Metadata, The Frame of Reference For Future Survey Research

Raymond B. Fagan, Securities Automation Industry Corp. (SIAC)
80 Woodcliff Ave., Woodcliff Lake, NJ, 07679

Key Words: Metadata, Data Warehouse, Multidimensional Indexes, Frame of reference, Survey Research

A) Metadata, The Last Dimension of Survey Research, has shown us how three governmental statistical agencies are improving the quality and reliability of their statistical data. Daniel W. Gillman et al. (1996) discussed the use of new standards, technology and equipment to increase customer access. Gordan Priest (1996) reviewed some current integration problems including disharmonies, stove-pipe analysis and contradictory output. Customers demand varies now due to a shift in technology (the Web/Internet) highlighted by Priest. Bo Sundgren et al. (1996) presented "The Swedish Approach to Adding Quality to Official Statistics."

As a business user, my focus will be on how these governmental agencies should assist commercial research. Specifically, what is the relevance of their metadata in the commercial environment?

B) Metadata must provide the frame of reference for all future studies. The definition of metadata depends upon the user's view or prospective. Three different approaches to metadata were presented. Each had a specific frame of reference summarized as follows:

1) The administrator's - customers will not understand how to use the data; therefore, it is necessary to teach them how to use it.

2) The researcher's - the quality of the study's data has not been defined enough to avoid repetition of the study. Clear definitions are required to appropriately use the data.

3)The standards committee member's - define everything about the data according to standard "x," so that the customer will understand it.

Although their concerns are legitimate, they do not necessarily represent current business priorities. The corporate user knows how to interpret the data, but the statistical data generated may not always be of relevance in commercial applied research.

I will present the metadata requirements from an information services (I.S.) viewpoint "*Why do we need it.*" From a commercial perspective, metadata must answer the five questions (*who, when, from where, why and how*) according to Tannenbaum (1994). More recently, Kimball (1996) stressed that each specific view must be defined at the lowest row level that is accessible to the company or researcher. Metadata must provide the customer with sufficient information to allow a non-study participant to use the dataset to establish reliable correlation's and/or multidimensional indexes.

Morris (1996) indicates *synchronization,* maintaining the frame of reference between the operational data and the decision support data, will become a major issue in implementing data warehouses. My colleagues' papers clearly indicate that customers' demands are now being heard. Customers will no longer tolerate poorly defined statistical data that cannot be easily manipulated or correlated. Furthermore, they will not pay for the statistical data that cannot be efficiently used. Implementing data warehouses has become a major computer industry problem. Metadata has now become a long term financial and a customer service issue.

Business needs metadata that provides unique keys to identify a study's data so different datasets can be properly compared. The frame of reference or context; plus, the multidimensional views (indexes)that a data warehouse requires for performance must be defined by metadata. Future researchers can only investigate the different viewpoints on a study's data if the multidimensional indexes have been created.

Metadata provides the customer with a frame of reference so he or she is able to:

1) Understand the data

2) Correlate it to his own variables

3) Implement a new index from a table-join (integration) of two different studies data

4) Uniquely define the context of the study's data versus that of the previous studies.

Metadata must create both the keys and the pathways to existing legacy data for future availability. Yet, the contextual information within a data warehouse repository must also be capable of adaptation to multiple locations, to hardware and software platforms and to new theoretical hypotheses. Current efforts are just scratching the surface. Repositories are not yet capable of defining multidimensional indexes for their legacy data. Refreshingly, government agencies are now providing new tools for their customers to facilitate data scrubbing and data mining (extraction's).

C) Metadata Limits (Assertions and Constraints): The other functional requirement of metadata is to define both the production oriented characteristics and the end user dissemination constraints for each study's data. Priest's paper addressed reporting bias, confidentiality and suppression rules that affect statistical data. To prevent data misinterpretation, the study's limitations and validity constraints must be clearly identified at the row level.

D) Moving toward a Metadata Repository: The future is now! Customers do not want to look through statistical data to figure out what they have or want. In the last three years, many fortune 500 companies in the United States have started major data warehousing efforts to mine existing transactional legacy data to obtain new competitive marketing advantages (Bruno, 1995).

Future researchers must apply new survey guidelines to their related metadata efforts (Paul Hessinger, 1995). Such guidelines require that other researchers must be able to: 1) *leverage the existing metadata to support cross_functional analysis* by providing links (HYPERTEXT or HTML) to alternate source materials until a repository is available.

2)*Refocus on data administration and standards* to scrub the existing stovepipe legacy study's data so that "what if analyses" can be performed.

3) *Adopt object-oriented modeling techniques* so each data object reported is truly independent. Objects must be defined with all the methods (protocols) and procedures (methodology) linked to it.

