The IASSIST QUARTERLY represents an international cooperative effort on the part of individuals managing, operating, or using machine-readable data archives, data libraries, and data services. The QUARTERLY reports on activities related to the production, acquisition, preservation, processing, distribution, and use of machine-readable data carried out by its members and others in the international social science community. Your contributions and suggestions for topics of interest are welcomed. The views set forth by authors of articles contained in this publication are not necessarily those of IASSIST.

Information for Authors

The QUARTERLY is published four times per year. Articles and other information should be typewritten and double-spaced. Each page of the manuscript should be numbered. The first page should contain the article title, author's name, affiliation, address to which correspondence may be sent, and telephone number. Footnotes and bibliographic citations should be consistent in style, preferably following a standard authority such as the University of Chicago press Manual of Style or Kate L. Turabian's Manual for Writers. Where appropriate, machine-readable data files should be cited with bibliographic citations consistent in style with Dodd, Sue A. "Bibliographic references for numeric social science data files: suggested guidelines". Journal of the American Society for Information Science 30(2):77-82, March 1979. If the contribution is an announcement of a conference, training session, or the like, the text should include a mailing address and a telephone number for the director of the event or for the organization sponsoring the event. Book notices and reviews should not exceed two double-spaced pages. Deadlines for submitting articles are six weeks before publication. Manuscripts should be sent in duplicate to the Editor: Walter Piovesan, Research Data Library, W.A.C. Bennett Library, Simon Fraser University, Burnaby, B.C., V5A 1S6 CANADA. (604) 291-4349 E-Mail: USERDLib@SFU.BITNET

Book reviews should be submitted in duplicate to the Book Review Editor: Daniel Tsang, Main Library, University of California P.O. Box 19557, Irvine, California 92713 USA. (714) 856-4978 E-Mail: DTSANG@ORION.CF.UCLEDU

Title: Newsletter - International Association for Social Science Information Service and Technology  
ISSN - United States: 0739-1137 Copyright 1985 by IASSIST. All rights reserved.
Metadata and User Interfaces: Promises and Problems

by Steven R. Howe  
& Robert J. Graham  
University of Cincinnati

Like most of us, I was certainly impressed the first time I encountered a well-designed interface for using data on CD-ROM. The dramatic improvements in user interfaces that accompanied the introduction of CD technology were motivated by the vision of the analyst interacting with the metadata. The early successes that we have seen have prompted us to ask what more might be accomplished with better metadata.

The goal for the use of metadata and the development of user interfaces should be nothing less than permitting everyone from the novice to the expert to function independently at a desktop machine. To achieve this goal, at least three problems need to be addressed.

First, we have untold numbers of studies for which the metadata needed to produce interfaces are either not available, or they are available but the demand for a custom interface is not sufficiently high to make it feasible to produce one. This problem is not restricted to our existing collections; researchers continuously produce new data sets, and relatively few might ever be the focus of a secondary analysis.

Second, interface developers have tended to develop products that are based on the way we use documentation in print. For example, an interface that permits reviewing the data dictionary may not allow the user to see clearly the skip patterns among the questions without resorting to a look at a printed copy of the questionnaire. The situation is analogous to the early days of the automobile when the body was made to look like a buggy while demands for new technology specific to motoring had to emerge slowly with experience.

Third, there are serious problems with CD technology related to network accessibility, ease of copying, and processing speed. How many of us are confident that we will be using existing CD technology ten years from now for our data storage needs? We need to ask ourselves if we are inadvertently making technological commitments when we make investments in metadata and user interfaces.

Metadata and New Directions for Interfaces

Before I present what I see as a solution to these problems, I want to spend some time talking about how I would like to see user interfaces evolve. My remarks are geared largely to the tasks faced by the secondary analyst of survey data, although I am open to the idea that librarians, students or scholars may need metadata for purposes that go beyond those which I discuss below. My objective is to demonstrate that the problems I have briefly sketched have two root causes in common: metadata is hard to compile; and user interfaces are tied to particular structures for metadata.

Screen Studies

There are certain questions that must be answered before the analyst even makes a decision about whether or not to investigate the use of a particular data set. Who conducted the study? When? What was the population? The sample size? What was the purpose of the study? What content domains are covered by the data? Given our goal of making the analyst working at a desk-top machine self-sufficient, how do existing interfaces rate?

The information needed to screen studies has traditionally been published in catalogs of holdings. It may be relatively easy to put exactly the same narrative material onto CD/ROM as into print, but users of these different media may not be equally well-served by having the information organized in the same way. I suspect that most of this type of information currently distributed on CD/ROM gets printed out and read instead of being reviewed interactively.

I have never tried to construct a questionnaire that would capture all of the information about a study that might be needed, but it is hard to imagine that all of the things I might care to know could be adequately represented in a set of numeric variables and text fields. Just glancing through the print documentation for the 1990 Census reveals detailed information about the geographic hierarchy, formulae for calculating standard errors, a discussion of data editing, and
more. However, it is clear that if none of this information is available in fields that can be accessed and manipulated under program control, I will get none of what I might need to know.

To the extent that user-friendly interfaces increase the amount of secondary analysis performed, we will have more and more naive users. I expect that this will increase the need for narrative material. Consider, if you will, whether or not a new user of Summary Tape File 3A on CD-ROM from the US Census Bureau would even know that they were working with sample data. The fact that nearly anyone can use the GO software means that the need for general information about a study is greater than it was in years past.

The introduction of macro languages such as SAS and SPSS meant that the researcher did not need to be expert enough with FORTRAN to string together subroutines from the IMSL library. However, it also meant that the volume of statistical analyses skyrocketed. The strides that have been made in the last 30 years in terms of improving access to and documentation for secondary data will have to be matched over the next decade if the promise of easy access to secondary data is to be fulfilled.

Plan Analyses
In planning his or her work, the secondary analyst needs to know what variables are in the file and have access to the alphanumeric strings that describe the variables and, in the case of categorical data, the values of each variable. Missing value codes are critically important and frequency distributions are often useful. From a functional perspective, most of these needs have been fulfilled by printed data dictionaries, at least for most studies. Because these needs have been fairly well defined, most of the user-interfaces for data products on CD-ROM handle these tasks about as well as the printed data dictionaries, which is to say they provide the minimal amount of assistance possible.

The traditional limitations of most software packages in handling this information has been criticized by Grant Blank in a paper delivered at the 1992 Computing in the Social Science conference, and quite appropriately so. We may go even further in our criticism by noting some of the features user interfaces could include that would provide the analyst with something more than an electronic copy of a printed document.

- A hot key could pull up a window displaying more detailed information about the variable, or usage notes (the CD/ROM for STF3 will do some of this)

- A parent/child function could allow the analyst to see the branching question that controls whether or not the current question was asked and the questions skipped if the current question is answered in a certain way.

- Variable selection could be facilitated by automatically identifying variables that are critically important for file matching operations or weighting operations. Going beyond the capabilities that are built into the Census Bureau's EXTRACT software, I can envision smart interfaces that prompt you to consider certain variables if specific others have been selected.

- As survey researchers become more sophisticated with question-wording or context experiments, we need for the interface to reproduce different questionnaire versions.

Even among the best interfaces for CD-ROM data products, we see an unfortunate tendency to reproduce the paper documentation on screen instead of providing a tool tailored to interacting with the metadata.

Conduct Analyses
Once the analyst has planned his or her analyses, metadata can be exploited via user interfaces to facilitate analysis. Some excellent examples of interfaces that work well in this respect include the EXTRACT software for use with US Census Bureau products and the interface for High School and Beyond, from the National Center for Education Statistics. Each of these products allows one to select variables and output data files or documentation, or both. The interface for High School and Beyond will produce an SPSS command file for accessing those variables that have been picked by the analyst. Because I view interfaces as doing a better job in this respect than in others, I will only enumerate the key functions the analyst might need:

- Variable subsetting, or stripping the file down to include only selected variables.
- Production of a software command file to facilitate accessing raw data files, written in the macro language of the user's choice.

- Production of an output data set, with case selection capability, including random subsetting for analyses that will employ cross-validation. Ideally the analyst could have several choices for the format of the file.

- Production of an output data dictionary to document the reduced data set.

- Help in rectangularizing data from hierarchical files or restructuring time-oriented data structures.

I purposefully do not include analytic functions in this list for two reasons. First, I believe strongly that users should be encouraged to work in a statistical package with which they can develop some expertise over time. Second, the table-generating capabilities of interfaces such as that I have heard described for use in conjunction with the 1990 Public Use Microdata Sample files from the US Census Bureau strike me as terribly limited in terms of their capabilities.

Resolve Analytic Problems
Anomalous, unexpected or possibly incorrect results will almost always arise in the course of performing a secondary analysis. Many of the resources the researcher requires to understand these potentially problematic results are the same ones the researcher requires to screen studies and plan analyses. Often, these problems are subtle and require the ability to examine questionnaires, hear instructions, or examine individual records. Some of these functions will require multimedia capability to display facsimiles of documents or play sound recordings. However, the major source of assistance with these types of problems is the same data that we need to screen studies.

The Problems
Earlier I alluded to three problems that must be addressed by data organizations. The problems may be summarized as follows:

- Metadata will range from complete and essentially perfect for large scale studies designed for secondary analysis to incomplete and imperfect for small studies that are archived with little thought given to their use as secondary data resources.

- As secondary data resources become easier to use, more naive users will avail themselves of the opportunities to perform secondary analyses, leading to increased demand for friendly, and perhaps even smart, interfaces.

- The tenuousness of the future of CD-ROM makes it incumbent upon us to ensure that our metadata can migrate from platform to platform.

All of these problems can be solved, in large part, by the same basic approach: develop software products to compile metadata and create user interfaces from the compiled data sets. What I envision is an interactive program that builds a metadata structure from three sources:

- Information entered by the researcher (or even the secondary analyst) in response to prompts (title, author, year, etc.).

- Textual information in external ASCII files that describes the study and the use of the data. Each file would have a topical theme (e.g., sampling, instruments). Some of these may be deal with standard topics and others can address topics of unique concern.

- Enhanced data dictionary files created by statistical packages such as SAS or SPSS. Indeed, the entire data dictionary portion of the compiled metadata file could be read in from an external file created by these packages if procedures such PROC CONTENTS and the DISPLAY DICTIONARY procedures were enhanced to include frequencies and better variable and value labeling.

I would also anticipate that the program would allow the researcher to review and edit the entire collection of metadata, adding variable notes, information about question flow, etc.
A key part of my proposal is that the compiling program and the associated interface that would access the metadata structure would be tolerant of any amount of missing data. Investigators could do as much as spend days (weeks) using all the options of the program, or as little as running the enhanced version of PROC CONTENTS. In either case, the output of the compiling program would be a metadata structure the user interface could access.

I will close by noting that as long as a software developer bundled the compiling program with an interface that would act on the resulting metadata structure, the community of data users would have a tool for producing interfaces in a cost-effective fashion even without standards for metadata.

1 Presented at IASSIST/FIDO '93Edinburgh, Scotland
Codebooks in the World of Networked Data Library Services

by Richard C. Rockwell
Executive Director
Inter-university Consortium for Political and Social 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 screens as 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 each codebook, 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 such as 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 to the 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 standards 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 will have 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 have to 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 put itself 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 marginalia 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.

Spring/Summer 1993
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 of data 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.

Preparation Machine Readable Codebooks at the Zentralarchiv
Actual Situation and General Perspectives

by Rolf Uher and Rolf Rontgen
Zentralarchiv fur Empirische Sozialforschung at
the University of Cologne

Abstract:
In the first part of the paper we will show, how the different parts of the original documentation at the ZA will be processed to end up in a full-text machine-readable codebook on disk and on paper. The second part presents a "checklist" of tasks, which should be considered if we think about a new - ideally IFDO-wide - production-line for codebooks. This "checklist" is also meant to be an offer for the ongoing discussion, and should be expanded by ideas from interested colleagues.

A. Introduction to the First Part
In the mid-seventies the ZA began to process the machine-readable codebooks along the OSIRIS format using most of the original OSIRIS tools. This production-line substituted a system of programs developed at the ZA in the late sixties. The "philosophy" however standing always in the center of this strategy was:

1. to have a complete documentation for a study which includes all the information which a secondary analyst might need to interpret the data.

2. to start a text base for information retrieval purposes, regarding the fact that there will be an immense growth in the amount of relevant information (here: full-text retrieval on variable-level within a pool of studies/codebooks).

The OSIRIS format might be called "old-fashioned" or "unflexible" - and it is true for some cases - but there is no alternative format for it right now, in which complete question-text, complete answer-categories, archive-comments, general introductions, notes etc. can be processed. Generally spoken the "unflexible" format has the advantage that you can transfer an OSIRIS-codebook to nearly any other format which considers the variable-structure of a flat data-file.

