

Aesthetics, Exposure, and Impact: Distributing Public-Use Data on the World Wide Web

David Robinson and Laura Kline
Carolina Population Center
The University of North Carolina at Chapel Hill

In disseminating public-use data on the World Wide Web (Web), distributors of data face a tradeoff between embracing new technology and holding fast to the traditional standards of their field. Advances in technology are changing not only the way data distributors do their job, but also the audience for their work. Since anyone with a modem and an Internet browser can access data sets on the Web, using the Web to distribute data substantially increases the number of potential data users. The Internet seems to be the wave of the future, but it is growing so quickly and the ultimate ramifications of its innovations are so difficult to predict that it is hard to know how much things will change during the next few years. A data distributor's first response might be to wait and see what develops. However, significant advantages in terms of exposure and experience can be realized by getting on the Web now. By doing so, one sends a signal of prescience and technological sophistication--not only to other research institutions, but also to the larger Internet community. Moreover, with the current effort to develop secure transactions protocols on the Internet, there is every reason to believe that tomorrow's technology will be even better suited for data dissemination. What data distributors need is a balanced approach to today's Web that allows for anticipated growth. In what follows, we present a framework for developing a Website based on our experience with the Russia Longitudinal Monitoring Survey (RLMS). Within this framework, we address some of the benefits and difficulties of data dissemination on the Web.

The home page for the RLMS project was the Carolina Population Center's (CPC) first attempt at designing a home page explicitly for the distribution of data sets. The decisions we made and the criteria by which we measured ourselves are probably obvious to people who have visited our

site. We approached the Web as if it were a space in which we could boast about our study, tell the story of our data, and offer our data to others. It might also be clear from looking at our site that we are not data distributors by training. Since we have no formal training in disseminating data, we developed our own standards for documentation and data distribution. The development of our own site was also a result of the way that we approached the task of designing a home page. Our goal of building a Website to disseminate data involved both constructing a Website and preparing data sets. Instead of addressing each separately, we allowed these two activities to play off one another.

Thinking about the different stages of our project gave us critical insight into the tasks of building a Website and preparing data sets. We realized that it would be useful for us to approach these activities in a series of phases. Considering these phases helped us organize our tasks and guided us in anticipating our future needs on the Web. The framework we used is well known to those who have participated in a software development project--it involves iterations of a two-stage cycle. The first stage is a planning and development stage. In the second stage, the plan is executed. Subsequent iterations of the cycle mirror the initial two stages. The third stage is like the first in that planning and development are involved, yet it brings to bear information learned in the prior two stages. Likewise, the fourth stage executes the plans of the third, and so on, as long as innovation occurs, errors are uncovered, and users' expectations change. Recognizing the inherently sequential nature of our framework and the manner in which technological change was likely to impact it, we endeavored to fully utilize the personnel available to us in each stage. Moreover, optimal utilization of personnel helped us avoid problems that can arise from the interdependence of the two phases.

Our framework is designed to reduce errors and produce quality results. However, measuring the success of our approach or of any Website is difficult. Profit, in the standard economic sense, is revenue minus cost; in the realm of developing a Website and providing data to researchers, this definition is certainly not robust enough. Some data distributors might place a high value on the

convenience and ease of use of their site; others might stop simply at making their data available to the public. Regardless of how profit is defined and whether or not it is defined in explicit terms, every site has an associated measure of profit or success.¹ Developing a Website requires making commitments in hiring personnel, purchasing hardware and software, and determining a focus. The viability of these commitments is jeopardized when uncertainty like technological change strikes. The fact that technological change could reshape the hardware and software paradigms that define the Web at any given point in time could make it difficult to recoup the investment involved in Stage 1 planning and decision-making.

Stage 1 - Planning

The first stage of preparation and exploration was crucial to the development of our Website. In terms of the Website itself, we wanted to make our site interactive enough to encourage people to read our documentation and to use the data sets. In terms of disseminating data, we wanted to employ features of the Web that could make accessing the data and documentation easy and quick. The first step in planning for both goals was to determine personnel needs and to contact the people who could help us.

