
DIGITAL ACCESS TO SCHOLARSHIP AT HARVARD
DASH.HARVARD.EDU


HARVARD LIBRARY
Office for Scholarly Communication

An Introduction to the Dataverse Network as an Infrastructure for Data Sharing

Citation
King, Gary. 2007. An introduction to the dataverse network as an infrastructure for data sharing. Sociological Methods and Research 36(2): 173-199.

Published Version
doi:10.1177/0049124107306660

Permanent link
http://nrs.harvard.edu/urn-3:HUL.InstRepos:4215067

Terms of Use
This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA

Share Your Story

The Harvard community has made this article openly available.
Please share how this access benefits you. Submit a story .

Accessibility

Sociological Methods & Research
Volume 36 Number 2
November 2007 173-199
© 2007 Sage Publications
10.1177/0049124107306660
http://smr.sagepub.com
hosted at
http://online.sagepub.com

An Introduction to the Dataverse Network as an Infrastructure for Data Sharing

Gary King
Harvard University, Cambridge, Massachusetts

The author introduces a set of integrated developments in Web application software, networking, data citation standards, and statistical methods designed to increase scholarly recognition for data contributions; to put some of the universe of data and data-sharing practices on firmer ground; and to facilitate the public distribution of persistent, authorized, and verifiable data, with powerful and easy-to-use technology, even when the data are confidential or proprietary. The goal is to solve some of the political and sociological problems of data sharing via technological means, with the result intended to benefit both the scientific community and the sometimes apparently contradictory goals of individual researchers.

Keywords: *informatics; data sharing; data archiving; virtual hosting; statistical analysis*

Introduction

The infrastructure underlying the world of printed books and articles is powerful, easy to use, and remarkably influential. The persistence of the

Author's Note: I thank Merce Crosas for directing our software development team; our programmers, including Leonid Andreev, Wendy Bossons, Isabelle Chopin, Gustavo Durand, Ellen Kraffmiller, Akio Sone, and Robert Treacy; and Micah Altman and Sidney Verba for many years of collaboration on our earlier Virtual Data Center project and for suggestions on this work. This work was supported by the Library of Congress (PA#NDP03-1), the National Institutes of Aging (P01 AG17625-01), and the National Science Foundation (SES-0318275, IIS-9874747). For helpful comments on earlier versions of this article, I thank Caroline Arms, Peter Bol, Wendy Bossons, Merce Crosas, Jeremy Freese, Sunshine Hillygus, Dan Ho, and Kevin Quinn. Further information is available at the Dataverse Network project home page, http://thedata.org. Please address correspondence to Gary King, Harvard University, Institute for Quantitative Social Science, 1737 Cambridge Street, Cambridge MA 02138; e-mail: king@harvard.edu.

content of print documents is ensured by simple procedures (copies are made, printed, and stored in public libraries all over the world), and community standards like interlibrary loan agreements and systematic cataloging methods make finding them easy, even if published long ago. The only technology required to extract information from this material is a set of eyes (and knowledge of the language), and no special procedures need to be followed for books or articles published even hundreds of years ago. Verifying that you have indeed found the exact book or article listed in a given citation is easy. Various information technology innovations have been layered on this system, such as digitized books and articles, computer search techniques, cross-library common catalogs, citation services, and others.

In contrast, the world of quantitative data[1] is not remotely as organized or secure. A number of highly successful international archives have formed in specialized areas, and they have managed to obtain and preserve many of the larger data sets that have been produced. However, in many scholarly fields, the data used in large fractions of published books and articles do not exist in public archives and often cannot be located anywhere. Data sometimes exist on individual researchers' Web sites, without professional backups, off-site replication, plans for format conversion and migration, or professional cataloging. Sometimes URLs (uniform resource locators, or browser Web addresses) are given, but they often do not last for long. Data created more than 5 to 10 years ago may exist in defunct storage media or the inaccessible formats of old statistics or database packages, operating systems, or hardware.

Data citation practices differ across fields, among archives, and across and even within individual publications. Data are sometimes listed in the references, sometimes in the text, and only occasionally with enough information to guarantee future access to the identical data set. Data referred to may no longer exist, may not be available at all, or may be available only from the author or with his or her approval (which is sometimes forthcoming only if the author thinks he or she won't be criticized). Data listed as available from the author are in any event unlikely to be available for long and will not be accessible after the author retires or dies. Copy editors have few fixed rules for citing data, and often no rules at all. Replicating published tables and figures, starting from the original data, is often difficult or impossible (see Dewald, Thursby, and Anderson 1986; Fienberg, Martin, and Straf 1985; King 1995, 2003, 2006).

In increasing numbers, researchers have been responding to this situation by attempting to adhere to the replication standard—"Sufficient

information exists with which to understand, evaluate, and build upon a prior work if a third party can replicate the results without any additional information from the author" (King 1995: p. 444)—by making publicly available a *replication data set* for each of their empirical articles or books. Replication data sets include the original data and any other information needed to reproduce the numerical results in a published work. This type of information is especially important in fields where rerunning the experiment from scratch is infeasible. Increasing numbers of scholarly journals have been supporting this movement by adopting rules that encourage authors to make replication data sets available or that require them as a condition of publication. And scholars in a variety of other fields have encouraged the support of related agendas, including sociology (Freese 2007 [this issue]), economics (Anderson et al. 2005), medicine (Bachrach and King 2004), psychology (Johnson 2001), education (Schneider 2004), engineering (Whitbeck 2005), earth sciences (Board on Earth Sciences and Resources 2002), life sciences (Board on Life Sciences 2003), and others (Klump et al. 2006).

Although many debates among scholars have occurred over the nature of replication requirements, and some opposition remains, the trend in most fields is strongly in favor of some version of these rules. This should be no surprise. After all, sharing data with other scholars is perfectly rational for authors, even in the most narrow, self-interested sense: Articles with accessible data are cited *twice* as often as otherwise equivalent articles that do not provide data access. Journal editors who require authors to make data available are also acting rationally in advancing the success of their publications: Articles in journals with replication policies that make data available are cited *thrice* as frequently as otherwise equivalent articles without accessible data (Gleditsch, Metelits, and Strand 2003). And with scholarly impact a key measure of the success of an academic, the citation count is close to the coin of the realm (with an effect measurable even in salary; Diamond 1986). These results make a great deal of sense because the contribution of an individual scholarly work is much greater if other scholars can validate it and build on it to produce new knowledge. Yet the advantages of data sharing to the scientific community as a whole are likely to be considerably greater than even these substantial individual benefits. Of course, merely because an activity is in our individual or collective interests does not mean that we will partake. What is needed for authors and editors to participate is not only that they understand the benefits, but also that the task be sufficiently simplified and

the efforts distributed so they are not distracted from their primary goals of authoring and editing.

