Supporting End-users of Statistical Information: The Role of Statistical Metadata in the Statistical Knowledge Network[1]

Carol A. Hert[2] and Stephanie Haas[3]
Syracuse University, University of North Carolina-Chapel Hill
(cahert@syr.edu; stephani@ils.unc.edu)
Project URL: http://ils.unc.edu/govstat

1. Abstract

Metadata is integral to many processes in the Statistical Knowledge Network (SKN). Recent efforts such as the Semantic Web (Berners-Lee et al., 2001), the joint work of ISI and Columbia University (Ambite et al., 2002), and ours (Marchionini, et al., 2003) express the importance of metadata in supporting integration of information across multiple sources. This paper adds to those efforts by highlighting two aspects of using metadata to help users "find what they want and use what they find." The first is to identify integration challenges that users experience and metadata to resolve them. The second is to establish an architecture that supports metadata acquisition and use throughout the SKN.

2. Introduction

To illustrate metadata issues, consider the scenario of a student who needs to find out how the economies of her county and state compare to that of the United States as a whole. Table 1 shows some statistics that she would find on March 3, 2003. The student needs to understand what statistics might help, as well as understand the statistics retrieved. She must begin to understand what makes the numbers different, whether she can compare 1999 with 2003 data, whether seasonally-adjusted numbers can be compared to

Economic Indicator	Yakima County	Yakima MSA	Wash. State	United States
Unemployment rate	11.1% (2000 Census) 12.7% (Jan. 2003 prelim, not seasonally adjusted) (Washington State website)	12.1% Dec. 2002 not seasonally adjusted (BLS)	6.2% (2000 Census) 6.6% Jan. 2003 prelim. seasonally adjusted (Wash. Website) 7.4% Jan 2003 not adjusted (Wash. website) 7.0% Dec. 2002 seasonally adjusted (BLS)	5.8% (2000 Census) 5.7% Jan 2003, seasonally adjusted (BLS)
Poverty Rate	21.9% (1999, for families, from 2000 Census data)		7.3% (1999, for families, from 2000 Census data)	9.2% (1999, for families, from 2000 Census data)

Table 1: Example Economic Indicators for Yakima, Washington State, and the U.S.

[1] Our grant (NSF EIA 0131824) relies on the help and resources of the Bureau of Labor Statistics, Census Bureau, Energy Information Administration, National Agricultural Statistics Service, National Center for Health Statistics, National Science Foundation Science Resources Studies, and Social Security Administration. Our project team consists of Stephanie W. Haas (University of North Carolina (UNC)), Carol A. Hert (Syracuse U.), Gary Marchionini (UNC), Catherine Plaisant (University of Maryland-College Park (UMD) and Ben Shneiderman (UMD). Sheila Denn (UNC) and Jenny Fry (UNC) were instrumental in conducting the user study.
[2] School of Information Studies, Syracuse University, Syracuse, NY 13244-4100
[3] School of Information and Library Science, University of North Carolina-Chapel Hill, CB #3360, 100 Manning Hall, Chapel Hill, NC 27599-3360

non-adjusted, and so on. The scenario expresses opportunities for technical integration (gathering and coordinating the data/metadata and identifying discrepancies) as well as intellectual integration issues.

3. Integration Challenges From The Users' Perspectives

The scenario above is from the team's study of users engaged in integration tasks. The study's [4] goals were to identify what aspects of integrating data present challenges as well as behaviors used during integration. During winter 2002-2003, we observed 24 users engaged in tasks having the potential for exposing these challenges[5]. Tasks involved using data from multiple agencies, usually combining at least two topical or geographic components. From our observations we identified user behaviors as well as user problems.

Users attempt to compare statistics:

- From multiple geographic units at different granularities (city, state, region for example).
- From geographic units that a user judged identical (such as a county and a Metropolitan Statistical Area (MSA).
- Generated when there are definitional differences in concepts and variables (such as two different variables that both provide a statistic for unemployment rate.)
- Across time (such as using time-series data or when statistics are from different years).
- Generated from different sources or found in different sources.
- That are index values.

