WP. 19
ENGLISH ONLY

**UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS**

UNECE Work Session on Statistical Dissemination and Communication
(12-14 September 2006, Washington D.C., United States of America)

Topic (iii) How to present metadata

UNDERSTANDING THE ROLE OF METADATA IN FINDING AND USING STATISTICAL INFORMATION

Invited Paper

Submitted by University of Washington, Tacoma; USA[1]

I. THE IMPORTANCE OF METADATA IN UNDERSTANDING STATISTICAL INFORMATION

1. It is understood that metadata availability can enhance retrieval processes, improve online information organization and navigation, and facilitate user understanding of online objects. Our work focuses on developing an understanding of how and when users utilize metadata in order to model the resultant metadata requirements to support electronic access to and use of statistical information. Over the last several years, we have conducted a series of users studies on which we report here.

2. Metadata play a central role in the finding and interpretation of statistics. Metadata, such as the units used to report the data, the time period to which the data refer, the question the interviewer asked to generate the data, the sampling and non-sampling error of the statistics, and so on, are all critical to understanding the meaning and potential usage of a given statistic. It may also be necessary to understand the underlying concept that a statistic is intended to represent, the reason a particular statistic was produced, the history of the instruments used to produce the statistic, and so forth.

3. We have conducted three systematic investigations into usage of metadata and are currently engaged in a fourth. These are studies of:

1. Metadata requirements for understanding tables based on understanding the uncertainties users experience during table usage (Hert & Hernández, 1999).
2. Metadata requirements in integration tasks (Denn, Haas, & Hert, 2003).
3. One integration activity, the making of comparisons, investigating types of comparisons made and the rules experts employ while making comparisons (Hert, 2004).
4. The key metadata elements used in selecting potentially useful information objects.

The overall purpose was to inform the development of metadata infrastructures, such as XML schemas, to be used to support statistical information dissemination and use processes.

[1] Prepared by [Carol A. Hert, University of Washington, Tacoma, cahert@u.washington.edu; Sheila O. Denn, Simmons College, sheila.denn@simmons.edu].

II. STUDY ONE: USER UNCERTAINTIES WITH TABULAR STATISTICAL DATA: IDENTIFICATION AND RESOLUTION

4. The intent of the first study was to provide insight into metadata elements necessary to support usage of tables as well as understand the potential difficulties in attaining that metadata. This work provided a baseline categorization of types of metadata needed to enable users to resolve their questions associated with statistical tables. This project addressed the following questions:

- What questions and uncertainties do users have when investigating the statistical tables used in the study?
- What are the answers to these questions?
- To what extent is metadata available to answer the questions? What metadata would be necessary to answer these questions?

5. A total of 169 questions/uncertainties was identified. The categorization scheme had eight main categories: '*Definitions needed', 'Table structure', 'Rationales needed', 'Survey methodology information needed', 'What's the difference', 'Interpretation of table needed', 'Information currency information needed*' and '*User uncertainty is not clear*'. The majority of uncertainties (155 of 169 or 91%) are in the first four categories mentioned. These categories included uncertainties at several different granularities (e.g., for a term, about a survey's methodology, or for a specific table) and that covered the range from statistics in general, to survey methodology, to presentation.

6. The researchers attempted to find answers to all user questions. Only 59 of the 169 questions (35%) had an answer available online. The 59 include two questions that had supplementary information provided by an expert. A cluster of categories had rates of online answers ranging from 43% to 56%. These were: '*table structure*' (43%) '*survey methodology information needed*' (43%), and "*category inclusion information needed'* (56%). A second cluster ranged from 29% to 36%. These were '*term definition needed*' (29%), '*tool functionality'* (a sub-category of table structure) (29%), and '*rationales needed'* (36%). Forty-eight percent of the questions (81 out of 169) needed the assistance of an expert to answer and for 17% (29 out of 169) no answers were found or given by an expert.

III. STUDY TWO: METADATA NEEDS DURING INTEGRATION TASKS

7. The second study's goals were to reveal the kinds of issues that come up when users are completing integration tasks using statistical data, to identify the metadata elements associated with these issues and challenges, and from that information to begin to construct a metadata architecture that supported the metadata elements.

8. We described user interactions in terms of a set of key themes or "stories" that expressed a number of perspectives on the users' experiences and metadata usage. The themes were:

- User knowledge: People exploited domain knowledge, task knowledge, knowledge of statistical processes and agencies, and expectations about the availability and possible currency of information.
- Surveys and statistics: Granularity of data issues, different measure types, and manipulations (such as seasonal adjustments) all played a role in interactions.
- Interpretation of information: In interpreting individual statistics, expert users were sensitive to the date, units of measurement, associated footnotes, whether the data were preliminary, final, or revised, and so on. The comparability of dates, geographic units, data sources, variable definitions, and data collection strategies (survey vs. census primarily) all became important.
- Date issues: The timeliness, or currency, of the statistics that participants were attempting to use. The amount of difficulty that this issue caused for study participants was related to how much prior knowledge they brought to bear about particular agencies, their surveys and

censuses, and the frequency with which those surveys and censuses are conducted. Reconciliation of dates associated with multiple statistics was an important integration activity. This was difficult for the participants, because information about the intervals at which particular statistics are reported, and how quickly they are disseminated to the website after the reporting interval has passed, is not readily available.

- Geography: Three aspects were identified: 1) user knowledge of geography, 2) the available geographic granularity of a given statistic, and 3) the role of geography in integration activities.
- Navigation and information layout
- Terminology: Users' ability to understand or use specialized terminology or map it to more common concepts.
- Integration activities.

