Prospects for Distribution of DHS Datasets via the Internet

Trevor Croft and David Cantor
Demographic and Health Surveys
Macro International
11785 Beltsville Drive
Calverton, MD 20705-3119
croft@macroint.com
cantor@macroint.com

Paper prepared for
the Population Association of America meeting
in New Orleans, May 9-11, 1996

1. Background

The Demographic and Health Surveys (DHS) Program has been conducting surveys in developing countries for the past 12 years. A major output of this effort has been a rich body of datasets available to researchers around the world. To date, over 250 datasets have been made available for 70 separate surveys. DHS has created a data archive for the distribution of these data and over 7000 copies have been distributed worldwide.

1. Survey content

DHS surveys cover a number of topics, and usually include several sets of data. Typically, a survey will include a **household schedule**, an individual **woman's questionnaire**, and a **community service availability questionnaire**. Increasingly, a man's **questionnaire** is also being used in the survey, particularly in sub-Saharan Africa.

The standard woman's questionnaire covers the following topics: Background characteristics, Lifetime reproduction, Contraceptive knowledge and use, Maternity and breastfeeding, Immunization of children, Diarrhea, fever and cough in children, Anthropometric measures for children and their mothers, Marriage, Fertility preferences, Husband's background, and Woman's work status. Additional modules are often used to collect further information on certain topics, including Natural family planning, Social marketing, Sterilization, Pill use compliance, Maternal mortality, Causes of death, AIDS, and Woman's employment.

2. Data file types

For each survey, DHS produces a separate data file for each questionnaire type: household, woman's, man's, and service availability. These data are known as the **raw data**, although the data are really not raw. By the time a researcher sees the data file, it has been fully edited, complete data for dates of key events in the dataset have been imputed, and textual responses have been numerically coded. The data are, however, still in the structure used to collect the data, based on the questionnaire.

Data structures that are used in collecting data are often not the most convenient for analysis. The questionnaire and structure used are specific to each survey, although they are based on standard questionnaires. As DHS is a program designed to produce comparable data across countries, and to provide cross national comparative analyses of these data, a common data format was needed for all surveys. Under the first phase of DHS, a **standard recode format** was produced for the women's data. Under the second phase, a standard recode format was introduced for the household data, and under the third phase, a standard recode format is being used for the male data. Thus, in addition to the raw data files for each questionnaire type, there are also recode data files for each questionnaire format.

3. Data formats

The standard recode data, as well as the raw data are made available in four data file formats:

- **Flat files**, containing a single record, usually of more than 2000 characters in length, for each case in the data file. These data files are designed for use on mainframe computers.
- **Rectangular files**, containing a fixed number of records per case, with shorter record lengths. These data files are designed primarily for use with *SPSS/PC+* or *SPSS/Windows* and other PC-based software.
- **Hierarchical files**, containing a variable number of records per case, depending on the content of the case. These files are designed for use with packages supporting complex data structures. They are the standard format used by the Integrated System for Survey Analysis (*ISSA*), the package used for all data processing in the DHS program.
- **Transposed files**, which are a proprietary format used by *ISSA* in which the data matrix is transposed, to provide efficient processing of data. This is the prime format of data used with the package *EASEVAL*, although *EASEVAL* also reads the hierarchical format.

The flat, rectangular and hierarchical data file formats are all ASCII files, while the transposed file format is a binary file. All of the ASCII based formats can be used with any package capable of reading ASCII files, although processing with certain packages is easier if the most suitable file format is used.

4. Media for distribution

DHS has been distributing data from the project for more than ten years, and has supported a wide range of file formats. Currently, DHS supports data distribution on **diskette**, and on seven different types of **Bernoulli cartridges** for use with PC-compatible computers, and on several **magnetic tape** formats for mainframe computers. DHS also has the capability of producing custom **CD-ROMs** for large data requests. Data are distributed on diskettes and on some of the Bernoulli formats in a zipped format, due to the size of the data files (some data files exceed 100Mb in size, although most are considerably smaller at around 10-15 Mb).

5. Next step: the Internet

Data distribution is a major component of the DHS program, and DHS employs full-time staff whose sole responsibility is to manage and distribute data. With increasing numbers of requests for data, increasing numbers of datasets available, and the increasing size of the data files, the distribution requirements for DHS have increased substantially. This has led to some problems and delays in the distribution of data, and a considerable burden on DHS to continue to provide a responsive service to researchers.

Additionally, the Internet has appeared as a global means of communication. It provides possibilities unimagined a few years ago. With communication from the US to Australia taking seconds, and transfers of fairly large quantities of information over the Internet taking just minutes, new ways of distributing data are becoming available. For these reasons, DHS is looking at the possibility of distributing data over the Internet.