Organizing the Team

The tasks of designing and implementing the RLMS home page involved the help and expertise of several people, each of whom made significant contributions. Throughout the development stage, we consulted with the principle investigator of the project who shaped the scope of our efforts. The RLMS home page benefitted from the fact that the Webmaster for CPC had already developed home pages for the core of our research center; these pages ran off an in-house server. The Webmaster served as a good resource for us in answering questions and providing us with literature on Hypertext

¹This is especially true if it bids resources away from other activities going on in the center or institution in which the Website is developed.

Mark-up Language (HTML). We relied heavily upon the skills of our Unix programmer whose knowledge of Perl scripts converted our desire to track data users into a real solution and who helped automate and systematize our procedures. Furthermore, we benefitted from the help of a senior systems administrator who successfully negotiated with UNC's Office of Information Technology (OIT) for storage space for our data and documentation files on an anonymous File Transfer Protocol (FTP) server. In addition to editorial help from one of our senior programmers, many other co-workers gave us useful comments and constructive criticism on early versions of our home page.

Managing the Data

After enrolling the people required to make the project a success, it was necessary to determine how we could narrow our focus and make the enormous tasks of developing a home page and distributing data sets more manageable. The RLMS project is a longitudinal study which currently consists of six rounds of data collection at both the household and individual levels. One survey sample comprises the first four rounds of data collection, while a second sample includes the fifth and sixth rounds. Both samples are large; the first involves about 6,500 households, the second involves roughly 5,000. Given the size of our data sets, the task of making all the data available on-line is huge. Since we were given a relatively short amount of time in which to design and implement a home page, one of the first decisions we had to make was how much data to put on-line and in what form.

Given our time constraint and our inexperience in doing Web work, the project chose to make only the first round of data available and to provide those data in just one file format. Since the project had stored its data on four different computing platforms in the previous year, we had been using SAS xport files so that we could easily move files from one platform to another. Our familiarity with the SAS xport format and its usefulness in allowing our data to be read on different platforms and into other statistical packages made it a good choice for our data files on the Web. We hoped that the task of putting the remaining rounds of data on the Web would benefit from the lessons we learned getting

the first round ready. We also hoped that the experience gained from this initial effort would dissipate some of our beginners' naivete about disseminating data.

After making the decision to focus on Round I, we explored our data sets and devised a strategy for accessing them from our pages. The principle data sets for the RLMS project have thousands of observations and hundreds of variables. It would be a burden for people to have to download the entire household data set, for example, if they were interested only in household expenditures. We needed to come up with a plan for data organization on the Web that differed from what we were using in-house. Therefore, we divided our large data sets into smaller sections that matched the divisions by subject matter found in the questionnaires. For instance, in the individual data set, there is a section of employment questions grouped under the heading "work" that we kept together as a unit. In each of these smaller files we kept some basic identification variables so that each subset of the data stands on its own. Of course, people have the option of downloading all the subsets of files for a particular data set. If a user wants to recombine them into complete files like we use in-house, he or she can read our pages to find out which identification variables the files have in common.

In preparing data sets for distribution on the Web, it is important to weigh the considerations of the end user against how the data are used by in-house researchers in order to determine the best structure and format for the files. Originally, the names of our variables were largely a series of numbers referring to the question numbers in the questionnaires. Since the questionnaires changed between rounds, a particular question number and old variable name in Round I may refer to a completely different question in Round II. Although we were using the numeric variable names in-house, we felt that it would be much easier for people to use our data longitudinally if the variables were named mnemonically and corresponded across rounds. So, we changed the variable names to mnemonic names that attempt to capture the gist of the questions. At first, the changes in the data sets were available only in the Web data sets. For in-house purposes, we continued to use the old variable

names based on question numbers for a couple of months. Since people working on the project were already familiar with the numeric names, we knew that it was going to be a large adjustment for our research team to start using the new variable names. Eventually, however, we did modify our in-house data sets so that the variable names matched those on the Web.

