Metadata Systems for the U.S. Statistical Agencies, in Plain Language

Daniel Gillman (BLS), Peter B. Meyer (BLS), Francisco Moris (NSF/NCSES), William Savino (Census), Bruce Taylor (IES/NCES) 
(For SCOPE metadata team)

FCSM /CSPOS, July 10, 2020
Views presented by the authors do not necessarily represent the views of their agencies.
Outline

Goals: harmonization, interoperability, machine readability.
- Our agencies can do better at this, cooperatively. This is a primer.
- Basics of metadata – what you need to interpret a statistic: 11.3
- Metadata systems -- libraries, museums, data.gov, classification
- Guidance from US and international institutions
- Example projects
- Recommendations and takeaways
Metadata for tangible objects

- Metadata has a long history notably for libraries

- Library catalog systems
  - Authors, titles, when and where published, length, type, topics
  - Recent standards track separately books as concept/content and as physical items

- Museum catalog
  - For each item in the collection
  - Name, maker, where it was made, provenance, type, materials, size, dimensions, conditions, legal restrictions, location, and photos

- A smartphone’s components meet many metadata standards
  - “Since metadata are data, then metadata can be stored in a database…” (ISO/IEC 11179).
Definitions

- Metadata associated with a data set helps use, describe, interpret, and organize it.

- *Statistical metadata* are data used to describe *statistical objects*.

- Information used in this *role* are metadata.

- Metadata may include:
  - *data description*: variable names, units, frequency
  - *definitions, methodology*
  - *microlevel detail on collection or processing or paradata*.

A spreadsheet

<table>
<thead>
<tr>
<th>Month</th>
<th>Forecast</th>
<th>Sales</th>
<th>Variation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan 17</td>
<td>42,000</td>
<td>38,532</td>
<td>-3,468</td>
</tr>
<tr>
<td>Feb 17</td>
<td>45,000</td>
<td>41,934</td>
<td>-3,066</td>
</tr>
<tr>
<td>Mar 17</td>
<td>45,000</td>
<td>42,163</td>
<td>-2,837</td>
</tr>
<tr>
<td>Apr 17</td>
<td>45,000</td>
<td>43,050</td>
<td>-1,950</td>
</tr>
<tr>
<td>May 17</td>
<td>45,000</td>
<td>45,145</td>
<td>145</td>
</tr>
<tr>
<td>Jun 17</td>
<td>48,000</td>
<td>47,745</td>
<td>-255</td>
</tr>
<tr>
<td>Jul 17</td>
<td>48,000</td>
<td>49,623</td>
<td>1,623</td>
</tr>
<tr>
<td>Aug 17</td>
<td>46,000</td>
<td>52,539</td>
<td>4,539</td>
</tr>
<tr>
<td>Sep 17</td>
<td>45,000</td>
<td>47,324</td>
<td>2,324</td>
</tr>
<tr>
<td>Oct 17</td>
<td>45,000</td>
<td>44,700</td>
<td>-300</td>
</tr>
<tr>
<td>Nov 17</td>
<td>42,000</td>
<td>44,923</td>
<td></td>
</tr>
<tr>
<td>Dec 17</td>
<td>48,000</td>
<td>51,120</td>
<td></td>
</tr>
</tbody>
</table>

Total: 546,000

Spreadsheets contain a few metadata fields:
- *tab names*,
- *table names*,
- *column names*,
- *user comments*. 
Statistical metadata

Typical statistical objects

- Concepts (especially their definitions)
- Variables
- Value domains (allowed values) for variables
- Classifications systems, code lists, and individual categories
- Questionnaires and forms
- Data collection questions
  - Wording, Response choices, Flows (skip pattern)
- Instruments (implemented questionnaires)
- Sampling plans
- Estimators
- Processing
  - Editing, Coding, Allocation
- Data sets
- Tables and N-cubes

Statistics are conceptual not tangible

Statistics and data are related to **concepts**. Statistics have semantic relations to other values, e.g. percentages of something.

