

Defining a Metadata Structure to Support Integration

Sheila O. Denn

School of Information and Library Science, University of North Carolina at Chapel Hill, CB# 3360, 100 Manning Hall, Chapel Hill, NC 27599-3360. Email: denns@ils.unc.edu

Jung Sun Oh

School of Information and Library Science, University of North Carolina at Chapel Hill, CB# 3360, 100 Manning Hall, Chapel Hill, NC 27599-3360. Email: ohjs@email.unc.edu

Maria Cristina Pattuelli

School of Information and Library Science, University of North Carolina at Chapel Hill, CB# 3360, 100 Manning Hall, Chapel Hill, NC 27599-3360. Email: pattm@email.unc.edu

Introduction

In the GovStat Project (<http://ils.unc.edu/govstat>), one of our goals is to build architectures and interfaces that facilitate integration of data from across Federal statistical agencies. We have been engaged in a process of defining the metadata elements and structures necessary to support this integration. This poster outlines the user-centered, scenario-based process we have undertaken thus far, demonstrates examples of how metadata supports integration, and outlines the areas of our future work.

As increasing amounts of data are being served on the Web, there has been an increasing demand for Web facilities that can integrate data from across different sources into a unified presentation for the user. The linchpin of this kind of service is an information architecture that captures not only individual data elements and their associated metadata, but the relationships among and between the data elements as well. There is no standard process for creating such an architecture as of yet, but we believe our experiences in the GovStat project will be of interest to other information scientists engaged in building information architectures designed to support integration. We will also explore some of the unique challenges posed by statistical data in terms of information seeking and use, and how these have impacted our design work.

Scenario-based Design

The work of the GovStat project has been conducted with an eye toward creating a statistical knowledge network (SKN). The SKN would provide a public intermediary that would integrate data from different statistical agencies and render it to users in a seamless way (see Figure 1). We began our metadata efforts with the idea that in order to pursue the vision of an SKN, we needed a metadata schema that included the elements necessary for the integration of data and concepts within the different sub-domains

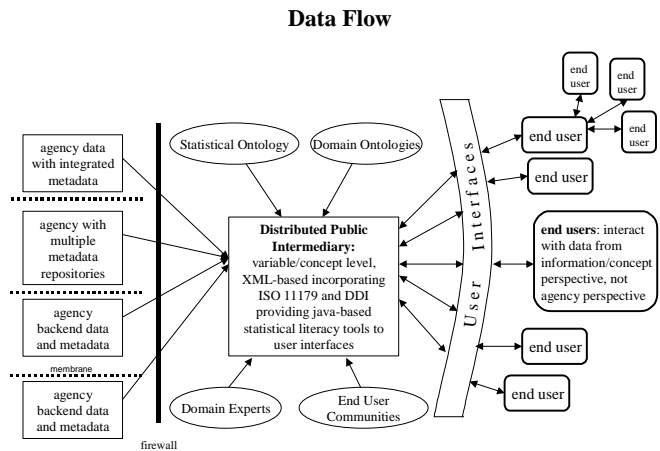


Figure 1. Data Flow in the SKN

represented by the information on government statistical websites.

We first undertook a metadata user study to determine what metadata elements are necessary for users both to find and to understand statistical information (Denn, Haas, and Hert, 2003). We reasoned that this would help us develop a minimal set of metadata elements necessary to integrate data from across the agencies.

We used the scenario-based design process described in Carroll (2002) and Rosson & Carroll (2001) to develop tasks for users to complete in the study. We concentrated on tasks that would require the user to gather information from more than one agency. Once we developed the initial set of candidate tasks, we asked professionals from within the Federal statistical agencies to validate the tasks and confirm that they were indeed representative of the kinds of information problems that users bring to statistical web sites. Once we had pared our set down to five tasks, we

identified resources, both within the Federal statistical agency websites and at other websites, that could be used to complete the tasks.

Design Process and Rationale

With the vision of the SKN in mind, we analyzed the results of the metadata user study and began designing our schema. Our goal is to enable users to find statistical data without prior knowledge of particular sources and to provide appropriate contextual information to allow users to interpret and use the data properly. Further, we want to provide sub-document level access and integration across documents and agencies. To this end, we first identified requirements and constraints for the schema from the perspective of both users and agencies. While taking a user-centric approach, we also understand that providing functionality for agencies to facilitate their acceptance of our schema is essential.

The metadata user study pointed up some issues that would be important to us as we designed the schema. Two of these issues are the ability to make comparisons, and the importance of context.

Often when users consult statistical data, they do so with an eye toward making comparisons using the statistics, whether comparisons between a single or related statistics across different periods of time, comparisons between different geographical areas, etc.

Context is especially important in the statistical environment because of the relative information density of statistical tables as opposed to straight text. In order to understand a number that is in a particular cell of a table, there is a great deal of context that must travel with it, including the row and column headings with which it is associated, the units in which it is measured, the population to which it refers, and other information about the way it has been measured or aggregated.

We want to be sure that our schema provides the elements necessary to support comparisons and to associate contextual information with the data.

