REPORTS AND COMMUNICATIONS

A Digital Library for the Dissemination and Replication of Quantitative Social Science Research

The Virtual Data Center

MICAH ALTMAN
LEONID ANDREEV
MARK DIGGORY
GARY KING
AKIO SONE
SIDNEY VERBA
Harvard University

DANIEL L. KISKIS
MICHAEL KROT
University of Michigan

The Virtual Data Center software is an open-source, digital library system for quantitative data. The authors discuss what the software does, how it provides an infrastructure for the management and dissemination of distributed collections of quantitative data, and the replication of results derived from these data.

Keywords: open-source, digital library, quantitative data, replication

Researchers in social sciences, and in academia in general, increasingly rely on large quantities of numeric data. The analysis of such data appears in professional journals, scholarly books, and more and more often, popular media. For the scholar, the connection between research articles and data is natural. We analyze data and publish results. We read the results of other analyses, learn from them, and move forward with our own research.

But these connections are sometimes difficult to make. Data supporting an article are often difficult to find and even more difficult to analyze. Archiving, disseminating, and sharing data is crucial to research but is often costly and difficult (Sieber, 1991). Consequently, our ability to replicate the work of others and to build on it is diminished. Researchers, university data centers, and students all face challenges when trying to find and use quantitative research data.

AUTHORS' NOTE: Supported by National Science Foundation award No. IIS-9874747.

Social Science Computer Review, Vol. 19 No. 4, Winter 2001 458-470

The Virtual Data Center (VDC) software is a comprehensive, open-source, digital library system designed to help curators and researchers face the challenges of sharing and disseminating research data in an increasingly distributed world (Altman et al., 2001). The VDC is also a first step toward better citation of data. Current citations of data are typically ad hoc, fragile, and shallow. Ultimately, digital libraries such as the VDC will serve to make citations more robust and research more replicable.

VDC FEATURES

The system provides five areas of functionality:

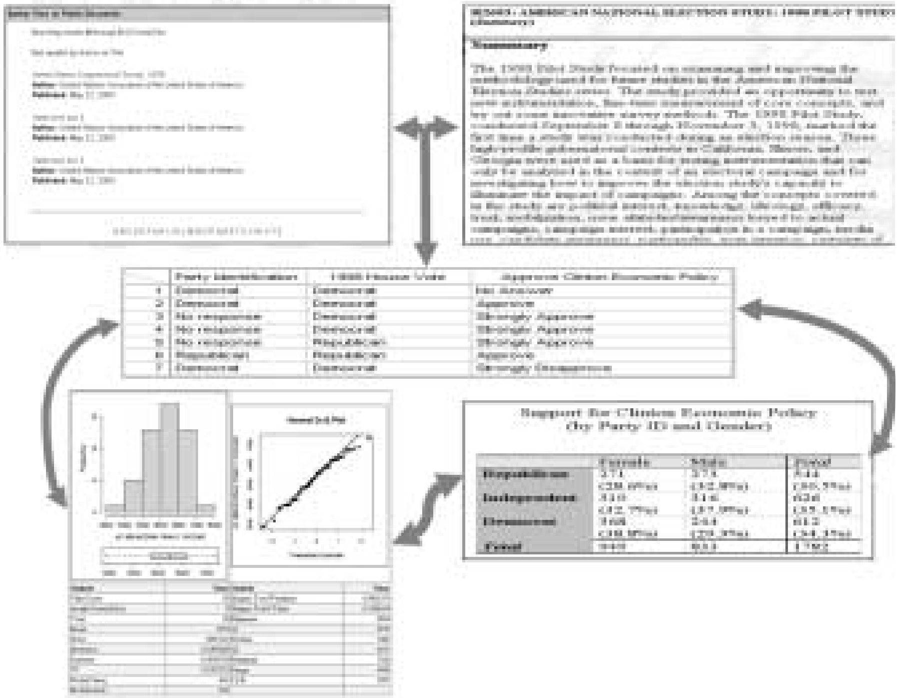
1. *Study preparation:* unique naming and conversion tools for multiple data and documentation formats and tools for preparing catalog records for datasets;
2. *Study management:* file-system independent data set and documentation storage, archival formatting, cataloging;
3. *Interoperability:* Dublin Core, MARC, and DDI (Data Documentation Initiative) metadata import and export and OpenArchives and Z39.50 query protocol support;
4. *Dissemination:* extract generation, format conversion, and exploratory data analysis;
5. *Distributed and federated operation:* location-independent name resolution, distributed virtual collections, federated metadata harvesting, repository exchange and caching, and federated authentication and authorization.

The VDC provides functionality for users, curators, and producers of data. For users, it enables online search, data conversion, and exploratory data analysis facilities. For curators, it provides facilities to create virtual collections of data that bring together and organize data sets from multiple producers. For producers, it offers naming, cataloging, storage, and dissemination of data.