One aspect: Statistics and datasets have **dimensions**, e.g. unemployment rate for young Hispanic males in PA.
Data sets and data.gov

- Documentation of a dataset is metadata
  - *Descriptive metadata* includes the methodology and year of data collection

- Data.gov lists Federal data sets
  - It shows information agencies share in a standard format on their own web sites
  - Data.gov’s Open Metadata Schema is in JSON format

---

**Dataset Fields**

See the [Further Metadata Field Guidance](#) section to learn more about the use of each element, including the range of valid entries where appropriate. Consult the [field mappings](#) to find the equivalent v1.0, DCAT, Schema.org, and OKAN fields.

<table>
<thead>
<tr>
<th>Field</th>
<th>Label</th>
<th>Definition</th>
<th>Required</th>
</tr>
</thead>
<tbody>
<tr>
<td>@type</td>
<td>Metadata Type</td>
<td>IRI for the <a href="#">JSON-LD data type</a>. This should be <code>dcat:Dataset</code> for each Dataset.</td>
<td>No</td>
</tr>
<tr>
<td>title</td>
<td>Title</td>
<td>Human-readable name of the asset. Should be in plain English and include sufficient detail to facilitate search and discovery.</td>
<td>Always</td>
</tr>
<tr>
<td>description</td>
<td>Description</td>
<td>Human-readable description (e.g., an abstract) with sufficient detail to enable a user to quickly understand whether the asset is of interest.</td>
<td>Always</td>
</tr>
</tbody>
</table>
Data dictionaries for variables

- Attributes typically included when describing variables include:
  - The concept a variable represents (say, marital status)
  - Value domain (<s, single>, <m, married>, <sp, separated>, <d, divorced>, <w, widowed>)
  - Datatype
  - Universe (say, adults in the US).

- Data dictionaries, variables, and variable attributes can be reused

- Goals: usability, interoperability, machine readability
  - A variable definition or data dictionary can be reused by URL/URI
  - Interoperability: Linked open data and RDF
  - Helps interoperability, inference, and prediction
Our data sets use classifications for discrete, qualitatively distinct groups

Example: Occupations
- Population Censuses 1850-2010 had detailed 3-digit occupation lists
- Many occupation category systems across time/place
  - SOC, O*NET, ISCO, HISCO, each of many countries, many versions and variants

A data observation can say: occupation 55.
- To interpret it, one needs to know which classification system it’s from.
- Want to compare it to observations across time, data sets
- Those are metadata issues.
- Crosswalks or concordances match categories
  - These are tables or decision trees; Machine learning can help
  - Classification management systems software can help track

There are too many other classification systems to name: Industries: Census/ACS/CPS, SIC, NAICS, ISIC; geographies; jurisdictions; illnesses and injuries; medical procedures; crops; types of schools, components of GDP, technologies in patents, ...
Metadata for surveys

- What question was asked to produce the variable in the final data set?

- Our **unemployment rate** comes from survey data
  - It’s a function of specific questions the respondent was asked
  - Are you in the civilian US population? Are you working? Hours? If not working: are you searching for work?

- To interpret income or earnings, after-tax, year, bonuses, stock options
  - DDI Life Cycle standards address these issues
  - NCHS’s Q-Bank
Storage and transmission of metadata

- Same formats can be used to store and transmit metadata
- For sharing data, machine-readable metadata should be sent along with it.
  - A Web API may send back data and metadata in this kind of XML. This one is from BEA.
  - The first 3 lines here have metadata so the client computer can interpret the rest as a table.

Example Return:

```xml
<BEAAPIT>
  <Request>
    <RequestParam ParameterValue="GETDATASETLIST" ParameterName="METHOD"/>
    <RequestParam ParameterValue="Your-36Character-Key" ParameterName="UserID"/>
    <RequestParam ParameterValue="XML" ParameterName="RESULTFORMAT"/>
  </Request>
  <Results>
    <Dataset DatasetDescription="Standard NIPA tables" DatasetName="NIPA"/>
    <Dataset DatasetDescription="Standard NI underlying detail tables" DatasetName="NITUnderlyingDetail"/>
    <Dataset DatasetDescription="Multinational Enterprises" DatasetName="MNE"/>
    <Dataset DatasetDescription="Standard Fixed Assets tables" DatasetName="FixedAssets"/>
    <Dataset DatasetDescription="International Transactions Accounts" DatasetName="ITA"/>
    <Dataset DatasetDescription="International Investment Position" DatasetName="IIP"/>
    <Dataset DatasetDescription="GDP by Industry" DatasetName="GDPbyindustry"/>
    <Dataset DatasetDescription="Detailed Regional Income data sets" DatasetName="RegionalIncome"/>
    <Dataset DatasetDescription="Detailed Regional Product data sets" DatasetName="RegionalProduct"/>
    <Dataset DatasetDescription="Retrieves various Regional datasets" DatasetName="RegionalData"/>
  </Results>
</BEAAPIT>
```
External guidance and constraints