Until there is no other solution which - at the same time - can convince us that it really works, it is a matter of reason to apply to a reliable tool.

B. ZA Codebooks
The ZA produces (machine-readable) codebooks as a standard documentation for some ongoing projects: ALLBUS (Germany's General Social Survey), ISSP (International Social Survey Programme), German Election Studies. The discussion about the EUROBAROMETER-codebooks is not finished yet. Besides that codebooks will be produced for single studies (e.g. Wohlfahrtsurvey, Health-surveys, Youth-surveys ...) or cumulations of study-series (e.g. Politbarometer), according to capacity in the codebook-department at the ZA and according to the expected user-demands.

C. Processing a Codebook - First Step
The first step producing a codebook is to make the questionnaire machine-readable. The general tool for this step is a text-processing software on the main-frame. The structure of this "raw-codebook" is already very near the OSIRIS format.

At the same time we are running tests with scanner and OCR software. This is a promising perspective for the future, thinking of a more "machine"-supported approach in producing full-text documentation.

D. Processing a Codebook - Further Steps
The second step producing a codebook goes along with the processing and cleaning of the data. The end of this step is a data-proofed documentation, which describes the (completely processed) data-set (and which is not meant to be a "data handbook" or something else). The tools used in this second step are partly original OSIRIS programmes (FBUILD, FMRG, etc.), some are IBM utilities and software, some are from the DDA (MERGET, SLABGEN etc.), and some are ZA software (CDBK-PRT etc.). The ZA tool CDBK-PRT can cope with tasks, which OSIRIS cannot solve:
- merge SPSS CROSSTABS output into the machine-readable codebook-file (e.g. for international comparative studies, cumulative files etc.)

- merge frequencies with more than 4-digits per answercategory

- can print as well crosstables as univariate frequencies in one variable

- considers layout-parameters like bold-printing, underlining, page-formatting etc.

E. Codebook Output
The "traditional" output from a codebook (OSIRIS format or whatever) was on paper. But the amount of workload, time, and paper for producing all the codebooks requested grew and grows beyond reasonable borders. On the other side more and more users ask for machine-readable codebooks on floppy-disks. Thus the (above mentioned tool) CDBK-PRT allows to output a codebook in different ways:

- print a codebook on the mainframe line printer (for the purpose of proof reading the documentation during the production process)

- print a codebook on a laser printer using the PRESCRIBE language (for the purpose of duplicating it on a photocopy-machine internally or at copy shops outside the ZA)

- print a codebook on a laser printer using the POSTSCRIPT language (for the purpose of duplicating it on a photocopy-machine internally or at copy shops outside the ZA)

- copy the codebook output to a disk file in the POSTSCRIPT language (for external users who have access to POSTSCRIPT and want to reproduce the whole or parts of the codebook at their place)

- copy the codebook output to a disk file in ASCII-format only with carriage-control characters for page-feeds. (for external users who want to reproduce the whole or parts of the codebook using a simple printing routine or who want to import the machine-readable documentation into a text processing software.)

ZA's experiences with the distribution of codebooks on disks and with the responses from the side of the users are not yet very systematically. But the fact that people who once received a machine-readable copy start to request codebooks on disk also for other studies is a promising development.

F. Changing Profiles
One critical point must be mentioned at this place: Once we have started to distribute codebooks on disks the feedback about the number of usages of the documentation will probably be decreasing even though more people might have access to the codebooks (e.g. in PC pools) than before. This means however that we need other arguments for the legitimation of the archival work to the funding organisations.

The profiles seem to change, as well the profile of our services as the profile in the structure of the demands of our users. (The question: Which side influences the other? should not be discussed right here.) Additionally services and demands become more and more international, which means at the same time that international (inter-archival) cooperation becomes more and more important.

G. Introduction to the Second Part
The second part of this paper now tries to define tasks and demands for "the codebook" in general (processing and formats), considering "tradition" and perspective. This list is, in the present form, a rather subjective collection of items but it is meant to be a help for the ongoing discussion and an offer for additional ideas.

Collection of items for a (new) codebook production-line (Questions: What is a codebook? What is it used for?);

information on the variable level: technical description of a data-set:

in terms of a "raw-data file" (position, width, deck etc.);
in terms of an SPSS (or other) system file (Variable name, label, definition of missings, decimal places, variable type etc.);

questionnaire information (question text, answer categories, other information from the questionnaire, comments from the archive) data information (frequencies - weighted or unweighted -, crosstabs, other statistics, graphics);

less redundancy in the codebook example: a list of "dummy-variables" could be arranged in a way that redundant information is left out and only the relevant frequencies are documented in formats like: tables, graphical presentation etc. to allow readers of the codebook to look at the relevant information at the first sight;

general information about the study:

preface (study description, index, list of variables, copyright information, how to cite a codebook)

appendix (notes explaining special variables, copy of the original questionnaire)

purpose of a codebook:

information for the (secondary) analysts, who work with the data

input for information retrieval

data-handbook for people, who don't use the survey data but only the codebook???

media:

codebook on paper (in-house production, printed at an editing house, copy-shop)

lay-out (fancy looking - or - plain printing routine?)

codebook on tape/disk/other media/via file-transfer (which printing-routine, form feed characters, comparability problems)

exchange format:

codebook on tape, disk, other media, via file transfer (with the same format in each participating archive, special software for processing, formatting, and producing/reproducing output on paper, on disk etc.)

codebooks in a general concept:

codebooks (as defined above) plus data plus general background information (aggregate data, maps...) plus retrieval-system plus print and other output options plus bibliography

on CD-ROM or whatever can be thought of as a "service package" containing everything which a user might need

producing codebooks:

tools for mainframe, PC (MS/DOS, OS/2) or workstations(UNIX, AIX..) to produce the codebooks in the exchange format

New ways of entering the past: The Netherlands Historical Data Archive

by Jeroen Touwen
Netherlands Historical Data Archive Reserve
Department of History, Leiden University

Introduction
Since information technology is rapidly developing into ever newer and nicer forms and shapes, all researchers sooner or later will face the computer. Thus in the Humanities, in particular in History, the interest in using the computer is still spreading and new techniques are applied in analysis, retrieval and storage of scientific data. To encourage and assist these efforts of historians in the Netherlands, a national data archive was established. The main aim of this data archive is to catalogue and archive machine-readable data. A second goal is to spread new techniques and fulfill the function of an intermediary in the world of information. This article introduces the activities and recent developments of the Netherlands Historical Data Archive (NHDA).  

It takes three wings to fly
The NHDA is a young historical data archive. Since the first steps towards her establishment in 1987, the NHDA developed into an institution with three main sections: data documentation, scanning and education. These three sections direct their attention to six major activities:

Data documentation
(1) documenting, archiving and disseminating historical data,
(2) documentation center with services on data location and historical computing

Scanning and OCR
(3) providing a Scan/OCR Laboratory,
(4) carrying out digitalization projects,
(5) offering courses in scanning,

Education
(6) courses in Historical Information Science for Ph.D.-students, staff and in a post-doctoral one-year-course.

The data archiving activities are carried out by the data documentation section, but the scanning and OCR forms a separate branch of the archive. In the Scan/OCR Laboratory researchers can use advanced equipment and software to scan and read old or damaged documents. This section is an addition to the back-bone activities of data archiving of the archive. The Scan-OCR laboratory also carries out external projects. Various institutions in the Netherlands, like the International Institute of Social History, have given assignments to the NHDA to make collections of historical material or inventories electronically available. We will focus our attention on several of these activities.

Historical Data Archiving
Archiving historical machine-readable data requires extra attention, and it is from this point of view that data archiving will be treated here.

A survey, conducted in 1989, reported on 131 historical data sets developed or used by historians in the Netherlands. Compared to a previous survey, held in 1987, that listed 81 data sets, this shows remarkable growth. It is obvious that in the meantime this figure will have grown considerably.

This shows that historians increasingly use the computer for organizing, retrieving and analyzing their material. To provide them with the opportunity of using existing collections of machine-readable data, in many countries historical data archives have been established, in analogy with social science data archives.

Furthermore, the need was felt for a coordinating institution, to serve as an intermediary between the users and data hosts, disseminating information to historians on specific data banks and information services and cataloguing special data collections, software, list-servers, and newsletters of related institutions. This type of 'information-brokerage' is a new but essential development in a world that is, due to new information technologies, to an ever larger extent providing giant masses of information to virtually every place on earth. Providing information on data banks, software and techniques to historians in the field is useful. An example is the coordination of what will become a national infrastructure of historical data banks in the Netherlands. This is one of the ambitious projects that are just underway. The purpose is to integrate three major Dutch data banks.
into one 'Dutch History Machine'. In this project, the NHDA only plays a small role. At several universities, historians that have been working on large data bases will put effort in this project. I am positive that in a few more years, universities all over the world will be able to access the Dutch History Machine through the Internet.

Of course, we think it is also important to collect and archive electronic data in the "old-fashioned" way (by storing them on our computer system), in the same way traditional paper archives will ever remain unavoidable.

Thus in 1987-1989 plans were made to establish a historical data archive in the Netherlands, that could exist independently next to (but still get organisational advice and support from) the Dutch social science data archive, the Steinmetz Archive in Amsterdam.

After a pilot-project in 1989-1990, in 1991 the NHDA was acknowledged as a national expertise centre, a government supported institution that serves a national interest. To quote the IAASSIST QUARTERLY of winter 1991, "in the Netherlands, a historical data archive is on the steps, however, funding is lacking". Since then, in several government financed projects the NHDA has been able to expand and develop necessary skills and specializations. Some of the projects, that were carried out (in the scanning and education departments) even earned back some money. Eight people are working with the NHDA on a part-time basis at the moment, and with advanced computer equipment both storage and access of data material can be effortlessly realised. Recently the NHDA attracted its first full-fledged data documentalist.

**Historical Data Set Description Scheme**

One of the first activities of the NHDA has been to develop a standard for cataloguing historical machine-readable data. This led to the Historical Data Set Description Scheme (HDDS): an adaptation of the Standard Study Description Scheme (SSDS) which is used in several social science data archives. Historical data sets have distinct features that need specific attention when archiving and cataloguing. Especially the historical sources on which the data sets are based, need to be documented carefully, since they often take a central position in the point of view of the researcher that collected the data.

The historical source material on which the data set is based, preferably needs to be documented *per file* (see figure 1). We regard a data set as consisting of one or more files, each of which is based on one or more sources. (Thus we tend to speak of a data set as the primary unit of cataloguing, cf. 'study' in social science data archives, see below.)

In a series of international workshops progress is made to reach an international standard for the description of

---

**Figure 1: The Historical Data set Description Scheme (HDDS) as used by the NHDA.**

Data model of the Historical Data-set Description Scheme.

```
          PI
         /   \
        CT   DS
         \   /
          PU  SO
          |   /
          FI  VA
          |   /
          CO

Explanation:

<table>
<thead>
<tr>
<th>Code</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>DS</td>
<td>Data set</td>
</tr>
<tr>
<td>FI</td>
<td>File</td>
</tr>
<tr>
<td>VA</td>
<td>Variable</td>
</tr>
<tr>
<td>CO</td>
<td>Code</td>
</tr>
<tr>
<td>CT</td>
<td>Contacts</td>
</tr>
<tr>
<td>PU</td>
<td>Publication</td>
</tr>
<tr>
<td>SO</td>
<td>Source</td>
</tr>
<tr>
<td>FI</td>
<td>Keyword</td>
</tr>
<tr>
<td>VA</td>
<td>Geographical area</td>
</tr>
<tr>
<td>CO</td>
<td>PI Person/Institute</td>
</tr>
</tbody>
</table>
```

---

*Spring/Summer 1993*
historical data sets.

To give an impression of the HDSS, we present its datastructure below. As the HDSS closely resembles the SSDS used in other archives, the total number of 97 descriptive fields (of which some have a repetitive character and need to be filled out more than once) will not surprise data archivists. When all these fields have been specified, the data set is catalogued in a complete and detailed way.

Are historians different from social scientists?
At certain points the HDSS is more extensive than the SSDS: such is for example the case in the section on sources. Other parts, such as the methods of sampling or questions dealing with a survey, get less attention in this scheme. For historical data sets one can consider the historical source as a starting point and the data file as a unit of description.