There are a number of pitfalls for a user on the way to an integration behavior. For example, if the user does not know the geographic location of an entity, it will be impossible to "click on the map" to find information on that entity. Lack of knowledge about whether a given statistic actually exists will prevent a user from using it in a comparison. Thus, understanding what problems users experience, especially in the context of integrating behaviors is critical. Typical problems included inability to find definitions, to determine what geographic granularities were available for a given statistic, to know whether statistics actually existed, and uncertainty about the utility of a statistic or comparability of many statistics.

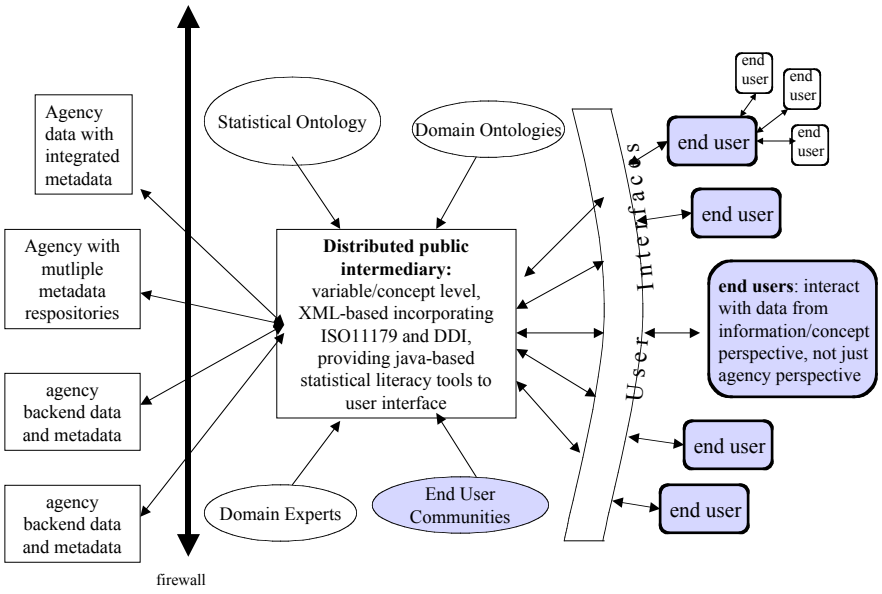
The behaviors and problems provide critical clues to the statistical metadata necessary in the SKN. Problems with meanings of terms can be resolved with definitional metadata, geographic relationships need geographical coordinates but also definitional information (how does Yakima county overlap with the MSA), questions about methodology need technical documentation, and so forth. In this study and a previous one (Hert and Hernandez, 2001), the team identified challenges to acquiring metadata. Agencies may have metadata but in hard-to-access formats, metadata may exist only in an agency analyst's head, or there may be multiple sources. Metadata to support integration activities is likely to be difficult to acquire in final form as it may involve metadata across agencies or be intertwined with specific expertise and knowledge. Addressing these problems requires an SKN architecture that supports extraction of metadata from multiple sources, the creation of additional metadata, and its incorporation into end-user tools.

4. Movement Of Metadata In The SKN: From Agencies To User Tools

Figure 1 shows the general architecture our project is using to express the SKN infrastructure and thus serves as the basis for metadata transfer efforts. The public intermediary provides functions connecting

[4] Study details including methodology will be available at our website: http://ils.unc.edu/govstat.
[5] As determined by pre-testing and discussions with agency personnel.

statistical agencies to end-users. For metadata, it provides the conduit by which agency metadata is extracted, pre-integrated (if appropriate) and transferred to specific user tools. The public intermediary might also include derived metadata such as a set of rules concerning appropriate comparisons for consultation when data with discrepancies are retrieved. Particular tools would incorporate selected metadata and be able to present them, via an interface, to users. Efficiency of metadata transfer and integration processes is facilitated by use of standard formats (such as ISO11179, DDI) throughout the



SKN.

