What Metadata When?
Ideas from the User Perspective

Ann Green
Social Science Statistical Laboratory
Academic Media & Technology
Yale University
ann.green@yale.edu

Information seeking process is a cycle

- Non-linear steps:
  - find
  - identify
  - select
  - obtain
  - use
- Users are dependent upon metadata in each step
- Interactive process of applying information and building knowledge

Types of questions

- Quick facts
- Summary data: a profile, a snapshot
- Intensive analysis

All are characterized by “relatively vaguely expressed information need” Sundgren

For each: define use of statistics and the process involved.

Type 1: Quick facts

- A single number or a hand full of numbers
- Use:
  - Transcribe or print
  - May want to put into spreadsheet
  - May want to generate a simple map
- Process: find source, locate data items, interpret adequacy, work through selection process. Little emphasis on use.

Type 2: Summary data

- A profile, a snapshot
- Use:
  - Summary statistics, map or graph
  - Integrate multiple sources
    - Composites
    - Comparisons
- Process: find sources, locate data items, interpret adequacy, work through selection process, integrate potentially disparate sources. Use can be constrained by Web tools and lack of metadata.

Type 3: Intensive analysis

- May involve complex data bases, analytical tools, and advanced methods
- Use:
  - Exploratory analysis
  - Visualization: graphics, mapping
  - Transformations, recoding
  - Analyzing complex relationships among related data items
- Process: find sources, locate data items, interpret adequacy, work through selection process, integrate potentially disparate sources, use a range of basic and sophisticated analytical tools, intense interaction and manipulation of statistics

Levels of user sophistication no longer determine level of use, type of question, or process of using statistics.

As users become more comfortable with statistics and Web interaction, expectations about finding information and interacting with it rise.

Types of questions are no longer discreet, finding a Quick Fact can be a jumping off point into compiling profiles and can lead to more demanding analysis. In the other direction, intense analysis relies upon, uses and produces basic facts.

Information **seeking** processes must be linked to an integrated **analytical** processes. No dead ends.

Users should not be constrained by their level of stat expertise, or their knowledge of data sources, or the structure of data sources

Old model:

The level of expertise, type of question, and structure of data sources determined the path of finding information. Information systems are based upon this model.

New model:

User interaction is not constrained by the level of expertise or type of question. Paths of finding information and the interaction of using information are dynamic and interrelated.

Users of all levels want to choose from multiple applications:

- visualization and graphics
- ability to slice and dice, compare and contrast
- want to see information over time and in geographic context
- want to integrate multiple sources

Why metadata matters at all of these levels

- Users have different levels of knowledge about data sources and about statistics
- Knowing the data source does not lead to intelligent use of the data
- Knowing statistical methods does not lead to intelligent use of the data

Metadata builds the bridge to intelligent use.
Metadata must support interpretation.
Metadata must frame information with context.
Not a matter of simply adding more information, but of efficiently adding context.

Decision making about data must be an interactive process with flexible metadata systems.

For what purposes do users need metadata?

Metadata for locating data

Metadata for determining fitness for use

Metadata for combining data

Metadata for documenting subsets

Metadata for graphs, maps, digital objects

When do users need metadata?

- Need full information at the right time, and throughout the information process
  - In discovery and evaluation
  - Within extract and analysis systems
  - In understanding relationships between data items
  - Output must have accompanying meta data
  - Output must know its sources and be able to retrieve further information or additional data items
- Need to know the history of how the data were produced and how the data have been compiled or restructured.
- Data must retain their source, context and descriptive information as they travel the Net.

Users need data + metadata linkages throughout the information process

- Information about data should be coupled with the data
- Metadata should have a natural presence with data, not an afterthought or awkward presentation
- Metadata should be considered as being integral to all data manipulation functions
- Metadata should trace usage of data through the history of a data analysis cycle
- Metadata should guide calculations and describe transformed data

12/01/01 3:27 UN/ECE Work Session on Statistical Metadata 15

The Current nature of data on the Web

- Broadcast of data to the desktop has lead to an overwhelming number of stand alone systems, duplication and variance in quality.
- Systems don't talk to each other.
- Descriptions of data are not written in language users can understand
- Descriptions of data are not linked to numbers
- We still aren't getting what we need: real time data from local, state, and federal sources.
- Local resources aren't integrated into the Web.

12/01/01 3:27 UN/ECE Work Session on Statistical Metadata 16

What users need to know

Context:

- source, authoring body, purpose, methodology, version

Content:

- topics, keywords, abstracts, related materials
- Concise understandable descriptions of data items

Geographic and temporal specificity

Measures of quality and reliability (Stats Canada)

Access restrictions

Strength of the DDI

- Local practices standardized
- compatible with other types of metadata, e.g. MARC, Dublin Core
- collaborative, cross-community development
- trained metadata professionals: archivists
- Supported by user oriented institutions: archives and data intermediaries in universities
- Based upon user oriented tools:
  - **Printed codebook contents: information about context and content**
  - **Archive descriptions and catalog records: information for identification and discovery**
  - **Data dictionaries, stat package setups, output from CATI/CAPI survey collection: item level information**
- Goal of non-proprietary format with archival longevity

Highlights of the DDI

- Convergence of archive, analysis, resource discovery, and production
- Promotes a non-proprietary neutral standard
- Committed to both content and format standards based upon standards in use and in development among data archives and data producers
- Robust, detailed, yet modular and flexible
- Meets requirements of preservation, it allows reconstruction of full documentation

12/01/01 3:27 UN/ECE Work Session on Statistical Metadata 19

DDI process has users at the center

- Stakeholders defined content
- Sensitive to user centric approach
- Practical and usable
- Convergence of data users, data producers and archives

12/01/01 3:27 UN/ECE Work Session on Statistical Metadata 20

Build and share standard meta data resources:

- Crucial to the development of an integrated social science infrastructure
- Capture meta data resources at points of data collection, distribution, and when value is added
- Need a unified effort based upon shared standards and goals
- Create complete and fully tagged XML DDI format
- Follow principles of interoperability and preservation

12/01/01 3:27 UN/ECE Work Session on Statistical Metadata 21

Defining a process for building the metadata infrastructure

Must involve data producers, systems developers, and data users

Problems:

- documentation is not being created from cradle to archive
- documentation is not produced for the purposes of data users

Practical steps to evolve metadata:

must be simple, straightforward, realistic

12/01/01 3:27 UN/ECE Work Session on Statistical Metadata 22

Who can make this happen?

- Data producers:
  - must see the value of complete meta data.
  - must be motivated to produce all the components of intelligent data.
  - need production tools that create metadata files as part of the data production process
- Data systems developers:
  - Need to work with users to deliver metadata at the right time and to efficiently add context for decision making.
- Data intermediaries (archives, libraries, consulting staff):
  - must articulate the need for metadata + data.
  - must articulate the need for the ability to integrate data.

12/01/01 3:27 UN/ECE Work Session on Statistical Metadata 23

The social science research environment: laboratories for exploring the presence of metadata

- Collaborative work across disciplines, organizations, and political boundaries
- Integration of data from multiple sources
- Web based experiments
- Spatial visualization
- Cradle to archive data management including metadata
- Teaching applications: including distance ed

12/01/01 3:27 UN/ECE Work Session on Statistical Metadata 24