In many social science data archives, the data are archived by STUDY: a study can consist of several data files, but the documentation is stored on the level of the study. Typical historical data sets, as constructed by scholars conducting historical research in archives, usually do not originate from a survey or sample, but from the historical source. We found that many data sets concerning a specific problem or research project consist of closely related files each representing a specific selection or section of a source. Also, several sources can have been combined in a file, especially when this file is generated by merging data in a later stage. Therefore, sources will be catalogued on the file level, but documenting a source one can refer to previously described sources in order to avoid double work.

Since we noticed before that historians tend to dislike ("unknown makes unloved") the language used in social science computing (or any computing, for that matter), we try to have plain language explanations at each field.

On the other hand, historians can learn a lot from what has been achieved in the social sciences. Noteworthy is that especially the social and economic historians, who are most related to the social sciences, use the computer a lot.

New media
One of the other issues we have been working on last year is how we can document and catalogue the various new media that are nowadays on the market. 'New media' refers to CD-ROMs, multi-media data collections, hypertext databanks (or hypermedia) and network sites or data hosts (listservers, discussion-lists, on-line data banks, etc). Strictly speaking, these are not all new media but merely new organisations of data. It does not matter whether the medium is magnetic tape or disk, or optical disk, or a remote data institution, although in the last case the physical storage of the data is different (it is someone else’s problem...).

- Multi-media applications
- Thematic CD-ROM collections including user interface
- Serial CD-ROM publications including user interface
- Hypertext systems
- Relational databases
- Document servers (FTP-sites and listservers, on-line databanks)
- Electronic discussion lists
- Text corpora (tagged or indexed)

In the historical field, not all of these new media are fully applied. We acquired some historical CD-ROM data collections. There are several network discussion groups that a historical data archive wants to keep track of or point out to its clients (HUMBUL, History, Humanist, IASSIST, and on various subdisciplinary themes). In the workshop this January, we exchanged ideas on this subject with the historical branches of the Essex ESRC data archive (Colchester) and the Danish Data Archive (Odense, Danmark). I will state a few preliminary conclusions that are based on these sessions.

In the case of certain media, like multi-media, hypertext applications or certain CD-ROMs, the data cannot be separated from the application, when archived or documented. In most data archives, as a principle both an original copy of the data and a copy in ASCII are kept. The application in which the historical source data is stored is part of the multimedia or hypertext data collections. One cannot extract the data from a hypertext system without essentially changing the data, or losing essential information. It is not worth the trouble to study all different source data that can be stored on a CD-ROM, when that CD-ROM offers a especially designed user interface to query and retrieve these data or their source-description. Thus, we could say, it documents itself.

For example, on the CD-ROM 'Dutch Printers Devices', a large collection of printers’ book-images, dating from the sixteenth and seventeenth century, are stored both as an image and in transcriptions on the compact disk (Koninklijke Bibliotheek, Den Haag, 1991). They are classified in an elegant special application (Iconclass Browser) that is based on the ICONCLASS classification system, an extensive classification scheme which gives hierarchical codes to art historical images. This CD-ROM
can be catalogued as a thematic data collection, thus combining media and data as one entry. At this point there is no use in extracting the data, converting them to a SAS, SPSS or OSIRIS format or archiving them in ASCII. For many of the media-dependent forms of information, like network sites, the same thing can be said. Thus many types of modern information collections in machine-readable form cannot be regarded as collections of raw data files, but on the other hand, to a large extent document themselves.

On the archiving of text-corpora, finally, the guidelines of the Text Encoding Initiative (TEI) on the use of the Standard Generalized Mark-up Language (SGML) should be adequate in supplying the essential information within the text corpus itself. Historical text-bases can contain all source-descriptions under the heading of "Citations", according to the SGML Guidelines.14

Referring to the distinction between a study description versus a data set description, one can state that data collections which are application-dependent can be regarded as one study. One could even say: a published data collection can be regarded and even catalogued as a book. Naturally the relevant question is not whether we store these data in a separate book catalogue, or in a selection of fields in the large data catalogue. However, we think it is absolutely vital to specify standards of doing this.

Do-It-Yourself
Another activity of the NHDA data documentation section that we can briefly introduce is our effort to develop a Do-It-Yourself module for researchers’ data cataloguing. In our view, researchers can be persuaded to document their own data when there is something to be gained for them. Many archives distribute long questionnaires in an effort to get as much documentation as possible from the depositor. Often the technical documentation that comes with a data set forms a bottle-neck in the cataloguing process, especially in the historical discipline, when in the seventies and early eighties the computer was used only scarcely.15

Traditionally, historians studied their sources with a pen and a piece of paper. Of course, this is still to a large extent the case. Many materials cannot be analyzed quantitatively. Thus, the phase of computer analysis is often only a small part of the project and the value of the documentation of materials and methods is often underestimated. Details of data editing procedures and codebook modifications have been lost all too often. In other cases, a lot of interpretation and standardisation was necessary before computerised analysis was possible. Especially in these cases, the documentation is needed to estimate the scientific value of the data.

A joke often heard in data archives is that if the researcher does not supply adequate documentation, the data sets will be stored in a special data bank: a dusty box under the table of the data documentalist. These possibly valuable or even unique data are lost for secondary analysis, neither can conclusions of these studies be checked or corrected by other researchers! In recent conferences and workshops attention has been paid to the side of the researcher in data documentation. The requirements for a DIY-routine can be seen in relation to this problem.16

Short output in standard form (2) can also be used as an appendix in publications, to state the data to which the publication refers. As H.J. Marker says in one of his earlier articles, a standard description of quoted data is useful. A personal organizer would help achieving these

---

**Figure 2: Do-it-Yourself program: Gains and requirements**

<table>
<thead>
<tr>
<th>Requirements program:</th>
<th>Requirements User:</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) easy to use</td>
<td>(1) A Personal Computer</td>
</tr>
<tr>
<td>(2) On diskette (runtime)</td>
<td>(2) Documentation of the data set</td>
</tr>
<tr>
<td>(3) Intelligent/stubborn questions</td>
<td>(3) A will to cooperate</td>
</tr>
<tr>
<td>(4) Surveyable results</td>
<td>(4) Time to collect and type</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gains User:</th>
<th>Gains Archive:</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Record on disk or paper of own data holdings</td>
<td>(1) Efficient supply of cataloguing information</td>
</tr>
<tr>
<td>(2) Short and long output in organized format</td>
<td>(2) Saves time and money</td>
</tr>
<tr>
<td>(3) Help in standardization of new projects</td>
<td>(3) Info for converting or restructuring</td>
</tr>
<tr>
<td>(4) Data archive service: data conversation</td>
<td>(4) Consciousness raising effects</td>
</tr>
</tbody>
</table>
standards by spreading a standard way of describing data materials for researchers.\textsuperscript{17}

In figure 2 we show the requirements for a Do-It-Yourself module for data cataloguing by researchers, and the profits of using such a system for both the researchers and the data archive. The gains for the archive of this kind of tool are obvious and hardly need to be mentioned. An efficient supply of cataloguing information for the data documentation is achieved, which saves time and money. However, the output of the DIY-program should be checked and completed. This can be done by the archive employee within the application, before it is appended to the main catalogue. Information for converting or restructuring the data, when this should be done, is now more readily at hand. And finally, not the least profit from this procedure, a nice and friendly documentation program will have consciousness-raising effects in the discipline. The whole aim of distributing guidelines, schemes and data catalogues to researchers is to convince them of the use of data documentation and secondary analysis of machine-readable data. The DIY-program will contribute to this goal.

Requirements versus gains
The first requirements of the program are obvious: the program should be easy to use and ought to be distributed on a diskette, preferably in a runtime version without other confusing software programs bothering the user. One puts the disk in the slot, types 'Install' and the program installs itself, etcetera. The third requirement (see figure 2) is essential to get the researcher to follow the rather complicated structure of the HDDS. The program should be intelligently structured and ought to lead the user through questions and subsections. It should be stubborn at certain spots, to persuade the user to fill out sections that he or she is inclined to skip. This can be done by showing reminders after the last question is completed or by blinking in striking colours when the user wants to skip sections dealing with important information (such as titles, software formats, and source-evaluation).

Of course the user needs to have access to a PC. This should not be a problem in 1993. Next the researcher must have access to the background documentation on the data set. Considering that the user is the research initiator, we of course hope that this requirement is met. But the program can also be sent around to different persons involved, who then each add their piece to the documentation routine. Consulting, adapting or extending the documentation can easily be done.\textsuperscript{18} Finally the will to cooperate and, last but not least, a certain amount of time are essential requirements. Of course these are crucial parameters, but, apart from making the program as efficient as possible, beyond our control.

We hope that depositors appreciate and value an organised record on disk or paper of their own data holdings. We offer this in short and long output.

\textbf{Figure 3: Sample screen of DOCIT!, the test version of the Do-It-Yourself data documentation module under development at the NHDA}

\begin{center}
\begin{table}
\begin{tabular}{|l|l|}
\hline
1 Original title [MEMO]: & Scheepsreizen naar de Oost \\
2 English title [MEMO]: & Shipping to the Indies 1700–1900 \\
3 Number of files: & 3 \\
4 Language: & 5 \\
5 Abstract / Summary [MEMO]: & This data set contains 24393 shipjourne \\
6 Rationale for the dataset [MEMO]: & Scientific \\
7 Striking characteristics [MEMO]: & Data are collected from 7 Dutch harbour \\
8 Accessibility: & 2 \\
9 Depositor requirements: & \\
10 A & 1 No restrictions \\
& 2 No restrictions to scientific use \\
& 3 Consultation with depositor required before use of data is advised \\
& 4 Written permission of depositor required for use of data \\
& 5 Special arrangements to be made with depositor \\
& 7 Other conditions \\
\hline
\end{tabular}
\end{table}
\end{center}

Press ENTER after completing this field

\textsuperscript{17} Rationale Language: English
\textsuperscript{18}
In the long run, one can expect that standards for quoting data materials will be reached and perhaps even that the publication of source data sets gets academic recognition. The short output in standard form can be used as an appendix in publications, to list the sources and data sets on which the publication is based. A personal organizer in this sense would help achieving this. Also guidelines on documentation could be distributed in this way.

While documenting, one is unpurposefully reviewing the data documentation and evaluating the experiences in standardisation, for example in the coding of variables. These insights will undoubtedly help the researcher in future projects. However, we should not overestimate the value of these benefits to the researcher: an experienced computer user will shrug his shoulders at this suggestion. Another aspect is that one can more easily refer to previous decisions on coding and standardisation when these are easily accessible.

The use of the data archive services, for keeping and preserving the data, converting them to new formats or media, protecting them from fires and floods, is of course a good reason to deposit data at the archive. The archive may require documentation before accepting the data, using the DIY program to ease the pain. Thus, these are all gains that either the user will realize himself as valuable, or that we think will be good for him or her.

The Scan/OCR Laboratory
In these times, when the cost of labour is quickly outgrowing the price of technology, one can consider using optical scanners and Optical Character Recognition software (that 'reads' the images and transforms them into computer data) to enter data into the computer. It stands to reason that in the humanities, with its often text-based research, a lot of interest exists in these techniques. One scans a book or a set of documents, which is like making a photo-copy, and sits back to have the computer translate the whole text into an ASCII or word-processor textfile. An international workshop, held at the NHDA in June 1993 and organized in cooperation with the Nijmegen Institute for Cognition and Information (NICI), gathered researchers from all over the world to exchange ideas on the application of these techniques.  

In the Scan–OCR–Laboratory the latest scanning equipment and Optical Character Recognition programs are applied to historical documents. Both historical source material and old bibliographies and archival inventories are scanned. Different Scanning and OCR software packages are on the market, applying several algorithms to recognize printed text. The Scan–OCR Laboratory specializes in data-entry of old, damaged materials: choosing from a range of different OCR-packages and applying special wordprocessor-macros, it tries to find a way to handle the distinct flaws of a document (like blotches, irregular type-setting, bad edges, complicated formats). In addition, the issue is studied how the material, when digitized, can be organized in such a way that searching and retrieving is optimally feasible.

The Scan/OCR Laboratory has conducted projects for the International Institute of Social History (IISH), the Institute for Netherlands History (ING) and the Netherlands State Institute for War Documentation (ROID).20 We will briefly state the conclusions of these projects.

The International Institute of Social History (IISH: Internationaal Instituut voor Sociale Geschiedenis) has a large collection of archival material and books on the national and international Socialist movements and social history. In the project carried out in commission of the IISH, three different types of data materials were tested for scanning and digitizing.

(a) The Biographical Dictionary of Socialism and the Workers' Movement in the Netherlands (P.J. Moertens, ed.) contains short biographies. Part I (of five volumes) was made machine-readable as a test project. The Scan/OCR Laboratory developed a small hypertext system in Freebase with these data.