Consider the following use cases: First, an undergraduate is writing a term paper on the 1996 U.S. presidential election; next, a graduate student in the School of Public Health is researching the epidemiology of heart disease in France; finally, a senior professor in the economics department is, for the first time, testing new models of the political factors affecting economic growth. At first look, all three users appear to have very different research needs. The student needs to find a single number—the percentage of women in the Northeast who voted for Clinton. The graduate student is attempting to extract a large subset of data from a larger study along with an accompanying geographic map of the data; the senior professor, meanwhile, needs to develop an extensive set of data comprising interrelated variables from dozens of data sets. Although the magnitudes of these research tasks are different, each researcher faces the same set of core tasks. These tasks all involve searching for a relevant data set, extracting an appropriate subset of the data, and constructing a summary of the data. (These tasks are illustrated in Figure 1.)

For the curator of the university data center and/or library, the VDC provides efficient and flexible dissemination of the collection. The curator can use the VDC system to make all data sets available online through a consistent set of user interfaces. The VDC also assists with preservation of the collection by converting datasets into XML (Extensible Markup Language), preservation-friendly formats and by separating the methods used to access data sets from the storage technology.

For producers, the VDC simplifies archiving, naming, and coordination with disseminators and end users. Through its implementation of the DDI specification, it ensures the information they used in creating the data is retained in the dissemination process and eventually delivered to the end users. In addition, the ability to attach persistent, unique iden-



	Party Identification	1998 House Vote	Approve Clinton Economic Policy
1	Democrat	Democrat	No Answer
2	Democrat	Democrat	Approve
3	No response	Democrat	Strongly Approve
4	No response	Democrat	Strongly Approve
5	No response	Republican	Strongly Approve
6	Republican	Republican	Approve
7	Democrat	Democrat	Strongly Disapprove

Support for Clinton Economic Policy
(by Party ID and Gender)

	Female	Male	Total
Republican	271 (28.6%)	273 (32.8%)	544 (30.5%)
Independent	310 (32.7%)	316 (37.9%)	626 (35.1%)
Democrat	368 (38.8%)	244 (29.3%)	612 (34.3%)
Total	949	833	1782

Figure 1. User Interaction With Prototype Virtual Data Center

tifiers to a particular dataset allows the producer to disseminate data through multiple archives while retaining citations that always point back to the original producer. When a producer serves objects to curators through the VDC, the system will ensure that the curator can access any objects for which they are authorized and that the curator always has an up-to-date copy of metadata and object. If multiple curators share collections, the VDC also guarantees that multiple copies of the same item are assigned the same identifier. That is, the system will "borrow" items from other curators (if authorized), reducing the load on the producer's server. Furthermore, multiple copies or the item are presented to the user as a single work with multiple locations, making it easier to find relevant works from that producer.

IMPLEMENTATION: BALANCING RESEARCH AND PRODUCTION REQUIREMENTS

Our implementation strategy emphasizes *open source* development and integration of the system into a production environment. The director of the Digital Library Initiative, Phase 2, of which the VDC is a part, noted the "unnatural separation" between the producers and consumers of digital libraries and called for a balance among research, application, content, and collections (Griffin, 1998). In keeping with this admonition, the VDC software system is not simply an isolated research project but is also a part of Harvard University's first generation production digital library system. VDC benefits from participation in an unusually large and decentralized library system, from cross-fertilization with Harvard's own digital library efforts (see Flecker, 2000), and from the heavy usage patterns of the Harvard research community.

The requirement that the system support production use in a decentralized environment has a number of implications. First, the architecture must be flexible enough to accommodate the administration of collections and their contents by multiple and independent curators. Second, the system must be accessible by standard Web browsers without special configuration. Third, the system must support the protocols and standards currently in use in library environments. Conversely, as a first-generation production system, there is much in the way of architecture, implementation strategy, and features that remains to be discovered.

VDC CORE ARCHITECTURE

The VDC digital library borrows core concepts from Arms, Blanchi, and Overly (1997) and from NCTSRL (Networked Computer Science Technical Reference Library) (Davis & Lagoze, 1999) and extends these in innovative ways to support services on digital objects, complex collections, distributed authentication authorization, and deep citations. Moreover, the objects stored in the VDC system differ in important ways from these earlier systems.

The basic object managed in the system is the study. Each study comprises a meta-data object and a set of associated data objects. The metadata object follows the DDI standard (Ryssevik, 1999) and contains all of the structural metadata for that study as well as the descriptive meta-data for the corresponding (abstract) intellectual work. The associated data objects consist of text files (usually for supplementary documentation), MIME (Multipurpose Internet Mail Extensions)–typed BLOBs (Binary Large Objects), and/or structured quantitative databases. The metadata object acts to document the study and to tie the associated data objects together.

These objects are managed with a set of cooperating services. The core of the architecture supports four services: the UIS (User Interface Service), the repository service, the name resolution service, and the index service. (See Figure 2. Core components are shown in white.)

The UIS is the gateway to the system and coordinates access to the other components. The UIS supports two user interfaces: one for the end users of the library and another for the curators who manage the collections. Both are accessed through a standard Web browser.

