A Functional Approach to Documentation and Metadata

by Stephan Greene[1]
University of Maryland,
College Park

Introduction

The continuing struggle for documentation standards for social science data is reflected not only in the purely archival context, but also in the context of the varied research that makes active use of archived data. The documentation of archived social science data is, of course, a central component in archival practice. Documentation serves to describe archived data for those who gather data in archives, and functions as an aid to the cataloging, and subsequent retrieval and dissemination, of data. For social science researchers, the end users of data, documentation functions as the primary vehicle for gaining an understanding of the nature of raw data. This understanding is critical if researchers are to be successful in their secondary analyses of the data.

The work discussed in this paper is motivated by the desire for documentation to remain useful as the data it describes is transformed through analytical procedures. It addresses some traditionally ignored user-based issues of data manipulation and integrity, how those issues imply structural weaknesses in methods of data documentation, and how the consideration of these unmet requirements can guide the design of improved documentation. This functional approach, emphasizing considerations of practical data use, leads to a rethinking of traditional documentation from that of a generally static reference document to a dynamic entity more appropriately considered a form of metadata. The correct provision and understanding of metadata, any information that adds meaning to the base data (McCarthy 1982), is increasingly susceptible to compromise as the base data move from static printed tables, as often seen in codebooks, to the dynamic environment of the interactive exploration and analysis tools now appearing on computers in a variety of settings. It is within the context of this dynamic functionality that the work described here is most relevant.

In this paper, I will describe my preliminary implementation of the use of metadata processing in data derivation in an electronic historical atlas of demographic data. The implementation experience has suggested useful concepts regarding documentation and metadata. In addition, I will describe a system for producing machine-readable documentation for social science databases developed at the Swedish Social Science Data Service at Göteborg University . This system acts in many ways as an archival response to some of the extant problems of documentation, as it seeks to address both the functional requirements of data users and their varied software tools, and the archival needs of an archival agency. Finally, I will discuss some approaches I have been exploring for further addressing the problems of documentation targeted in this research. The disciplines and their aspects that may be useful are primarily techniques of knowledge representation in artificial intelligence, elements of traditional library science, and considerations of data theory and dimensional analysis. The eventual goal is a formal model of social science data and its analysis which can directly guide the design of documentation and metadata.

A Preliminary Implementation

As the use of micro computers continues to increase, so does the number of computer-aided studies of numeric social science data. In addition to standard commercial spreadsheet, database, and statistical packages, unique data management and analysis applications tailored to specific purposes, which might generically be called demographic or social science data information systems, are beginning to emerge (Miller and Modell 1988). Any thoughtful consideration of the nature of computer-aided studies of social science data must acknowledge that these environments are permeated with numerous opportunities for misinterpretation and incorrect manipulation and analysis of data (Conroy 1994; Greenstein 1994).

The humanities and social science research communities have identified the need to proceed with caution in computer-aided data analysis (Greenstein 1994). To date, these researchers have only their own expert knowledge to guide them in the rational manipulation of data. Yet even the most skilled expert is subject to generating erroneous data, simply by making a small typographical error resulting in an incorrect variable reference. Moreover, the use and potential abuse of social science data now extends beyond the research establishment. Social science data is now accessed in forums such as dial-up public access Internet sites and other consumer-oriented on-line services, which provide simplified access on inexpensive machines (Conroy 1994). The potential for uninformed use of data is thus amplified. The underlying structural reasons for the pitfalls of data use deserve increased attention.

One specific area which threatens the integrity of social science data is in the derivation of new variables from the existing variables of a social science database. Such procedures are typically performed with statistical packages.

The Great American History Machine (GAHM), an interactive historical atlas, is one specialized software product that exhibits the problems associated with data derivation. GAHM provides a browser with 200 years of United States census and election return data at the county level. The program supports the arithmetic combination of basic count variables such as census population counts into derived variables such as rates. Basic descriptive statistics and a choropleth map can be displayed for each basic or derived variable (Miller and Modell 1988). In this particular application, as in many others, it is entirely the user's responsibility to be sure that the derivations he or she performs are correct. Typical errors include dividing one rate by another, or adding a monetary value expressed in thousands of dollars to a monetary value expressed in millions of dollars without an equalizing conversion. The possibilities for error are as numerous as the possibilities for the derivation of new and interesting variables.

