Thoughts on Inter-operable Metadata for Data Quality, Transparency, and Reproducibility

CNSTAT Panel Study on
Transparency and Reproducibility of Federal Statistics for NCSES

Francisco Moris, Ph.D.

www.ncses.nsf.gov

21 May 2019

What to expect

- Take-aways
- Definitions & Inter-operability in U.S. Data Policy
- Challenges and opportunities
- Scenarios with possible outcomes
- Questions for the Panel

Take-aways

- **Inter-operable** metadata – metadata that can be used across surveys & IT platforms – & **selected core** [cross-agency] metadata

    has the potential to contribute to **quality and transparency goals** reaffirmed in recent data policy.

- Metadata are also important for **dissemination and data integration** with administrative & other Big Data.

- However, developing inter-operable metadata (within or across agencies) is challenging and remain an **elusive goal that can benefit** from the work by this Panel.

Definitions

- "Metadata is defined to be **data that defines and describes** other data". "Since metadata are data, then metadata can be stored in a database…" (ISO/IEC 11179). Focus of presentation: statistical metadata.

- Metadata types by content:
  - data description: variable names, units, frequency; sectors (~labels in tables)- '*structural metadata*'
  - concepts, definitions, methodology - '*reference metadata*'
  - microlevel detail on collection or processing - '*paradata*'

- by format: textual 'documentation' vs. machine-readable (e.g. coded metadata fields or variables)

- by scope: metadata about microdata, aggregate estimates, datasets, linkages, etc.

A spreadsheet
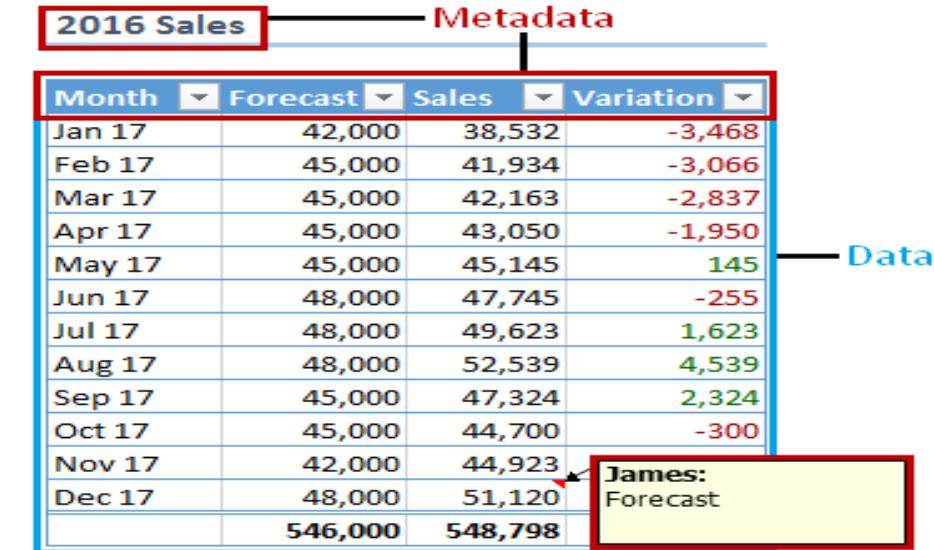
2016 Sales			Metadata
Month	Forecast	Sales	Variation
Jan 17	42,000	38,532	-3,468
Feb 17	45,000	41,934	-3,066
Mar 17	45,000	42,163	-2,837
Apr 17	45,000	43,050	-1,950
May 17	45,000	45,145	145
Jun 17	48,000	47,745	-255
Jul 17	48,000	49,623	1,623
Aug 17	48,000	52,539	4,539
Sep 17	45,000	47,324	2,324
Oct 17	45,000	44,700	-300
Nov 17	42,000	44,923	
Dec 17	48,000	51,120	
	546,000	548,798	

James: Forecast

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

Definition & Stages or degrees of Inter-operable Metadata

Inter-operable metadata: able to be shared or used across surveys & IT systems.

Three stages or degrees:

Harmonization: identifying
selected core metadata
e.g., Common Metadata
Vocabulary

