Metadata and User Interfaces: Promises and Problems

by Steven R. Howe [1]
& 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 restructing 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 multi-media 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 couldbe 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