In this article, we respond to this challenge by introducing a set of integrated developments in Web application software, networking, citation standards, statistical methods, and other technology that is intended to put some of the universe of social science data and data sharing on somewhat firmer ground. The plan described here is intended to produce progress for replication data sets associated with specific publications as well as for data sets collected separately from or prior to related books or articles.

Our specific focus has been on social science data, but most of the solutions are generic and may apply more broadly. Throughout, the goal is to offer technological solutions to political, social, economic, and intellectual problems holding back progress in science and technology due to data-sharing issues. For editors, we show how the plan can be implemented so simply that it can be delegated entirely to a copy editor or clerical staff and does not even use much of their time. For authors, we provide a simple Web application to accomplish all relevant tasks, supported behind the scenes by a wide range of services and without any local software installation required. The benefits are not only simplicity; journals, authors, archives, and data producers will each receive substantially more credit in terms of visibility on the Web and formal scholarly citations than they do now for the data they make available and articles they publish.

We begin the following section by setting out the special requirements needed for effective data sharing. The next section then describes the solutions we have developed and are making available.

Requirements for More Effective Data-Sharing Infrastructure

We outline here eight requirements that, if met, should substantially improve the data-sharing infrastructure and greatly increase the productivity of the scholarly community. For ease of exposition, we focus mainly on the role of journals and authors in making replication data sets available. But, as should be clear, the same infrastructure would also directly benefit other types of data collections either not associated with specific publications or that become associated well after the data were made available. Similarly, book publishers, graduate departments supervising dissertations, and even undergraduate programs with senior theses would benefit from the infrastructural requirements described here when data sets are created in the

course of these written works. We also discuss all these other applications in the third section.

Recognition

The central role of the scholarly journal in helping authors to make data associated with its published articles available, and of the author in creating the data, must be publicly recognized in a more visible way than now occurs. Citation credit should be apportioned both for the original article and separately for the data. Journals ought to be able to unambiguously and visibly brand their replication data archives as their creation, even while using the services of large professional archives and others to make this task easy. Authors should also have their own replication data archives with their own contributions on their own Web sites, also without having to construct these themselves.

Public Distribution

Data distributed with an article should become accessible to the scientific community without having to obtain permission from the author for each use. Having to obtain permission from an author to read a published article, such as by agreeing ex ante not to criticize it in print, is so obviously unacceptable it no longer occurs. The same should be true for data. Science requires the transmission of information through public means, not private agreements.

Authorization

Although free and open access to data is ordinarily preferable, it is neither always feasible nor necessary for the purpose of guaranteeing the public distribution of quantitative information. Those who wish to access the data can reasonably be asked to fulfill whatever authorization requirements the original author needed to meet in order to acquire, distribute, and archive it in the first place. This may include signing a licensing agreement (such as agreeing to respect confidentiality pledges made to research participants), signing the equivalent of a guest book, being at an institution with a membership to the right archive (like the Interuniversity Consortium for Political and Social Research [ICPSR] or the Roper Center), or even paying a fee. And different requirements may apply to different portions of a data set.

Validation

Editors, their designees, and future researchers need to be able to check that a specific set of data exists, even if they have not met the authorization requirements. Just as a copy editor can ensure that a citation to a foreign language source is appropriate, the editorial staff should not have to be experts in the substance or methods used in the article and should not have to download and examine the data themselves. Editors need to be able to verify the data's existence and identity, even if highly confidential, and without necessarily being able to see the data.

Verification

Journals and researchers need to ensure that the data associated with each published article are frozen and cannot be changed without detection. Future researchers also need to be able to verify that data they obtain are in fact the data the author originally made available. Somehow, journals must ensure data authenticity into the future, even if at some point the data are converted to a new format. Thus, we need to be able to verify that a data set that has been transformed from SPSS to Stata to R, from a PC to a Mac to Linux, and from 8-inch magnetic tape to 5.25-inch floppies to a DVD actually contains information content identical to that the author originally made available. Similarly, updates and new versions are desirable for some purposes, but the original version associated with the article must remain available.

Persistence

Researchers decades from now need to be able to find the data, access it, validate that it is the data set associated with the article in question, and verify that the data set contains the same information from it that the author originally provided. Journals and the scientific community need to follow some procedure that gives us confidence that these facts will remain true into the indefinite future, no matter what changes will come in methods of data distribution and network access, data storage formats, statistical and database software, operating systems, and computer hardware.[2]

Ease of Use

These tasks must be simple and easy to use and follow. At present, neither editors nor authors employ professional archivists and cannot be

expected to maintain the software and hardware infrastructure necessary to follow professional archiving standards. Journals ought to be able to rely on the infrastructure developed by existing archives or others to provide these services to them easily and automatically.

Legal Protection

How many journals running replication data archives have engaged legal counsel to consider the potential liability they may have about the way they accept and distribute data? Most journals merely post on their own Web sites whatever data authors submit, with no check by journal staff, no internal review board (IRB) approval, and not even any signed testimony by the author that distributing the data would violate no laws. Publishers have well-honed procedures for dealing with copyright and liability issues for printed matter, but these standard copyright assignment forms do not cover acquiring and distributing data off the printed page. This is especially true if the data have not been checked for material that may be proprietary, defamatory, insufficiently de-identified, obscene, or otherwise potentially illegal to distribute. Of course, journals should not be expected to employ lawyers or deal with IRBs, so this problem needs to be addressed in a way that does not put them, their university, their publisher, or any organization associated with the journal in legal jeopardy. Archives, on the other hand, are well set up to deal with these issues, so journals and other data collectors ought to be able to take advantage of all the work these archives have done and the infrastructure they have in place.

The Dataverse Network as a Proposed Solution

The technology used by the Dataverse Network project includes Web application software, which is a program that can be used by anyone with access to a Web browser and without any specialized software installed on one's own computer. The Google search engine is an example (http://google.com). The Web application software is *hosted* (i.e., installed) on computers managed, owned, or under the control of the host, not the user (e.g., the Google company, not Google users, own the computers that run Google search software).

