BOF: Developing an Integration Infrastructure for The Statistical Knowledge Network: Metadata and Standards
Convener: Carol A. Hert, Syracuse University

Advance Description of the Session: A number of the digital government projects are addressing issues related to the integration of statistical information from multiple sources. Mechanisms for storing, retrieving, transmitting, and integrating that information are highly dependent on metadata that describes, explains, and provides context for the statistical information. These metadata include ontologies, terminological systems of all kinds, technical documentation about methods, etc. There are a variety of standards in play relevant to these efforts: DDI (Data Documentation Initiative), ISO11179 for metadata registries, RDF, terminological system standards, etc. The BOF would look at which metadata aspects and standards various projects are working with, discuss issues of coordination of these efforts, and begin to strategize on how to draw on expertise in the statistical agencies to inform ongoing decisions in this area.

Meeting Report

Contrary to the technical focus of the BoF's call for participation (listed above), the discussions of the participants generally reflected issues relating to the socio-cultural and organizational contexts in which metadata exists. The group considered organizational constraints in providing metadata, strategies for enabling organizations to identify their capacities for capturing and utilizing metadata, and building business cases for metadata efforts. There was a general consensus among the participants that metadata will benefit agencies, practitioners and researchers.

Researchers wanted better access to data and to metadata. The statement of this need led to a discussion of why this is often so challenging for agencies. It is often difficult to locate and inventory existing metadata. There is sometimes a disconnect between organizational entities that produce metadata and those that may need it making it difficult to trace the path. Agencies may be concerned that their metadata are not "perfect" and thus may be unwilling to disseminate them. Personnel in agencies may not be rewarded for attention to metadata quality resulting in a disincentive. The cost of gathering and maintaining metadata may also be prohibitive for many agencies. And, as one participant commented, "the devil is in the details," it may be easy to see the need for better metadata but less easy to actually work with the extensive and complex sets of metadata that exist.

There is a need for metadata at multiple levels of granularity. Documents and data are produced at multiple levels of sophistication and detail, but most of the attention so far has been at the survey/dataset level.

Some discussion of standardization of various metadata occurred. However, this seems unlikely given the decentralized nature of the US statistical system and the necessity for different concepts and variables (similar though they may be) across the agencies. Even

such “standard” concepts as those relating to demographics are challenging to standardize when one considered all the possible contexts in which such concepts are useful.

Agency personnel who were present stressed the need for a business case that would convince organizations to undertake expanded metadata activities. The potentially huge cost of metadata management is a huge deterrent when agencies can not tie that investment back to cost savings in survey processes. It was suggested that the model employed by the GIS community might be useful. The GIS community was able to make an argument that the lack of metadata was making it hard for them to share data—and they demonstrated how much it cost to get the geospatial data and then they had to agree to a small set of common elements (about 25). Tying the business case to government wide efforts was also suggested, in particular, the Federal Enterprise architecture management initiatives (more information at www.feapmo.gov). States also have similar initiatives underway. Another strategy (for statistical agencies) might be to tie the business case development to quality efforts. Organizations such as the National Academy of Sciences (the component engaged with statistics) and the American Statistical Association might be useful partners in this regard. Also mentioned were the Office of Intergovernmental solutions and Pi-square-2 (a joint practitioner and research community). We grappled with how to define and quantify the benefits in a way that can be measured and justified, either from the bottom-up by showing benefits to the individuals or small groups, or top-down by showing how it helps satisfy agency mandates.

An important aspect of the GIS example above was the “starting small” approach that was taken. While it is desireable to eventually have a full-blown statistical metadata system, it probably is easier to sell small concrete examples first. A number of the participants thought that starting small would be a very useful strategy for pushing metadata initiatives. Several indicated that one approach would be to start first with metadata elements and then derive survey instruments, etc. This technique is being used for some activities at the National Center for Health Statistics, according to one participant.

Another organizational aspect was mentioned by the participants from the Center for Technology in Government at the University of Albany. They have a document that enables an organization to assess its capability to create a metadata system and also describes the ideal and how you move to that ideal. This discussion moved into strategies for how organizations could create metadata. The Albany participants mentioned funding students to do summer work creating metadata in various agencies. The question was raised of what possible public-private ventures might be formed to support metadata creation efforts.

While the discussion was largely non-technical, there was some discussion of the use of natural language processing techniques to support extraction and development of metadata to populate ontologies and respositories. Eduard Hovy is currently pursuing this type of work.

The participants also considered some "what next" ideas? There was some feeling that there may be value in bringing together people interested in metadata issues across domains) to look at the organizational issues and models for organizational development that might be applicable across domains. The statistical agencies may want to consider the development of a small set of metadata elements. Marshall DeBerry at FedStats expressed some interest from FedStats in such a project. Continued work on demonstrating the value of metadata and articulating the business case for metadata efforts was also recognized as a pressing need.

Attendees:

Carol A. Hert, Syracuse University, cahert@syr.edu
Sheila Denn, University of North Carolina-Chapel Hill, dens@ils.unc.edu
Larry Jackson, University of Illinois-Urbana Champaign, lsjackso@uiuc.edu
Theresa Pardo, University of Albany – Center for Technology in Government, tpardo@ctg.albany.edu
Donna Canestraro, University of Albany – Center for Technology in Government,, dcanestr@ctg.albany.edu
Bill Kules, University of Maryland, wmk@cs.umd.edu
Eduard Hovy, University of Southern California/Information Sciences Institute, hovy@isi.edu
David F. Pinto, CIIR, University of Massachusetts-Amherst, pinto@cs.umass.edu
Vijay Venkataraman, University of Maine, vijay@umit.maine.edu
Mousumi Chatterjee, Renesselaer Polytechnic Institute, chattm2@rpi.edu
Richard Fujimoto, Georgia Tech, fujimoto@cc.gatech.edu
Marshall DeBerry, FedStats, mdb@fedstats.gov
John Gawalt, National Science Foundation/Science Resource Studies, jgawalt@nsf.gov
John Bosley, Bureau of Labor Statistics, bosley_j@bls.gov
Rick Swartz, Census Bureau, richard.w.swartz@census.gov
Stephen Powers, Babson College, spowers1@babson.edu