The UIS is implemented as a set of Java servlets, each of which encapsulates access to particular services and objects. Each object or service is itself described in XML, and XSL (Extensible Stylesheet Language) is used to render the object.

The repository stores and manages digital objects and the administrative metadata (such as the object's owner or last time of access) associated with them. A repository access protocol allows for maintenance and hiding the details of their storage (currently a SQL [Structured Query Language] database) from the rest of the system. The repository itself treats every object as a MIME–typed BLOB. All knowledge about complex objects (objects that cannot be rendered by a browser without preprocessing) is encapsulated inside the UIS.

The NRS (Name Resolution System) manages identifiers for each digital object. Each distinct intellectual work stored in the system is assigned a unique identifier. The NRS uses URN (Uniform Resource Name) methods (Daniel, 1997; Moats, 1997) to resolve each identifier to a repository (or set of repositories) that stores a copy of that work.

The IS (Index Server) manages indexing and searching (queries) of the descriptive metadata associated with each object. Index servers act with a large amount of independence—they are assigned sets of identifiers that they are responsible for indexing. In addition, the index servers asynchronously resolve the identifiers to a repository, retrieve the metadata component of these objects, and build indices based on this metadata.

Together, these four services provide the core of digital library functionality. To support specialized services on these objects and to support distributed operations, we introduced a

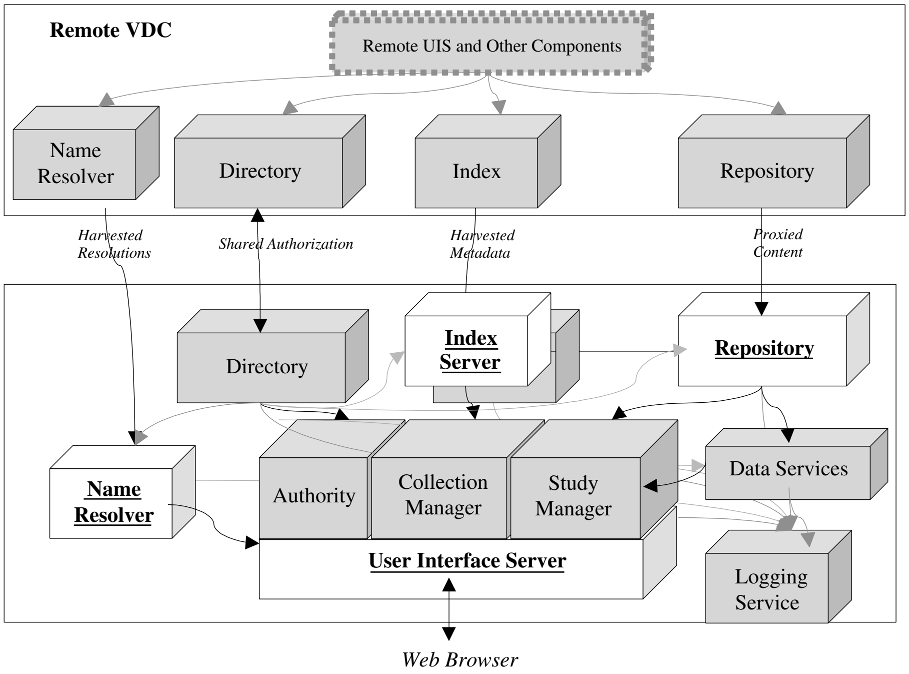


Figure 2: Simplified Representation of VDC Architecture
NOTE: VDC = Virtual Data Center.

number of other services (see Figure 2; new services are shown in gray). We discuss these new services in the next section.

MULTIPLE MODES OF DISTRIBUTION: DISTRIBUTED COMPONENTS AND FEDERATED LIBRARIES

One of the more innovative aspects of the VDC is its support for multiple models of distributed operation. Each VDC library comprises a set of interoperating components that can function independently and that can be distributed across systems. In addition, the VDC supports "virtual collections" within a VDC library that bring together studies indexed in multiple index servers and stored in multiple repositories. Moreover, unlike the NCSTRL system, independent VDC libraries can cooperate together in a federation to share individual studies or entire collections.

In the simplest form of distributed operation, components are distributed across multiple systems or networks, but only a single administrative unit is formed (see Figure 2, bottom portion). To support this scenario, the VDC architecture introduced three new components to the Arms–NCSTRL architecture.

1. A directory services component provides a central registry for all other components in the administrative unit. This service enables components to locate each other and allows components to be added or removed from the system dynamically.
2. A centralized logging facility provides an interface where each of the distributed components can record events that occur in the process of servicing a request. This supports debugging of the system and auditing of the system usage.

3. A data services broker coordinates reformatting, subsetting, aggregation, and analysis of study objects. This component supports the UIS in performing services on data objects for users. Because the UIS consults the broker about what services can be performed on each object, it is possible to add other types of services without modification to existing components.