Web application software, which must be physically hosted somewhere, can also be used to provide *virtual* hosting facilities so other Web sites can provide the same facilities as the centralized host but branded in the style of the local Web site and without any software installed locally.

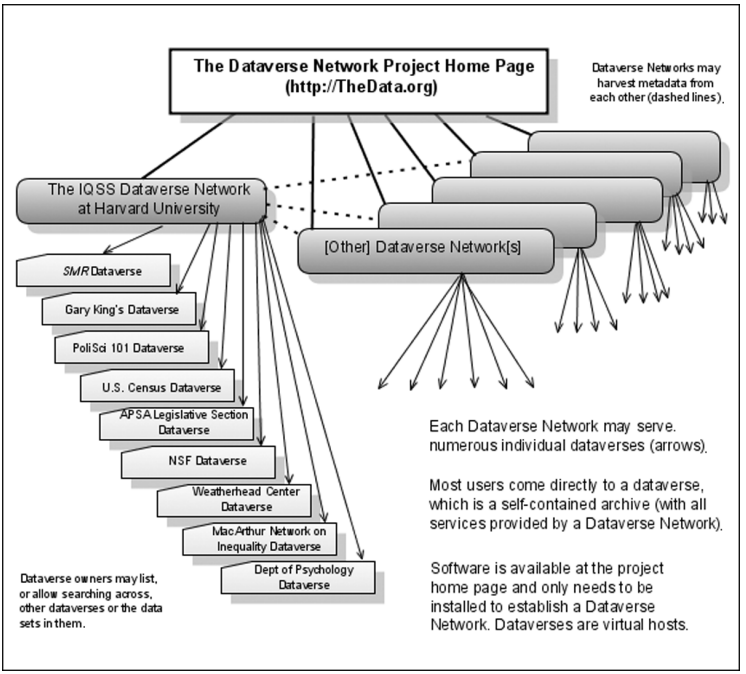
For example, Google provides virtual hosting facilities so that searches can take place from other people's Web pages, with the Google search results returned in the same style as the site from which it originated; the owner of the Web page does not have to install all of Google's software (even if it were possible). For example, a virtual host for Google searches exists at my home page (http://gking.harvard.edu). If you type something into the Google search box there, search results are returned with all the menus and styles of the rest of my Web page; however, if you look closely (at the address bar of your browser), you will see that the URL of the search results page is not the URL of my home page, but instead is that of a Google server, which is the physical host. To set up this virtual host, I entered my Web site style information at a centralized Google Web site and then changed a few lines of HTML on my own Web page. I did not need to install any software on the computer that hosts my Web site. The Dataverse Network project makes extensive use of a more sophisticated version of virtual hosting technology.

The centerpiece of our proposed approach to meeting the requirements set out in the previous section is a Web application called the Dataverse Network, which we developed. Although Dataverse Network software is open source and freely available, the vast majority can install virtual hosts or use those hosts and so do not need to install any software locally. This method of delivering services makes available sophisticated technology in a way that is extremely easy to use by a wide range of people and organizations.

The Dataverse Network project is organized into three levels, illustrated in Figure 1. Users will benefit from all three but most will interact only with the third level:

1. The Dataverse Network project home page, a single Web site listing the individual Dataverse Networks and providing access to software, documentation, and related project information and software tools (see http://thedata.org).
2. Multiple individual Dataverse Networks, each of which represents a physical installation of our software on computers it owns or controls, by an organization with its own data repository, and that hosts dataverses owned by itself or others of its choosing. For example, the Institute for Quantitative Social Science [IQSS] Dataverse Network can be found at http://dvn.iq.harvard.edu/dvn.
3. Many individual dataverses, each of which is a virtual host providing a self-contained archive, physically hosted and served by a

Figure 1
Structure of the Dataverse Network Project



Note: IQSS = Institute for Quantitative Social Science; APSA = American Political Science Association; NSF = National Science Foundation.

specific Dataverse Network. A dataverse requires no local software installation but is entirely under the control of and branded in the style of the Web site of the journal, author, university, or other dataverse owner. Yet all software and hardware installations, management, archiving, preservation, and other hassles are taken care of by the organization installing the relevant Dataverse Network. The first page of a dataverse can be as simple as a list of data sets, a hierarchical organization of data sets in one or more research areas, or a list of data sets organized by a journal's volumes, issues, and articles. Linked pages from the main dataverse page, which are also virtually

hosted, include numerous services such as citation information, documentation, subsetting, translation, and statistical analysis services.

For example, a dataverse with my data sets can be found by clicking on ''data'' at my home page. Similarly, a dataverse with the holdings of the Murray Research Archive can be found by clicking on ''find data'' on its home page, http://murray.harvard.edu. The dataverse in each has the exact look and feel of the home page but, if you look closely, you will see that the URL is different, which means it is served from elsewhere and is why no local software or hardware installations are necessary.

Dataverse Network software is new and written from scratch with the latest technology (see the appendix for details), but its design builds on two decades of experience with prior systems. These include most recently the Virtual Data Center (Altman et al. 2001), which we developed, ran, and used to serve data to the Harvard University and MIT research communities (at the Harvard–MIT Data Center, which is part of the IQSS at Harvard University). Precursors to the Virtual Data Center we built ranged from stand-alone software guides to local data up to and including pre-Web programs to FTP cataloging information to other sites across campus automatically at designated times. Each improvement in automation has dramatically increased data usage among our faculty, staff, and students.[3]

We now explain how this all works in sequential steps, from the perspective of journal editorial staffs, authors, future researchers, teachers, scholarly organizations, granting agencies and research centers, major research projects, universities, departments, libraries, data archives, and confidential data producers. (As most of these sections build on those that come before, they are intended to be read sequentially.)

Scholarly Journals

To fulfill the requirements above, a journal using the new system needs only to change one procedure and a few lines on its Web site. The procedural change involves asking the copy editor to check that any data set used in the article is properly cited and to include this in the journal's standard ''instructions to contributors.'' Copy editors already verify that references to print sources are cited according to their standards, so this rule would be implemented in the course of their normal activities. The one-time Web site changes are also easy to implement, require no software installation, and should take only a few minutes. We now describe the data citation and Web site changes.

