Codebooks in the World of Networked Data Library Services

by Richard C. Rockwell[1]
Executive Director
Inter-university Consortium for Political and and
Socila Research

Twenty years ago data archives and data libraries moved out of the world of punched cards to the world of magnetic tape. A similar migration is underway today, with a less clear migration path ahead of us. Some see a bright future for network transmission of data and for remote mounting of disks over the network, others see diskettes (of increasing capacity) and CD-ROMs continuing to play an important role for many years, and still others see new varieties of magnetic tape coming to dominate data distribution media. What does seem clear is that, whichever media are used, the machines on which most researchers will soon be doing their work will not be mainframes even minicomputer mainframes. In the place of interactive systems built around mainframes and associated clusters of terminals will be variations on the theme of the distributed computing environment, built around some combination of powerful desktop computers and workstations, file servers, and compute servers. These systems will be connected by local Ethernet-speed networks and those LANs into national and international networks operating at increasingly high speeds. Researchers will have come to expect the network to deliver a wide variety of services directly to their desktops.

Distributed computing has already had many effects on the research process, one of the most prominent of which is rising levels of intolerance for inconvenience, delay, and clumsy service. This is particularly true among researchers who have learned to use the Internet or other networks for something beyond electronic mail. Many social science researchers know that it is theoretically possible for them to obtain a data set over the network from anywhere in the world, often in a matter of minutes; to store those data on a local hard disk; and to begin analyzing those data immediately. They know this is possible because they see it now being done by their colleagues in the natural sciences and even in the humanities. They also know again, because they see it being done that it is possible to logon to remote computers and through X-Windows use software and data sets resident elsewhere as if one were a local client of the distant computer, with that computer using one's desktop screen as the display device. They know that complex documents and data bases can be transmitted over the network and that these files can be searched and readily displayed. And they expect that all these services will become faster, more intuitive, more effective, and cheaper next year than today, and that new services will be continually added.

I conjecture that all data library services are facing a rising demand from the research community to "get with it" in the networked world, and that impatience is rising among our users that we have not blazed a trail into the networked world. We probably have little time remaining to make the necessary moves, because many of us (as well as many of our customers) are losing the mainframes on which we relied for the production and use of older media. Within the past year, several major university members of ICPSR have become incapable of using reel-to-reel tape. This transition thus has a number of implications for data archives and data libraries, one of which I will discuss today: the documentation of data. Old hands in the data archive movement will recognize a debt for this proposal to Ralph Bisco, who thought far beyond his time. Some will also note how much easier it would be to move in this direction as a start-up organization, without an existing archive of thousands of reels of tape to deal with and continuing needs for service from users.

Documentation in a distributed computing environment
When data were distributed on reel-to-reel tape and manipulated on a mainframe, it was practical to provide documentation in hard-copy form. The mails could carry printed documentation as readily (and as slowly) as they could carry tape reels. Researchers could consult this documentation in a central campus data library and make copies as needed. The process by which the researcher was connected to the data took days or weeks. Lots of paper was involved in the process, and expenses for documentation were becoming an increasingly significant portion of archival budgets.

If network distribution of data or remote mounting of disks becomes the norm for data distribution, data library services will be compelled to find alternative ways of distributing documentation. It is rather difficult to stuff a book down the network. A situation in which the researcher can obtain the data over the network but must wait days or weeks to obtain hard-copy documentation by mail would clearly be unacceptable. Sending documentation by fax would be impractical and expensive for large documents.

Hard-copy documentation is more feasible if data are distributed on diskettes, CD-ROMs, or mag tapes, but the financial implications of this strategy are considerable: distribution of documentation in machine-readable form is extremely cheap compared to hard-copy. ICPSR spends about 5,000 for each of the printed codebooks it prepares for the Eurobarometer series, and about 16-18,000 for the documentation of a major study such as the 1992 American Election Study. Duplication costs for non-printed hard-copy codebooks are high and rising. There is an inventory problem, and this involves such mundane things as space rental. In addition to financial considerations, from the researcher's point of view it may be more convenient to have the documentation bundled with the data on the distribution medium. In any event, the data and the documentation must travel together.