(b) The Bibliografie der Social-Politik, by Josef Stammhammer, Part I (of two volumes) was scanned. The IISH has thirty meters' worth of this kind of bibliography. An interactive search-utility to make these accessible would be an aim well worth to pursue! The items were scanned and read with ProLector and the resulting ASCII-file was tagged for author-fields, title-fields, etc. Thus each item was structured and the data could be converted into a data-base. A very complicated WordPerfect macro was even used to separate the lemma's!

(c) Tests were carried out with the scanning of newspapers: both the originals and newspapers on microfilm were tried with several OCR-software programs. For several of these tests, the NHDA did not have the right equipment (e.g. 'A0'-scanners that can scan complete newspaper pages). In these cases digitizing was tested at commercial companies.

Formats, non-Latin fonts, vulnerability and printing quality were all evaluated in these tests. In a general conclusion, we can state that the use of scanning techniques, including correction, turned out to be faster than keying by hand in all instances. Only the digitizing of newspapers turned out to show many problematic aspects of the material. Although in principle the tech-
Scanning and OCR-equipment cannot yet be used as a photo-copying machine. It takes expertise and special treatment of each collection of source materials to scan them. However, even in comparison with manual data keying in low-wage countries, automatic document entry is cheaper. Software and equipment have to be carefully selected, and a break-even point should be calculated after testing a selection of the material, to determine what quantity can be scanned for a certain price. In the near future, the developments in scanning and reading of micro-film are closely watched. This will open another wide area of heuristic and archival material to be entered into the computer, which, as for now, cannot yet be done.  

**Data services and education**

The last major goal of the NHDA that we will mention here, after data documentation and scanning, is “organizing courses on historical computing”. Even if the “trend” in historiography in the early nineties seems to be towards a narrative approach of the discipline, many historians are fascinated by both computer techniques in data analysis and the computer as a heuristic tool. There are many scholars and students that want to learn various skills that open new possibilities for their research.

A successful initiative has thus been the organization of a one-year postgraduate course on Historical Information Processing, the Data Bank of Urban and Regional History (DABURH), now in its third year. In this full-time programme (including a six week stay in Leicester University and a two month external traineeship), sixteen graduate students learn how apply the computer in Humanities research: in analysis, reports and presentations. In addition the NHDA organizes short courses on historical computing for Ph.D.-students.

**Conclusion**

In a brief overview I have tried to highlight the main activities in 1992-1993 of the Netherlands Historical Data Archive. Both on the subject of data documentation and on the subject of data-entry (scanning and OCR), the NHDA has been conducting research in a number of projects. This approach creates the possibility to start cataloguing and documenting Dutch historical data sets on one hand, and providing services and expertise on modern information techniques on the other. Thus, in addition to the many ‘history machines’ that we already have, the NHDA hopes to offer ‘new ways into the past.’

**Annotation**

1. This paper was originally presented at the IASSIST/IFDO Conference in Edinburgh, 25-27 april 1993, at the session ‘Data libraries - the new and the reformed: delivering data for secondary analysis’. 

Techniques could be used, integrated processing of these was not yet feasible.  

In the tests conducted for the Netherlands State Institute for War Documentation (RIOD: Rijksinstituut voor Oorlogsdocumentatie), the emphasis was on the quantity of the material that had to be digitized. The NHDA did tests to consider the efficiency of converting inventory lists of the DOC-II archive. The RIOD has 5.000 typed pages of these lists, of which some are stencilled, others are photocopies or laserprints. These lists enable the retrieval of circa 20.000 archival documents. A selection of the documents was scanned with both ProLector and the Kurzweil 5200 system. Break-even points were calculated for the use of both systems, including the tagging of the data, which had to be done with macros on the bases of page lay-out and type-fonts. The text-files had to be corrected by hand, which was very time-consuming. Also the capacity of the sheet-feeder of the scanner turned out to be a bottleneck in the conversion process. In the end product, field indicators (tags) specified the number of the box, of the folder and the number of the document itself, as well as the description of the document. The RIOD will develop the retrieval interface themselves and merely wanted to know if scanning was feasible. However, it should be noted that after a test-pilot, one never can predict the real time of processing the whole collection.

For the Institute for Netherlands History (ING: Instituut voor Nederlandse Geschiedenis) the NHDA scanned a selection of the serial publication 'Repertorium van Tijdschriften en Artikelen betreffende de Geschiedenis der Nederlanden' (Repertorium of magazines and articles on the History of the Netherlands). This heuristic tool is one of the most frequently used bibliographies in historical research in the Netherlands. It consists of 27 volumes, published over the years 1940-1988, in which over 115.000 publications are catalogued, not regarding the older volumes published between 1863 and 1940. Since fonts, text lay-out and page formats changed over time, a series of different tests had to be carried out. Only the volumes that appeared after 1987 have been composed with the use of a data-base program, so for these last two years a simple computer conversion can be carried out.

In this project the alternatives for a retrieval tool were also studied. An on-line catalogue or a CD-ROM publication are possibilities, but it should be kept in mind that before the actual data entry takes place, decisions on retrieval and search-interfaces should be made. For example, when in the final product the distinction between books and articles should be available, this distinction should in one way or another be noticed by the OCR-program and stored in the data-base.


8. In Leiden (January 1993) and Odense (June 1993) workshops were organised in which participated the Danish Data Archives, the ESRC Data Archive, the Netherlands Historical Data Archive and delegates of Russian data archives in Moscow (June 1993). These workshops resulted in a special workshop on Historical Data Archiving at the VIIIth International Conference on History and Computing, Graz, 24-27 September 1993. These papers was published in a preliminary workshop-volume: H.J. Marker (ed.), Historical data archives on common ground, draft version, Danish Data Archive (Odense 1993). An official publication is being prepared.


10. On the percentages of historical sub-disciplines in the use of historical data sets, see N.S. van Hall en W. van Tellingen, Nederlands Historisch Data Archief III: Historische Databestanden 1990 (Amsterdam, 1990), p.11. Of the 131 historical data sets in described this volume, 82 have a social-economic component. On British and other European data sets, one can refer to K. Schurer and S.J. Anderson (ed.), A guide to historical data files held in machine readable form (London 1992) pp. 1-8. This guide contains brief descriptions of more than 376 British historical 'studies' (of which 114 are held at the ESRC Data Archive) and another 350 data sets from Austria, Danmark, Germany, Norway, the Netherlands, Hungary, Sweden and the United States.

11. The authors note a striking range and variety in the transcribed source material, but notice a main body of census records and parish registers. They do not specify a main thematic aspect, but do notice the special historians' preference to mixed alpha/numeric files, in contrast with social scientists, who are often engaged in the purely statistical analysis of social survey data.

12. See also H. Tjalsma's contribution, 'Describing "other" data material', in the forthcoming workshop-volume: H.J. Marker (ed.), Historical data archives on common ground, draft version, Danish Data Archive (Odense 1993).

13. For example, the NHDA acquired the Spanish archive of incunabular 'Admyte' on CD-ROM, and the Dutch CD-ROM "Dutch Printers Devices", which contains pictures and descriptions of historical printers' symbols. Other thematic historical CD-ROM collections are CETEDOC (early christian latin texts) and Patrologia Latina (more latin texts), Perseus (pictures and data on the Antiquity) and the forthcoming CR-ROM of the Medieval and Early Modern Data Bank and the Research Libraries Group (containing medieval and early modern Western European price series and exchange rates of different contemporary currencies).


16. For example, the IASSIST '93 Conference workshop on 'Creating Documentation Guidelines for Data procedures' convened by S. Dodd and B. Winstanly talked extensively on writing and disseminating new guidelines.


Managing metadata: issues and approaches

by Joanne Lamb
Centre for Educational Sociology
University of Edinburgh

Introduction
Over the last few years there has been an increasing interest in metadata and the role it plays, both for cataloguing purposes and to give secondary analysts better understanding of the data they are using. The terms 'metadata' and 'metainformation' are currently used in a variety of contexts. This paper examines some of these contexts and discusses the kinds of metainformation that are relevant to different kinds of data and to the different uses to which that data might be put. Drawing on experience gained in examining data gathered by social science surveys (micro-data) and aggregated data produced by official statisticians (macro-data), it discusses the issues involved in the initial capture and maintenance of these various types of metadata. Particular attention is given to ways of using metadata to inform cataloguing systems and on-line information, and to the interfaces between various metadata holdings. The paper concludes by considering the impact of a more rigorous demand for metainformation on the data capture process.

Background
The Centre for Educational Sociology (CES) is a Research Centre of the UK Economic and Social Research Council, situated in the University of Edinburgh. We have been collecting survey data since the 1970s and have been using metadata in survey processing since the mid 1980s. The initial impetus to develop the use of metadata came from a need to rationalise the documentation of a series of related large and complex surveys collected between 1976 and 1985. In 1990 we became members of a European Commission shared-cost project, with partners from Belgium, England, Luxembourg, and Spain, to construct an interface to statistical information.

Metadata and the role it plays
We first need to determine what metadata and metainformation are. At a superficial level, metainformation is 'information about information', and metadata is 'data about data'. This is not as frivolous an explanation as it may seem. Information about information can be seen as a never ending hierarchy, a point that will be discussed later. The METIS User guide\(^2\) has a diagram which succinctly illustrates the relationships between these concepts (Figure 1).

However, we need to be more precise than this if we are to consider the use of metadata. To my mind the basic characteristics of metadata are as follows: it describes data that exists either physically or conceptually; it is

![Figure 1: Diagram illustrating the relationships between metadata and metainformation.](image-url)
stored in a computerised form; and it is relevant to a particular task, providing aid in either the processing or the understanding of data. This is a minimalist definition that is deliberately all embracing, but the emphasis is on computerised data and on the fact that the metadata is either used by a computing system, or presented to a user to help in the use or interpretation of the data.

Having accepted that we want to use this metadata, we come to two further questions. How can we make it useful, and how do we represent it in computerised form in order to make it useful?

**Contexts for metadata**

Before answering these questions, we want to look at contexts for metadata. By context we mean the circumstances under which the metadata is to be considered, i.e. its relevance to a particular task, providing aid in the processing of some data, or the understanding of some data. To illustrate this, we will look at some particular examples of the use of metadata and then identify some categories that are used.

**Examples of metadata**

This section is composed of a list of some users of metadata, together with a short summary of that usage. The list is incomplete, and the examples have been chosen to show the range of applications using or discussing metadata.

A. The ESRC Data Archive has a remit to make economic and social data accessible for analysis, and therefore need to describe data at a macro level. Their Bibliographic Information Retrieval Online (BIRON) system allows the user to search for studies in terms of subject matter, and they are also interested in providing more detailed metadata for the datasets that they distribute for teaching and research.

B. The ESRC Research Centre on Micro-Social Change in Britain at the University of Essex conducts the British Household Panel Study which is an annual survey in which all adults living in a representative sample of British households are interviewed. Part of the remit of the Centre is to promote the use of panel data, and to this end they have designed an on-line documentation system for a complex longitudinal survey.

C. Statistics Canada have been using metadata to drive their system for distributing files and software to customers on CD. This is one of the best examples of Government Statistical Offices taking a marketing approach to delivering usable, understandable data to their users.

D. In 1987, in an ESRC funded project, the CES constructed a 'documentation database' (DocDb). The focus of this project was on tracking the way that questions and associated variables had changed over time in a number of surveys which needed to be combined for analysing trends.

E. The EISI project, was funded by Eurostat under the DOSES initiative. The object is to help official statisticians use unfamiliar data. We approached the problem from the users' view, and produced an online guide to the existing paper documentation produced by official statistics offices. A further development could be to access the statistical data itself.

G. Because of their special needs, Geographic Information Systems (GIS) and their metadata have been the subject of considerable study. A conference on Metadata in the Geosciences was held at the University of Loughborough in December 1990.

L. The Edinburgh University Data Library is also concerned with macro metadata. A part of the University Computing Service, the Data Library has strong links with the Library and has played an active role in the development of catalogue standards for computer files.

M. The METIS user guide is a UN publication providing a formal definition of metadata for statistical information. It was produced in 1989 by the Statistical Computing Project for the United Nations Development Programme and the Economic Commission for Europe. The aim of the METIS group was to work out procedures for describing existing data within statistical information systems, to develop a tool for the users of the system, and to develop a tool to serve the needs of statistical information management systems.

O. There is also interest in metadata in the medical profession. An article in the Journal of Occupational Medicine in December 1991 outlines, among other things, good practice for archiving epidemiology studies.

S. Metadata is also used in survey processing, in all steps from questionnaire design to documentation and archiving. In CES we have identified metadata that is necessary during the survey processing as well as metadata that is used to describe the resulting datasets.