(Because services are provided through a local broker, these services can be applied to studies copied from other parts of the federation. This is possible even where the federation member providing the study does not support these services locally. A discussion of this federated model follows.)

These three components, along with the repository, index server, name resolver, and UIS (described in the previous section) cooperate to form a single "library" service. Users see one point of presence, and one set of administrative rules is maintained there for the VDC library unit.

The second mode of distributed operation adds an additional dimension of flexibility: Multiple independent VDC systems can be "federated" together to share collections. In this mode, LDAP (Lightweight Directory Access Protocol) referrals are used in each local directory server to point to the directory server of cooperating VDC libraries. The harvester in each library then uses these referrals to locate remote indices. It harvests metadata from the indices to replicate indexing and name resolution information across local name resolvers, enabling any member of the federation to find copies of studies stored in a neighboring repository. When such studies are requested by local components, a proxy is then used to retrieve and cache copies from neighboring repositories.

Authentication and authorization is also federated. Each VDC library maintains authority over how studies in its repositories are accessed. Distributed authentication works as follows: A user can log in from any UIS in the system but must identify their home VDC in the process. The user is then redirected to their home VDC for authentication, which, if successful, supplies signed credentials that identify the user as a member of that institution. Finally, the user is redirected back to the UIS where they originally entered.

When an authenticated user makes a request of the system, this request must then be authorized. Authorization is role based—a user is mapped to multiple roles based on his or her localized profile of attributes (e.g., membership status). Each study is assigned to one or more logical access classes, as determined by its curator. To authorize an operation on an object, the system looks for a {role, class, operation} entry in the local VDC authority table.

In a federated context, access to an object is always determined by the VDC owning that object. The owning system authorizes each request by (a) identifying a remote user's home library from their credentials, (b) mapping that user's attributes to a set of roles at the home library, (c) mapping the home roles to a set of roles within the "owning" library, and then (d) searching for a {role, class, operation} entry in the local authority table.

This process is analogous to how brick-and-mortar libraries function: Guest borrowers can present library cards from a cooperating institution, signaling that they are authentic members of that institution. The local library then assigns them, for example, "guest faculty" status and authorizes access to its materials on this basis. The result is that content from a group of libraries is made available to the users of each library while each library maintains complete control over how its collections are accessed and over the authentication of its patrons.

VIRTUAL COLLECTIONS

In addition to federation, VDC also supports a complementary way of creating and organizing distributed collections—the "virtual" collection. Virtual collections give the curator an opportunity to directly mediate between users and sets of studies. By creating a virtual collection, the curator identifies a body of logical content and how that content should be represented to the user, regardless of whether that content is owned by the curator. Virtual collections are managed by the collection service component with the support of the metadata harvester service.

Figure 3 shows the architecture for virtual collections. Each virtual collection comprises a specification of the content of that collection and a set of views to be applied to that content. The content of a virtual collection comprises a set of queries that are run against a set of local index servers or other collections. Views are then overlaid on the content to provide navigation and display. In addition, the harvesting component runs asynchronously to gather metadata from remote index servers and remote collections, which is then replicated in a local index server and can be used as content for local collections.[1]

For example, consider a curator who wishes to create a virtual collection of studies about Argentina. The content selection rule would specify a set of index servers or other collections in which relevant data are likely to be found as well as a query for all studies in those servers that have Argentina listed in the coverage metadata attribute. Thus, the content of the virtual collection is not fixed—as new studies are added to repositories, they are indexed by the index servers and dynamically incorporated into the virtual collection.

The curator would also designate a set of views that should be applied to this content. He or she might create their own views or use views already supplied by the VDC system. Some examples of views include the following:

- A simple search interface, which allows the user to search Dublin Core fields.
- A recent additions list that filters the query results by creation time and shows the 100 most recently created studies.
- A thematic outline that shows the content as organized by LOC (Library of Congress) subject classification (if available) or other controlled vocabulary.
- An author index that determines the list of authors by analyzing the collection content itself (as opposed to using a controlled vocabulary, as above) and groups the studies in the collection by author.

These views are not static renderings of particular content but rather are logical descriptions of how the content should be searched, navigated, and displayed. So, curators can reuse views and apply them to different virtual collections with different content wherever the metadata attributes used by the view are present in the virtual collections.

Virtual collections are flexible and powerful because they make use of multiple layers of distributed services. Index servers can index items in distributed repositories, and harvesters use standard protocols to gather indices from remote index servers and remote collections. This means it is possible to create virtual collections that extend across, or even beyond, a federation.