To begin to solve this problem, a preliminary assessment of common data types and attributes was gleaned from a selected sample of census and supplemental data sets resident in the GAHM database. A prototype facility for the support of unit control and error checking in the derivation of new variables from existing variables was also implemented. Metadata in support of unit control was identified and entered for a subset of the GAHM database. This enhanced data subset was then used to develop and test the prototype. The data structures and procedures for handling variables in the C-language code for GAHM were augmented to incorporate the new metadata. Similarly, the code that implements GAHM's existing abilities for data derivation was augmented to process the metadata.

GAHM's data derivation is accomplished by an expression evaluator. Through a simple point-and-click interface, users can combine variables from the database within an arbitrary algebraic expression. Like a simple calculator, the expressions may utilize addition, subtraction, multiplication, division, exponentiation, unary minus, and several additional functions like square root. It is the generality and ease of use of this expression facility that make it at once powerful and problematic. It is, unfortunately, quite easy for the user to commit errors in unit matching when combining variables algebraically. For example, users often may sum variables expressed in entirely different units. For a metadata structure to be useful for identifying mismatched units in syntactically correct expressions, it seems intuitive that similar generality is required of the metadata processing used to ensure semantic correctness. Thus, the method by which generality is achieved in basic data derivation was used as a model for the implementation of the use of metadata in data derivation.

The existing GAHM expression evaluator is implemented as a sort of mini-programming language within the program. Internally, the program uses a data structure known as a last-in-first-out queue, or a stack (Wulf, et. al. 1981), for the manipulation of the operands in an expression. The operands on the stack consist of constants, database variable values, and intermediate values occurring as the expression is computed. Metadata was incorporated into this scheme by creating a second, parallel stack for metadata data structures. Thus, in compiling a data stack to represent an expression, general "compile-time" checking of *syntactic* correctness of the expression can be performed. As the evaluation of syntactically correct expressions is carried out, the metadata stack provides for general "run-time" checking of *semantic* correctness.

Under this general scheme, each variable in the test database was specified with the following items of metadata: scale, unit type, unit, and weight. Options for the scale field were ratio, interval, ordinal, and nominal (Stevens 1946). The initial implementation of the use of this metadata in data derivation concentrated on the four basic operations: addition, subtraction, multiplication, and division. In cases where meta-operation routines for checking semantic correctness detect a possible error, a warning is displayed to the user describing the possible source of a problem. The user is given the option of aborting the procedure, or proceeding with the calculation despite the possible error. If a (potentially) erroneous calculation is carried out, the program tags the resulting unit as "not determined." The existing GAHM interface was modified slightly to include display of the resultant unit value to the user as part of its data description display. Greene (1994) discusses the implementation, as well as some relevant technical database issues, in much greater detail.

Results

There are two primary results from this experimental implementation. First, rudimentary support for some common derivations in social science data manipulation was achieved. Derived units were automatically and correctly expressed to the user. Nominal data was treated correctly for this first time in the GAHM application. Second, from a programming viewpoint, the use of a metadata stack parallel to that of the base data stack proved to be an elegant mechanism for providing general semantic analysis of the data.

Most important, however, are the broader implications of these results. The solution, as implemented thus far, suggests an approach that would address the broad issue of managing social science data in a dynamic environment: *metadata must closely follow the data it describes through any transformational procedure* . Metadata must *guide* the application of transformational procedures, and then *describe* the transformed data. It must undergo parallel procedures, tightly coupled with the procedures applied to the base data, that result in transformed metadata that both reflects the transformation of the base data and provides guidance for further transformations of the data.

The principle of integrated metadata has direct implications for social science database management. Social science databases, once compiled, consist of primarily static data. Its associated metadata is also essentially static. If analytic tools are to offer dynamic data manipulation facilities, they must do so for both types of data. Plans for the design and development of software tools that aim to improve semantic support through the use of metadata will benefit greatly by considering both metadata and metadata processing as integral to all data manipulation functions. It should no longer be sufficient to provide information as to the contents of a social science database with a printed document, or even a machine-readable version of that document. This information should be coupled with the database as a structured metadatabase that can be processed in tandem. An integrated, structured metadatabase is dynamic documentation. It is more than an on-line codebook, and it is more than a context-sensitive help system. It is a knowledgeable link between data and the software manipulation of data.