Our data management decisions were complicated by the fact that the project's data management needs were being analyzed and carried out in tandem with those of the Website. Putting data sets on the Web in a different format than what is used in-house certainly allows for potential confusion and creates extra work for data managers. Had our files been smaller and easier to use, we might have kept them in exactly the same format on the Web. However, we felt that there would be real advantages in the long-run for us and for all users of the RLMS data if we made changes to the data sets before releasing them to the public.

Page Design

Once we determined which data sets we would distribute and the format in which we would distribute them, we could begin thinking about a possible structure for our Website. Although the initial reasoning behind creating a Website might be to disseminate data, a home page is also a point of primary introduction to a research study for anyone using the Web. As the point of first contact, the home page itself should be appealing and inviting. Thus, it was critical that we employ state-of-the-art features and that we view our site on as many browsers as possible to experience how others might see our pages. This approach stands in direct contrast to that espoused by some who believe that the best strategy is to keep the HTML simple. We argue that people are much more likely to want to visit a high quality home page. A few years ago, having a page on the Web was a novelty in and of itself. In order to stand out on today's Web, it is becoming essential to incorporate background colors, meaningful links to other sites, search capability, and compelling graphics. We wanted to make our site so appealing that when people saw our pages, they immediately wanted to read our source code to

figure out how we created them.

Record Keeping

Another goal of our project was to create a home page with which users interact. This was achieved in part by making our Website inviting. The pages were designed to attract the general Web user's attention. However, we also wanted feedback from the people who visited our site. To accomplish this, we included a mail back option so that people with questions or comments would have an easy way to contact us via email. We realized that not everyone would use this option and we were curious about all the other visitors who remained silent. Thus, two initial design features of our site are a counter of people who access our home page² and a file on our Unix server in which we store the names, addresses, affiliations, and email addresses of those who have requested FTP instructions for our data.

Instead of offering immediate access to our data files, we require that people who are interested in FTP-ing the data sets fill out an email form in order to get the instructions. The receipt of the email form at CPC initiates a Perl script which then returns an email message describing the FTP process and listing the names of the files to the originator of the request. Since the only way for the user to get the FTP instructions is to send us a correct email address, we can keep track of the origination of data requests. At present, we have no way of knowing whether those people actually FTP the data because our campus computing center does not report FTP attempts. We may be able to get this information in the future. However, since people are free to copy the data sets from a friend or colleague, we will never be able to know how many people are really using our data.

Through our address file, we can generate an email list in order to notify current users of updates to existing data or of the availability of new data. We can also use the list to answer questions

²We use a counter designed and maintained by George Carmichael of the University of Alberta's WebSupport Group.

from funding agencies about the extent of our data dissemination (to the best of our knowledge). In our pages, we explain that users will be informed of changes and updates through mass mailings, so it is to their benefit to register with us. The user does not have to check our page periodically for error reports since we will send out a mass mailing if there are any serious developments. Our address file also takes care of the accounting procedures that data distributors have long employed to record who is using their data. By taking advantage of Perl scripts and on-line forms, we achieve a similar level of record-keeping electronically.