ENABLING DEEP CITATIONS IN ACADEMIC JOURNALS

A fundamental goal of the VDC project is to increase the replicability of research by providing a foundation for "deep citation" of quantitative data. The principle that references to data and data analysis be specific enough to support replication of the research is widely

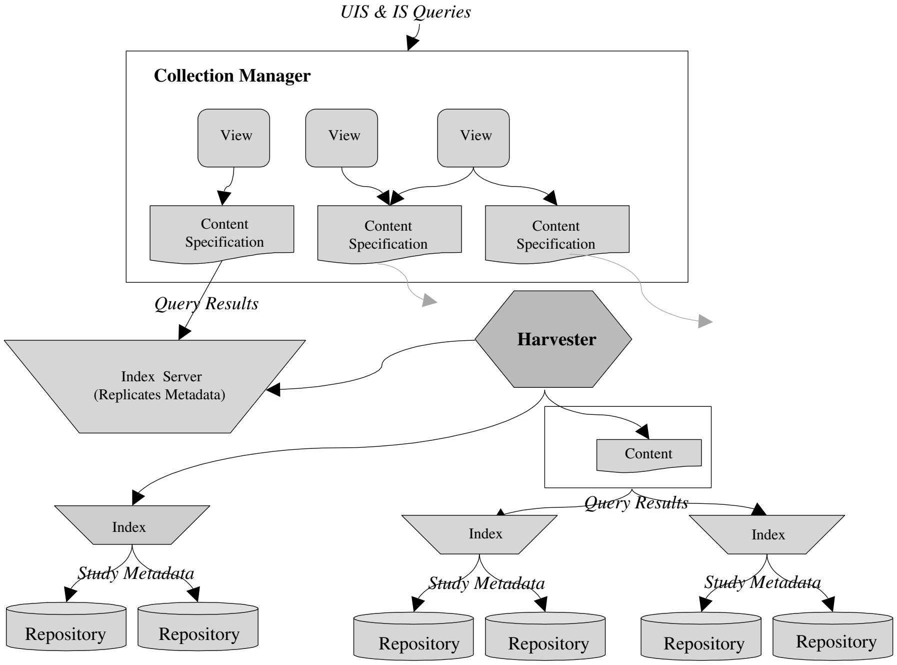


Figure 3: Virtual Collection
NOTE: UIS = user interface service, IS = index server.

accepted as a critical part of publishing (King, 1995). In many other fields, the premier journals, such as *Science*, *Nature*, *American Economic Review*, *Journal of the American Statistical Association*, *Journal of the American Medical Association*, *Social Science Quarterly*, *Lance*, *American Political Science Review*, and the various journals published by the American Psychological Association, explicitly document in submission guidelines that the provision of important data on which the article is based is a condition of publication.[2]

Currently, however, citations (and other references) to shared data are generally insufficient to uphold this principle. Citations to quantitative data as used today, however, suffer from three fundamental problems. First, in published articles, citations to data are ad hoc: They are far less specific and systematic than references to textual sources. No standard citation format exists to describe what data sets were used in a research article. Although data used in research articles are sometimes discussed in footnotes, they are almost never included as a formal citation in a bibliography of references.

Second, citations to data are typically fragile: They fail to answer the most basic questions needed for location and verification of the data over the long term. Which version of the public data set, exactly, did the author use? And, is the version of the data used accessible from a data archive (or library or publisher), although perhaps in a different format, still intellectually equivalent to the version used by the author? Moreover, whereas large public archives have been tremendously successful in obtaining large data collections, many research articles are based on small sets of data that are either collected by the researcher or derived from multiple sources in ways that are not rigorously documented.

Third, citations of data are usually shallow. Citations of data rarely contain precise information about the portion of the data set used and the manner in which it was manipulated. Even where it is possible to find the relevant data set, it is seldom possible for any reader to easily and unambiguously traverse the chain of empirical evidence amassed to support the conclusions published: What are the full details of calculations on the data? How did the analyst move from publicly available data to numerical results? What coding rules were used for which cases? Were variables recoded, and if so, how? How were multiple data sets merged or subsets of the data selected? How were missing values treated in the analysis?

Moreover, datasets have a fundamentally different "functional granularity" (Paskin, 2000) than journal articles—it is essential for a citation of data to enable the researcher be refer exactly to particular data elements used and transformations made upon those elements so as to be able to reproduce published results (Williams, Bunn, Moore, & Poole, 1998).

Several brief examples, drawn from our experience as researchers and data librarians and from personal communication with administrators of data collection projects and archives, illustrate this point. As for ad hoc use of study numbers, many major data archives, such as those of the U.K. data archive, the ICPSR (Inter-University Consortium for Political and Social Research), and the ROPER (Roper Center for Public Opinion Research), use acquisition numbers to informally identify their systems. Often, researchers use these study numbers in an attempt to uniquely identify the data set used in an article. Unfortunately, even during the past 10 years, the limitations of this practice have been revealed.

- A major archive renumbered its acquisitions, invalidating, or rendering ambiguous, many previous references.
- In some cases, a dataset published by a third party (e.g., the U.S. Government or Gallup) is disseminated by multiple archives. When this happens, each archive typically assigns the data set a different acquisition number. Thus, references using these study numbers appear to refer to different data but actually implicate the same intellectual object.
- In some cases, a publisher withdraws data from the data archive. This invalidates the study number, although the data set may continue to be available from other sources.
- When a cumulative research study is extended (e.g., with another wave of data), the previous study number may be "deaccessioned" and a new one assigned. Again, although the relevant data continues to be available, the citation to it becomes invalid.
- Researchers distribute slightly (sometimes substantially) modified versions of data sets that also exist in archives but refer to these in publications as they would the original.