**A typography of metadata**

The examples listed above have been analysed and their use of metadata has been classified into various categories which are shown in Figure 2. The categories are listed below, with identifying letters showing which example institutions used each category. The categories are listed in descending order of frequency of occurrence.
This table a number of questions. There is a broad consensus on 7 of the 30 items listed. However, there is also a long tail of about half of the items which have only one or two mentions. We have to ask ourselves why this is the case. Are these items solely of specialist need or is it the case that they are more difficult to capture? Are these concepts that have been identified in theory, or have they been put to practical use in a working system? There is now a history of some twenty years of metadata usage, and more recent publications have begun to look at a theoretical approach. Svein Nordbotten 13, Bo Sundgren 14, David Hand 15 and K. A. Fröschl 16 have all recently written papers which concentrate on a conceptual approach rather than a pragmatic one.

When is metainformation relevant?
Having put the study of metadata in context, we return to the practical questions facing practitioners, and ask what relevance this metadata has to them. In response to the question ‘Who needs metadata?’ I would argue that all users of data do: the producers, the users and the ‘brokers’, i.e. the archivists and librarians. However, different kinds of users have different needs. We also have to ask what metadata is needed, what its function is, and how the metadata is obtained. I would submit that its function is threefold, to aid documentation, to improve the production of data, and to inform all users of the information relevant to their particular task.

Producing the data
In the following section, I am concentrating particularly on survey data, and have identified four groups or sections who contribute to producing final data. At each stage the data may be available to end users (i.e. people not involved in the production process and who therefore know nothing about the data). Each of the groups contribute to the metadata by supplying some information about the data that is useful to the end user.

Figure 3 represents this process. The data is processed by one section after another, and the metainformation about the data is extended by each section. The designer conceptualises the survey and causes the raw data to be collected. The IT department computerises the data at the micro level. The statistician aggregates the data at the macro level, and the analyst defines printed tables.

Figure 4 represents the information about the data.
gathered at the micro level. The first group is identified by the Designer. This person or team identifies the research question, designs a suitable instrument for collecting relevant data, and has knowledge of the background to the problem. The contribution can be summed up as sample design, questionnaire design and purpose of the survey.

The second group is identified as the IT department, but includes the administration of the survey, coding and data capture as well as design and identification of datasets. This group provides information on practical aspects of the survey process, response rates, coding notes, anomalies discovered, type of analysis package, codebook information, availability and whereabouts of data. This information is summarised as codebook, physical description of the data set and technical report.

Figure 4 represents the information about the data gathered at the macro level. The group represents by the statistician aggregates the data, defines indicators and derived variables and puts it in a form suitable for analysis at the macro level. This activity includes the creation of time series and the merging of several surveys.

The analyst defines what information should be made available in published form. This includes the selection of indicators, the design of the tables and the identification of footnotes, i.e. information that must be published alongside the tables to enable them to be interpreted correctly.

Representing the metadata
So far we have discussed some examples of uses of metadata and have seen that a crude classification system can be applied to it. We have also seen that some classes are more widely utilised than others. We then looked at the process of collecting survey data and identified the kind of information (metadata) that is associated with the four main steps in the process. We now need to consider how this information can be captured in a usable computerised form. In order to do this effectively we need to establish a common framework. This means some more fundamental work on the nature of metadata itself.

At the beginning of this paper 'information about information' is described as an infinite hierarchy. The potential amount of information is overwhelming. It is therefore important to bring order to the confusion. Metadata needs to be classified, ordered and associated with its function. Only then can we maximise its potential. To do this we need to draw on the cataloguing skills of librarians, and also on the concepts taken from software engineering. We need to consider essential models. The Object Oriented paradigm which associates data and purpose is a useful way of approaching the problem. In summary, we need to categorise metadata by use and by meaning. We also need to consider how rapidly each
category of metadata might change.

**Maintaining metadata**
The preceding section discussed theoretical and conceptual questions associated with metadata. This section looks at the more practical issues. We highlight a number of questions about the nature of metadata and the implementation of systems using it.

If we accept the notion of an infinite hierarchy of information at the intellectual level, we still need to examine the concept pragmatically. We need to consider if there are practical reasons for distinguishing between data and metadata, for example in relation to existing analysis packages, and whether there is an intuitive difference suitable for the kind of data we are handling, i.e. is there some kind of metadata (e.g. unit of measurement) which humans expect to see ‘closer’ to the data than others.

A second consideration is that of physical storage. Should data and metadata be stored in one system, or are the structures and uses such that they are better held separately? If they are stored separately, how do we ensure that the data and metadata are kept consistent?

There are practical questions concerned with the definition of a dataset. Is there such a thing as a definitive dataset? Should the data be considered independent of the statistical package in which it is held? We also need to consider metadata for related datasets and suites of datasets. Are we describing physical or conceptual datasets? Can we have subsets that are valid studies? Can we merge datasets into valid studies? There are special problems associated with time series and longitudinal datasets. For example, how do we describe changes in the real world?

Next we need to consider upgrades and versions. We need to know how to handle modified or restructured data. What happens when we add new derived variables? Should a dataset be static or dynamic. If we resolve these problems, we need to consider how the documentation can reflect these decisions.

**Conclusions**
In conclusion I first want to reflect that metainformation is probably more difficult and more expensive to capture than the data it describes. It is also the case that metainformation is generated at all points in the system. For these reasons, metadata essential to all users should be identified, and captured once at the most suitable point in the process. Having captured this expensive commodity, metadata should be made to work. It should also be held in a flexible structure so that it can be transferred between systems.

Finally, there is a great deal more thinking to be done on the nature of metadata and how it can be used to ensure that the data we all use is accessible and can be interpreted with the minimum of error. This means that time and resources need to be given to the theoretical and conceptual questions that are unresolved. The study of metadata needs to be seen as a valid intellectual activity in its own right and not only as a by-product of a particular statistical system. Only then will we be able to implement standards which can have sufficient validity to be widely accepted in the heterogenous and fast moving world that data libraries are trying to serve.

1 Presented at IASSIST/IFDO 93 Conference held in Edinburgh, Scotland. May 1993.


4 The BIRON System: the Archive's on-line catalogue and subject index. ESRC Data Archive University of Essex Wivenhoe Park, Colchester, Essex CO4 3SQ, UK.


7 Expert Interface to Statistical Information, EEC DOSES programme project no B34, 1990-1992


UKBORDERS: online access to digitised boundaries for use with UK Census

by Peter M Burnhill and Stephen J Solomon
ESRC Regional Research Laboratory for Scotland
& Edinburgh University Data Library

Introduction
This is a report of work in progress in the design and commission of a UK Boundary Outline and Reference Database for Education and Research Study (UKBORDERS). The UKBORDERS Project aims to provide online networked access to the digitised boundary outlines of census areas bought by the Economic and Social Research Council (ESRC) and the University Funding Council's Information Systems Committee (ISC) on behalf of UK higher education.

In addition to matters to do with data description and system design we also address issues relating to census geography, focusing on the three UK Population Censuses carried out one dismal Sunday in April 1991, and this lends an element of cross-national comparison. We identify four basic geographic systems which, we conjecture, underlie census geographies in all modern democracies: common placename geography, postal geography, electoral (and, thereby, administrative) geography, and the geography of the map-making surveyor. Historically, the feudal and ecclesiastical system of parishes and townlands provides a fifth census geography, one which, in the UK, has assisted comparison across time. The area geography chosen for census collection need not feature as a geography for publication of results from census: the Scottish census office broke the link between Enumeration District (ED) and output area for its 1991 Population Census.

UKBORDERS contains a 'library' of digitised boundary outlines for use in thematic mapping and in geographic information systems (GIS). The reference database component in UKBORDERS has emerged as critical in making the data facility an effective means for researchers to exploit boundary outlines. This reference database incorporates and makes explicit the various sets of spatial relations for the UK's myriad geography. Once established, this can assist census users exploit the spatial characteristics of the Census more fully, providing, for example, corresponding spatial contiguity matrices and retrieving appropriate area identifiers for Census small area statistics.

UKBORDERS is based on a networked Sun workstation and makes use of three different software systems: a geographic information system (to handle general management, the graphical-user interface and the boundary data); a relational dbms (for the spatial relationship between place, postcode, grid-reference, etc.); and a text-management system (for handling the 'actionable' metadata and associated catalogue information).

The ESRC has grant-funded the Regional Research Laboratory for Scotland at the University of Edinburgh to carry out the UKBORDERS Project as part of the ESRC/ISC Census Initiative. Funding began in November 1992. Edinburgh University Data Library has experience in providing a wide range of data facilities and it is hoped that network access to UKBORDERS over the UK Joint Academic Network (JANET) will be provided in late 1993.

Digitised boundary outlines
Digitised boundaries are used in automated catographic (mapping) software to produce thematic maps of the geographic variation in data from census, survey and many other data sources. They are also used in various spatial operations in the growing number of GIS software packages.

Display of census results through maps is a very effective means to provide a statistical summary, showing both the general level of incidence and measure of spread through geographic variation, across country, region or whatever. What is important is that each researcher can obtain the set of digitised boundaries appropriate to the research purposes and data in hand. This requires a shared understanding of census geography as well as the facility to make such retrievals from a database.

At a technical level, digitised boundary outlines are sets of computer readable co-ordinates that define areas. Each set is generally stored as linked segments (vectors) and define complete areas (polygons). Digitised boundaries vary in the accuracy with which the co-ordinates represent 'ground truth', an accuracy ultimately constrained by the number of co-
ordinates used in 'digitising' a given line and the accuracy of the source maps or remote images from which boundary information is derived. The greater the precision of digitisation the larger the resultant data set. There are several ways to format this co-ordinate information and their relation to ground truth; software products vary in their input format requirements, some demanding accompanying topographical information.

For many purposes, especially thematic mapping for display of finished statistical analysis, a researcher may not require the full precision available in a given digitised boundary dataset. Digitised boundaries which have been 'generalised' lose some of the minute detail available but are often more appropriate for thematic mapping. On other occasions, especially in GIS operations, the fullest extent of available precision may be critical.

Census geography
All census data are intrinsically spatially-referenced and commonly relate to bounded areas. Concern about confidentiality of information pertaining to identifiable individuals is only one reason why information is aggregated into areas. First, census information is collected on an area basis in order to allow efficient and accurate data collection by teams of enumerators assigned to cover non-overlapping areas. And, second, much of the reason for carrying out a census, rather than a set of efficient sample surveys, is because a census yields more accurate small area estimates, albeit only for a single time-point. The bounded areas for census are therefore both space and time specific; the data of census and of boundary line should be comparable.

Census geography in the UK
In the United Kingdom of Great Britain and Northern Ireland (UK) there are three census offices: the Office for Population Censuses and Surveys (OPCS) for England and Wales, the Census Office of the Northern Ireland Department of Health and Social Services and the General Register Office (GRO) for Scotland. Each is headed by a Registrar-General who, following an Act of Parliament, is directed by an Order of Council to carry out a Population Census. Inevitably, for the countervailing reasons of tradition and innovation, each does something different. In the UK, therefore, there were three separate, but concurrent and non-overlapping, Population Censuses carried out in 1991, each having a different variants of census geography. In practice, there was a great deal of liaison between the three Offices especially on the content of the three census forms which was very similar, with language variants (Gaelic and Welsh) and some local variations in question wording.

The greatest difference between the three Census Offices was their use of geography. Fortunately, each of the three census geographies can be related to four underlying types of reference systems: the common placename, the postcode, the electoral (and thereby the administrative) area, and the system of referencing by map-making surveyors, which in the UK is (are) the National Grid(s). From a superficial look at censuses carried out elsewhere it would seem that these four reference systems have widespread applicability use in census geography.

We take the address to be the most basic unit in census geography. Arguably, the lowest entity in the census is the individual person, linked to family and household. Nevertheless, the address provides the best means to spatially-reference a household, which is itself defined in terms of communal living at an address. This can be referenced in different ways, some more formally than others. The oldest is the common placename such as Cardiff (a city in Wales), Chelsea (a residential area in London) or Nine Mile Burn (a small hamlet to the South of Edinburgh). More recently, greater use is being made of the conventions of the Royal Mail which reference addresses in terms of a postcode unit (eg AB10 5NL or G8 3TW), using the geography used to deliver letters and parcels.

Political accountability through democratic vote provides the third system, whereby addresses are grouped by the Registrar of Electors and the Boundary Commission into the Polling Districts of the Electoral Wards used to elect Members of Parliament and to elect Councillors to County, District and City Councils; this geography then taking on an administrative role for the delivery of (local) services.