To leverage existing efforts in metadata development and to promote interoperability, we surveyed and consulted several metadata schemas and standards. Among others, the Data Document Initiative (DDI) (<http://www.icpsr.umich.edu/DDI/index.html>), a metadata standard for the social science community that has been broadly adopted in the statistical world (especially in data archives), provided a starting point for modeling the overall structure of our metadata schema. However, its emphasis on microdata and standalone datasets does not fit our needs for representing inter-related entities for a variety of statistical objects including tables, news releases, reports, etc.

In order to address those issues mentioned above, as well as providing context and making comparisons possible, we

need to represent and interrelate concepts and values in a systematic way. ISO/IEC 11179 (<http://metadata-stds.org/11179/index.html>), an international standard for the definition, naming and registration of any kind of data elements, served as a core reference for modeling our metadata element structures and concepts.

Once we defined an initial set of elements, we have been involved in an iterative development-testing process. For each cycle, with a scenario that involves integration of data from across agencies at various levels of complexity, we marked up data from multiple agencies to see whether the metadata schema can capture the important aspects of data, to examine how well our metadata schema can support users in completing defined tasks, and finally to estimate how many man-hours it takes to mark up a typical document. During this process, a considerable effort has to be made to balance complexity with functionality. Unique characteristics of statistical data, such as the issues of comparison and context described above, call for a quite complex schema, which in turn requires a great deal of effort for proper markup. We recognize, however, that this complexity is an obvious barrier for agencies in adoption of our schema. In response to this challenge, we have developed a streamlined data model with minimal essential information, while retaining the ability to add on to the model to represent additional information that individual agencies might require.

Part of the process of designing a metadata schema to support integration has been to define for ourselves what integration really means in this context. To that end, we have devised a hierarchy of integration that defines the different levels of integration we hope to achieve, and expresses where our schema lies within that hierarchy (see Figure 2).

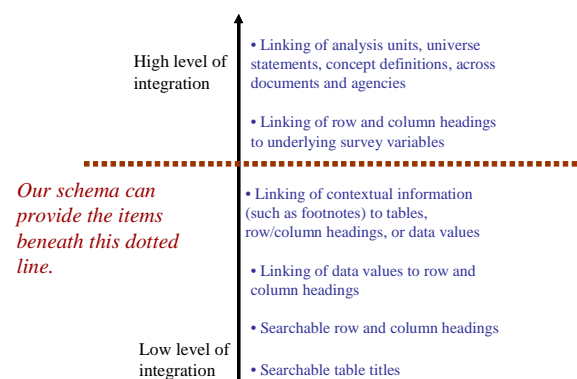


Figure 2. Hierarchy of Integration

Current Status and Future Directions

We have sought a balance between the schema's expressiveness and ease of use by testing the schema throughout the design process. We believe we have created a streamlined version that satisfies both requirements. On the one hand, our metadata schema appears to effectively enable access at the sub-document level and facilitate information identification and comprehension down to the level of individual data cells and footnotes. On the other hand, the design of the schema has been kept simple enough for facilitating agency adoption.

While the schema can function as a common denominator within and across agencies, it can also be extended and customized in order to accommodate specific representation requirements of individual agencies.

While a metadata schema provides one piece of a solid intellectual foundation for integrating information for users, there must be effective user interfaces built on top of the metadata structure to display this information to users in an efficacious way. We are in the process of exploring what constitutes an effective, metadata-driven interface for information integration, with a focus on supporting the comparison and contextual issues identified above.

The ultimate goal of our project is to transfer the technology developed to the statistical agencies. In the case of our metadata work, we must face the challenge of communicating the value and potential of the schema to the agencies so that it will be adopted and incorporated into the agencies' work flow.

Manual markup of documents can be a tedious and time-consuming activity. We are now in the process of designing a markup workstation with the aim to provide a functional and friendly interface for markup, as well as to semi-automate markup of some of the simpler elements. We are also in discussions with individual agencies to explore ways to leverage agency data production processes to support some level of automated markup.

The schema we have developed appears to be a promising standardization tool for data representation, access, and exchange. As for statistical data integration, the schema supports and facilitates aggregation, alignment, and reconciliation of data values. More advanced data integration capabilities, as indicated at the highest level of our hierarchy of integration represented in Figure 2, appear to be beyond the capacity of a metadata schema alone. The disparity of data sources, the variety of collection methods and the heterogeneity of vocabularies are major barriers to further information integration at the user search level.

The research streams described above suggest a vision toward which we would like to work. That vision would include an overarching knowledge structure that formally defines and encapsulates the meaning of the concepts used in the domain and the relationships between these concepts. By associating these meanings to the data instances, and

thus enabling mapping based on semantics, we believe we can successfully address and reconcile disparities in data representation and effectively further the integration process.

REFERENCES

- Carroll, J.M. (2002). *Making Use: Scenario-based Design of Human-Computer Interactions*. Cambridge, MA: MIT Press.
- Denn, S., Haas, S.W., and Hert, C.A. (2003). Statistical metadata needs during integration tasks. In: DC-2003 (September 28 – October 2, 2003, Seattle, WA). Accessed online at http://www.siderean.com/dc2003/301_Paper50.pdf.
- Rosson, M.B., and Carroll, J.M. (2001). *Usability Engineering: Scenario-based Development of Human-Computer Interaction*. Morgan-Kaufman.