Global Access to Data Resources:

Where's the Metadata?

by Mark A. Carrozza & Steven R. Howe *

Howe and Graham (1993) proposed that "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." They identified three problems that needed to be addressed:

- Metadata must be transportable from platform to platform.
- There will be pressure on interface designers to make interfaces ever smarter, as more and more naïve users access metadata.
- Metadata will vary in quality, depending largely upon whether the research team intended the study to be available for secondary analysis.

The purpose of this paper is to reassess where the social science community is with respect to the above issues. Throughout this paper, we will be careful to distinguish between studies intended for use in secondary analysis and other studies. One of our themes is that tremendous progress has been made over the past five years with respect to data sets intended for use by secondary analysts. In contrast, very little progress has been made with respect to the problem of making metadata available for the tens of thousands of other studies published each year in the social sciences.

Transportability

Of the three issues identified by Howe and Graham (1993), the greatest amount of progress has been made in terms of the transportability of metadata (and data). This is not to say that the problems have all been resolved, but it is now possible to imagine a future in which transportability is a non-issue. While researchers have been able to routinely transmit error-free data around the world for the past six or seven years, it has only been in the past two years or so that the problem of data-storage has been solved.

The University of Cincinnati has recently purchased an HP 330FX Optical Storage Jukebox to store its social science data collection. The Jukebox, with 330GB of direct online storage, is connected to a Windows NT file server that It is also easy to forget the fact that the recent success of Java promises that access to metadata via the Web can, in theory, be unfettered by operating system or platform differences.

Interface Development

The UC system compares very favorably to almost any other data archive in terms of access to secondary data. However, as these kinds of storage devices become more commonplace, there have been no comparable improvements in the quality of user interfaces to access data sets. With few exceptions, user interfaces have not progressed appreciably in the last five years. The University of Michigan has developed impressive web sites for analysis and extraction of data from the General Social Survey and the American National Election Study. The Bureau of Economic analysis has marginally improved user access to the Regional Economic Information System (REIS) CD-ROM with the release of a Windows interface. Unfortunately, The Bureau of the Census GO/Extract combination and the National Center for Health Statistics SETS software have remained essentially unchanged over the last several years.

ICPSR also seems to be staking out a position that is distinct from that of industry. The ICPSR Data Documentation Initiative is moving in a direction away from that of software developers such as Microsoft and SAS (Microsoft and SAS are both members of the Meta Data Interchange Specification Initiative).

While there have been modest advances in the ways that statistical software packages such as SAS and SPSS permit the analyst to make use of metadata in working with a set of data, packages have by and large remained stagnant in the amount and types of metadata they support. Most packages do a very poor job of supporting any type of metadata beyond what can be considered "data definition metadata" (i.e., variable labels, value labels, missing value definitions, etc.), and even with respect to these kinds of metadata, the packages' capabilities are nearly identical to what was available a decade ago, although more of this information is available in point-and-click interfaces.

Perhaps most importantly, none of the major packages have

produced any revolutionary new tools for capturing metadata that archivists will need for bibliographic purposes or that secondary analysts will need for planning their work. As just one illustration of the type of metadata sorely needed but impossible to capture in these packages is information about skip patterns. On the one hand, it must be acknowledged that software package designers must feel frustrated at the lack of standards for metadata in the user community. On the other hand, both SPSS and SAS did at one time pace the user community in terms of promoting better and better data definition features.

Variability in Metadata Quality

As just noted, producing metadata for a set of data in 1998 is not remarkably different than in 1968: someone involved in the process of research data management has to do a lot of typing. As a result, the metadata available for a study varies tremendously in quality, ranging from very good for large, government-sponsored efforts such as the census to very poorly for the student who has never been taught the fundamentals of research data management.

There is, thus, a sharp distinction at present between the accessibility of data resources designed for secondary analyses and virtually all other ones. Data sets collected and prepared for the user community as secondary resources are increasingly available via the Web and are slowly becoming more and more accessible to end user as the social science community learns what constitutes a useful interface. As metadata standards become better established and cataloging tools become more sophisticated, we can expect the pace at which these studies are made available to accelerate. Ironically, the pace at which we are losing primary research data is probably increasing. More and more research is published, and we would guess that smaller percentages of it are being archived.

The Future

Our common goal should be nothing less than to create metadata and user interfaces that allow the community of data users to access and process secondary data. Metadata standards, although varied and at times painfully subject specific, have emerged. Our most popular data management and analysis applications, however, continue to lag in meeting the needs of the social science community.

Our recommendation for solving the user interface problem is unchanged from five years ago. We suggest an interactive program shell that allows both the researcher and the end user to:

- Enter information that documents the bibliographic record of the study, including study title, principal investigator, year, funding source, related studies.
- Create topical files that detail study information – including topics such as sampling, copies of instruments, relationships between study data sets, calculation of weights and standard errors, definition of terms, documentation of calculated variables or fields, and originating hardware and software platforms.
- Develop data definition and data manipulation structure – including definition of elements and element formats, complete labeling information, descriptive statistics, and free-field explanatory notes.

A well-developed system would allow the researcher or other person responsible for data documentation to either create a default minimal metadata collection that would provide facilitate subsequent file access, or create very detailed documentation with all study specifics stored as part of the metadata collection.

We also need to work harder at promulgating research data management standards and encourage professional associations, journal editors and funding agencies to require the archiving of research data.

References

Howe, S. R. & Graham, R. G. (1993) Meta-data and user interfaces: Promise and problems. International Association for Social Science Information and Service Technology Quarterly, 17, 4-7.

*Presented at IASSIST/CSS 1998 Conference New Haven, CT., May, 1998. Mark A. Carrozza Institute for Policy Research Steven R. Howe Department of Psychology University of Cincinnati (513) 556-5077 mark.carrozza@uc.edu steven.howe@uc.edu

IASSIST - CAPDU

17-21 MAY 1999
TORONTO CANADA

The International Association for Social Science Information Service and Technology (**IASSIST**) and the Canadian Association of Public Data Users (**CAPDU**) announce their joint 1999 conference, "*Building bridges, breaking barriers: the future of data in the global network*".

The conference will be held May 16-21, 1999 on the University of Toronto campus in Toronto, Ontario and will address issues of computing and information services in social science research, teaching, and data management.

This is IASSIST's 25th annual conference, and the ninth CAPDU conference.