- Metadata is in U.S. laws and regulations
- The UNECE family of metadata standards (GSBPM, GSIM)
  - Statistical business process model
  - Information models
- FAIR principles for scientific data
  - Findable, Accessible, Interoperable, and Reusable

It’s helpful not to rebuild from scratch; adopt standards implementation that meet guidance already

US laws specify metadata

- National Archives and Records Act of 1934
- Freedom of Information Act of 1967
- Privacy Act of 1974
- Paperwork Reduction Act (PRA) of 1995
- Digital Accountability and Transparency Act (DATA) of 2014
- Geospatial Data Act of 2018
- Information Quality Act
- Executive Order 13859 on Maintaining American Leadership in Artificial Intelligence (2019)
- Financial Transparency Act of 2019
- Grant Reporting Efficiency and Agreements Transparency (GREAT) Act of 2019

More on next slide
Recent U.S. laws and regulations

- OMB Directive M-13-13
  - Defines data.gov and its standards
- Federal Data Strategy
  - Guiding principles that encourage harmonizing federal data
  - Notably: reuse
- Evidence-Based Policymaking Act
  - Make harmonization of data easier for policy conclusions
  - Open government, open machine-readable formats
  - Encourages Web APIs
  - Codifies CIPSEA law
Example systems, dictionaries, projects

- FGDC – Program effort to develop geographic data standards
- NIEM – For interoperable data used for security, defense, public safety, justice, intelligence, and emergency management
- GIDS – Software to generate diverse Census questionnaires for the Economic Census, which differ for each industry

Non-Federal:
- ICPSR’s DDI codebook and thesaurus
- Wikidata, Schema.org, SDMX, JSON-stat, . . . many more
ROI on metadata systems is not all in terms of money.

**Costs include:**
- Coordination of subject, survey, and IT specialists
- within and across agencies

**Benefits: Metadata helps**
- Facilitate use of our data directly
  - Help integrate and interoperate from other sources/agencies
  - Simplify questions users pose to us
- Retain organizational knowledge
  - Help address risks and costs of obsolescence of code & data; transparency
- Develop future systems
  - To reuse survey questions, definitions of variables and classifications (DDI Codebook)
- Conduct research – ours and others
  - Statistical agency staff need to be involved
Recommendations and takeaways (1)

- **Reuse** established terminology, classifications, metadata schemes
  - Plan and share with other statistical agencies
  - Saves time and achieves interoperability and comparability
  - Meet standards and FAIR principles

- **Build** small; think big
  - Implementations help shape standards and vice versa
  - Identify opportunities and stakeholders for metadata systems
  - Try tools from other subject matter
  - Partner with external services working with Federal data
    - E.g. Google, Statistics USA, IPUMS
**Recommendations and takeaways (2)**

- **Learn** metadata tools
  - Engage with experts; attend conferences
  - Connect to professional groups and international institutions
  - Know the lingo – our glossary may help
  - Be aware of metadata standards for relevant subjects
    - DDI, FGDC/NGDA, NIEM, SDMX, GSIM, GSBPM
    - See tools and guidance at Data.Gov and Federal Data Strategy Action Plan
  - Take training; we can develop training together

- **Participate** in and advocate machine-readable metadata
  - Statistical agencies can enhance data.gov’s Open Metadata Scheme
    - Standardized data dictionaries, seasonal adjustment tag, classification management
  - Use other metadata standards relevant to statistics
Contact

Any questions?

What else should we know about metadata issues?

Contact for the SCOPE Metadata team
Dan Gillman, team lead, Gillman.Daniel@bls.gov