Integrated metadata could and should be used to trace the history of a data analysis cycle. Users should be able to trace links to their original data long after they have been working with measures derived from it. Providing these functionalities should contribute to the development of a “metadata culture” in which data referenced in an apparent vacuum is unacceptable. While users should always be able to override the suggestions of any error checking mechanism in data manipulation, as they can in the experimental case just described, the fact that they must explicitly do so may help force more rigorous defenses of these deviations from mainstream methodology. Integrated metadata can help prevent unintentional errors, and perhaps it can also improve the climate in which decision making from social science data takes place.

There are many limitations, however, to the experimental solution done in GAHM. Additional work lies in the development of a more complete treatment of error checking in data analysis. A litany of concerns must still be accounted for, as described in Greene (1994). The logic of semantic analysis of data and data manipulation is complex, and a more theoretically grounded approach is needed for a comprehensive solution. Archival strategies, then, must remain flexible. It is not yet possible to commit to any one approach to documentation, as data users are currently performing a wide variety of tasks related to social science data.

An Archival Response

The Swedish Social Science Data Service in Göteborg, Sweden, has developed a documentation system called A-Side (Archival System for Interoperable Data Exchange). This system, a UNIX application written in C and using the X11 Window System, produces a family of machine-readable, variable-level documentation formats. The initial process of generating machine-readable documentation with A-Side can be quite resource intensive if no electronic text is initially available. However, once the data is entered, many possibilities arise.

A-Side’s primary output resembles a traditional OSIRIS codebook. Introductory study information is followed by variable descriptions along with residuals and other detailed information pertaining to each of the variables and its supporting raw data. The OSIRIS output file is the archival format that no user will ever see. It is a highly structured, or tagged, ASCII-only (and thus neutral) format. Subject to the life of the storage medium used, and the survival of knowledge of ASCII, the format and meaning of these files can always be understood with careful study, even without the benefit of knowledge of OSIRIS formats. In this context, the rigid structure of the old OSIRIS format is an asset.

More importantly, the format is structured enough to allow for the relatively easy authoring of small utility programs to generate other formats. The A-Side system has recently integrated a number of these utilities and can now, with the invocation of a single menu option, produce HTML codebooks, SPSS setups, as well as several other rich-text formats for printing and on-screen display. The production of SGML formatted files, or files in any other format not yet conceived, should be relatively easy to implement, given the current status of the system. Internally, the A-Side system maintains a great deal of variable-level metadata, and other, richer, primary output formats can conceivably be generated by it. Thus the family of supportable formats is open-ended.

The approach of the A-Side system is characterized by *interoperability* and *sustainability*. Serving diverse needs requires the ability to interchange data and easily generate formats for various purposes. The system is currently positioned to serve a diverse user community, and should scale up well to serve future needs yet to be defined. It is sustainable in that it will always generate a neutral archival format that can be easily adapted to new software and new formats. The system is best viewed as a means to an end, or rather, to many ends, though it does indeed perform some useful core functions for generating documentation. But its most compelling feature is its ability to facilitate the use of data and metadata with other software for more substantive purposes.

Formal Documentation Design

Keeping in mind the idea of integrated metadata, while remaining positioned for whatever formatting requirements the future might bring, we can begin to think about what will improve documentation, and how we can design these improvements. For documentation to become a structured, integrated, and dynamic complement to data, it must be formalized. I believe there are several approaches that may prove to be useful in this effort. The first is that of enumerative taxonomy in the traditional sense of library science. The description of social science data can benefit from an enumeration of the kinds of things such data addresses. Some form of “off-the-shelf” classification should be available to data documenters. This would include some elements of authority control, such that we might begin to see easier data interchange and integration. The provision of a comprehensive characterization of data to which documenters might appeal can ease the process of documentation, and might help support the increased generation of documentation by primary investigators themselves. Geo-spatial data management is ahead of social science data management with respect to authority control and data interchange. Geo-spatial data managers may appeal to entities such as place-name authorities, and data interchange standards are already established.

Second, the realization of functional, operational metadata as documentation can benefit from current work in the development of ontologies for knowledge sharing, a research movement in knowledge representation in artificial intelligence. As used in artificial intelligence, an ontology is a formalized declaration of domain contents. Unlike taxonomic approaches, an ontology specifies domain contents with a *canonical basis* (Sowa 1984), which structures the content more deeply by constraining the types participating in different relationships, in effect creating a concept system rather than merely a list of instances and attributes. An envisioned ontology of social science data would formalize descriptive concepts relating to the entities typically described by social science data, as well as the more interpretive concepts of data and measurement theory, scaling techniques, statistical methods, and dimensional analysis.