Data citation procedures. The copy editor should ensure that articles adhere to the standard for the citation of quantitative data developed in Altman and King (2007), which we briefly summarize here; the software we describe below will create all elements of this citation automatically for a given data set.

A real example of a citation to a replication data set (used in King and Zeng 2007), as it would appear online, is

> Gary King; Langche Zeng, 2006, "Replication Data Set for 'When Can History Be Our Guide? The Pitfalls of Counterfactual Inference,'" hdl:1902.1/DXRXCFAWPK UNF:3:DaYlT6QSX9r0D50ye+tXpA== Murray Research Archive [distributor]

where the authors, year, title, and distributor are easily readable by humans, and the two other components—beginning with hdl: and UNF:, separated by blank spaces, and breaking between lines when necessary without a dash—are technology-based and machine readable. The first component is hdl:1902.1/DXRXCFAWPK, a "unique global identifier," which is a string, unique among all such names, that by international convention and specific institutional guarantees, permanently identifies the data.[4] This identifier is designed to persist even if URLs are no longer used or if the Web itself is replaced with something else. For convenience in the medium term, the citation also includes a URL that appears for the online version only by hotlinking the identifier (which is why it appears underlined in the citation above). The URL makes it easy to retrieve the information using a Web browser today. In print, the URL would be included as an additional element in the citation.[5]

The citation also includes a "universal numeric fingerprint" (UNF), which is a string of letters and numbers computed from the data set that uniquely identifies its content, UNF:3:DaYlT6QSX9r0D50ye+tXpA== .[6] UNFs have four especially attractive features of use. First, the UNF algorithm uses cryptographic technology to ensure that if any portion of the data set is changed, the identifier will also change. Tricking the algorithm so that a change in one or more numbers in the data does not alter the UNF is for all practical purposes impossible (and has never been accomplished even for hypothetical examples).

Second, the UNF is calculated from the content of the data, not the form in which it is stored. Thus, no matter what format conversions the data have gone through to deal with changes in statistical packages, operating systems, database programs, computer hardware, storage media, and

access methods, a researcher decades from now can recalculate the UNF from the downloaded data set and can use it to verify that the content of the data is identical to that which the author originally provided.

Third, knowing only the UNF, a copy editor can be confident that he or she will from then on be referencing one specific data set that cannot be changed. Thus, the UNF is like the ultimate summary statistic, such that if two data sets have the same UNF, they must also be the same. Such is not true for means, variances, or other commonly computed statistical summaries.

Finally, the UNF has noninvertible cryptographic properties, which guarantee that learning the UNF of a data set conveys no information about the content of the data. This means that with access only to the UNF, the copy editor can still be confident that the citation refers to one specific, unchanging data set, even if he or she is not permitted to see the data. Authors can also take advantage of this property of UNFs to distribute the full citation, including the UNF, even for the most sensitive confidential or proprietary data, all without any risk of disclosure. UNFs thus provide complete protection for the author and any of the author's agreements with research participants or data providers at the same time as they provide complete protection for the journal and the scientific community. The technology eliminates the grounds for disagreements due to divergent interests: Data can be validated and verified without violating human subjects and other rules.

Web site procedures. Two procedures are required of the journal's Webmaster.[7] First, the Webmaster needs to sign on to a Dataverse Network Web site and create a dataverse for that journal. A dataverse is a self-contained online data archive specially designed in this case for the journal, with all the services of a professional archive but without the hassle, expense, hardware, software, or legal issues involved in creating an archive from scratch. To accomplish this, the Webmaster would merely copy some HTML or other Web style information from the journal's Web site, and all the Web pages in the new dataverse would then have the same look, feel, and branding as the rest of the journal's Web site. For most journals, the main page of their dataverse would offer an expandable list of year, volume, issues within volumes, articles within issues, and data sets corresponding to specific articles. The final step is to link to this new site from the journal's main Web page.

For visibility and branding purposes, anyone viewing a journal's dataverse would interpret it as being part of the journal's Web site, but it

would actually be hosted and served by an existing archive and the Dataverse Network project. This means that the journal can take advantage of all the infrastructure of the Dataverse Network and any professional archive to which it is linked. The editors thus do not need to worry about installing and maintaining local software or hardware. At the same time, all the recognition for the contribution of establishing the journal's dataverse and all its associated studies would remain with the journal and its sponsoring scholarly organization, publisher, and authors.

The journal's dataverse is not a single Web page, but rather a collection of pages dynamically generated via the latest technology, that provide a large and growing array of data-related services to the journal, authors, teachers, librarians, and future researchers. The journal's dataverse offers the public a simple list of archived data sets, organized as the journal sees fit, but in most cases sorted hierarchically by year, volume, issue, and article. The same journal dataverse provides protected facilities to allow authors of accepted articles to upload data and create a citation. Other researchers will then be able to search or browse for available data sets. Users can view the citation and, with proper authorization, access the data (and many other features we describe below). These and other features will make the journal and its articles more visible and more valuable to the scientific community by increasing the rate at which the journal's published work is read, cited, verified, and used in subsequent articles.

Authors

The journal's dataverse offers authors of accepted articles a secure facility to upload data and associated documentation. After uploading, the author is issued a formal citation to these data, including a unique global identifier, UNF, and the other components, to be used in the author's article and to be cited in other publications. We also recommend that authors add data citations to their curriculum vitae in a separate section, just as they do now for books, articles, software, and other contributions.

When authors upload data and documentation, they will also enter associated *metadata*. Metadata is "data about data" and may include information such as the article for which the data were collected or are otherwise connected with, how the data were obtained, and keyword summaries of its content. It may also include more detailed information such as variable names and labels, value labels, missing value codes, data structures, and other types of extensive machine-readable documentation.[8] Some standards for minimal metadata fields may be required by journals

or particular dataverses, but in general the more metadata the author provides, the more services the Dataverse Network will be able to offer other researchers. These services include descriptive statistics, graphics, more sophisticated statistical analyses, format conversion, and others.

Authors can also be given the option of establishing their own dataverse on their own Web pages. The main page of the author's dataverse would likely be a list of all data they have made available, including replication data from any journal they published in and any other data sets they created. To do this, they would merely need to follow the same two simple procedures that the Webmaster for the journal follows, described above. The lightweight nature of Dataverse Network technology means that having a journal dataverse and researcher dataverse with some overlap, by referring to some of the same data sets, requires almost no additional administrative overhead, storage, or effort.