2. Current DHS Internet Activities

1. Where is DHS?

DHS has been using the Internet for more than four years. Initially, access was limited to e-mail. In addition to each staff member having an e-mail address, DHS has two specialized addresses: **archive@macroint.com** for the DHS data archive and **reports@macroint.com** for publications.

In the spring of 1995, DHS established its first World Wide Web site. Since then DHS has expanded its Web activities and continues to develop and extend the Web site. DHS can be found at **http://www.macroint.com/dhs/** [Note that the last slash is needed to ensure access to links from the home page.] At the time of writing, the DHS Web site has received well over 5000 visitors.

2. What is available?

To date, the **DHS Web site** is purely informational, but it is an ever changing site to which we will be adding further information on a regular basis. The current pages include:

- Brochure providing a description of the DHS program and its goals.
- Survey status giving the status of each DHS survey.
- Fact sheets providing key figures from each survey.
- Press releases describing key findings from each survey.
- Newsletter presenting articles and information from surveys and the DHS program.
- Publication information including complete lists of publications and how to order them.
- Data Archive information and how to order data.
- Information on other related projects at Macro, including nutrition chart books, and the India NFHS surveys.

3. How is it supported?

All of the DHS Web information was originally placed on the Internet using a SCO UNIX server. Since then, Macro International (the DHS parent organization) has obtained a new Web Server with greatly enhanced capabilities. The new server utilizes WebSite 1.1 software operating on a Compaq Proliant 1500 running under Windows NT. WebSite 1.1 is a 32-bit multi-threaded World Wide Web server that combines power and flexibility with ease of use. It utilizes an intuitive graphical interface and works under Windows NT and Windows 95.

The WebSite software offers a number of important features for the management of a web site. In addition to providing for basic document retrieval, it supports the use of image maps with "hot regions" for selecting links, and automatic URL fix-up (for example, filling in the trailing slash on the URL name if the URL points to a directory).

In addition, the WebSite software supports *CGI* programs, allowing for fill-in forms and document-based queries, and provides database integration with *Visual Basic* and *Access*. The software also provides security features with username and password access control, and IP address and host name filtering.

3. Proposed Internet Activities

DHS is in the process of planning its further Internet activities, and is considering a number of possibilities. Below we briefly describe the major items we intend to make available via the Internet.

1. What will be available?

1. DHS data

DHS is planning to eventually make all of the DHS datasets available over the Internet for researchers to download. This will be performed over a period of time, primarily due to the large volume of DHS data available. Data will be accessible via File Transfer Protocol (FTP), either through the DHS Web pages, or directly through an FTP connection. Data will be stored in zipped files, together with the appropriate machine readable documentation, such as frequency distributions, and *SPSS* and *SAS* description files.

DHS will also make the individual recode documentation for each country available as Web pages that can be viewed through a Web browser. DHS data alerts, which document problems that have been found with data files, will be made available in the same manner.

2. India NFHS data files

Data from the India National Family Health Survey (1992-93) is the first to be made available through the DHS Web site. Beginning in May, 1996, users will be able to download NFHS data in *SPSS/PC* format for all states participating in the NFHS and for the country as a whole. Data are available in zipped archive files for both household and individual questionnaires.

The zip file also contains the SPSS dictionaries. Registered users will also be able to download the documentation for these datasets in either ASCII or *WordPerfect 6.1* format. From the same Web page, users can also download the *Select* utility for easily subsetting their *SPSS/PC* datasets. Data from the NFHS is distributed free of charge, after users have been registered.

3. Report text

It is expected that the text of DHS reports will also be made available over the Internet, again for downloading via FTP. There are three components to DHS reports: text, tabulations, and graphics. Reports may be provided with all three components in one file, or the components may be broken into three separate files, or possibly, text and tables will be in one file, with graphics in a separate file. DHS will be looking for feedback as to which form would be the most suitable for researchers.

4. Online newsletter

DHS has begun experimenting with a prototype online newsletter. Tentatively named *DHS Discoveries*, the electronic newsletter would be available in two versions. One would be a hyper-linked version on the DHS Web site, where users could read summaries of the latest findings from DHS, complete with charts and graphics. The other would be a text only version through an electronic mailing list, supported by a mailing list manager such as *Listserv*. This would be delivered to subscribers via their e-mail system approximately every 4 weeks, and would contain the same information as the Web version, but would be available to those who do not yet have Web access. Users could automatically subscribe and would then keep receiving information until they unsubscribe. Each time this newsletter would be delivered, recipients would have an opportunity to receive some of the latest articles, publications or findings from recent DHS surveys. It is envisioned that this type of electronic newsletter would be delivered to up to 2000 subscribers, free of charge. This would be especially important to subscribers in parts of Africa and Asia, where Web service is not yet presently available.