Hard-copy documentation fails another test as well: one of the principal services afforded by the networked world is the ability to do a full search of the contents of many different archives. If a substantial part of the documentation (say, the codebook) is not available in computer-readable form, these search services will not work to their fullest. Researchers will be less well-served than they will have a right to expect.

The current design of the International Directory Network primarily a project of the European Space Agency provides one example of full access to documentation. IDN maintains a four-tiered structure for documentation, all of which tiers are searchable through the network. At the top is the directory, modeled on the concept of the Yellow Pages, which provides orienting information to collections of data sets. ICPSR might have some 30 directory entries in such a structure, one of them pointing to a collection of data sets on "Mass Political Behavior and Attitudes: Historical and Contemporary Electoral Processes." Underneath this directory entry would be an inventory of the some 160 studies included in this directory entry, consisting of the study descriptions for those studies from the Guide to Resources and Services. For each element in the inventory the third level of documentation (confusingly called the "guide" in IDN terminology) would consist of study documentation (codebook, questionnaire, sample description, etc.). The fourth level would provide direct access to the data, in a "browse" facility that would permit researchers to explore the data to determine if the data meet their needs. This service will be implemented around the world, so that the researcher can simultaneously search the contents of archives on three or four continents.

The IDN concept is essentially a partial implementation of hyper-text and has been designed around the needs of a particular community, the remote-sensing community. Whether or not that system will well serve social science must be evaluated. There are incomplete alternatives available, including the GIDO system being developed by the Swedish national archives and the Isis system employed by the German national archives, as well as more general systems such as WAIS, Gopher, World Wide Web, and Mosaic. Whichever of these systems or others we will eventually adopt remains uncertain. However, it is, I think, high time for the data library services community to begin planning for a new era in documentation standards and methodologies. I lay out below my preliminary thinking on principles for documentation, ways of implementing those principles, and how the researcher might use the new facility.

Principles for electronic documentation of social science data

In the past we have tended to treat data inventories, variable indices, codebooks, marginals, and questionnaires as separate entities, often providing electronic access to one but not to another part of the documentation system. Further, the process for retrieval of data has been separate from the search and documentation system. A driving principle for the future would seem to be that search facilities, access to documentation, and data distribution must all be integrated into a single system. The barriers separating codebooks, questionnaires, and data should be eliminated. We should also consider providing additional documentary elements, such as bibliographies, core articles, fundamental tables and graphs, and advisory notes.

This system will have to work in a variety of environments, ranging from mainframes and minicomputers to UNIX, MS-DOS (Windows and not Windows), and Macintosh systems. It will have to support the continued distribution of reel-to-reel tapes and tape cartridges as well as diskettes, CD-ROMs, and several forms of network access. It will have to facilitate the use of paper as well as the use of screen displays. It cannot be dependent upon any peculiar operating system or hardware configuration. It cannot be bound to the printed page or to an image of the printed page. It cannot presume that it exists solely for the purpose of documenting CD-ROMs or solely for FTP. That is, the documentation system must be functional in all currently competitive computing environments, and in both electronic and physical media environments.

Codebooks must display attractively on screens, pop-up boxes, desktop printers, line printers, and typeset books. The machine-readable codebooks that we now distribute do not display very well on screens, and as a result many members of the research community find that mode of access unacceptable. They usually print copies of what we send them (or ask us to print them), and the result is often an unattractive, bulky, hard-to-use document. ICPSR members have sometimes had considerable difficulty printing our machine-readable codebooks in a usable form. We ought to strive for device independence so that the documentation will display in the same way no matter what the display device (and do so easily). This involves more than just ensuring adaptability in some form to any existing device; the capability to display documentation well and efficiently on any device is a design criterion.

We need to give much more attention to the readability and attractiveness of documentation. Our best codebooks today resemble those we were producing 20 years ago, when line printers were the only display devices that we had. As a result, they have an enormous amount of white space. They use only the typographic abilities available on the line printer, principally meaning the use of all-caps for some text. The information content of a printed ICPSR codebook, page for page, is very low, meaning that we are printing thick books that could be thin and more usable (cheaper too). This becomes a special problem when we use screensas display devices, because the current design is almost antithetical to optimal design for screens. Had we started with the screen as the display device, I think codebook design would have been radically different. Besides eliminating excess white space, we need to use typographic tools such as italics, bolding, underlining, and boxes so that users can more easily locate the information they are seeking.