A particularly useful example of an ontology is the Engineering Mathematics ontology (called EngMath) developed by Gruber and Olsen (1994). This ontology supports modeling of an engineering perspective of the physical world. This ontology can serve in some respects as a substrate for an ontology of social science data, which will exhibit some mathematical similarity to engineering. Mathematics, as used by both engineers and social scientists, is intended to represent quantitative phenomena. Given that, many of the specific design elements of EngMath are directly applicable to an envisioned ontology for social science data. EngMath formalizes, among other things, conceptualizations of *physical quantity*, *physical dimension*, and *units of measure* . The fact that these three concepts are separated is the key design innovation of this ontology. Physical quantities attempt to represent quantifiable aspects of the real world. Briefly put, “quantifiable” means that the measured entities “admit of degrees” (Ellis 1966) rather than being a yes-or-no attribute, in a qualitative sense. The ability to be quantified also implies the ability to be algebraically manipulated.

The separation of physical quantities from the units of measure used in their expression is a useful abstraction in that physical quantities are fundamental notions, while units of measurement are merely matters of convention. Distinguishing the conceptualization of units of measure from the conceptualization of physical quantities allows for the expression of physical quantities without committing to a particular system of units, of which there are several in common use. More importantly, the distinction between the two concepts supports the straightforward conversion from one conventional system of units to another, which in turn supports the comparison of similar physical quantities that are expressed with different units of measure.

Dimensional analysis, developed as a technique for analyzing the behavior of physical systems, informs the evaluation of mathematical operations on physical quantities. The separation of dimensions from quantities in the engineering mathematics ontology supports direct application of the principles of dimensional analysis. That physical quantities are characterized by physical dimensions is what distinguishes them from abstract numeric entities. The distinct conceptualization of physical dimensions provides many advantages. Useful algebraic or comparative operations on physical quantities must exhibit *dimensional homogeneity*. For example, it is meaningless to compare a measure of mass against a measure of length, or to add a measure of time to a measure of temperature. The distinct notion of physical dimension allows the enforcement of dimensional constraints independent of particular instances of physical quantities or units of measure.

EngMath provides a good example of a formalization of domain contents. A unit conversion program has been written, using EngMath, to facilitate the interchange of data about physical quantities in engineering. EngMath thus helps solve problems with engineering data that are similar to the problems of managing the manipulation of social science data. There are limits, however, to the degree to which mathematics, as formalized by EngMath, will support the mathematics of social science data. The limits appear primarily in the area of dimensionality. Engineering mathematics enjoys general agreement among its users with respect to basic dimensions such as mass, time, length, and temperature. More complex notions of dimensions, such as force, are expressed in terms of the basic dimensions. In social research, there is little agreement on what to measure, and the dimensions of humans and their artifacts and activities are less well defined. Within the formal abstract algebra used in the engineering ontology, important social science quantities, such as persons or dollars, have as their dimensions the “identity dimension”, or in the terminology of dimensional analysis, they are *dimensionless*. Research in human geography and social science data theory (Haynes 1975; Jacoby 1991) has identified the need for further investigation into the current limits of dimensional understanding. Dimensionless count data constitute a significant proportion of available social science data.

Beyond the specific example of EngMath, emerging design principles for ontologies (Gruber 1993) specifically meant for knowledge sharing dovetail nicely with the requirements of structured metadata as documentation for social science data. The first design principle is that of *clarity*. The formalism required in ontology design will enforce clarity in the definitions of the concepts of social science data. Greater clarity and specificity in the collection and dissemination of social science data will always be welcomed.

The principles of *monotonic extendibility* and *compartmentalization* are related guidelines for ontology design. An understanding of the purposes for which a conceptualization will be used should inform the design process. Users of an ontology should be able to extend it for their own purposes monotonically, that is, without requiring changes to the base ontology. Compartmentalization supports the monotonic approach to extendibility. Where it is possible, an ontology should be broken down into component ontologies. This allows users to select those components that are useful, without being forced to inherit those that are not. As users extend ontologies for their own purposes, their extensions should likewise be compartmentalized. Other users may then access extensions, with the same benefits. Given the diversity of methods in data collection and analysis in social science research, an approach that emphasizes a basic core conceptualization that is agreeable to most potential users, and that can be extended as needed without alteration, is inherently appealing.