Map makers and other surveyors have their own system, the National Grid(s), alphanumeric co-ordinates, registered to (true) North and used in preference to the latitude/longitude system, which divides Britain and Ireland into 100 km squares; the two systems are administered by the Ordnance Survey (GB) and Ordnance Survey (NI).
Four Geographic Systems Underlying UK Census Geography

<table>
<thead>
<tr>
<th>common placenames</th>
<th>postal geography</th>
<th>electoral geography</th>
<th>surveyor’s geography</th>
</tr>
</thead>
<tbody>
<tr>
<td>fuzzy hierarchy</td>
<td>strict hierarchy</td>
<td>mixed hierarchy</td>
<td>metric hierarchy</td>
</tr>
<tr>
<td>city/region</td>
<td>Postcode Area (egMT9)</td>
<td>County/Region Parliamentary Constituency District</td>
<td></td>
</tr>
<tr>
<td>town/village</td>
<td>Postcode Sector (eg MT9 5)</td>
<td>1km Square</td>
<td></td>
</tr>
<tr>
<td>neighbourhood/hamlet/locality dwelling</td>
<td>Postcode Unit (eg MT9 5NL)</td>
<td>Ward</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>10 metre reference</td>
<td></td>
</tr>
</tbody>
</table>

National Grid(s)

Inter-censal comparison is often frustrated by changes in census geography, especially in the enumeration district. Of the four underlying geographies mentioned, only the National Grid(s) offers complete stability over time. Common placename usage varies on an ad hoc basis (although it can be very robust). The postcode system has been used to assist the analysis of change over time but the Royal Mail issues new postcode units (and alters boundaries of old ones) as operational needs dictate. The Boundary Commission periodically reviews the spatial requirements of the Representation of the People Act, the results of which may then disturb the electoral Ward boundaries. Occasional local government reorganisation is more thorough-going in its disturbance. This too highlights the need to include temporal referencing in the metadata that UKBORDERS held on each boundary outline datasets.

The censuses carried out in Britain during the Nineteenth Century made use of a fifth geography, the ecclesiastical system of parishes, reflecting an earlier form of political accountability, but also reflecting the fact that the clergy were used as enumerators in the earliest censuses and statistical accounts. In Britain parishes have lost relevance for the (population) census offices but remain an important output area for the annual agricultural censuses. The townlands system in Northern Ireland is an analogous historical system which, though not strictly comparable to the parochial system, assists comparison across longer periods of time.

The UKBORDERS facility sets out to exploit the four basic system of geography and their inter-relation in order to offer users of census data a coherent view of the variety of digitised boundaries available. The relation between the address and these four complementary reference systems can be shown in diagrammatic form.
Superficially, census geography looks most straightforward in England and Wales. OPCS defined some 130,000 Enumeration Districts for the 1991 Census, and these EDs were both a unit for collection of information through enumerators and for the publication of the resultant small area counts. These EDs were different from the set used for the 1981 Census, although, as in 1981, the 1991 EDs can be aggregated to the local government Wards which were current near to Census Day. There is only an approximate relation between the 1991 EDs for England and Wales and postcode geography. This inexactitude is significant since it means that researchers cannot link census data to many other sources of data a straightforward manner. In order to offer remedy, OPCS has created their own 'part-postcode units' whenever
a proper postcode unit straddled the boundary of an ED, and is providing an Index relating EDs to these postcodes and postcode units, the latter including a population count to assist weighted combination of areas.

In Scotland, GRO made direct use of the postcode geography. As in 1981, the postcode unit featured explicitly in the definition of the Enumeration Districts used for the collection of census information. The Scottish EDs cross-cut local government Wards, although they nest approximately into postcode sectors instead.

For 1991, GRO took the view that the input area (the ED) need not be the area used for output of the census counts: this was crucial. GRO then decided to base its Output Areas (OAs) on the 1981 ED geography, although it allowed some of the larger 1981 EDs to be split or, where necessary, combined. (The average size of the Scottish OA is much smaller than the average ED for England and Wales.) This will assist comparisons across time, especially since it had re-issued data from the 1971 Census using this 1981 ED geography. GRO has also published a 1990 edition of the Postcode Directory for Scotland, incorporating all postcode units extant at December 1990. This can be used to indicate the relation between postcode units and Scottish Output Areas.

The identifying code within each Census 1991 ED for England and Wales includes the relevant codes for the ‘standard’ higher order geographies. For example, EDs have a four-part alphanumeric code which identifies the County, District, Ward and the Enumeration District: 09LLAA01 is structured as CCDwEE. The identifying code within each Census 1991 Output Area for Scotland has a comparable structure, providing the relevant codes for Region, District and, to an extent, also to postcode sector.

The Census Office (Northern Ireland) chose yet a different approach to census geography: they went further and applied the (Irish) National Grid reference to each household included in the Census. The areas chosen to publish census output could therefore be independent of the system of EDs used to collect census information. The choice of output area was therefore wide open. Nevertheless, and perhaps unfortunately, the ESRC is funding a project in the Census Office (NI) which will lead to the publication of data using the 1991 ED geography as the smallest ‘small area’, although there will be a variety of other geographies available at higher levels: (pre-1992) Wards, postcode sectors and the 1 km and 100m grid-square output already scheduled. The townlands would be another contender. The Irish National Grid is a different system of co-ordinates from that used in Great Britain, designed to allow a good fit to the whole of Ireland; it too is square and oriented to (true) North, and is, therefore, at an acute angle to the National Grid used in Great Britain.

**Consortium purchase for academic access**

The Economic and Social Research Council (ESRC) and the (then) University Funding Council Information Systems Committee (ISC) purchased two sets of digitised boundaries associated with the 1991 Population Census. The outlines for the 1991 EDs used in the 1991 Population Census for England & Wales were purchased from the EDLINE Consortium, and were commercially produced after the Census was taken. The ESRC and ISC are purchasing outlines of the postcode units from the Scottish Census Office, the GRO. GRO had digitised these for use in conducting the 1991 Population Census for Scotland. The equivalent digital product for Northern Ireland is being purchased by the same consortium joined by the Department of Education for Northern Ireland. As discussed above, these will be at the 1991 ED level, created after the Census was taken: they are not (yet) available.

In the 1990s, therefore, the digitised boundary sets for the whole of Britain available to academics are at much greater resolution than were the Ward-level and Postcode Sector-level boundaries in the 1980s. The 1991 map outlines are, therefore, much larger in number, require more attention to storage and retrieval, and are much more versatile.

These are important research resources, especially useful given recent advances in the accessibility of thematic mapping and GIS software, in part brought about by the ISC purchase arrangements for ARC/INFO 5, and by the availability of low cost mapping packages for desktop computers.

**The UKBORDERS Project**

The UKBORDERS project aims to provide a facility for the academic researcher to access the mix of digitised boundaries, indexes and other geo-referencing directories now becoming available. It complements the other census-related projects funded by the ESRC, and will allow a wide range of census users make sense of the myriad geographies used in the UK Censuses.

Ideally, the target user group for the facility will encompass the many classes of staff and student who wish to use the
digitised boundary outlines and the Census statistics but who do not have extensive expertise in computing or automated mapping. In practice, we must phase the development of functionality in the system. We have given priority to meeting the needs of researchers who are computer-literate and spatially-aware and who have some experience of statistical mapping. We regard these as being the initial target user group for the facility as we believe that this will yield the biggest productivity gain for Census usage. Subsequent generations of the facility will give priority to two very different clientele, offering benefits both to the general quantitative researcher, who may not be expert in either thematic mapping or computer use, and to the expert user of geographical information systems (GIS).

The project has four components:

1. **data accession and quality assurance** - This involves the receipt and cataloguing of digitised boundary outlines for Great Britain and Northern Ireland as these become available to the UK academic community. As part of the UKBORDERS Project, attention is given to the 'quality assurance' of the postcode unit boundaries for Scotland. A separate project at the University of Manchester was charged with the task of subjecting the ED boundaries for England and Wales to a comparable test of 'quality assurance'.

2. **boundary outline library** - The creation of a 'library' of digitised boundary outlines for Great Britain and Northern Ireland is the main objective for the project. This will include the postcode unit (PCU) and Output Area (OA) boundaries for Scotland and the Enumeration District (ED) boundaries for England and Wales and all the 'standard' administrative areas in each Country and Province of the UK recognised in the small area statistics published by the three Census Offices (Standard Areas). In time, boundary outlines for other, non-standard areas will be added.

3. **access software** - Easy-to-use access software will allow these boundary outlines to be browsed, selected and retrieved in a format suitable for use with GIMMS, ARC/INFO and a number of other mapping and GIS packages. As the majority of researchers will want to extract only a (relatively small) subset of the total number of available boundaries (often for one or more of the Census 'standard areas', and often to do so only for thematic display) attention is given to procedures for sub-set selection and for generalising the detail in the selected boundary dataset.

4. **geo-reference directory** - The creation of a reference database will allow researchers to identify the codes for the census small areas contained within a given area. This lies at the heart of the project. These may be the standard administrative areas (eg Rushmoor District) or may be defined by some other key geographic grouping (eg defined by a list of postcode units or National Grid co-ordinates). This geo-metadata database will relate 'places' and areas to the codes for Census (small) areas through postcodes, various indexes and the National Grid - including attention to the estimation of 'weights' where Census areas cross postcode unit boundaries as they do in England and Wales.

**Data for UKBORDERS**

In the UK there are digitised boundaries for the 150,000 (approx.) census output areas. These and 130,000 unit postcodes used to define the Scottish Census areas, are being integrated in a library of boundary outlines together with the corresponding indexes and directories that link these census geographies to the four underlying geographies identified earlier.

Data, in the form of indexes and directories, are the key to the reference database, being incorporated in a relational database which defines the set of relations between different geographic reference terms: placenames, postcodes, EDs, OAs, Wards, etc. The National Grid provides the underlying metric and the postcode unit (and in the case of England and Wales, the 'part' postcode unit) acts as a basic building block. The Data Library has experience in using software to provide access to geographic reference directories, especially the Postcode Directory for Scotland through PCGET (a front-end facility for an application of ORACLE created in 1987); analogous facilities providing access to the Postcode Address File (UK), the Postzone File (GB) and to various placename gazetteers and indexes.

The ED boundaries form the most detailed level of boundary outline for England & Wales available through UKBORDERS. These are to be supplied as a set of one metre resolution co-ordinates, in the 'generic' format specified by ESRC, and organised in County files. This 'generic' format for digitised boundaries was especially formulated for ESRC/ISC to facilitate the conversion of the boundaries to a wide range of mapping packages, and a special project is being funded at the Census Dissemination Unit, University of Manchester, to carry out such an exercise. Using the generic format, each county file starts with 10 lines of descriptive text. This text includes the full text name of the County, a copyright notice, date and release number of the file, history of changes and the current length of the file in
lines. An indication of the accuracy of the data (in terms of the accuracy standards listed above) is also included.

The criterion set is that 95% of digitised points have to fall within +/- 2 mm of the corresponding point on the base map when plotted and overlaid. Errors of +/- 4 mm are deemed to be acceptable for the remaining 5%. These requirements imply the following accuracy:

<table>
<thead>
<tr>
<th>Base map scale</th>
<th>'Real World' 95% accuracy level</th>
<th>100% accuracy level</th>
</tr>
</thead>
<tbody>
<tr>
<td>1:1250</td>
<td>+/- 2.5 metres</td>
<td>+/- 5.0 metres</td>
</tr>
<tr>
<td>1:2500</td>
<td>+/- 5.0 metres</td>
<td>+/- 10.0 metres</td>
</tr>
<tr>
<td>1:10,000</td>
<td>+/- 20.0 metres</td>
<td>+/- 40.0 metres</td>
</tr>
</tbody>
</table>

In Scotland, the unit postcode boundary data form the most detailed boundaries available through UKBORDERS. These were also bought by ESRC/ISC and were supplied to the project by the General Register Office (Scotland) in GIMMS format.

The digitising of the boundary data for Northern Ireland has yet to be completed. Once completed, the NI Small Area boundaries will be delivered to the UKBORDERS project for quality assurance and entry into the UKBORDERS library of boundaries.

An Index of Place Names (IPN)
Our use of the common placename is novel, but potentially very rewarding. An Index of Place Names (IPN) relating to the 1991 Population Census is not due to be published until 1994. However, the IPN relating to 1981 is available, in printed and computer-readable format, for both England and Wales and for Scotland. Placenames change slowly and so much of the information in the IPN remains relevant. We intend that it should play a key role in making census boundaries and counts accessible to a wide range of researchers.