A. Integration Activities

9. We captured data about the process of integrating information, as well as barriers to successful integration. The most important integrating activities were 1) making comparisons 2) noting discrepancies (between data, in presentation approach, etc.) and/or reasoning about that difference and 3) manipulating statistics (e.g., mathematical, exporting to spreadsheets). Of these, the comparison activity appeared to be the core integrating activity. We identified a number of comparisons common across our participant pool; these types were:

- Comparison across geographic units,
- Comparison when there are definitional differences across concepts and variables,
- Comparison across units of time,
- Comparison across different sources (websites, surveys, censuses, reports, etc.),
- Comparison across index values.

10. We identified barriers to the successful integration of statistical information. These were:

- Lack of definitions or source information
- Lack of user knowledge of appropriate strategies (e.g., using time series data, types of calculations to perform)
- Lack of user knowledge about usage of index values, statistical activity purpose and approach
- Interface design problems (such as scrolling row and column headers)
- Inconsistent data across sources
- Inconsistent interfaces
- Inability to determine whether data wanted for comparison are available
- Lack of domain knowledge
- Lack of knowledge of how to handle domain terms such as inflation, seasonal adjustment
- Terminology differences

IV. STUDY THREE: THE METADATA REQUIREMENTS TO SUPPORT STATISTICAL COMPARISONS

11. To further understand comparisons and how metadata supports users making comparisons, we undertook a third study. The study's intent was to identify the 1) types of comparisons made or attempted by users and 2) the information that experts bring to bear on the making of those comparisons.

12. The experts identified the following comparisons as being relevant to end-user usage of their statistics:

- Comparison across geographic units,
- Comparison when there are definitional differences across concepts and variables,

- Comparison across units of time (data collected/reported at different times, different aggregations (e.g., quarterly, monthly), and different reference periods),
- Comparison across different sources (websites, printed sources, etc.),
- Collection approaches (survey vs. census, household vs. establishment, etc.),
- Comparison across index values,
- Terminology comparisons (same word-different concepts, same concept-different words),
- Comparisons with deflated vs. real dollars, corrected vs. uncorrected data, preliminary vs. final release data, seasonally adjusted data vs. non-adjusted data,
- Comparisons involving data with differing confidence intervals,

13. The experts also articulated some common questions or aspects of the data that they need to understand before attempting a comparison. These are framed as questions to ask below:

- Has there been a change in classification of the variable over the time period of interest?
- Are the data final?
- Are the data being compared from different time periods?
- Are the data being compared from different geographic units with the associated rule of thumb: get units geographically close to unit of interest?
- Is the comparison among index values?
- Is the comparison among measures that already have a comparison built in (e.g., percent change, percent distribution, rates, ratios)?
- Is the comparison among variables that are commonly conceptually confused (e.g., occupation and industry, employment rate and unemployment rate)?

These questions provide guidance as to necessary metadata elements for facilitating comparisons.

V. STUDY FOUR: KEY METADATA ELEMENTS IN ASSESSING RELEVANCE

14. The motivation for our current study starts with the fact that metadata creation is generally an expensive and labor-intensive process. Thus one wants to prioritize metadata creation by: 1) assigning the metadata elements that are most often used (for a particular task such as information discovery) and/or 2) assigning greater or lesser amounts of metadata based on the "value" of a given entity.

15. Identifying relevant items (after retrieval) is a bridging step to statistical information use. Thus facilitating a user's ability to judge whether items are potentially useful is a precursor to use. In this study we ask the question:

> What metadata elements enable a specific type of user (to be designated) to assess potential relevance of an entity to a specific task?

16. The basic logic of the study is to ask users to select, from a set of entity representations, those entities that seem relevant for the task at hand. By relevant, we mean "fit for the task at hand," and we will refine that further prior to undertaking the study. By asking users to discuss what "cues" they used to make the judgments about selection, we learn about what metadata elements are useful. In this study, we will ask them to make judgments, then look at the selected entities themselves and re-judge the potential relevance. We expect that in discussing their judgments in the second step, they will point to aspects of the entities that indicate what constitutes value in addition to expressing more about the metadata "cues". We intend to report on results at the meeting.

17. All of the work outlined above has bee undertaken with an eye toward building a model of statistical metadata from the perspective of the end user of statistics, rather than from the perspective of the producer of statistics. As statistical agencies focus on making their data more accessible to the end user population, these efforts are important in providing a basis for prioritizing both the metadata elements that are provided as well as the statistical information objects about which metadata will be

provided. This is especially an issue for retrospective assignment of metadata to pre-existing statistical information objects.

VI. ACKNOWLEDGEMENTS

18. Funding for this work came from National Science Foundation In-Collaboration Grants EIA 0131824 and EIA 0129978 and from the United States Bureaus of Labor Statistics and Justice Statistics (2005-BJ-CX-K016). The opinions expressed are those of the authors.

VII. REFERENCES

19. Denn, S.O.; Haas, S.W; & Hert, C.A. (2003). Statistical metadata needs during integration tasks. 2003 Dublin Core Conference: Supporting Communities of Discourse and Practice—Metadata Research & Application. DC-2003: Proceedings of the International DCMI Metadata Conference and Workshop (Sept. 28 – Oct 2, 2003: Seattle, WA). pp. 81-90. Available at: http://www.siderean.com/dc2003/301_Paper50.pdf. Accessed 11/3/2004.

20. Hert, C.A. (2004). The Metadata Requirements to Support Statistical Comparisons. Govstat Technical Report. Available from the author.

21. Hert, C.A. and Hernández, N. (2001). User Uncertainties With Tabular Statistical Data: Identification And Resolution: Final Report to the United States Bureau of Labor Statistics (Purchase Order #B9J03235). Available at http://ils.unc.edu/govstat/fedstats/uncertaintiespaper.htm Accessed 11/4/2004.