A dataverse for an individual researcher would replace the lists of data sets some scholars make available on their own Web pages and would meet the requirements listed in the previous section. After all, individual scholars are also not professional archivists, and the data sets on their own pages are not likely to persist in readable formats or perhaps at all when the scholar changes computers, switches Web addresses, moves universities, retires, or dies. Depositing data directly in archives, such as the ICPSR, would solve this problem, but this happens too infrequently, even when required by granting agencies. Part of the problem is that some scholars perceive that their work would be insufficiently recognized and instead seen as a product of the archive. Regardless of its accuracy, the perception does seem to be widespread in some fields. Fortunately, the problem can be fixed, as this particular use of Dataverse Network technology should lead to proper recognition for the authors' hard work and indeed more visibility and scholarly citation credit than they would have by listing data on their own page or shipping it off to an archive. At the same time, they will also be making a larger contribution to the scholarly community by sharing data and guaranteeing its persistence through the support of a social science archive. We intend that a not insignificant side effect of these developments will be a substantial increase in the holdings of our public data archives.

Future Researchers

Data sharing is only of use if there are others to share data with. From the perspective of the (future) researcher who may be able to build on or

learn from existing data, the benefits can be classified in terms of data access and statistical analysis.

Access to a specific study. A dataverse sponsored by a journal or researcher should make citing and finding data considerably easier for future researchers. A scholar can read an article or book, see the data citation, and use it to access the same data as were used in the article. The data should be able to be found no matter how much time has passed since the original publication and no matter what format and physical location the data set is stored in. That each dataverse is part of the larger Dataverse Network ensures that data in one dataverse will be easy to locate from anywhere on the network and indeed anywhere on the Web.

Once the data are located, the researcher can compare the UNF stored with the data to a new computation of the UNF from the data downloaded to verify that the data has identical content to the original. This, of course, can all be done automatically. The citation (and the Web page summarizing the available metadata) will list the UNF constructed when the data set was originally uploaded. Another UNF can then be created by the software at any time, online, and without having to download the data. (Of course, authors may also make available through the network the computer code used to produce their tables and figures.)

Researchers can also use the dataverse's *forward citation* abilities to check whether the original author or others have subsequently updated or incorporated the original data or created a new data set somehow related to the original. New versions will have their own citations, with links back to the original. Standard Web search techniques can also be used with the data set's unique global identifier to find articles or books that used the original data.

Access to the universe of data. The dataverses created by journals, authors, or others are branded according to the wishes of the editors and authors and are hosted virtually at their own Web sites. However, dataverses are not necessarily isolated from each other. Instead, they each belong to a specific Dataverse Network that can organize and display them for browsing or searching. Dataverse Network software includes powerful search and advanced search features that allow one to find data by looking across all data sets, across those within a specific Dataverse Network, or only within a specific dataverse. The search capabilities also enable one to search for a specific dataverse that offers a specially organized view on some relevant area.

In addition to search capabilities, each dataverse also offers detailed customizable choices to dataverse owners for providing a hierarchical organization of available information. Each hierarchical organization of data sets, and even of other dataverses, provides a convenient overview of information easily browsed by users, as it is conceptualized by the dataverse owner.

Any Dataverse Network may harvest metadata from any other (willing) Dataverse Network so that users who come to a dataverse in one network can search across all the studies in any other included network. In addition, a Dataverse Network includes facilities to easily harvest metadata and data from providers who do not run Dataverse Network software directly.

Many dataverses and several Dataverse Networks are now in the process of being set up.

Data analysis. When sufficient metadata have been included with a data set, researchers with proper authorization will be able to take advantage of many services available to study these data. (Those without authorization will be told how to obtain it.) These services enable users to subset the data by observation (such as women 18–24 who voted for Bill Clinton in Massachusetts in 1992) and by variable (such as vote preference, unemployment status, and partisan identification); run descriptive statistics and graphics; launch sophisticated statistical analyses, all online; translate the data to a format readable by one's favorite statistical software, database program, or spreadsheet package; and download the data.

The Dataverse Network Web application also incorporates Zelig statistical software (see Imai, King, and Lau 2006, 2007, http://gking.harvard.edu/zelig), which provides a common framework for statistical analysis and software development and a unified syntax for using a large and growing number of statistical methods implemented in the R Project for Statistical Computing (http://www.r-project.org). The advantage of this arrangement is that any method in Zelig can also be run by a user via a graphical user interface (GUI) without knowing R, as part of any dataverse. These models can all be run online, without having to download data or install, learn, or run one's own statistical software.

This development is designed to shorten the time it takes for a new statistical innovation to make it into the hands of applied users. New statistical innovations are typically published in professional journals, the method having been implemented in code created by and often available only to the author. At one time, and to some extent even now, others can

use the new methods only after waiting years to see if some commercial software company decides to reprogram the method from scratch. The development of R has greatly shortened the time from development to use by enabling many articles to come with an implementation of the method in an R package that others can use, at least in principle. R represents a tremendous advance in statistical computing, publicly available code, and community standards, but since it is a command-line-oriented programming language it is still accessible only to relatively sophisticated users. Matters are not made easier by the huge diversity across these packages in program syntax, discipline-specific jargon for documentation, substantive examples, mathematical notation, and quantities of interest.

Zelig used within R drops the bar for users considerably since all packages used through Zelig require the same three commands, in the same syntax, all with directly parallel notation and documentation. The Dataverse Network takes this one step further by using Zelig facilities to build a GUI as part of any dataverse. Since the GUI is automatically generated, the time from statistical innovation to use by an applied researcher is considerably shortened. To include a package in Zelig requires writing only a few bridge functions, and as soon as Dataverse Network software is recompiled (which takes place centrally without user involvement), the new method will automatically be part of any dataverse's GUI for statistical analysis. Whereas Zelig enables sophisticated statistical analysts to use a diverse array of statistical methods quickly after they are created, a dataverse enables a much wider range of applied researchers to take part.

As dataverse services become more useful, researchers should demand that authors improve the metadata offered with their data sets. Some of this demand may also be met by teachers and others who can create value-added data sets with the same data and better metadata.

Teachers

A dataverse is sufficiently easy to create that it is convenient to use it to organize and present existing data in a particular way, even when new data is not part of the picture. For example, teachers can easily create a dataverse comprised of a menu of data sets handpicked for students in a specific class. This dataverse could be used to offer students choices from which they could write a class paper or to convey to them the data resources in the field. Collections like these can be created with a selection of specific studies or as a combination of live searches for studies that meet a set of predefined criteria.