Examples of the fragility of references include the following:

One of the largest continuing data collection efforts in political science uses the CIESIN (Center for International Earth Science Information Networks) geographic correspondence database Web site (Geocorr) when aggregating their data for release to the public. Because neither the Geocorr database nor the methodology behind its creation has ever been explicitly published, when the Geocorr Web site is updated, it may become impossible to reproduce exactly the particular aggregation rules used in previous studies.

The Bureau of Labor Statistics' Current Population Survey, which is widely used by economists, often corrects or updates a particular survey after its initial release. Often, these (admittedly minor) updates are available only online, are not announced, are not tracked by other disseminators (such as ICPSR), and do not change the version number of the data set. It is almost a certainty that the data obtained by a researcher who tracks a reference in a published article back to the CPS (Current Population Survey) Web site will be at least marginally different from the data used by the original author(s) in the published article.

The Poole-Rosenthal congressional roll-call voting scores have been widely used in recent years in analyses of congressional member behavior, and ICPSR archives the data used in the original 1984 study of Poole and Rosenthal and several subsequent years (Poole & Rosenthal, 1984). However, data covering an additional decade, along with corrections to the original voting data in ICPSR's collection, appear only on the Web site of one of the authors, and it has moved several times since it was created.

In regard to the shallowness of current references, consider the following:

In the course of our own research in only the past few years, we have found it productive to replicate more than a dozen research studies (Altman & McDonald, 1999; King, Honaker, Joseph, & Scheve, 2001; King, Tomz, & Wittenberg, 2000; King & Zeng, 2001). Because we had chosen articles in which the data were derived from public sources, the replication process should have been simple and straightforward. But our experience was that many barriers to replication persist. In practice, we find that current replication policies are not sufficient, in part because of the imprecision with which data are cited. Although some authors provided online related materials, typically, we were still unable to reproduce results without contacting authors about missing data management details or supplementary data.

In one replication we attempted recently (Altman & McDonald, 1999), the original author, who had preserved the data from his study, could not successfully reproduce that data from current, supposedly identical sources.

These examples illustrate the range of ad hoc, fragile, and shallow ways in which data are cited at present within the social sciences. Our findings are not isolated, however—similar problems have been reported in other fields (Feigenbaum & Levy, 1993; McCullough & Vinod, 1999). The reliable exchange of research data is vital to the advancement of the social sciences (Ceci & Walker, 1983; Sieber, 1991), and the requirement that another researcher be able to understand, evaluate, build on, or reproduce research without any additional information from an author (King, 1995) is a fundamental principle of scientific research.

In contrast to the system of ad hoc study numbers now in use, a citation to a dataset will enable a particular data set originally obtained in the holdings of one distributor to be found in the holdings of another. In addition, this standard would enable the researcher to precisely and uniformly describe the variables and observations extracted from the dataset to support each particular published result (e.g., tables and figures) appearing in the journal article. This standard would also, as much as possible, allow common transformations and recodings of data to be recorded in a standardized way. Moreover, this standard would specify protocols for versioning datasets. This will allow researchers to consistently recognize substantive changes to a dataset (e.g., when data is adjusted) and consistently ignore nonsubstantive changes (e.g., when data is converted from SPSS to SAS format), and citations referring to "old" datasets will be able to be correctly mapped to newer versions, where possible (e.g., when a new year of data is added to a dataset and the previous study number is abandoned.)

The VDC system takes a substantial first step toward improving citation to data. First, by providing every study in the system with a persistent, location-independent identifier, links to a study will remain valid when repositories are relocated[3]—they are more robust. Second, by converting every study into a canonicalized XML format, we also contribute to the robustness of the citations because the study then becomes insulated from changing statistical software. Third, by developing a syntax for specifying the subset of the study used in an analysis and embedding that in the URL (Uniform Resource Locator) with the persistent identifier, we support citations that are "deeper" than those currently used to refer to data. Fourth, by providing these functions with an open-source system (see as follows) that anyone can examine or use, we take steps toward standardizing the current ad hoc system.

For deep citations to become a reality in academic publishing, however, many technological and institutional mechanisms must still be developed. Naming conventions, registration methods, and citation protocols for publishing and citing datasets must be refined, standardized, and recognized. Moreover, citations to data must be integrated with the systems and databases currently used to manage citations to journal articles.

CREATING AN OPEN-SOURCE INFRASTRUCTURE FOR SCIENTIFIC RESEARCH

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). Even 10 years ago, digital library development faced similar challenges. Now, however, infrastructure and standards are developing rapidly, and we believe that open source now provides a strong base for scientific research and infrastructure.