Policies about Updates

Although our address file makes it easier for us to inform users of data changes, we felt that we needed to establish reasonable goals for updating data files. Since our study is longitudinal, we have the burden of matching identification variables across rounds. By releasing only Round I, we were able to forestall the inevitable task of looking at mismatched observations. However, after looking at records across the first four rounds of our study, it may be necessary to release an edited version of Round I that incorporates necessary data changes. Due to the labor intensive aspects of data cleaning, we decided that it was reasonable to assume that we would release new versions of data files no more than once a month. Based on our updating policy, we employed a convention for naming files that incorporates the month and year of a particular version of a file: the first eight characters describe the contents of the file, four characters denote the date, and the final three indicate that the file is a SAS xport file. For example, r1hhrost.1095.xpt is a section of the Round I household file that contains household roster information, was released on the Web in October of 1995, and is in SAS xport format. Although people who download the files to a DOS operating system must shorten the file names, it was important for our internal record keeping to use more than the DOS standard of eight-dot-three character file names. By using an eight-dot-four-dot-three naming convention, we supply the user with a cue for naming the data set on the target system: remove the middle four characters and a standard

SAS xport file for DOS appears.

Documentation

In addition to making data available, a data distributor also compiles all relevant documentation, thereby imposing structure on the data, which makes them easier to understand and analyze. Whether done by the distributor or the research institution from which the data set originated if it differs from the dissemination site, thoroughly documenting a data set takes time. This has a distinct payoff: the more time spent on processing and documenting the data, the less likely it is that they will be riddled with misunderstandings, omissions, miscodings, and errors. But the idea of making numerous pages of documentation available and required reading before a user can access data flies in the face of the current push to develop glossier, faster, and more interesting pages on the Web. We needed to establish reasonable goals for our study in terms of the quantity of documentation available on the Web.

In determining our documentation needs, we realized that we should include several types of documentation. We wanted to provide general information about the structure of the study as well as about the personnel involved. We also saw a need for a metadata section, that is, a section of data about our data. In addition to these types of documentation, we needed to make available the questionnaires, the actual survey instruments translated from Russian into English. For the RLMS project, the questionnaires also serve as codebooks. Since the questionnaires are long and have complicated formatting, we chose not to make hypertext versions available. Instead, we converted our word-processed files into PostScript files so that they could be downloaded and printed on a PostScript printer. We left the questionnaires intact instead of dividing them into smaller files to match the data files that we created for the Web. We did so for a couple of reasons. First, upon browsing the questionnaires, people might find that they are interested in obtaining additional sections of our data. Secondly, we wanted to maintain the integrity of the questionnaire files, as some researchers are

interested in the order in which questions are asked.

Other documentation-related issues that we considered were how to cite the study and how to get people to read the documentation. We took two approaches to the issue of citations. We included the name of the study and a listing of its collaborating agencies on the bottom of each home page. We also dedicated a short section of our on-line documentation to an explanation of how to cite the study. Furthermore, we wanted to encourage the user to read the documentation and to take it seriously. The more complex the data set, the more the user needs to be encouraged to read and download the documentation. For our study, we identified our metadata page as crucial for anyone interested in using our data. Thus we chose to force all inquiries about data and questionnaires through that page.

Since we did not approach our task as full-time data distributors, we aimed to place limits on the amount of user service we would provide after getting our site on-line. We realized that we would not have much time to answer questions because our efforts were supposed to be concentrated primarily on getting our site on-line. After establishing the site, we were to return to our regular jobs as applications programmers. We hoped that the maintenance of the site would be relatively minimal and that our major efforts in the future would be to prepare additional data sets and to update our pages, not to provide user service. Our goal, then, was to include enough documentation so that we would not have to explain ourselves further by email or the phone, without providing extraneous material that would cause people to become discouraged with our site.

Stage 2 - Implementing

Since we devised a common look and feel to our pages in Stage 1 as well as an efficient means of getting at the data, implementing Stage 2 was relatively simple. We spent some time scanning images, turning existing documentation into HTML, and preparing data and other documentation for storage at one of UNC's central Unix machines.

We reduced the amount of time spent creating the Website by converting existing documents

into HTML rather than by writing all new pages for the site. By leveraging existing information, we not only saved time, but we also provided the public with critical project information through primary documentation: the user can read pages from grant proposals, interviewer instructions, and supplements to survey instruments.