Example entries from IPN (E&W):

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Burnham Green</td>
<td>Lo</td>
<td>Bucks</td>
<td>South Bucks</td>
<td>325-1</td>
<td>SU9383</td>
<td>-</td>
</tr>
<tr>
<td>Burnham Green</td>
<td>Lo</td>
<td>Herts</td>
<td>East Hertfordshire</td>
<td>533-1</td>
<td>TL2616</td>
<td>-</td>
</tr>
<tr>
<td>Burnham Market</td>
<td>UA</td>
<td>Norf</td>
<td>Welwyn Hatfield</td>
<td>532-1</td>
<td>TL2616</td>
<td>-</td>
</tr>
</tbody>
</table>

Example entries from IPN (S):

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Buckie, Mill of</td>
<td>Ag</td>
<td>19</td>
<td>Rathven</td>
<td>AB5 2</td>
<td>290</td>
<td>NJ 4264</td>
<td>-</td>
</tr>
<tr>
<td>Buckie West</td>
<td>DW</td>
<td>19</td>
<td>St.Ninians</td>
<td>FK6 5</td>
<td>473</td>
<td>NS 7484</td>
<td>3,653</td>
</tr>
<tr>
<td>Buckieburn Buittle</td>
<td>CP</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>521</td>
</tr>
</tbody>
</table>
# Comparison of content between IPN (E&W) and IPN (S):

<table>
<thead>
<tr>
<th>code</th>
<th>IPN (S)</th>
<th>IPN (E&amp;W)</th>
</tr>
</thead>
<tbody>
<tr>
<td>coverage</td>
<td>Scotland</td>
<td>England and Wales</td>
</tr>
<tr>
<td>reference date</td>
<td>1981</td>
<td>1981</td>
</tr>
<tr>
<td>entries</td>
<td>settlements</td>
<td>settlements</td>
</tr>
<tr>
<td>number of entries</td>
<td>9,000</td>
<td>62,000</td>
</tr>
<tr>
<td>National Grid ref on OS map</td>
<td>1 km</td>
<td>1 km</td>
</tr>
<tr>
<td>entry typology</td>
<td>eight</td>
<td>eight (nine)</td>
</tr>
<tr>
<td>LGR</td>
<td>Region</td>
<td>Co</td>
</tr>
<tr>
<td>IA</td>
<td>/Island Auth,</td>
<td>District</td>
</tr>
<tr>
<td>LGD</td>
<td>LG District,</td>
<td>/London Borough,</td>
</tr>
<tr>
<td>Town</td>
<td>Town</td>
<td>New Town,</td>
</tr>
<tr>
<td>Loc</td>
<td>Locality,</td>
<td>Urban Area,</td>
</tr>
<tr>
<td>Ag</td>
<td>Agricultural Community,</td>
<td>Locality,</td>
</tr>
<tr>
<td>RED</td>
<td>Regional Electoral Division,</td>
<td>Parish,</td>
</tr>
<tr>
<td>DW</td>
<td>District Ward</td>
<td>/Welsh Community,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Urban Area Sub-Division</td>
</tr>
<tr>
<td>(1974) County ID</td>
<td>n.a.</td>
<td>Y (54)</td>
</tr>
<tr>
<td>LG District ID</td>
<td>Y (56)</td>
<td>Y (403)</td>
</tr>
<tr>
<td>Regstn Dist ID</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Civil Parish ID</td>
<td>Y</td>
<td>n.a.</td>
</tr>
<tr>
<td>Postcode sector</td>
<td>Y</td>
<td>Y (not for places w/o boundaries)</td>
</tr>
<tr>
<td>Population count</td>
<td>(not for places w/o boundaries)</td>
<td>(not for places w/o boundaries)</td>
</tr>
<tr>
<td>1981 Pop Present</td>
<td>University of Edinburgh, in Data Library</td>
<td>University of Edinburgh, in Data Library</td>
</tr>
<tr>
<td>title proper</td>
<td>GRO(S)</td>
<td>OPCS</td>
</tr>
<tr>
<td>publisher</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The Census Offices may be updating these Indexes. Some of the classifications used in the IPN will change. The 'locality' classification in the 1981 IPN for England & Wales relates to areas of continuously built-up land, separated from other built land by a given distance. In the decade following the creation of this classification, considerable building work has reduced the amount of 'unbuilt' land. This may mean that the 'locality' classification has to be altered. In addition to this change in classifications over time, the classification schemes used in the Scottish Index and the IPN for England & Wales also differ. This means that the database will have to cater for both temporal and spatial differences in classification schemes.

Of the two counts of population, 'population present' (actually present at a given location on the night of the census) and population resident (number of people who live at that location), the population count provided in the IPN is the former ('population present count'). The 1991 version of the IPN should, perhaps, include the alternative methods of measuring population as an additional item of information.
**Features in UKBORDERS**

We assessed the required features for UKBORDERS by examining the tasks that users would require of an on-line system providing access to the Census boundaries. Accessibility over the UK Joint Academic Network (JANET) was among the main requirements for the system, so too was the capability to serve multiple users. The latter highlighted issues to do with acceptable response times and, therefore, processing speed, which in turn raised issues concerning disk space requirements. We believe that the language in which the user informs the system about the area of interest and characteristics of the boundaries of interest are of great importance. We therefore identified the need to establish terms, common to user and facility, and to incorporate these in the user-interface.

We identified a number of tasks for software.

- **register** The boundary files have been acquired for academic research and teaching. A registration scheme is required which will include verification of authority to use the data.

- **browse** These tasks include the ability to browse through information about the boundaries available through UKBORDERS - to allow users to assess for themselves the type of ‘ready-made’ boundaries held (their accuracy, their generalisation level, their current level of aggregation) and the higher-order boundaries that could be created in straightforward fashion through aggregation.

The boundary data are catalogued using fields appropriate to computer files in the extended form of AACR2 (Chapter 9) and ISBD:CF). This has already been done for the District, Ward (England and Wales), postcode sector (Scotland) boundaries held in the Data Library for use with the 1981 Census statistics.

- **extent** The ‘area of interest’ refers to the extent of the geographical area of interest to the user: eg Edinburgh city, the County of Devon. These areas may be referred to by placename. Other methods for specifying the area of interest include specifying a bounding rectangle of O.S. national grid co-ordinates, specifying specific geo-unit codes for extraction (such as postcode units or ED codes), defining an irregularly shaped study area within which all boundaries of the specified constituent geo-unit will be extracted.

A geo-unit is a discrete geographic element, such as the area defined by a postcode unit. Two types of geo-unit are defined. The **target** geo-unit(s) are the boundaries of principal interest to the user - the actual level(s) of boundary outlines to be retrieved. The target boundaries may already exist in the system ‘ready-made’ or have to be ‘custom-made’. The **constituent** geo-units, are the building blocks from which the target geo-units may have to be ‘assembled’ or ‘custom made’.

- **select** Most user enquiries will be satisfied with a limited number of boundary outlines sets. These will already exist in UKBORDERS, either actually stored ready for extraction, or ‘virtually’ so, dependent upon the cost benefits of storage and computation. These are regarded as ‘ready-made’ boundaries.

- **retrieve** The object of UKBORDERS is to allow users to retrieve a dataset of boundary outlines, one which is relatively easy to select, extract and take back for entry into the user’s chosen software.

The user should not be unwittingly exposed to the complexity of the actual computational procedure required to take a user from initial registration to the extraction of the required boundaries and transfer of a dataset to a local computing environment. Moreover, we anticipate having users with a greatly varying level of expertise with computer mapping. The user-interface therefore should be “user-friendly”, in the sense that it is easy to understand and easy to use, and in the sense that it introduces the user to terms appropriate to thematic mapping of census information using the digitised boundary outlines.

- **view** The ability to allow users to see boundaries visually ‘on-screen’, prior to their extraction exploits the inherent spatial character of the boundary outlines. This assists users retrieve exactly what they want; providing the target geo-unit required, at an appropriate level of detail and for the area of interest.

UKBORDERS will provide two interfaces, both with a help facility. Full exploitation of the extensive graphic facilities provided by UKBORDERS will require a display device with X-window capability or equivalent emulation. However, all the functionality necessary for accessing the required boundaries will be in a ‘non-graphical interface’ requiring...
access via VT100 emulation. The graphical user interface will provide enhanced usability via the menus, button selections and other widgets now expected.

We have also sought a simplicity of language and operation in the interface: for example, we regard each retrieval as producing one single boundary layer. This could be a simple layer, comprising only one level of target boundary outlines, or a complex layer, comprising two or more levels, one nested within the other. Only one such layer, simple or complex, may be retrieved at a time. Where two or more boundary systems result in intersecting outlines, we have thought it safer to require users to undertake two or more retrievals, producing two or more layers.

A retrieval of boundaries at a single geo-unit level, such as the Scottish Output Areas, is regarded as a simple layer. A retrieval of boundaries for postcode units nested with Scottish Output Areas (or alternatively, EDs nested within Wards in England and Wales) would be regarded as a complex layer. Because Ward outlines and postcode sector outlines cross-cut (intersect) their extraction would require two retrievals, resulting in two datasets.

- **assemble** In addition to the retrieval of ‘ready-made’ boundaries, UKBORDERS will also allow users to retrieve ‘custom-made’ boundaries. These boundaries will be formed by using the links in geo-unit identifiers and in the spatial indexes. The custom boundaries will be assembled from their constituent geo-units: eg output area boundaries being assembled from their constituent postcode unit boundaries. The constituent geo-unit is, therefore, the building block and is the smallest unit of accessible boundaries.

A user wishing to assemble the boundaries for a target geo-unit (eg Wards) will require the system to generate boundaries from their constituent geo-unit (eg EDs) using the identifying code within each Census 1991 ED for England and Wales.

If users prefer to assemble their own custom boundaries, it will be possible to retrieve the boundaries for their own area of interest at the constituent geo-unit level together with an index for assembling these to the target geo-unit level.

- **generalise** The boundaries will be held on the system at the level of detail supplied. The accuracy standards enforced for the boundary digitising may mean that this level of detail will be too great for some applications involving use of the boundaries. A facility to generalise the lines to a more appropriate level of detail will be made available.

- **convert** Users of the system will expect to be able to use the boundaries in a wide range of mapping packages or GIS. The boundaries should therefore be made available in a range of file formats, suitable as input to a number of software packages. This will greatly enhance the usability and use of the 1991 Census boundaries.

It is worth remarking here that the task of converting the boundary outlines from the format in which they were delivered (GIMMS and generic) to ARC/INFO is not simple. This is largely for two reasons. First, some boundary areas are like lakes within another boundary area, and this can cause some confusion in the formatting of the co-ordinate sets. The second, related, point is that many mapping packages, including GIMMS, provide spatial information about the topology of complete boundary areas (polygons) which GIS packages, such as ARC/INFO, find ambiguous. This also applies to information in the ‘generic’ format, itself an elaboration of the GIMMS format.

- **transfer** Once the required boundaries have been selected and the required file format specified, a file separate to those held by the system is created. This file holds a single layer of boundaries at the target geo-unit level for the area of interest in the required file format. Additional files for other layers, Census data or index information may also be created.

Files may be delivered to the user’s own workspace on their own local machine. This is achieved via ftp (file transfer protocol) or email (electronic mail). As stated, the more a set of boundaries has been generalised, the smaller will be the resultant dataset, the easier will be its transfer and its subsequent use on a given software/hardware environment.

**System Design**

The main requirement for UKBORDERS was that the system should handle the large number of digitised boundaries covering England & Wales, Scotland and Northern Ireland. The phased delivery of these datasets required a flexibility to handle additional loads of data over time. The inherent ‘vector’ nature of the digitised boundaries meant that the
software used to handle them had to be based on a 'vector' data model, and be capable of handling associated attribute data.

The diverse number of user-tasks, outlined above, for UKBORDERS to meet, and the wide range of users anticipated, meant that the software selected should house facilities for a “user-friendly” user-interface. The geographical character of the boundaries and their use in mapping made the provision of graphical capabilities an obvious feature to have. An ability to display the boundaries on a terminal screen with a GUI (graphical user interface) would greatly enhance the usability and approachability of the system for even the most novice of computer users.

The need for a ‘vector’ data model combined with a graphical user interface that could be easily tailored to meet the demands of both novice and experienced users led to the decision to create the system as an application of Arc/Info, a leading GIS product. The capability of GIS software products to relate information to the spatial dimension makes them particularly applicable to a wide range of geographic projects. The Arc/Info GIS is particularly useful to UKBORDERS as it provides the ability to generate tailor-made graphical user-interfaces, to store large volumes of vector information and related attribute data and has its own programming language AML (arc macro language) for creating applications.

The hierarchical nature of many of the boundaries and their potential to reference these to other (hierarchical) geographic systems made the provision of a relational database essential. The large number of indexes required to relate common placenames to a number of different registration systems (National Grid, postcodes) confirmed the need to store these indexes in a relational database management system (rdbms).

