Statistical Metadata and Statistical Registries

Daniel Gillman

U.S. Bureau of Labor Statistics

CNSTAT Panel on Transparency and Reproducibility in Federal Statistics

Caveat

- Opinions in this presentation are due to the author and do not necessarily reflect official the policies of the US Bureau of Labor Statistics.

Metadata – A Short History

- Philip Bagley, 1968
  - *Extension of programming language concepts*
  - Book – published by NBS in 1969
- Jack Myers, 1969
  - But not published
  - Founded the Metadata Company
    - Trademarked the term Metadata
    - Sued others for its use

Metadata – A Short History

- Bo Sundgren, 1973
  - Statistics Sweden
  - *An Infological Approach to Databases*
    - PhD dissertation, published
  - 35 year influence on international statistical community
- All 3 efforts independent

Metadata – A Short History

- 1980s
  - Lawrence Berkeley Lab
    - CODATA project
    - About statistical data
    - Led to OSIRIS, which led to DDI
  - UNECE Statistical Metadata Working Group
    - Called METIS
    - Bi-yearly meetings
    - National Statistical Offices
    - Heavily influenced by Sundgren
    - Led to GSBPM, GSIM, and others

What Is Metadata?

- "Data about data"
- But, are all metadata about data?
- Consider the Dublin Core Metadata Initiative
  - OCLC – Online Computer Library Center
    - Dublin, OH
  - Dublin Core Metadata Element Set
    - 15 core elements
    - Plus 40 additional qualifier terms

What Is Metadata?

- Dublin Core
  - Used in many digital libraries
    - Museums and Book libraries
- Requires expanded view of metadata
  - "Data about some resource"
- Note, data are not always metadata
  - Metadata is a role; not an absolute
- Better definition
  - 'Data being used to describe some objects"

Telephone Call Example

- Telephone company knows
  - Who called whom
  - Whether connection is made or why it failed
  - Length of connection
- Metadata if describing a telephone conversation
  - Or lack thereof
- Suppose need to discover networks of contacts
- Then, data can be used to construct graph
  - Caller
  - Receiver

Statistics

- What do we want to describe? Examples are …

- Variables
- Value Domains
- Data
- Classification schemes
- Code lists
- Questions
- Questionnaires
- Instruments

- Sample design
- Weighting
- Data transformations
  - Editing
  - Classification
  - Allocation
  - Estimation
  - Disclosure control

Statistics

- Metadata supports 3 main uses:
  - Discovery
  - Understandability
  - Usage
- Advantage to expanded definition of metadata
  - No need to pre-classify
    - E.g., structural, conceptual, administrative
    - Produces multiple classifications
  - Just select kinds of objects to describe

Statistics

- Sometimes metadata do not directly describe data
- Examples –
  - Q-Bank NCHS (https://wwwn.cdc.gov/QBANK/Home.aspx)
    - Describes questions
    - Research supporting their use
  - C^2 Metadata (http://c2metadata.org/)
    - "Continuous capture" of metadata
    - Automates documenting data transformations
      - From statistical packages

Metadata Management

- Metadata are data
- Managed in database
- Usually called a repository
  - Probably due to historical requirements
    - Often document management
- Metadata managed in 2 main types
  - Passive: text, documents, not machine processable
  - Active: machine processable
- Current practice – more active metadata

Metadata Management

- Statistical metadata repositories
  - Varying scope
  - Some not about data
  - Many about more than data
    - Question banks
    - Classification servers
    - Data set catalogs
    - Corporate data dictionaries

Metadata Management

- Statistical metadata repositories
  - Many NSOs in
    - Europe
    - Canada and Mexico
    - Australia, New Zealand, and South Africa
  - have broad scope systems in place
    - Broad – covers the range of statistical activities
  - International Household Survey Network
    - DDI application (http://www.ihsn.org/)
    - Managed by World Bank
    - For developing countries

Metadata Management

- Statistical metadata standards
  - Describe statistical data and processes
    - GSIM – Generic Statistical Information Model
    - DDI – Data Documentation Initiative
  - Describe statistical life-cycle
    - GSBPM – Generic Statistical Business Process Model
  - Enhance dimensional data transfer
    - SDMX – Statistical Data and Metadata eXchange

Registries

- Registry versus Repository
  - Registries include
    - Repositories (they manage something)
    - Administrative functions (they carry out some mandate)
- Example of registry
  - US state motor vehicle licensing
    - Repository – metadata about each vehicle
    - Registry – management procedures implementing law

Registry Standards

- ISO/IEC 11179 – Metadata registries
  - 6 part standard on data
    - Semantics
    - Structure
    - Registration
  - Part 6 – registration
    - Contains procedures for maintaining data registry
      - Applicable in general as well
    - Includes quality criteria for metadata
    - Includes metadata life-cycle management

Registry Standards

- ebXML registry (http://www.ebxml.org/)
  - Electronic business XML
  - Developed by OASIS
    - Organization for the Advancement of Structured Information Standards
  - System for implementing a registry
  - Borrowed ideas from ISO/IEC 11179
- SDMX adopted ebXML registry for its registration needs

Registries in Statistics

- Examples
  - BLS Longitudinal Database
  - SDMX global registry
  - (https://registry.sdmx.org/FusionRegistry/)
  - Population registers in Europe
- Many more

Integration

- Metadata systems built independently
- 13 principal US statistical agencies (and others)
  - Each with own budget
  - Each with own business requirements
- How could we build an integrated metadata system?
  - Need principles
  - Need standards
  - Need interoperability

Basic Principles

- Web interface
- Standards
  - Statistical for metadata model and exchange
  - Internet for access and communication
- Common terminology for mapping content
  - Examples: BLS taxonomy, Census table of contents

Standards

- Effective standards are developed through
  - Consensus driven process
    - General agreement with no sustained dissent
  - Open
    - Any stakeholder has the opportunity to participate
  - Balanced
    - Stakeholder community adequately represented
  - Fair
    - No stakeholder has control or advantages over others
  - Transparent
    - Development process open to anyone for inspection

Standards

- What are they?
- Sets of provisions agreed upon by consensus
- Provisions
  - Requirements
  - Recommendations
  - Instructions
  - Statement
- Conformance – How to comply
  - Satisfaction of all requirements
  - Some requirements are implied or in other provisions

Standards

- ISO, DDI-Alliance, OASIS
  - All have effective procedures
- Consequences
  - Standards developed this way
    - Seen as fair
    - Gain wide acceptance
    - Level the playing field

Interoperability

- Definition
  - Ability of separate systems to work together without human intervention
- Two main kinds
  - Syntactic – ability to communicate and exchange data
  - Semantic – ability to understand exchanged data
- Sufficient condition
  - Standards conformance

Conformance

- Implications
  - Internal systems
    - No need to be conformant with metadata standards
    - Developed to optimize individual agency needs
  - External interface
    - System must LOOK conformant to outside user
      - Another system or person
    - Based on agreed common metadata standard

Questions?

Contact Information

Dan Gillman
Mathematical Statistician
Office of Survey Methods Research
www.bls.gov/osmr
202-691-7523
Gillman.Daniel@bls.gov