E) Statistical Metadata Repository: Gillman et al. (1996) discussed how the U.S. Bureau of the Census (BOC) plans to address reporting within their proposed statistical metadata repository (MDR) for both macrodata and microdata. Macrodata refers to three categories: systems, applications and administrative. Microdata refers to the actual study characteristics. The current modeling effort using Open Workgroup Repository (OWR) from Manager Software Products (MSP) was also identified.

To clarify how his metadata will be defined, Gillman et al. (1996) reviewed the Statistical Design And Survey Methodology Metadata Content Standard *(SDSM)* of US Government. The national and international standards also being used in the repository development were discussed. The key standards involved were:

1) Reference Model for Data Management (RMDM) ISO 10032:1995(E)/Command Manipulation Language (CML) 1995

2) Spatial Data Transfer Standard, 1994, ANSI/ISO Z39-50

3) Specifications for Data Elements, ISO 11179, (Part 1-6) Standards Committee X3L8.1996.

4) Information Resource Dictionary System (IRDS) FIPS Publication 156, 1989.

Data Access and Dissemination System (DADS) and Federal Electronic Research and Review Extraction tool (FERRET) query tools were discussed relative to their use in scrubbing the metadata for the available legacy datasets. But, these legal standards currently do not address the requirements of business data in marketing research, business plans and SEC reporting. For example, the omission of the legal chain of evidence rule (i.e. who has possession of this data from its origination to the present time), inhibits legal use of survey data. Several metadata parameters that business needs were not clearly discussed.

F) The Customer's Metadata: Statistical agencies must provide metadata in order to prosper in the future. These agencies must be able to answer the following business questions:

1) *How do I find the specific data?* Agency metadata must be cross referenced by non-technical key words. There must be time, geographic, political, business and economic cross references for each data item object presented.

2) *Where is it?* The location where the business may access the data desired is also critical to the user. This metadata must cross reference not only the geographical address or physical location but it also must provide the electronic "cyber" location for the data. At the least, the universal resource location (URL) of the host computer must be provided. Preferably, a contact or E_mail address where the data may be obtained should also be given.

3) *What does it contain?* The description of the dataset is probably the most difficult to develop because it is dependent, upon the perspective of the user. Different attribution maybe required by different end users due to their perspective or viewpoint. The traditional statistical variables such as time, place, sample size and methods must also be given. In addition, the multiviews or alternate frames of reference for the same dataset will also be a significant benefit in future surveys.

4) *How do I access it?* The customer's initial requirements are how to obtain the actual data. Three levels of metadata are necessary to answer this question:

a) *System: (How is it stored?)* What is the currently available electronic storage media? In what operating system, is it (DOS, Windows and UNIX)? Under what graphical user interface (GUI) and in what file format (EBIC, ANSI, DOC, PDF, etc.) is it provided?

b) *Application (What software was used to prepare it?)* What information is available and how was it produced?

c) *Administrative* (What is unusual about this data?) Was anything done to data to conform to governmental regulations, etc.?

5) *What should I know about the survey's resultant data?* Why was this survey done? What was the purpose of the study (output oriented, internal use, end user request)? Was any bias shown in the study's design?

G) Conclusions: The issues and concerns of the business user must also be included in the new agency metadata in order to address the needs of the commercial world.

In summary, metadata must contain:

1) Non_technical terms.

2) Alternate unique multidimensional keys.

3) Hypertext links to the study's methods and procedures.

4) Disclosure of any confidentiality or suppression rules.

5) The study's purpose and design limitations.

H) References:

Bruno, D., (1995), Platinum Technology Inc., "Moving Legacy Data into the Warehouse," Edge Magazine: Extra, 1995.

Gillman, D. W., Appel, M.W., and La Plant, Jr., W. P., (1996), "Statistical Metadata Management: A Standards Based Approach," 1996 Joint Statistical Meetings, Chicago, 1996, (draft)

Hessinger, P., (1995), Platinum Technology Inc., "Organizing and Leveraging Metadata," Edge Magazine: Extra, 1995.

Kimball, R., (1996), The Data Warehouse Tool Kit, John Wiley, New York, 1996.

Morris, H., (1996) International Data Corp., "Blueprint for a Data Warehouse," Computerworld, June 3, 1996. (White Paper).

Priest, G.E., (1996), "In Search of Data Integration: No Matches Found," 1996 Joint Statistical Meetings, Chicago, 1996, (draft).

Sundgren, B., and Dean, P.,(1996), "Metadata: Quality Element in Official Statistics-The Swedish Approach," 1996 Joint Statistical Meetings, Chicago, 1996, (draft).

Tannenbaum, A., (1994), Implementing a Corporate Repository, "The Models Meet Reality," John Wiley, New York, 1994