

What You Need to Know – Too *Standards and Interoperability*

Dan Gillman

Bureau of Labor Statistics
FCSM Metadata Workshop
Washington, D.C.
14 September 2018



Outline

- Standards in General
- Interoperability
- Case for Standards
- Data Integration Scenario
 - ▶ Discovery
 - ▶ Data dictionary
 - ▶ Methodology
- Overview of Statistical Metadata Standards



Standards

- Many standards development organizations (SDO)
- Open standards built by a process that is
 - ▶ Consensus-driven general agreement w/o sustained dissent
 - ▶ Open any stakeholder can join
 - ▶ Transparent process available for inspection
 - ▶ Fair everyone has same rights
 - ▶ Balanced stakeholders represent user community
- Includes ISO, W3C, NISO, DDI Alliance, UNECE



Standards

- Caveat -
- Many SDOs, many standards
 - ▶ “Standards are great. There are so many of them!”
 - Karsten Rasmussen
 - ▶ “Standards are useless; look at the second S!”
 - Adrienne Tannenbaum



Interoperability

- Interoperability – ability of one system to work independently with some or all of another system
- Applied often to computerized systems, but also to data
- Data interoperability – ability to use data from another source without help from that source
- Implies extensive metadata are available



Interoperability

- But, metadata are data, too
- Data interoperability must include metadata interoperability
- Does this require the metadata have metadata?
- Shared metadata model needed
 - ▶ Standard
 - ▶ Technical specification
- Minus that, data problem is just repeated

Standards – Why?

- Reduces or eliminates design steps
- Increases chances for interoperability
 - ▶ Standards neither necessary nor sufficient
- Building systems – claims of conformity
 - ▶ Conformance – Satisfaction of all requirements
 - ▶ Systems can be built independently
 - ▶ Allows system builders to achieve interoperability

Standards – Why?

- If your metadata system conforms to a specification
 - ▶ I can build a system to read your metadata automatically
 - ▶ I can write metadata in a format you can understand immediately
- But, if I use a different specification, then
 - ▶ I have to translate your metadata into my specification and vice-versa
 - ▶ May not be easy
 - ▶ With 13 principal statistical agencies (minus OMB),
 - Possible translations: $(13 \text{ choose } 2) = 78$
 - This is too complex; Need cooperation

Standards – Why?

- Adopting standards greatly reduces this problem
- There's still the problem of the second S
 - ▶ There may be many standards to choose among
- Let's try to make sense of this problem
 - ▶ Standards developed to solve certain problems – Scope
 - ▶ Don't use them beyond their scope



Standards Illustrated

- Through a data integration scenario
- Illustrate metadata “content” standards
 - ▶ Focus on what can be described
 - ▶ Not on how to build a system
- Overview, not detailed descriptions
- Include some about the groups developing the standards

Scenario

- “America’s Safest Cities”
 - ▶ by Zack O’Malley Greenburg
 - ▶ 26 October 2009 *Forbes Magazine*
- Rank cities by “livability”
 - ▶ Workplace fatalities
 - ▶ Traffic fatalities
 - ▶ Violent crimes
 - ▶ Natural disaster risk

Scenario

- Rank MSAs based on
 - ▶ Numerical ranking for each measure
 - ▶ Sum of rankings
- Questions
 - ▶ Can we find and understand relevant data?
 - ▶ If so, where? how?

Scenario – Discovery

- Natural to ask if data can easily be found through search
 - ▶ Quick answer – No
 - ▶ Google searches not entirely successful
 - URLs provided for relevant web sites
 - Relevant data sets, no
 - Still had to search web sites to find data
- Discovery is a very hard problem
 - ▶ Guarantee to find all resources on a particular subject??

Scenario – Discovery

- Another solution – data set registry or catalog
 - ▶ Think – library card catalog
 - ▶ But – on line
- Look at Data.Gov
- Many other catalogs in existence
 - ▶ Museums – Smithsonian Museum of Natural History
 - ▶ Libraries – Library of Congress

Discovery (Catalog) Standards

■ Relevant standards

- | | |
|---------------------------------------|--------------|
| ▶ Project Open Data Metadata Schema | Data.Gov |
| ▶ Dublin Core Metadata Initiative | NISO, ISO |
| ▶ MARC – MACHine Readable Catalog | NISO, ISO |
| ▶ ISO/IEC 11179 – Metadata registries | ISO |
| ▶ DCAT (Data Catalog Vocabulary) | W3C |
| ▶ DDI (Data Documentation Initiative) | DDI Alliance |