For another example, some teachers may wish to create especially high-quality metadata for one or more existing data sets so that students can analyze the data set online, perhaps with statistical methods chosen by the instructor, with minimal effort. For example, if they do not exist within the data set already in the Dataverse Network, the instructor might add labels to variable values, among other things, so that the detailed text documentation becomes mostly unnecessary or can be replaced with a much shorter version. This procedure thus follows and automates aspects of the successful SETUPS (Supplementary Empirical Teaching Units in Political Science; http://www.icpsr.umich.edu/SETUPS) exercises created by the ICPSR.

A dataverse created for teaching is analogous to a syllabus, which is also one of many possible alternative views of an existing set of scholarly material. And just as with the increasingly common practice of establishing online archives of syllabi, a dataverse created for teaching can also be shared with others, if the instructor chooses to do so. This facility thus enables the scholarly and teaching community to capture some of the efforts of individuals so that we might all improve our teaching, and it also rewards individual teachers by allowing them to publish their dataverses within a Dataverse Network and to receive additional visibility and recognition for doing so. We have even developed a standard way of giving a formal citation to a dataverse so recognition for this type of teaching or curator contribution can be maximized as well.[9]

Scholarly Organizations

Similar to a teacher creating a dataverse with links to all the relevant data sets on a particular subject, a scholarly group organized around a substantive or methodological theme may wish to create a dataverse to highlight the data contributions in its field or of its members. For example, sections of the American Sociological Association or the American Political Science Association could easily establish such dataverses. In addition, the associations themselves could establish a dataverse to organize and present a centralized view of the dataverses of its member sections.

Granting Agencies and Research Centers

Granting agencies and some research centers routinely fund data collection projects. They may wish to create a dataverse to provide easy and

organized access to data in their field, just like teachers and scholarly organizations. They may also wish to establish a dataverse to keep track of and advertise the data contributions their funding has produced and to guarantee that these data sets will have the widest possible distribution.

In addition, many of these organizations, such as the National Science Foundation and the National Institutes of Health, support the data-sharing movement by legally requiring data to be made public after a fixed time period under specified circumstances. However, much data that are supposed to be made available by grantees are still available only from the authors directly, with all the impermanence and lack of public distribution entailed, and some not at all. To fix this problem, these organizations could easily require grant recipients to include in their final reports a formal citation to any data set that was supposedly created. The granting organization could then decide whether to recognize claims that data were created in applications for future funding if no such citation were provided. These data citation requirements would be easy to administer, as they would be directly parallel to those required in final reports for article and book publications.

It is important to understand that data citations can be created and included in a public dataverse even for data that are highly confidential or that qualify for a specified embargo period, so long as all the proper rules for authorization and access are clarified and encoded in the dataverse. As long as the full citation is available and made public when the data set is created, the data set's existence and identity can be assured, and the rules for access can be established clearly and for all to see. Requiring data to have a formal citation and to be included in a dataverse does not constrain granting agencies or research centers from imposing any policies they may wish to require for access and authorization.

Major Research Projects

Scholars with major research projects that produce data will probably wind up including replication data from their project on each journal's dataverse in which they publish an article, and the principal investigators and other participants in the project may of course wish to have their own dataverse of the data from this project and perhaps others. Often, however, the goal of the investigators is to give these projects a separate identity. So for this purpose, creating a project dataverse that offers an organized view of all of the data sets created by that project will be useful.

A project dataverse offers an especially attractive way of communicating the contributions of the project to the scholarly community, university deans, granting agencies, and other prospective funders. The dataverse interface includes facilities to list news, notes, and plans so that the research project can communicate what data are coming in addition to what they have already made available.

Universities, Academic Departments, Data Centers, and Libraries

Major institutions like universities, often represented by departments, data centers, or libraries, need to provide access to research data for their faculty, students, and staff. Most pay for many subscriptions to data sources and memberships in international archives. Since there are many such data providers, universities must find ways of organizing all the resulting information for their faculty, students, and staff. Often this is more of a long list of available resources on a Web site rather than a unified interface. The Dataverse Network offers universities a way of organizing and presenting with a common interface data sets from a large and growing array of sources.

For this purpose, these institutions can offer a single dataverse with a hierarchical organization of many of the resources available to their community. Alternatively, they may wish to install an entire Dataverse Network themselves, with individual dataverses devoted to coherent subject matter groupings or academic disciplines or departments within their community. We discuss creating an entire Dataverse Network in next subsection.

The Dataverse Network also has infrastructure to identify who is a member of these institutions, such as by IP address, a link into university pin servers, passwords, or other authentication devices now in use. This will enable a university or other institution to buy a subscription to some data service and offer it to only their members. If they wish, they can also make the metadata available more widely so others can see what they have and inquire about access rules.

Data Archives

A dataverse requires no local storage, hardware, or software and thus requires no local IT staff. Those responsibilities lie with the Dataverse Network of which the dataverse is a member. Each Dataverse Network is sponsored by some organization or institution. The Dataverse Network, which serves out the individual dataverses to local sites, represents an actual

installation of our Dataverse Network Web application software. Data archives can easily be set up via our dataverse technology, but then the actual physical location of the data sets (the bits themselves) would lie with a Dataverse Network installed elsewhere. Although most of the users of this technology described above would benefit by not having to manage any of that locally, their data would still have an archival footprint. Archives, of course, are in the business of keeping and managing the bits and as such will often want them local. For some legal or IRB purposes, keeping the bits local is required. Archives that install Dataverse Network software also have the ability to create dataverses for their constituents or others.

Whether an archive chooses to use a dataverse to manage its holdings, or to install a copy of the Dataverse Network software and manage it and the necessary hardware too, Dataverse Network technology provides the key services a data archive might need to help acquire, manage, preserve, and distribute a collection of data sets along with the associated documentation and metadata. This includes persistent identifiers, citations, organization and distribution facilities, facilities for storage, archiving, cataloging, preservation formatting of data and metadata, translation, online analysis, distributed authentication and access control, remote repository caching, and virtual collections of remote objects.