5. User forum

DHS is proposing the establishment of a DHS user forum, where any researcher can post messages for DHS staff or other researchers to react to and comment on. This forum would probably be moderated, at least initially, to place some control on the messages that are disseminated to researchers who subscribe to the forum. The forum would be run through the same mailing list manager facility as the online newsletter. It would give researchers the opportunity to hear more about how other researchers are using the DHS data, ask questions when they have difficulty using the data, provide problem reports when they find errors in the data, and to provide input to DHS staff concerning data users needs.

6. FAQs

DHS plans to set up a section of its Web pages for frequently asked questions (FAQs). These would hopefully be questions that are received through a variety of media, particularly the user forum, and for which DHS believes there are a considerable number of users who would

benefit from the answers to the questions. There are already several pages linked under the DHS newsletter pages that contains some frequently asked questions. These will be expanded into a new FAQ section.

2. How will it be supported?

1. FTP

The DHS program plans to utilize FTP (File Transfer Protocol) for sending data files from DHS to data users. FTP allows a person to transfer files between two computers, generally connected via the Internet. If a system has FTP and is connected to the Internet, you can access large numbers of files available on a great number of computer systems. This can either be done through the World Wide Web or by direct FTP connection.

When using FTP, the remote user invokes a program, called a 'client' to connect to a machine that holds the files, called a 'server'. Many computer systems throughout the Internet offer files through anonymous FTP. This means that you can access a machine without having to have an account on that machine (i.e., you don't have to be an official user of the system). These anonymous FTP servers contain software, documents of various sorts, and all sorts of other information.

Anonymous FTP is a facility offered by many machines on the Internet. This permits the user to log in with the user name 'anonymous' or the user name 'ftp'. When prompted for a password, the user types his or her e-mail address -- it's not necessary, but it's a courtesy for those sites that like to know who is making use of their facility. Some sites require a valid e-mail address, others don't.

The DHS server will require FTP users to be registered before they will be allowed to download files. Potential users will need to complete a registration form, providing information on their name, location, affiliation and a description of the kind of research they plan to do using the data. They will also need to select a user name and password, which will need to be provided each time they access the FTP part of the DHS Web site. All of this information will be stored in an offline database. DHS will notify users who have applied for data access, with an E-mail or phone call, letting them know that their application has either been accepted or rejected. If the user's application has been accepted, the data user will be granted access to the FTP site, and they will be able to download DHS data. For researchers with access to the Web, the registration form can be completed interactively over the Internet. For researchers who do not have access to the DHS Web site or whose Web browser does not support forms, registration will also be possible via e-mail.

Each time a user attempts to access the FTP site, they will need to type in their user name and password which will be checked for authentication through a query of the database. If the password matches for that user, they will be granted FTP access. When the user downloads a DHS file or files, the downloaded information will be appended to another database for tracking purposes.

It is not certain at this time whether the e-mail return notification for authorization will be an automated procedure or a manual one. It will most likely start out as a manual operation. However, if a great many individuals request access, DHS may look into an automated authentication e-mail generating system.

2. World Wide Web

DHS would also make the FTP gateway available through its World Wide Web pages, where users could select the files to download after filling out several *CGI* (common gateway interface) forms. It would also be necessary to verify user authentication, by way of a security form. The steps to gain access to the desired files would look something like the following:

1) Access DHS Web Page
2) Complete User Registration Form and Submit
3) Registration and Authentication at DHS (24 hour wait)
4) Receive Confirmation of User Password from DHS via E-mail
5) Access FTP Server (via DHS Web page or directly via FTP connection)
6) Provide User Name and Password
7) Select Country, File Type and File Format (e.g. Bolivia, Individual Recode, Rectangular)
8) FTP File to User

These steps are essentially the same as for direct FTP access, and in fact users could use either the Web pages or the direct FTP access interchangeably. Essentially the Web pages would just be a more intuitive interface to the data than the direct FTP connection.

Data documentation and data alerts will also be accessible at the DHS Web site, both as Web documents displayed by browsers and as downloadable files in *Word Perfect 6.1* format.

The DHS Web site will also support the FAQ (frequently asked questions) section and the electronic newsletter, *DHS Discoveries*. This newsletter will be available not only over the World Wide Web, but also through e-mail subscription, supported by the mailing list server.

3. Mailing List Server

DHS plans to use electronic mail as a means of communicating with researchers, and will establish a mailing list manager, such as *Listserv*, on the DHS Web site. This facility will support a number of activities, including

- the distribution of the text only version of the electronic newsletters to registered users.
- the distribution of documentation updates and data alerts to users of specific datasets.
- the establishment of a moderated user forum for users of DHS data.