We need to provide for codebooks to be used directly as data documentation by a wide variety of statistical packages. It should be unnecessary to prepare one data definition file for SAS, another for SPSS, another for OSIRIS, and still another for NSDStat. The OSIRIS programming leader, Bill Connett, has already made a commitment to adopt our new standard codebook in the UNIX implementation of OSIRIS, if doing so is at all possible. We need to ensure that it is possible and make it commercially attractive for statistical software houses to implement the standard. And we need to ensure that systems suchas NSDStat that provide pop-up documentation windows on a question-by-question basis will be able to use the codebook for their sophisticated displays.

If codebooks are to be integrated into a system that includes directories, inventories, and data, and if this entire system is to be searchable over the network, simple flat text files will not suffice. Codebooks must be structured text documents, with directly-accessible entries ("access points" or "attribute sets") for study titles, sample information, variable names, variable labels, full question text, full code descriptions, marginals, missing values, and notes. This structure must support an extensive search capability, so that it is easy and efficient for the researcher to locate needed information accurately and quickly. A hyper-text design virtually seems mandated for this system.

The search facility should have a degree of information or intelligence about social science built into it. For example, it should have thesauri that permit it to identify a data set as containing information about "income" even if the data set documentation only contains the term "salaries" and "wages." Equivalent terms (such as "environmental attitudes" and "attitudes about the environment") should be treated as equivalent by the search facility without requiring the user to construct complex Boolean expressions. Grammatical transformations should be handled by the system, not the user; for example, plurals and changes of nouns to adjectives should be transparent tothe search process. Ideally, there would be a minimum standard set of terms used by all data archives in describing their contents, so that the ratio of hits to misses and false hits is kept as low as possible. As in past attempts to implement standardlists of terms, compliance will be a problem but perhaps less of a problem than before because the benefits of compliance will be so clear.

The researcher needs to have direct access to the actual questionnaires, not just to the codebooks. Because of the graphic complexity of questionnaires and ancillary documents such as flash cards, in many cases the questionnaire must be provided as an image rather than as an ASCII file. Our new standard should provide for the incorporation of bit-mapped scanned-image data.

It would be sad to design a codebook standard around the simplest kind of data and then find that it does not adapt to more complex data sets and must therefore be discarded. Our standard should generalize, from the first, to data sets of all sorts, including aggregate data, hierarchical data, contextual data, time series data, and textual data. It should not assume that researchers are capable of working only with flat or non-hierarchical files. Consideration should be given to documenting inverted files and relational data bases, as well as conventional structures and dynamic data bases.

Documentation should be designed so that it is applicable internationally, at least within the family of European languages utilizing a common alphabet. This probably means adoption of the new ISO standard character set, in place of

ASCII, so that characters not used in English can be represented. Or it means adopting a standard such as SGML that internally defines the character set.

Furthermore, the standard should be capable of being implemented worldwide on hardware that it is reasonable to expect users to have or to acquire. A system that requires that the user purchase a high-end workstation, for example, would not be acceptable. This means that it should utilize only standard, off-the-shelf hardware. On the other hand, while some system services ought to be available to users with lowest-end hardware, it would be highly undesirable to design the system around the limitations of the bottom end of computing equipment. It is perfectly reasonable to assume that users willhave access to 386-class machines with hard disks and graphics displays, for example. Without at least such a machine, users will effectively not be able to function in the new networked world.

The system ought to be built upon a foundation of commercial support for software as well as hardware. The archives cannot be responsible for developing, disseminating, or supporting software. What we offer should be compatible with familiar desktop tools such as word processors and with Internet tools such as WAIS or its successors; we should not haveto write our own display software, search programs, etc. We cannot afford to do that programming; furthermore, we are just not as good at it as are the commercial houses with their millions of customers and enormous financial resources. Further, it would be impractical to require that each potential user of our services install a special program or a special interface on the local computer. It would be far better for tools that are freely available on the Internet or are otherwise widely dispersed to be all that the user needs to search our documentation, display it, and then retrieve the data. If we can possibly avoid it, ICPSR will not putitself into the position of developing and supporting proprietary software for use by the research community.