Another design principle is *minimal encoding bias*. This principle can also be thought of as the *parameterization of convention* (Gruber and Olsen 1994). Many descriptive elements, such as units of measurement or natural language titles, are merely matters of convention, rather than basic concepts. The use of conventional terms is discouraged. This design principle helps support the goal of inter-agent communication and knowledge sharing, as it encourages conceptualization of core concepts independent of the conventions typically used to describe them. This applies directly to the needs of social science data, where data sets collected by different agencies differ radically in their conventions, such as their methods of naming variables, or the systems of units used to report data. Systems of measurement, which are conventions, must be parameterized. Doing so will support efforts toward data integration and interchange.

Another principle of ontology design is *minimal ontological commitment*. This principle stresses designing the weakest conceptualization possible to support the purposes for which it is designed. For example, to the extent that it is possible, formalized descriptions of data should exclude interpretive elements. Doing so maximizes the number of researchers that will agree to use the descriptions, as such descriptions would not require commitment to particular interpretations.

The principle of minimizing commitment works hand in hand with compartmentalization, which also reduces the degree of commitment required by any single ontology by encouraging small, modular ontologies. The minimization of ontological commitment may begin to address some of the structural problems that have plagued the search for generally acceptable methods to describe social science data. Controversies abound over the design, collection, interpretation, integration, and analysis of social science data. To the extent that it is possible, an ontology should support data interchange without requiring absolute methodological harmony.

Conclusion

Modern, dynamic data access and manipulation presents new challenges for the processes of collecting, describing, disseminating and analyzing social science data. The experimental solution developed in the context of GAHM shows promise, and suggests some broader principles that may guide the design of data description in a dynamic environment. Archival documentation strategies must be flexible and adaptable, and the A-Side system attempts to meet current needs, while remaining positioned to meet future needs once they become more clear. Documentation standards are difficult to define given the diversity to be found in all aspects of social science data management. Forms of documentation and methods of documentation production that support interoperability and interchange of information will provide the most benefit in the long run.

The design of new forms of documentation that begin to meet some of the concerns of data description outlined in this paper can benefit from considerations of taxonomy and ontology. The goal is to formalize, as much as possible, the contents of the domains of social science data. The resulting data models are then available to guide the generation of documentation. The design of formal conceptualizations of social science data can begin with a comparison to the treatment of engineering mathematics. Such a comparison, guided additionally by consideration of social science data theory, as well as dimensional analysis as applied to social science data, will help to isolate and eliminate weaknesses in the handling of social science data and lead to formalized ontologies, and thus similarly formalized metadata and documentation, for social science data.

References

Conroy, Cathryn. 1994. People, by the numbers. *CompuServe Magazine*. October, 32-36.

Ellis, B. 1966. *Basic Concepts of Measurement*. London: Cambridge University.

Greene, Stephan. 1994. Metadata for Social Science Data Derivation: A Preliminary Approach. Manuscript.

Greenstein, Daniel I. 1994. *A Historian's Guide to*

Computing. Oxford: Oxford University Press.

Gruber, Thomas R. 1993. Toward principles for the design of ontologies used for knowledge sharing. Technical report KSL-93-04, Knowledge Systems Laboratory, Stanford University.

Gruber, Thomas R. and Gregory R. Olsen. 1994. An ontology for engineering mathematics. Technical report KSL-94-18, Knowledge Systems Laboratory, Stanford University.

Haynes, Robin M. 1975. Dimensional analysis: some applications in human geography. *Geographical Analysis* 7: 51-67.

Jacoby, William G. 1991. *Data Theory and Dimensional Analysis*. Newbury Park, CA: Sage.

McCarthy, J. L. 1982. Metadata management for large statistical databases. In *Proceedings of the Eighth International Conference on Very Large Databases*. Saratoga, CA: VLDB Endowment.

Miller, David W. and John Modell. 1988. Teaching United States history with the Great American History Machine. *Historical Methods* 21(3):121-134.

Sowa, John F. 1984. *Conceptual Structures: Information Processing in Mind and Machine*. Reading, MA: Addison-Wesley.

Stevens, S. S. 1946. On the Theory of Scales of Measurement. *Science* 103(2684):677-680.

Wulf, William A. et al. 1981. *Fundamental Structures of Computer Science*. Reading, MA: Addison-Wesley.

All scale types except ordinal actually occurred in the test database.