Figure 1: The Components of the National Statistical Knowledge Network

A Statistical Interactive Glossary (SIG) illustrates this how metadata might flow through the network. The SIG is an enhanced glossary of common statistical terms, supported by the GovStat ontology (Haas et al., 2003), whose goal is to help users understand statistical terms and concepts in the context in which they are encountered. Explanations of terms thus incorporate entities and statistics from the users' current work context. The glossary is limited to terms users are likely to encounter. The content of the explanations may include definitions, examples, brief tutorials, etc. in a variety of formats including text, text plus audio, still images, animation and interactive presentations.

Based on the glossary framework, one can specify the necessary metadata, the source of that metadata, and standards that might support structuring and transmitting that metadata (Table 2). On examination, one can see some informational/metadata components (term definition, term name) that map to existing standards but these are not sufficient. SIG must include agency-produced metadata as well as derived metadata (such as the GovStat ontology). The type of analysis shown in Table 2 enables one to track metadata as it is drawn from agencies (presumably in standardized formats), is integrated with other metadata (the ontology, for example) and then migrated to specific tools. It promotes exploitation by end-user tools of the system(s) of metadata currently maintained by agencies and other components of the SKN. Similarly, it highlights gaps in existing metadata, indicating where metadata must be sought from other sources, or created from scratch.

Information/Metadata Needed	Task(s) using this metadata	ISO1179 or DDI map	Comments	Source of metadata
Term name (s)	Search, presentation, and anchor linking presentation	ISO11179 data element name (for variables)		Agency content, GovStat ontology
Definition	Provide content	ISO11179 data element definition (variables)		Agency documentation, statistics experts, statistics texts, GovStat ontology
Examples, demonstrations, etc.	Provide Content	None within ISO11179 or DDI	Values: Audio, video, static text/graphic; Links to more specific agency documentation	Agency content supplemented by designer
Context specificity level of definition (e.g. statistic, table, agency)	Provides specific explanations	None within ISO11179 or DDI	user control—needs context information	User's current webpage or table, GovStat ontology
Format	Type of presentation	None within ISO11179 or DDI	user control	Current user interaction or preset preferences, computing capabilities, GovStat ontology

Table 2: Sample Metadata Requirements for Statistical Glossary

5. Conclusions

Statistical metadata are essential to the development and utilization of the SKN. Our work has shown situations in which metadata are necessary and we are prototyping tools to support users in these situations. Our investigations are providing a model to guide our metadata work and beginning to specify approaches to moving metadata in the SKN. In the next two years, we will further specify the model, provide interface tools based on the model and to enhance the model's generalizability to the entire SKN.

6. References

Ambite, José Luis, Yigal Arens, Eduard H. Hovy, Andrew Philpot, Luis Gravano, Vasileios Hatzivassiloglou and Judith Klavans (2002). "Simplifying Data Access: The Energy Data Collection Project". *IEEE Computer* 34(2): 47-54.

Berners-Lee, T.; Hendler, J.; Lassila, O. (2001) The Semantic Web. *Scientific American*. May 2001. Available online at http://www.sciam.com/article.cfm?colID=1&articleID=00048144-10D2-1C70-84A9809EC588EF21 (accessed 3-5-03).

Haas, S.W.; Pattuelli, M.C.; and Brown, R. T. (2003). Understanding Statistical Concepts and Terms in Context: The GovStat Ontology and the Statistical Interactive Glossary. Dg.o. 2003.

Hert, C.A. and Hernandez, N. (2001). User Uncertainties With Tabular Statistical Data: Identification And Resolution: Available at http://ils.unc.edu/govstat/fedstats/uncertaintiespaper.htm

Marchionini, G.; Haas, S.W.; Plaisant, C.; Shneiderman, B.; Hert, Carol A. (2003) Towards a Statistical Knowledge Network. Dg.o. 2003