3. Logistic Issues

1. Data access restrictions

Under DHS sub-contracts with governments of each of the individual countries, the data from the country is the property of that country. Data are not public domain, but agreement has been reached in the sub-contracts that the data be widely available for researchers to analyze. Strictly speaking, permission is needed for researchers to use DHS data, and permission is given on a project by project basis. However, for the vast majority of situations, it has been agreed that DHS can distribute data to bona fide researchers without first contacting each country for permission. In a few cases, there are restrictions on access to the data, requiring specific permission for each request to be sent to the country whose data is to be used. Due to these requirements, open public access is not possible, and users will be required to register to use the data.

According to DHS sub-contracts with each of the countries, DHS must track all of the users of the data for each country, according to the project for which the data are to be used, to allow countries to know how their data are being used. DHS will maintain a database system tracking user's information, along with user names and passwords for accessing the data, and track all datasets that are downloaded.

2. Charging for datasets

DHS has had a policy of cost recovery for its data distribution activities throughout the life of the program. DHS currently charges $200 per survey for data to cover the cost of the media, the preparation time and the cost of packaging and mailing of data. It should be noted that the charges are not for the data, but merely to help cover the cost of distribution. With the advent of data distribution over the Internet, the issue of charging for datasets needs to be reviewed.

The major benefit to charging for datasets, other than to cover some costs, has been to discourage people from requesting data that they are unlikely to make use of. Most researchers who request data are keen to make the fullest use of the data. While there is much less rationale for charging to cover the costs of distribution of datasets over the Internet, the secondary benefit of discouraging users who are not serious about using DHS data should not be ignored. This is particularly important, in relation to concerns at DHS about bandwidth.

If DHS were to decide to charge for datasets, the next issue would be how to implement a charging system. There are three solutions to this that spring to mind: payment in advance, credit card payments over the Internet, and electronic cash. None of these solutions are ideal. If payment in advance is required, the major advantage of the Internet -- instant access -- is lost. Credit card payments and electronic cash also have problems: firstly, Macro's system is not a secure network for financial transactions, and secondly, these technologies are essentially available only in the developed world and not at all in the developing world.

Assuming that DHS were to charge for access to data, the final issue is how much to charge. Is $200 still a valid amount to charge for data, or should the amount be cut to a lower

figure of $50 per survey, say? Would cutting the price be unfairly discriminating against researchers who don't have access to the World Wide Web, and who, presumably, have less resources than those with access? Is it worth the expense of setting up a system to collect the money?

No decision has been made concerning charging for datasets over the Internet.

4. Hardware issues

The DHS Web site, proposed FTP accessible data files and DHS mailing list manager are all located on Macro International's Web server. The WebSite software runs under Windows NT on a Compaq Proliant 1500 with an online storage capacity of 4 gigabytes. Macro used Sprint as its Internet service provider and utilized a frame relay type connection operating at 56Kb per second until recently. In the past few weeks, Macro switched to DIGEX as its Internet service provider and installed a T1 type connection which will increase the bandwidth throughput capability by a factor of 24 to 1.5Mb per second.

Of course the success of this proposed FTP operation depends not only on server hardware capabilities at Macro/DHS, but also on client or user hardware factors. Some issues to think about are: Do the potential users have access to a computer and modem with an Internet connection (especially users in developing countries)? Phone lines in developing countries often cannot support baud rates greater than 9600 which slows down file transfer. Even with the necessary hardware a user can often face other hurdles in successfully completing FTP. The amount of network 'traffic' at either the server site or with the network service provider can cause problems in connecting or can drastically reduce the speed of file transmission. In developing countries, there are often older phone lines which generate more 'noise' which can also slow down rates of transmission or cause a disconnection. All of these issues raise concerns about how useful or viable Internet data distribution will be, but it is believed that all of these problems will disappear in the next few years, particularly as more and more developing countries have full Internet access, including access to the World Wide Web.

The authors have performed several test downloads using a compressed data file of 1.8 MB in size. With a 28.8K baud modem connection using America Online as the Internet network service provider, the file transfer took 11 minutes and 40 seconds. Downloads over a system that is connected directly to the Internet take less time, but downloads over a 9600 baud modem, over noisy telephone lines may require as much as an hour for the same size file. While this file size is about average, there are certain data files that are considerably bigger and may require as much as 12 hours in an extreme situation (assuming that the connection can be maintained for that long).

5. Conclusions

The prospects for DHS being able to distribute data over the Internet are encouraging. While there are a number of issues still to resolve, and this paper has only touched on a few, DHS should be in a position to provide data over the Internet in a few months. The experience working with the India NFHS data, which is due to be made available over the Internet in late

May 1996, will help to focus DHS efforts and will provide valuable insight into how the data will be accessed and used.

In addition, the other features that DHS plans for the Internet should be an important resource for researchers. It is hoped that these features will provide a complete service. Any suggestions on the services DHS hopes to provide and any feedback on the ideas presented in this paper will be appreciated.