In the VDC project we have built on the work of numerous other open-source projects instead of "reinventing the wheel." The *R* statistical language (Ihaka & Gentlemen, 1996) is used extensively in the data services component; the PostGres database system (Momjian, 2000) is the basis for our current repository component; the Apache Web server (Laurie, Laurie, & Denn, 1998) and Apache's Jakarta-Tomcat servlet engine provide a foundation for the user interface server (as well as other components); and OpenLDAP (`http://www.openldap.org/`) is used as the basis of the directory service and extensively in the distributed authentication and authorization components. Building on these existing projects enabled us to focus our implementation efforts on the innovative architectural aspects of the system, such as distributed virtual collections, rather than, for example, the nuts and bolts of protocol implementation or file storage.

An open-source development strategy has additional advantages. First, exposing the source code to a wide community of programmers makes it more likely that bugs will be spotted promptly and fixed. Second, the code can be adopted by those who find it useful and so will continue to progress after the project has ended (Raymond, 1999).

ACCESS TO THE VDC SOFTWARE AND ITS CONTENTS

It is our goal that the VDC be of use as part of the infrastructure for doing scientific research. Thus, to support the academic norms of openness and accessibility associated with research data, we are in keeping with Lessig's (1999) assertion that the code supporting the fundamental infrastructure for citations must be open. Our source code is open source and freely available to anyone for use, examination, and modification, forever (see `http://TheData.org`).

Although the code itself is freely available, some data served through the VDC system will not be so. We think there is great value in making the research data itself available freely, and we will make research data that we have produced freely available through the system and are encouraging our partner sites to do the same. We believe that there is even greater value in making the metadata available freely so that others can, at least, discover that a particular collection is in possession of something of interest, even if the data itself is restricted. We also realize, however, that there are often compelling reasons why some data cannot be disseminated publicly. Anything placed within the VDC software is subject to the licensing restrictions imposed on it by the producer and disseminator. Thus, our system supports flexible distributed authentication and authorization mechanisms to give the disseminator complete control over which users can access which parts of the data and what they can do with it within the system. For example, the curator of a data collection can make all data in the system publicly available or restrict access to certain data classes of local or federated users, such as users from ICPSR member institutions. The curator can also apply fine-grained restrictions on particular variables or even restrict how metadata is shared. Although, again, we encourage curators who use the VDC to make their collections and the metadata for them publicly available where possible.

CONCLUSION

Researchers, archives, and casual users all face many challenges when trying to find and use quantitative research data—and the replicability of research suffers as a result. Digital libraries can help to overcome these challenges. By providing a portable software product that makes the process of data sharing automatic and standardized, we believe that the VDC will help researchers and data archives to meet the challenges of sharing and using quantitative data and take the first step toward support for deep citations of quantitative data.

NOTES

1. The NCSTRL collections component (Davis & Lagoze, 2000) corresponds to the content specification in our architecture. The NCSTRL does not have functionality that directly corresponds to VDC views and does not support the harvesting of index servers.

2. 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.

3. Because URN (Uniform Resource Names) are not resolvable within standard Web browsers, we wrap the URN in a PURL (Persistent Uniform Resource Locator) (Shafer, Weibel, Jul, & Fausey, 1997) so that it can be accessed using widely available browser technology. Because PURL's are limited to a one-to-one resolution, one copy of the item (the copy contained in the VDC node that published the work originally) in the system is designated to be the canonical copy for purposes of PURL resolution.

REFERENCES

Altman, M., Andreev, L., Diggory, M., Krot, M., King, G., Kiskis, D., Kolster, E., Sone, A., & Verba, S. (2001). An introduction to the Virtual Data Center project and software. In, *Proceedings of the first ACM+IEEE Joint Conference on Digital Libraries*. New York: ACM.

Altman, M., & McDonald, M. (1999, July 17). *The robustness of statistical abstractions: A look under the hood*. Paper presented at the annual Society for Political Methodology meeting, College Station, Texas.

Arms, W. Y., Blanchi, C., Overly, E. A. (1997). An architecture for information in digital libraries. *D-Lib Magazine* [Online serial], *3*(2). Retrieved May 1, 2001, from the World Wide Web: `http://www.dlib.org/dlib/february97/cnri/02arms1`

Ceci, S., & Walker, E. (1983). Private archives and public needs. *American Psychologist*, *38*, 414-423.

Cowan, R. S. (1997). *A social history of American technology*. New York: Oxford University Press.

Daniel, R. (1997). *A trivial convention for using HTTP in URN resolution* [Online] (Internet Engineering Task Force, No. RFC 2169). Retrieved from the World Wide Web: `http://www.ietf.org/`.

Davis, J. R., & Lagoze, C. (2000). NCSTRL: Design and deployment of a globally distributed digital library. *JASIS*, *51*, 273-280.

Feigenbaum, S., & Levy, D. M. (1993). The market for (ir)reproducible econometrics. *Social Epistemology*, *7*, 215-292.

