu

Education

Ph.D.	Princeton University,	1959
M.A.	Princeton University,	1957
B.A.	Harvard College,	1953

Positions

1984–	Carl H. Pforzheimer University Professor, Harvard University
1972–	Professor of Government, Harvard University
1984–	Director of the Harvard University Library
1981–84	Associate Dean for Undergraduate Education, Faculty of Arts and Sciences, Harvard University
1983–84	Clarence Dillon Professor of International Affairs
1968–72	Professor of Political Science, University of Chicago
1968–72	Senior Study Director, National Opinion Research Center
1964–68	Professor of Political Science, Stanford University
1960-64	Assistant and Associate Professor of Politics, Princeton University

Related Publications

Voice and Equality: Civic Voluntarism in American Democracy [with Kay L. Schlozman and Henry E. Brady], Cambridge: Harvard University Press, 1995.

Designing Social Inquiry: Scientific Inference in Qualitative Research, with Gary King and Robert O. Keohane, Princeton: Princeton University Press, 1994.

Elites and the Idea of Equality: A Comparison of Japan, Sweden, and the United States, Cambridge: Harvard University Press, 1987.

Equality in America: The View from the Top, with Gary R. Orren. Cambridge, Harvard University Press, 1985.

Introduction to American Government, with Kenneth Prewitt and Robert Salisbury. New York: Harper and Row, Sixth edition, 1991.

Other Selected Publications

"What Happened at Work Today? Gender, Work, and Political Participation," *Journal of Politics* (Forthcoming), November, 1998 (with Nancy Burns and Kay L. Schlozman).

"Rational Prospecting: Recruiting Political Activists," *American Political Science Review* (Forthcoming), December, 1998 (with Henry Brady and Kay L. Schlozman).

"Knowing and Caring About Politics: Gender and Political Engagement", *Journal of Politics*, November, 1997 (Kay Schlozman and Nancy Burns), 1051-72..

"The Citizen as Respondent: Survey Research and American Democracy"
American Political Science Review, March, 1995, 1-7.

"Beyond SES: A Resource Model of Political Participation"
American Political Science Review 89, 271-294 (Kay Schlozman and Nancy Burns), 1995, 373-89.

Recent Collaborators not Listed Above
none

Advisors
Gabriel Almond

FACILITIES, EQUIPMENT & OTHER RESOURCES

FACILITIES: Identify the facilities to be used at each performance site listed and, as appropriate, indicate their capacities, pertinent capabilities, relative proximity, and extent of availability to the project. Use "Other" to describe the facilities at any other performance sites listed and at sites for field studies. USE additional pages as necessary.

Laboratory:

Clinical:

Animal:

Computer: **The extensive computer, data, networking and hardware facilities will be made available to the VDC project. Furthermore, the Harvard-MIT data center is a participant in the NSF funded Internet II project, and this VDC project will showcase the high-speed networking of Internet II.**

Office: **Offices will be provided for the PI's, post-docs, and associated graduate students.**

Other: ____________________

MAJOR EQUIPMENT: List the most important items available for this project and, as appropriate identifying the location and pertinent capabilities of each.

OTHER RESOURCES: Provide any information describing the other resources available for the project. Identify support services such as consultant, secretarial, machine shop, and electronics shop, and the extent to to which they will be available for the project. Include an explanation of any consortium/contractual arrangements with other organizations.