Stage 3 - Planning Revisions

We are now at the point of revising past work. In reviewing usage statistics as well as the battery of questions from would-be users that we have accumulated since last October, we have tried to devise a series of changes that will improve our pages as well as breathe new life into them. Of the four types of changes to be accounted for at this stage--initial errors, modifications in light of usage statistics, modifications to account for project evolution, and modifications to account for new technology--we plan to address the first three in our initial revision stage, with a view towards addressing the fourth in subsequent revisions.

Initial Errors

By mid-December, the volume of questions regarding the use of SAS xport files was sufficient to warrant the design and maintenance of a Frequently Asked Questions (FAQ) page. On it, we provide examples of how to convert SAS xport files to other formats and how to use them on various platforms. We also thought that the high volume of questions regarding the use of PostScript files resulted from user unfamiliarity. Thus we set forth to add documentation regarding PostScript files to our FAQ page. In doing so and in investigating the problem further, we discovered that our questionnaires were in Level 2 PostScript format which is unreadable by Level 1 printers. We are currently testing packages like Adobe Acrobat and CommonGround for a way to allow our users easy printing downloadable documentation. Unlike the PostScript format, these packages have the added advantage of enabling users to view downloaded documents on-screen.

The Impact of Usage Statistics

We have far more visitors to our home page than we have requests for FTP instructions. In one sense, this supports our initial hypothesis: many people interact with the Web page as a means of familiarizing themselves with our research and not as a means of acquiring data from our project. Despite the lengthy log of visitors, few provided us with unsolicited criticism or commentary. As stated above, however, we did receive a large number of questions relative to the number of people actually using our data. Given that we were reasonably unsuccessful in anticipating the types of problems and questions that users would have had, there is no way for us to know whether we were attracting lots of people with tangential interests or whether we were discouraging people with direct interest in our data. We asked ourselves if we could have converted more visitors into data users through a different approach. This has led us to consider the following series of modifications.

In our current revision stage, we are aiming to provide the maximum appeal to those who want simply to visit all of our pages, while simultaneously providing an easy approach to those whose sole interest is in accessing the data. In order to do this, we plan to mirror the basic page structure of our site. One set of pages will be graphics-intensive, glossy, and densely loaded with information, while the other set will contain a text-only path to our data. The two paths to the data will be identical from the point of view of providing crucial documentation in the form of summaries, explanations, and metadata for our public-use data sets; they will differ in the extent to which they utilize the advanced features of HTML to provide an entertaining and educational introduction to our project.

In short, where we once had a single webbed path to an order form through which it was possible to obtain FTP instructions to the data sets, we will soon have two paths. Both paths will provide the same substantive information needed to inform researchers about the development and construction of our data. One path will appeal to the person with a vague interest in Russia, whereas the other path will appeal to a researcher--perhaps in Russia--who, for either technical or personal

reasons, does not want to spend time on extraneous material.

In keeping with the view that visual cues are essential to presenting information persuasively, we are revamping our pages with new backgrounds. Instead of maintaining the same background color throughout the site, our next release will use different colors at different places along the way. The text-only path to the data will be one color, the high-graphics path will be another color, the metadata will be a third color, and the order form will be a fourth color. Our intention is to cue visitors to where they are in the progression from initial page to ultimate page through the use of background colors and styles.

Project Evolution

Finally, and most importantly, we are investigating new ways of distributing the data. While the method of requiring users to exchange an email address for FTP instructions suits our needs well, we worry that even the smaller data sets are too large and cumbersome for some users. We want to provide a novel solution to the common complaint that the data sets are time-consuming to download and too large to store on some machines. To solve this problem, our Unix programmer is currently developing a system that combines HTML, Perl scripts, and SAS to allow users to create their own data sets. Our new method will enable users to select the variables they wish to include in data sets created especially for them from tables containing variable names, labels, and summary statistics. Each personalized data set will then be created and shipped to the staging platform where it will reside for a pre-specified length of time. After the data set is shipped to the staging platform, the user will be notified via email of its successful creation and the fact that it will be accessible for a limited time.