Different archives in the same field are in some sense competitors, but they are operated by community-minded individuals and associations, so they also work closely together. Dataverse Network technology makes it even easier for them to interoperate with each other, including facilities such as harvesting metadata from each other, building common catalogs, and cross-archive searching, browsing, and presentation. For example, a group of the major social science data archives have joined together under the Library of Congress in the Data Preservation Alliance for the Social Sciences (DataPass), and they have a common catalog organized as a dataverse for new data acquired as part of the project.

Of course, any archiving scheme depends on the persistence of the institution where the archival footprint resides, and ultimately on its business model for survival. Dataverse Network software is designed to be installed by these existing institutions so that new institutions need not be built especially for this purpose.

Producers of Highly Confidential Data

Dataverse Network software gives dataverse owners total control over how their data are accessed and who accesses it. Present options include

making the data set fully open to the public, requiring users to sign a guest book so their identity is known, requiring users to sign click-through licensing agreements, or giving access only to specific authenticated individuals who have received appropriate authorization. We also have plans to allow financial payments for data, among other methods.

However, some data is so confidential, secret, or sensitive that it can never be distributed by automated means. Other researchers may be able to access such data only when in the physical presence of the original author or data producer, when in a secure facility, or if they meet other stringent conditions. Data with national security secrets, private sexual information, personal financial or medical information, evidence of criminal behavior, or other data collected on human subjects may fall into this category.

The question is what the best practice should be for authors who write reports with data such as these, and for data producers. Our suggestion is that the data producers set up a dataverse (or Dataverse Network), ingest their data into the system, and create a citation. In almost all situations, the *existence* of the data is not a secret, so the metadata can almost always be made available through the Dataverse Network, no matter how secret the data are. Each data set, even if highly confidential, can have a metadata description page so others can understand the nature of the data. For some, the full documentation can also be included.

The key, however, is that the full citation for these data, including the UNF, be made available. Making the UNF available guarantees that data producers, and anyone writing articles based on the data, are using a specific identifiable data set, where updates to the data are verifiable by a change in the UNF used. Although the general public will not be able to verify the results or build on the analysis by reanalyzing the same data, someone in the future (or the original authors) may gain access and be able to do so. Verifying that the data cited are actually part of a specific, known set of information will then be highly valuable. This is a reasonable standard even for individual researchers working on their own, tracking data sets under their full control. Since UNFs alone reveal no information about the original data, they and other metadata can be published even if the data cannot.

Concluding Remarks

We conclude with a brief summary of the benefits of Dataverse Network technology, in terms of the eight requirements we set out in the second section for an effective data-sharing infrastructure.

Using Dataverse Network technology guarantees that data producers, archives, and journals will receive more recognition, both for their scholarly articles and books and for their data. This recognition will occur through formal citations, including new data citations, more citations to their printed publications, and virtually hosted dataverses branded to match their own Web sites. Although data are branded as locally as anyone wishes, the data set itself is available in preservation format accessible entirely in public via rules known to all.

Data can be distributed through the Dataverse Network publicly even if they require special authorization for use due to their confidential or proprietary nature. Journal copy editors and future researchers can validate the existence of a specific data set that cannot be changed, even if they do not have permission to access it, or if they have permission but do not have the skills to extract and analyze the data. This may enable the data replication movement to spread to areas of science where researchers seem to hide behind privacy concerns instead of adopting the widespread and growing scientific norms of data sharing. The worst area seems to be public health and medicine, where researchers have legitimate and legally relevant concerns about privacy but where finding a way to increase data sharing has the potential to result in millions of people leading longer and healthier lives but many other problem areas also exist.

Through the universal numeric fingerprint that is part of the data citation, future researchers, journal editors, and others can verify that a data set, downloaded decades after the article associated with it was published, contains the identical content as was used in the article, even if the data has been transformed over the years many times between different hardware, operating systems, statistical software, and storage media. The persistence of the data is guaranteed by a unique global identifier as part of the data citation, standard preservation formatting offered by Dataverse Network software, and a strong link to an active archive with professionals who know how to carry out these tasks and an institution to back them up. The virtual hosting nature of dataverse technology means that work can be shifted to professionals, while still giving full power to end users to manage their own places in the universe of data. This covers even legal protections, which are regularly studied and updated by the archives for which each Dataverse Network is associated so that authors, journal editors, and others can focus on their expertise rather than these crucial but, for them, subsidiary issues. Dataverse Network software also automates some of the tasks currently performed by IRBs and lawyers.

The Dataverse Network project is related to the institutional repository movement, although the rich metadata often available with data sets allow statistical analyses and other features not often available to the typical repository (Lynch 2003). (We are also in the process of building formal software links and protocols to interoperate with existing repository software, such as DSpace.) Dataverse Network software is open source, free, and available to all. One can build modules that snap into it to provide facilities or services we do not yet offer. And, as the appendix describes, if worse comes to worst, the software is owned by all and anyone may use it to build on what we have done or to take the project in a different direction.

Appendix
Software Details

Dataverse Network software is open source, available at http://thedata.org, and licensed under Affero, a version of General Public License (see http://www.affero.org/oagpl.html). This license guarantees that the software is available now and all future versions of it, including any versions that may be created from it by others, will remain available. Web applications have a tremendous advantage in terms of ease of use, cost, installation, administration, and management. However, they also pose a disadvantage since, if the Web service that users rely on vanishes or the underlying software changes in a direction users dislike, users typically have no recourse. In contrast, with Dataverse Network software, users are fully protected since the software is owned by the community and in fact also by you. Under the license, if you do not like the direction we take the software in the future, you have full access to all versions of the source code and a permanent license to use it and build on it as you see fit.

Dataverse Network software is written under the Java Platform, Enterprise Edition (Java EE) 5, using the latest Java technologies, including Enterprise Java Beans (EJB) 3 and Java Server Faces (http://java.sun.com/javaee). It runs on top of the Glassfish Application Server (https://glassfish.dev.java.net). We use PostgreSQL for database software (http://www.postgresql.org), but other databases can easily be used instead, such as Oracle or MySQL. The data analysis component uses R (http://www.r-project.org) and Zelig (Imai, King, and Lau 2006, http://gking.harvard.edu/zelig) for statistical computing.

Notes

1. By "quantitative data" I mean systematic collections of measurements that can be or are intended to be machine readable. They can be produced by any person or group, so long as they are systematically organized and described. Much quantitative data is numerical, but it can also include digital representations of text, audio, video, satellite or medical imaging, or other sources.