Flecker, D. (2000). Harvard's library digital initiative: Building a first generation digital library infrastructure. *D-LIB Magazine* [Online serial], *6*(11). Retrieved May 1, 2001, from the World Wide Web: `http://www.dlib.org/dlib/november00/flecker/11flecker.html`

Griffin, S. M. (1998). NSF/DARPA/NASA digital libraries initiative: A program manager's perspective. *D-Lib Magazine* [Online serial], *4*(7). Retrieved May 1, 2001, from the World Wide Web: `http://www.dlib.org/dlib/july98/07griffin.html`

Ihaka, R., & Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, *5*, 299-314.

King, G. (1995). Replication, replication. *PS: Political Science and Politics*, *28*, 443-499.

King, G., Honaker, J., Joseph, A., & Scheve, K. (2001). Analyzing incomplete political science data: An alternative algorithm for multiple imputation. *American Political Science Review*, *95*, 49-69.

King, G., Tomz, M., & Wittenberg, J. (2000). Making the most of statistical analyses: Improving interpretation and presentation. *American Journal of Political Science*, *44*, 341-355.

King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis*, *9*, 137-163.

Laurie, B., Laurie, P., & Denn, R. (Ed.). (1998). *Apache: The definitive guide*. Sebastapol, CA: O'Reilly and Associates.

Lessig, L. (1999). *Code, and other laws of cyberspace*. New York: Basic Books.

McCullough, B. D., & Vinod, H. D. (1999). The numerical reliability of econometric software. *Journal of Economic Literature*, *37*, 633-665.

Moats, R. (1997). *URN syntax* [Online] (Internet Engineering Task Force, No. RFC 2141). Retrieved from the World Wide Web: `http://www.ietf.org/`

Momjian, B. (2000). *PostgresSQL: Introduction and concepts*. Reading, MA: Addison-Wesley.

Paskin, N. (2000). E-citations: Actionable identifiers and scholarly referencing. *Learned Publishing*, *13*(3), 159-166.

Poole, K. T., & Rosenthal, H. (1984). The polarization of American politics. *Journal of Politics*, *46*, 1061-1079.

Raymond, E S. (1999). *The cathedral and the bazaar.* Sebastapol, CA: O'Reilly & Associates.

Ryssevik, J. (1999, September 22-24). *Providing global access to distributed data through metadata standardization. The parallel stories of NESSTAR and the DDI*. Paper presented at the UN/ECE work session on statistical metadata, Geneva, Switzerland.

Shafer, K., Weibel, S., Jul, E., & Fausey, J. (1997). *Introduction to persistent uniform resource locators*. Dublin, OH: OCLC Online Computer Library Center.

Sieber, J. E. (Ed.). (1991). *Sharing social science data. Sharing social science data*. Newbury Park, CA: Sage.

Williams, R., Bunn, J., Moore, R., & Pool, J.C.T. (1998). *Workshop on interfaces to scientific data archives* [Online] Pasadena, CA: California Institute of Technology, Center for Advanced Computing Research. Retrieved May 10, 1999, from the World Wide Web: `http://www.cacr.caltech.edu/isda`.

Micah Altman (Ph.D., California Institute of Technology) is the director of the Virtual Data Center project, which aims to promote the sharing of research data by building open-source software tools. He is also the associate director of the Harvard-MIT Data Center and a postdoctoral fellow in the Department of Government at Harvard University. Dr. Altman has published (or has in press) more than 10 articles, datasets, and reviews on the topics of redistricting numerical accuracy, and digital libraries.

Leonid Andreev is a systems programmer for the Harvard-MIT Data Center and a member of the Virtual Data Center project team.

Mark Diggory is a software engineer on the Virtual Data Center project team.

Gary King is a professor of government in the Department of Government at Harvard University and a founding member of the center for Basic Research in the Social Sciences. He has been a Guggenheim fellow, a visiting fellow at Nuffield College, Oxford University, and is currently a fellow of the American Academy of Arts and Sciences. He has authored and coauthored 50 journal articles and four books in political methodology and other fields of political science. He is currently president of the Society for Political Methodology and serves as director of the Harvard-MIT Data Center.

Akio Sone (Ph.D., Syracuse University) is a postdoctoral fellow at the Harvard-MIT Data Center and a member of the Virtual Data Center project team.

Sidney Verba is a Carl H. Pforzheimer University Professor. He is the author and coauthor of a number of books on American and comparative politics and of articles on those subjects. He is a member of the National Academy of Sciences and a fellow of the American Academy of Arts and Sciences and has been a Guggenheim fellow and a fellow of the Center for Advanced Study in the Behavioral Sciences. His current research interests involve the relationship of political and economic equality, mass and elite political ideologies, and mass political participation. He is also the director of the university library.

Daniel L. Kiskis (Ph.D., University of Michigan) is an assistant research scientist at the University of Michigan and a member of the Virtual Data Center project team.

Michael Krot is a graduate student in the University of Michigan School of Information Science and a member of the Virtual Data Center project team.