Institute of Education Sciences
National Center for Education Statistics

METADATA AND PARADATA: INFORMATION COLLECTION AND POTENTIAL INITIATIVES

NISS Expert Panel REPORT ◦ November 2010

EXECUTIVE SUMMARY

Metadata and/or paradata accompany federal statistical agency data files to describe or define the data elements and the collection and processing of these data. Practices vary across the Statistical Community of Practice (SCOP). Distinctions between the two terms are not well-defined, and there are no generally accepted standards in use by all the federal statistical agencies. Therefore SCOP commissioned the survey of current definitions and practices in current use that make up this report with the objective of laying the groundwork for development of standardized definitions and practices.

The goals for this report are therefore to provide the necessary background information for developing and implementing standardized definitions and best practices for metadata and paradata within the US federal system. In particular, this report

1. Reviews existing practices and resources both nationally and internationally;
2. Inventories current metadata and paradata practices in federal statistical agencies;
3. Identifies key elements of metadata and paradata that would be useful to federal statistical agencies in development and implementation of standardized definitions.

The following definitions are proposed to SCOP:

- *Metadata*: Formalized data about statistical data needed to search for, display and analyze those data.
- *Paradata*: Formalized data on methodologies, processes and quality associated with the production and assembly of statistical data.

Note: *survey weights should be regarded as data*, even though calculation of them employs paradata about the design and conduct of the survey.

- *Markup Language:* a method for annotating text in a way that is syntactically distinguishable from that text and in consequence is computer processable, (e.g., HTML, HyperText Markup Language).

To date with respect to metadata and paradata, there are essentially two systems on which to build. Each is tied to a description language. DDI is based on XML (Extensible Markup Language); SCMS is based on UML (Unified Modeling Language).

A five-step process is suggested for the development and implementation of ICSP (Interagency Committee on Statistical Policy) agency-wide definitions and practices.

1. Agreement on definitions for data, metadata and paradata with *a concept of metadata that is completely independent of that of file structure.*
2. *Commitment to markup language-based metadata, and to markup language-based data*.
3. "Proof-of-concept" based on one or both of DDI and SDMX, consisting *of full markup-language-based metadata and data files for a modest-scale survey* including development of parsers that would put the data into common formats including development of parsers that would put the data into common formats.
4. Creation of an ICSP-wide *extensible metadata template for surveys*, together with tools for common tasks.
5. Creation of a *repository for metadata for surveys*.

Read the Full Report