2. As an example of the problem, my original "Replication, Replication" article refers readers to information "available via gopher or anonymous FTP from haavelmo.harvard.edu" (King 1995:452). Gopher and anonymous FTP are document and file transfer protocols that have now been largely replaced by the Web, and the Haavelmo server no longer exists in any form.

3. Whereas scholars once had to wait for 6 weeks for tapes to arrive from national archives and data analysis was effectively limited to faculty and advanced graduate students, sophomores and faculty alike now do online analyses from their dorm rooms and offices. The job of working in our data center has also become far more interesting. Instead of having to deal with long lines of faculty and graduate students (with their tenure clocks or dissertations beating in their chests) waiting their turn to scream at staff for their data, the only people who come to the data center in person now have interesting questions that the staff are eager to hear and happy to help with. And of course the result of all this also greatly facilitates a wide range of research across the university.

4. The header "hdl:" signifies the type of identifier, in this case a "handle" managed by the nonprofit Corporation for National Research Initiatives (CNRI; see http://www.handle.net); "1902.1" is the institutional guarantor (in this case the Institute for Quantitative Social Science [IQSS] at Harvard University), which serves as the handle's naming authority registered at CNRI and which commits to the persistence of the name and the connection with the specific data set; and "DXRXCFAWPK" is the unique local data set identifier. Other unique global identifiers are also allowed, such as Life Science Identifers or digital object identifiers.

5. The URL in this case is http://id.thedata.org/hdl%3A1902.1%2FDXRXCFAWPK, which indicates standard hypertext transfer protocol format, http://, and the institutional guarantor, thedata.org (in this case the Dataverse Network project at IQSS at Harvard University). Everything after the slash is the unique global identifier translated to follow standard URL syntax. These are details that need to be correct, of course, but authors and editors need not worry about them, as Dataverse Network software will construct them automatically. For details, see http://thedata.org/citations/unf.

6. The header "UNF:" identifies the string as a universal numeric fingerprint, ":3" refers to the version of the UNF algorithm (and so allows for future developments), and the remaining string is the data set's numerical fingerprint.

7. Some journals do not employ a Webmaster and instead use a Web site automatically created by their publisher. In this case, the work flow would differ, perhaps involving only a brief call to the publisher, in which case the Webmaster scenario in this paragraph merely illustrates how little technical work is needed.

8. We are fortunate to be able to take advantage of, and to support, the Data Documentation Initiative, an emerging international standard for defining social science metadata; see http://www.icpsr.umich.edu/DDI.

9. We do this by assigning the dataverse an author, date, title, unique global identifier, and URL. Although data sets are meant to be fixed, dataverses are live and may be changed over time. So just like a citation to a software project (or person), the object being cited with these citations is expected to evolve but nevertheless retain its essential characteristics.

References

Altman, Micah and Gary King. 2007. "A Proposed Standard for the Scholarly Citation of Quantitative Data." *D-Lib Magazine* 13(3/4). http://gking.harvard.edu/files/abs/cite-abs.shtml

Altman, Micah, Leonid Andreev, Mark Diggory, Gary King, Daniel L. Kiskis, Elizabeth Kolster, et al. 2001. "A Digital Library for the Dissemination and Replication of Quantitative Social Science Research: The Virtual Data Center." *Social Science Computer Review* 19:458-70.

Anderson, Richard G., William H. Greene, B. D. McCullough, and H. D. Vinod. 2005. *The Role of Data & Program Code Archives in the Future of Economic Research*. St. Louis, MO: Federal Reserve Bank of St. Louis Research Division.

Bachrach, Christine A. and Roslind B. King. 2004. "Data Sharing and Duplication: Is There a Problem?" *Archives of Pediatric and Adolescent Medicine* 158:931-32.

Board on Earth Sciences and Resources. 2002. *Geoscience Data and Collections: National Resources in Peril*. Washington, DC: National Academies Press.

Board on Life Sciences. 2003. *Sharing Publication-Related Data and Materials: Responsibilities of Authorship in the Life Sciences*. Washington, DC: National Academies Press.

Dewald, William G., Jerry G. Thursby, and Richard G. Anderson. 1986. "Replication in Empirical Economics: The Journal of Money, Credit and Banking Project." *American Economic Review* 76:587-603.

Diamond, A. M. 1986. "What is a citation worth." *Journal of Human Resources* 21:200-215.

Fienberg, Stephen E., Margaret E. Martin, and Miron L. Straf. 1985. *Sharing Research Data*. Washington, DC: National Academy Press.

Gleditsch, Nils Petter, Claire Metelits, and Havard Strand. 2003. ''Posting Your Data: Will You Be Scooped or Will You Be Famous?" *International Studies Perspectives* 4:89-97.

Imai, Kosuke, Gary King, and Olivia Lau. 2006. *Zelig: Everyone's Statistical Software*. http://gking.harvard.edu/zelig

———. 2007. "Toward a Common Framework for Statistical Analysis and Development". http://gking.harvard.edu/files/abs/z-abs.shtml

Johnson, David H. 2001. "Sharing Data: It's Time to End Psychology's Guild Approach." *Observer (American Psychological Society)* 14(8). http://www.psychologicalscience.org/observer/1001/data.html

King, Gary. 1995. "Replication, Replication." *PS: Political Science and Politics* 28:443-99.

———. 2003. "The Future of Replication." *International Studies Perspectives* 4:100-105.

———. 2006. "Publication, Publication." *PS: Political Science and Politics* 39:119-25.

———. and Langche Zeng. 2007. "When Can History Be Our Guide? The Pitfalls of Counterfactual Inference." *International Studies Quarterly* 51:183-210.

Klump, J., R. Bertelmann, J. Brase, M. Diepenbroek, H. Grobe, H. Höck, et al. 2006. "Data Publication in the Open Access Initiative." *Data Science Journal* 5:79-83.

Lynch, C. A. 2003. "Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age." *portal: Libraries and the Academy* 3:327-36.

Schneider, B. 2004. "Building a Scientific Community: The Need for Replication." *Teachers College Record* 106:1471-83.

Whitbeck, Caroline. 2005. "The Responsible Collection, Retention, Sharing, and Interpretation of Data." *Online Ethics Center for Engineering and Science*. http://onlineethics.org/reseth/mod/data.html

Gary King is the David Florence Professor of Government and Director of the Institute for Quantitative Social Science at Harvard University. His home page can be found at http://gking.harvard.edu.