This new method of data distribution will accomplish two things. First, it will provide a heightened level of interaction between the user and our Website. Maximizing the interactiveness of our site is an ongoing goal of our development team. Second, by trading off-peak processor time for storage space, it will allow our project to save on data storage costs. This savings will enable us to

lower the total cost of providing data to the end user.

In future revisions, we hope to take advantage of new technology that will enable us to further enhance the interactivity of our site. For instance, we are thinking about investigating the idea of having a Java-based site that will graphically present summary statistics from our data.

Thinking about Change

The framework presented in this paper has several important negative implications for data distributors when viewed in the broader context of the Internet and its rapid changes. On the one hand, market forces--factors like memory requirements, chip speed, the number of users, and the number of sites--are pushing technological innovation on the Web down a path that favors speed over accuracy. On the other hand, uncertainties about the rate of change and the direction of technological progress make it difficult to calculate the expected payoffs from investments--capital, human capital, or otherwise--in the Web. Furthermore, the breadth and scope of possibilities associated with presentation and the diversity of client-side software make it difficult to keep current of the newest methods and types of presentation.

With the Web offering immediate exposure and access, many people are getting into the data distribution business without having had formal training in the methods currently used in any sort of archiving field.³ In the past, researchers who were interested in obtaining public-use data sets presented their data requests in person, by mail, or by telephone to a staff member involved in data dissemination. This process generally included the acknowledgment of a professional identification or affiliation with a research institution. Although we require a user's correct email address before we release the FTP instructions, we do not know whether or not the user's other information, like name and institutional affiliation, is accurate. As a result, we have less control over who is using the data and our records are not as complete as they might be if our data were distributed in a more traditional manner.

³The authors are one example.

To illustrate the changes facing data distributors, it is helpful to imagine the evolution of gas stations over the past several years in much of the United States. In the golden age of data dissemination, data distributors ran full-service filling stations. You pulled up in your extra-long, gas guzzling--but very comfortable--Cadillac (mainframe), pulled out your gas card, and sat in your seat listening to the radio while the attendant took care of your needs. As the gas tank was filling up, he walked around the car, checking the tire pressure, the oil, and washing your windows for you.

With today's smaller cars and concerns for fuel economy and time savings, most gas stations are self-service. Now, you pull into the station in your two-door Honda, survey the row of pumps for the appropriate grade of gasoline, and pump it yourself. You might wave at the attendant to get the tank reset, but that is the extent of the interaction between the two of you. You do not even pay the attendant anymore--you just swipe your credit card in the machine attached to the gas pump itself.

Consequently, whether or not the tire pressure is checked, the windows are washed, or the oil gauge is read is completely at the discretion of you, the driver. No one will come running out of the service bay to admonish you for failing to use the squeegee on your windows. It is your fault alone if you have an accident because you could not see out your dirty windows or because you failed to inflate your tires to a sufficient level.

Unlike modern gas station attendants, data distributors *are* the custodians of automotive safety (i.e., research). Our reputation is affected whenever drivers have accidents, yet we cannot change the fact that we are operating self-service stations. Instead of going around the car and checking the fluids ourselves, we have to provide tools and instructions for the user to check his own tire pressure and oil level.

We can help ensure automotive safety in two ways. First, we can place the tools in the immediate view of drivers so that when they gas up their vehicles, they cannot help but notice the bucket of sudsy water for cleaning their windows. They also see the paper towels for wiping the oil

dipstick located next to the gas pump and they find the air hose for their tires in a convenient and natural place. The location and ease of use of these services make the driver want to use them.

Secondly, we can maintain brightly lit stations at which our prices and the available goods are clearly and attractively advertised. By doing this, we ensure that passersby on the Internet know our product and understand how we do business. If we do these things well, we will create happy, well-informed motorists who will enjoy their filling station experience in spite of the fact that they did it themselves.