Scenario – Discovery

■ Finding data – Discovery

▶ Workplace fatalities

- Bureau of Labor Statistics

▶ Traffic fatalities

- National Highway Traffic Safety Administration

Problem

- How do we know to select particular data sets?
- Are there others?
- Need data dictionaries to be sure



Scenario – Data Dictionary

■ Finding data – Discovery

▶ Workplace fatalities

- Bureau of Labor Statistics
- Data based on MSA
- Data given as number, not rate

▶ Traffic fatalities

- National Highway Traffic Safety Administration
- Data based on city, not MSA
- Based on rates

Scenario – Data Dictionary

- Data Dictionary – for statistical data
- Contains
 - ▶ Variables
 - or Measures
 - Code lists or Classifications
 - ▶ Questions
 - ▶ Maybe some methodology as well
- Description of variables needed at a minimum



Scenario – Data Dictionary

- Variables, Measures, Classifications – needed for
 - ▶ Selecting specific data sets
 - ▶ Using selected data sets
- Level beyond discovery
- Most discovery models don't account for this



Data Dictionary Standards

- ISO/IEC 11179
- DDI
 - ▶ Codebook
 - ▶ Lifecycle
- UNECE
 - ▶ GSIM (Generic Statistical Information Model)
- Inter-agency SCOPE/Metadata
 - ▶ Data dictionary specification

Scenario – Methodology

- Methodological issues
 - ▶ Questions
 - ▶ Sampling
 - ▶ Post-collection processing
 - ▶ Post-collection estimation
- These can affect analyses
- And there are standards to document these

Standards for Methodology

- DDI (Data Documentation Initiative)
 - ▶ Codebook
 - ▶ Lifecycle
- GSIM (Generic Statistical Information Model)
- GSBPM (Generic Statistical Business Process Model)



SCOPE/Metadata

- SCOPE - Statistical Community of Practice and Engagement
 - ▶ Group to leverage common practice among agencies
 - ▶ Reduce costs, Increase sharing
 - ▶ Formed inter-agency group on metadata
 - Produced first data.gov specification
 - Geared towards statistical data sets
 - Produced data dictionary specification
 - Variables, Measures, Code Lists, and Classifications
- SCOPE/Metadata
 - ▶ Meets bi-weekly
 - ▶ Needs more participants

ISO/IEC 11179

- <http://standards.iso.org/ittf/PubliclyAvailableStandards/index.html>
- First standard on metadata, model based, reusable metadata
- Operational needs for a registry or catalog
- Standard built in 6 parts
- Used as input to DDI, GSIM, SDMX, and SCOPE/Metadata
 - ▶ SDMX – Statistical Data and Metadata eXchange
- Freely available from ISO

GSIM and GSBPM

- Developed under UNECE
 - ▶ UN Economic Commission for Europe
 - ▶ Comprises Europe, Canada, and US
 - ▶ Statistical cooperative program is world-wide
- Statistical metadata standards under Modernization efforts
- Many countries involved, especially
 - ▶ Australia, Canada, New Zealand, US
 - ▶ France, Italy, Netherlands, Portugal, *Scandinavia*, Slovenia

GSIM

- <https://statswiki.unece.org/display/gsim/Generic+Statistical+Information+Model>
- Model of statistical information objects
 - ▶ 4 main sections
 - Conceptual, Structural, Business, Exchange
 - ▶ High level, conceptual model
 - ▶ No bindings – not directly implementable
 - ▶ Some effort to build implementable system (LIM)

GSBPM

- <https://statswiki.unece.org/display/GSBPM/Generic+Statistical+Business+Process+Model>
- Outline of statistical life-cycle processes
- Eight main phases
- Each phase has subparts
- Adopted by agencies to classify IT efforts and systems

DDI

- DDI Alliance - <https://www.ddialliance.org/>
- Consortium of data libraries, archives, producers, researchers
- Two threads
 - ▶ Codebook – data dictionary, not reusable metadata
 - ▶ Lifecycle – GSBPM-based
 - reusable, extensive methodology, includes Codebook
 - GSIM profile
- Both bound to XML, so immediately implementable
- University and commercial software available
- Yearly user conferences: NADDI, EDDI

SDMX

- <https://sdmx.org/>
- Managed by BIS, ECB, Eurostat, IMF, OECD, UNSD, WB
- For exchange of dimensional data
 - ▶ N-cubes, time series, other
- Based on XML, so implementable
- Complex learning curve
- Extensive installed base
- Yearly user conferences



Questions



Contact Information

Dan Gillman
(202) 691-7523
Gillman.Daniel@bls.gov