Unfortunately, the database part of Arc/Info (the INFO database) does not provide full relational database facilities. Its slightly rigid requirements for relating tables meant that it was not considered sufficient for a system involving large numbers of indexes. However, Arc/Info does provide the ability to connect to other databases. This includes the Ingres relational database management system, a fully relational dbms using the Standard Query Language (SQL) to process queries. Previous experience of SQL with Oracle meant that the learning curve for familiarity with Ingres would be fairly short for the project officer, who could also turn for advice to others in the University with considerable experience of Ingres; the University Geography Department had been undertaking another GIS project involving Arc/Info and Ingres to good effect, and the University’s Computing Service had considerable experience with Ingres applications. This helped confirm the notion that combining Ingres with Arc/Info would help in making UKBORDERS a success.

One of the more important qualities of UKBORDERS will be its ability to take users straight to their area of interest via placename. This requires extensive text-handling facilities. The large number of boundaries held coupled with the number of indexes, means that the project will generate a large amount of metadata. This ‘information about information’ is essential in order to keep track of the different data-sets held by the system. The ability to search this metadata would greatly assist users in obtaining the boundaries they require. The need for this facility served to highlight the requirement for good text-handling and text-searching functions.

The text-handling facilities provided by the Info and Ingres databases are limited. Neither lent itself readily to the task of providing good metadata searching. Conversely, no database assessed for use in the project provided good metadata facilities and good vector data handling and good relational facilities. We therefore identified the need for a third software system in addition to Info and Ingres.

The Data Library has considerable experience with one of the databases assessed for the metadata searching: BRS/Search. Therefore it was decided to use BRS/Search for the metadata tasks. This decision was facilitated by the fact that the design for the system had become greatly clarified at that stage. Arc/Info would be used to provide the overall control to the project via its AML programming language. In addition, Arc/Info would be used to hold the digitised boundary data (in fact, spatial data have to be held in Info if Arc/Info is used). The associated indexes and attribute data could be stored in the Ingres relational database and accessed via AML commands. AML also provides the ability to connect to other processes, in effect pausing the Arc/Info session. This facility enables BRS/Search to be used for the metadata searching.

When a user wishes to access the metadata, control is passed to the BRS/Search database. Results of the metadata search (i.e., defining the area of interest and the constituent and target geo-units) can be written to a file. On leaving BRS/Search and reactivating the Arc/Info session, the AML can access this file and then access the necessary boundaries, indexes and attribute data. If further fine tuning of the user’s request is required, then control can be passed back to BRS/Search and
results written out to and read from a file in the same way. This processing will appear seamless to the user as it will all be provided under an X-window interface.

The fact that UKBORDERS will be accessed by a wide number of users means that it must be robust; able to cope with illegal responses to prompts, incorrectly formatted data files and other ‘unexpected’ events. The system must also be able to cope with a number of different users accessing at the same time, without too detrimental an effect on processing speeds.

The requirement for multiple user access does not affect the selection of packages to be used. Arc/Info, Ingres and BRS/Search can all be tailored to cope with multiple users. Issues concerning processing speed have more to do with hardware. Fortunately, UKBORDERS is mounted on a Sun workstation and this provides a coherent upgrade path which will enhance the processing resources required to maintain acceptable response times under heavy user load.

The different geographies of England & Wales, Scotland and of Northern Ireland used at the time of the Census are volatile, that is, they are subject to change over time. So too are some higher-level areas to which census data may be aggregated. The system therefore had to respond to the corresponding alterations in boundaries, as new data at the lower levels of census geography, and as derived data at the higher levels of geography.

We anticipated that users would request new functionality from the system, and that we might be able to secure additional data resources to add to the functionality of the system. Such changes to the system are possible with the software that are being employed in UKBORDERS; the system can therefore be regarded as extensible both in terms of data-sets held and functionality.

Conclusions
Computer-readable data from the Population Census are largely spatially-specific, and their analysis and understanding require digitised boundaries. These too can benefit from a ‘data library’ environment, but their effective use requires attention to matters geographical.

Our first port of call is with the census offices themselves. It is reasonable that the geography used to define the Enumeration Districts, by which Census is carried out, should vary across census office and with each successive census. What matters more to users of Census, however, is the variation in the choice of small areas on which descriptive census counts are subsequently published in computer-readable form. That the area geography used for publishing counts and other summary statistics need not be the one used for collection purposes, especially given the possible and actual use of computerised address gazetteers, has been well illustrated by the practice adopted by the 1991 Population Census for Scotland. The Output Area need not be the ED.

Whatever the census geography, researchers will want to cross-relate census data to data drawn from other statistical sources, and defined using other geographies, typically those relating to common placename, postal geography, electoral geography (on which many service-delivery areas are also defined) and to the geography of the map-making surveyor. The UKBORDERS project has addressed these issues in order to allow a wide range of census users to make sense of these myriad geographies and provide access to the mix of digitised boundaries, indexes and other geo-referencing directories now becoming available.

The project is demonstrating the need to integrate different software packages. In this case these are a GIS, for the graphical-user interface and general management of the boundary data; a relational dbms, for the spatial relationships; and a text-management system, for handling the associated metadata and ‘actionable’ catalogue information. The choice of computing platform matters less although it must possess a ready, software-independent upgrade path, as well as be readily connectable to a wide area network for multiple access.

With the growth in computing and in the availability of both thematic mapping packages and GIS software, we believe that there is a need for a new type of library for digitised boundary outlines. Such a library will also enhance use of census data. Delays in the negotiations to allow purchase of the base boundaries have hindered progress. Nevertheless, we hope that Edinburgh University Data Library will be able to offer access to this UKBORDERS facility across the UK Joint Academic Network in late 1993.
1 This is a pre-publication draft based on a paper presented at IASSIST/IFDO'93, the Joint Conference of the International Association for Social Science Information Services and Technology and the International Federation of Data Organisations, in Edinburgh on 12-14 May 1993. Comments are welcomed.

2 This paper was presented at IASSIST/IFDO 93 Conference held in Edinburgh, Scotland. May 1993. All correspondence about this paper should be addressed to the authors at Data Library (EUCS), Main Library Building, Edinburgh EH8 9LJ, Scotland UK; email p.burnhill@ed.ac.uk

3 A Geographical Information System (GIS) has been described as “a computer system for collecting, checking, integrating and analysing information related to the surface of the earth”, P.A.Burrough Principles Of Geographic Information Systems For Land Resources Assessment (Monographs On Soil And Resources Survey No 12) Oxford: Clarendon Press 1986.

4 UKBORDERS has ESRC award reference number H 507/25/5101; RRL Scotland has ESRC award reference number A 504 /28/5008.

5 ARC/INFO is GIS software produced by Environmental Systems Research Institute of California, US.

6 We are grateful to the Registrar General and his staff for the ‘loan’ of these data during a hiatus in the negotiations between the ESRC/ISC and GRO.

7 GIMMS is computer mapping/GIS software distributed by GIMMS Ltd and created by TC Waugh, also a senior lecturer at the University of Edinburgh Geography Department.
IASSIST 1995

Partners for Access: Working together in a changing data environment
L'accès aux données dans un environnement en pleine mutation: un partenariat à développer

Quebec City, Canada May 9-12, 1995
21st annual conference of the International Association for Social Science Information Service and Technology

Preliminary announcement & CALL FOR PAPERS

The premier conference for professionals providing data services in libraries and archives, IASSIST 1995 focuses on the new opportunities for collaboration presented by the phenomenal growth of worldwide computing networks. Taking advantage of this new technology presents many challenges, and encourages data service providers to work together to ensure continued access to useful and high-quality data. The Program Committee is now soliciting papers on all issues relating to the provision of service for machine-readable numeric, textual and image data. Papers of special interest include the examination of merging social and spatial data (GIS) and the implications for data services. Another special interest is how changes in technology and an expanding clientele using data have an impact on our profession. Papers about the future role and function of data librarians and data archivists are being sought.

Other papers of special interest would include those focusing on data sources and research issues in global change, AIDS, poverty, or other comparable social research themes. Technical topics could include the uses of the Internet, UNIX applications in archives, migration from centralized computing to a distributed computing environment, or mass data storage issues. Library issues may cover bibliographic access tools, indexing standards or user services. Also of special interest are discussions addressing standards for data documentation and metadata. Papers addressing major barriers to access, such as national information policies as they relate to data, intellectual property rights (copyright), and confidentiality restrictions, are particularly welcome.

Conference Location:
Loews Le Concorde
1225, Place Montcalm
Quebec, QC G1R 4W6
Canada

Conference Site:
The perfect site for a conference, the Loews Le Concorde in Quebec City is located on the Grande Allee, an avenue characteristic of old Europe. Across from the hotel are the Plains of Abraham, a distinguished setting draped in a curtain of greenery offering a magnificent view of the St. Lawrence River. A brief walk from the hotel is Old Quebec, the only walled city in North America. Within Old Quebec, winding cobblestone streets are lined with colourful bistros and quaint boutiques. The conference hotel features 424 rooms and spacious suites. Each room has two telephone lines and a computer/fax connection. Hotel services include concierge assistance, secretarial services, and health club (heated pool in season.) The hotel is 20 minutes from Quebec
Conference hotel rates are $109 Canadian nightly for single or double occupancy. Reservations must be made before April 7, 1995.

LOCAL ARRANGEMENTS CHAIR:
Gaetan Drolet
Gaetan.Drolet@bibl.ulaval.ca
(418) 656-7970

PROGRAM COMMITTEE CHAIRS:
Walter Piovesan
walter@sfu.ca
(604) 291-5869
Chuck Humphrey
chumphre@library.ualberta.ca
(403) 492-9216

CONFERENCE INTENTIONS FORM
Please return this form by December 15, 1994 to:
IASSIST 1995 PROGRAM COMMITTEE
c/o Chuck Humphrey
Data Library
415 Cameron Library
University of Alberta
Edmonton, Alberta T6G 2J8
CANADA
or
E-mail: chumphre@library.ualberta.ca

NAME:
TITLE:
AFFILIATION:
MAILING ADDRESS:
ELECTRONIC MAIL ADDRESS:
TELEPHONE: FAX:

Below, check all that apply:

___ I intend to submit a paper on the following topic/title:

___ I would like to hold a panel/seminar/roundtable discussion on the topic of:

___ I am interested in presenting the following poster session/display:

___ Please send conference information to my colleague:

___ I will be willing to Chair a session.

___ Please keep me on the mailing list.
The International Association for Social Science Information Services and Technology (IASSIST) is an international association of individuals who are engaged in the acquisition, processing, maintenance, and distribution of machine readable text and/or numeric social science data. The membership includes information system specialists, database librarians or administrators, archivists, researchers, programmers, and managers. Their range of interests encompasses hard copy as well as machine readable data.

Paid-up members enjoy voting rights and receive the IASSIST QUARTERLY. They also benefit from reduced fees for attendance at regional and international conferences sponsored by IASSIST.

Membership fees are:
- Regular Membership: $40.00 per calendar year.
- Student Membership: $20.00 per calendar year.

Institutional subscriptions to the quarterly are available, but do not confer voting rights or other membership benefits.

Institutional Subscription: $70.00 per calendar year (includes one volume of the Quarterly)

Membership form

<table>
<thead>
<tr>
<th>I would like to become a member of IASSIST. Please see my choice below:</th>
</tr>
</thead>
<tbody>
<tr>
<td>☐ $40 Regular Membership</td>
</tr>
<tr>
<td>☐ $20 Student Membership</td>
</tr>
<tr>
<td>☐ $70 Institutional Membership</td>
</tr>
<tr>
<td>My primary Interests are:</td>
</tr>
<tr>
<td>☐ Archive Services/Administration</td>
</tr>
<tr>
<td>☐ Data Processing</td>
</tr>
<tr>
<td>☐ Data Management</td>
</tr>
<tr>
<td>☐ Research Applications</td>
</tr>
<tr>
<td>☐ Other (specify)</td>
</tr>
</tbody>
</table>

| Please make checks payable to IASSIST and Mail to: Mr. Marty Pawlock, Treasurer, IASSIST % 303 GSLIS Building, Social Science Data Archives, University of California, 405 Hilgard Avenue, Los Angeles, CA 90024-1484 |

<table>
<thead>
<tr>
<th>Name / title</th>
</tr>
</thead>
<tbody>
<tr>
<td>Institutional Affiliation</td>
</tr>
<tr>
<td>Mailing Address</td>
</tr>
<tr>
<td>City</td>
</tr>
<tr>
<td>Country / zip / postal code / phone</td>
</tr>
</tbody>
</table>