This commercial support is likely to come only if the standard that we adopt is widely adopted in other fields. Social science is not itself large enough to attract the needed level of commercial attention. We need to participate in shaping international standards and then ride on their momentum. For this reason, we need to examine emerging or existing international standards for documentation, such as SGML (Structured Graphics Markup Language), Z39.50 (the ANS Information Retrieval Service Definition and Protocol Specification for Library Applications, constructed under ISO 7498, the Open Systems Interconnection basic reference model), the MARC record and its descendants now under development at the Library of Congress, the standards already used by Isis or other commercial software, and any similar standards emerging from European efforts. In some sense, it matters far less which standard we adopt than that we adopt a standard that is larger than social science and that the standard have commercial interest.

It is critical that the standard be supported within the Internet. This means compatibility with existing Internet tools such as WAIS, Gopher, and/or the World Wide Web, or emerging tools such as Mosaic or CIESIN's information services tool formerly called Green Pages. These tools are rapidly increasing in power and reliability, and they have large resources behind their development. Social science can ride along on these developments, but it would be better for us to have a hand in shaping them.

One scenario for use of this system

Let me describe the session that I imagine a user conducting with this new system under the scenario that the Internet can carry the full load. She Telnets to a central access point and signs on to a directory search facility, probably using an X-Windows or (the forthcoming) Windows NT desktop interface. Using both free text search capabilities and the ability to form Boolean expressions, the user specifies the kinds of data in which she is interested. Upon command, the server accesses the directories of linked archives on three continents, reporting back that 11 different archives contain data on her topic. Using the same interface, the user searches or scans inventories of each archive, starting with study titles and moving to study descriptions when her interest is piqued. Having identified some six studies in which she is interested, within the same environment she searches or scans codebooks for each of these studies. She examines marginals to see if the data can support the design she wishes to execute. She views questionnaires. Finally, she does some preliminary tabulations or draws a scatter plot of a correlation. Perhaps three studies seem to meet her specifications.

She then places a request that those three studies be transmitted to a file server attached to her computer, along with the necessary documentation. The involved archives determine that a request for data from her is legitimate because of international data-sharing agreements. The archives automatically initiate FTP put processes to place the data on her file server, or they authorize her to issue a get command. (Or they permit her to remotely mount the disk containing the data set in a client-server configuration, soon to be implemented by SAS.) Within a couple of hours she has the data and documentation that she needs to do her research. Human labor, other than her own, has not been involved in this transaction. No paper has changed hands.

At her computer she displays the codebook as a hypertext document, clicking on a section of the questionnaire, expanding that to a single variable name, expanding that to a full question text, further expanding it to a set of value codes, then viewing marginals and notes. She prints on her desktop printer some portion of the codebook because she knows she will frequently use it; other portions she retains on the file server for future use. When she initiates an SPSS, SAS, OSIRIS, or NSDStat process, she simply points the program to the codebook for documentation, concentrating her attention on analytical commands.

Any questions about problems that she encounters can be addressed to a local campus data librarian. This professional has been electronically informed by the involved archives that the data have been transmitted to a user on the campus. This information will subsequently be used in reporting usage levels to the data librarian, and if accounting is involved, in billing for services. The data librarian has the ability to view on the library's screen what is being displayed on the user's screen and can directly assist the user in problem-solving.

When the user is finished with the data and/or documentation, she discards everything and frees up local disk space, knowing that she can as easily obtain the information a year or two later when she needs it again. Precious local disk space is thereby conserved, and the contents ofdata archives are not duplicated in miniature around the world. The 2,000 or more reels of tape with copies of ICPSR studies that once cluttered many campus computing centers are no longer needed. Tapes are primarily used as back-up media.

Providing this kind of service to the research community will not be easy and will not be cheap. It requires a substantial amount of research on standards and their implementation. It requires an enormous amount of work on existing documentation. The adoption of the standard across many data library services requires an unprecedented level of inter-archival cooperation. The whole process demands high levels of consultation with researchers and data librarians. For all these reasons, the goals sketched here may seem unattainable. I think that they are not unattainable, and that somebody will attain them before long and I hope it's us.

1. Prepared for the IASSIST/IFDO 93 Conference held in Edinburgh, Scotland. May 1993..