Harmonized

Quantitative
/Codified

Machine readable
XML, JSON, XBRL, SDMX
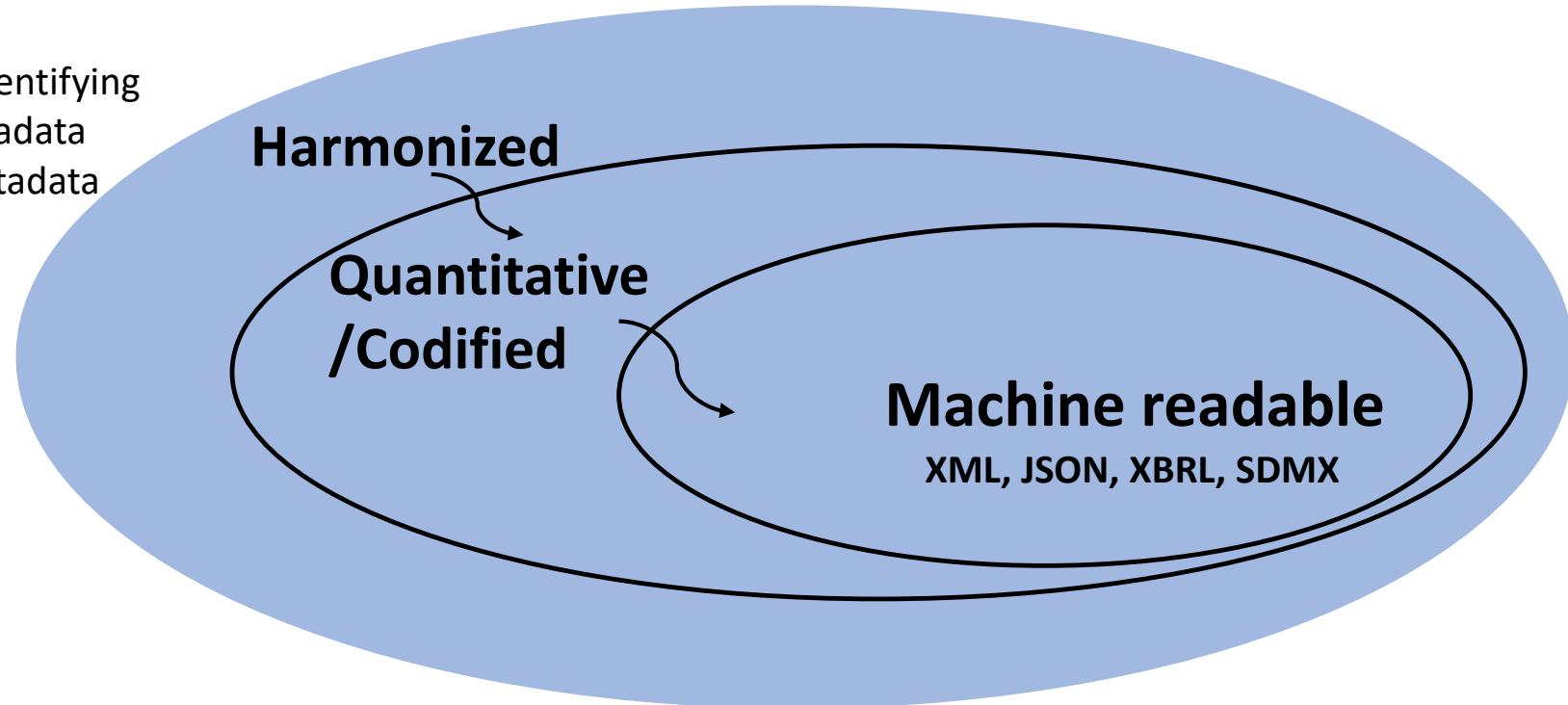
Stages or degrees of Inter-operable Metadata

Harmonization of statistical metadata is <u>perhaps the most difficult step</u> - needs consensus from
- relatively autonomous units and/or
- specialized communities of practice.

<u>But there are also examples</u> of harmonized metadata for Federal, commercial, and international data.

Examples of harmonized metadata

- Cross-agency metadata standards for geographic and geospatial data: FGDC (since 1990s) & Geospatial Data Act 2018 (PL 115-254)
- Convergence of voice, data, and video in telecommunications leading to today's mobile devices and emerging 'IoT'
- Eurostat, OECD, and Central Banks cross-country metadata tools for socio-economic data

Inter-operable data and metadata supports U.S. Data Policy…

1- OPEN Government DATA Act -Title II of Foundations for Evidence-Based Policymaking Act (2019)

"Federal Agency Responsibilities **to Make Data Open by Default**":

Open/transparent data: "Open government **data** asset means a public data asset that is—

(A) **machine-readable**; ["format that can be easily processed by a computer without human intervention"]

(B) available (or could be made available) in an open format;

(C) not encumbered by restrictions, other than IP rights…and

(D) based on an underlying open standard that is maintained by a standards organization".

Inter-operable data and metadata supports U.S. Data Policy…

2- OMB M-19-15 Improving Implementation of Information Quality Act (April 2019) – Cites M-13-13 & A-130

"Reproducibility requires… transparency about methods.., design…parameters…"

M-13-13: "…establishes a framework to …promote <u>interoperability and openness</u> [transparency]…. This includes <u>using machine-readable and open formats</u>, …and <u>common core</u> and extensible metadata."

A-130 Managing Information as a Strategic Resource (2016): "…information in a way that supports downstream <u>interoperability</u> among information systems and streamlines dissemination…in <u>machine-readable open formats</u>, using relevant data standards…includes standard extensible metadata..."

3- Federal Data Strategy (ongoing)

Practice 2: "Inventory Data Assets with sufficient …quality, and <u>metadata</u> to facilitate …<u>discovery, access, and use</u>." Practice 6: "Publish Data <u>Documentation</u>: …in publicly accessible <u>digital repositories</u>"

… metadata also supports integrating data from multiple sources

Recent guidance & reports on data integration

cover the role of metadata:

- OMB M-14-06 Guidance for Providing & Using Administrative Data for Statistical Purposes

- Data Quality Assessment Tool for Administrative Data (FCSM 2013)

- Federal Statistics, Multiple Data Sources, and Privacy Protection, Next Steps (NASEM 2017)

# Challenges and opportunities for statistical metadata

- Great agency-specific metadata tools and Federal research on metadata – as illustrated in presentations later today!
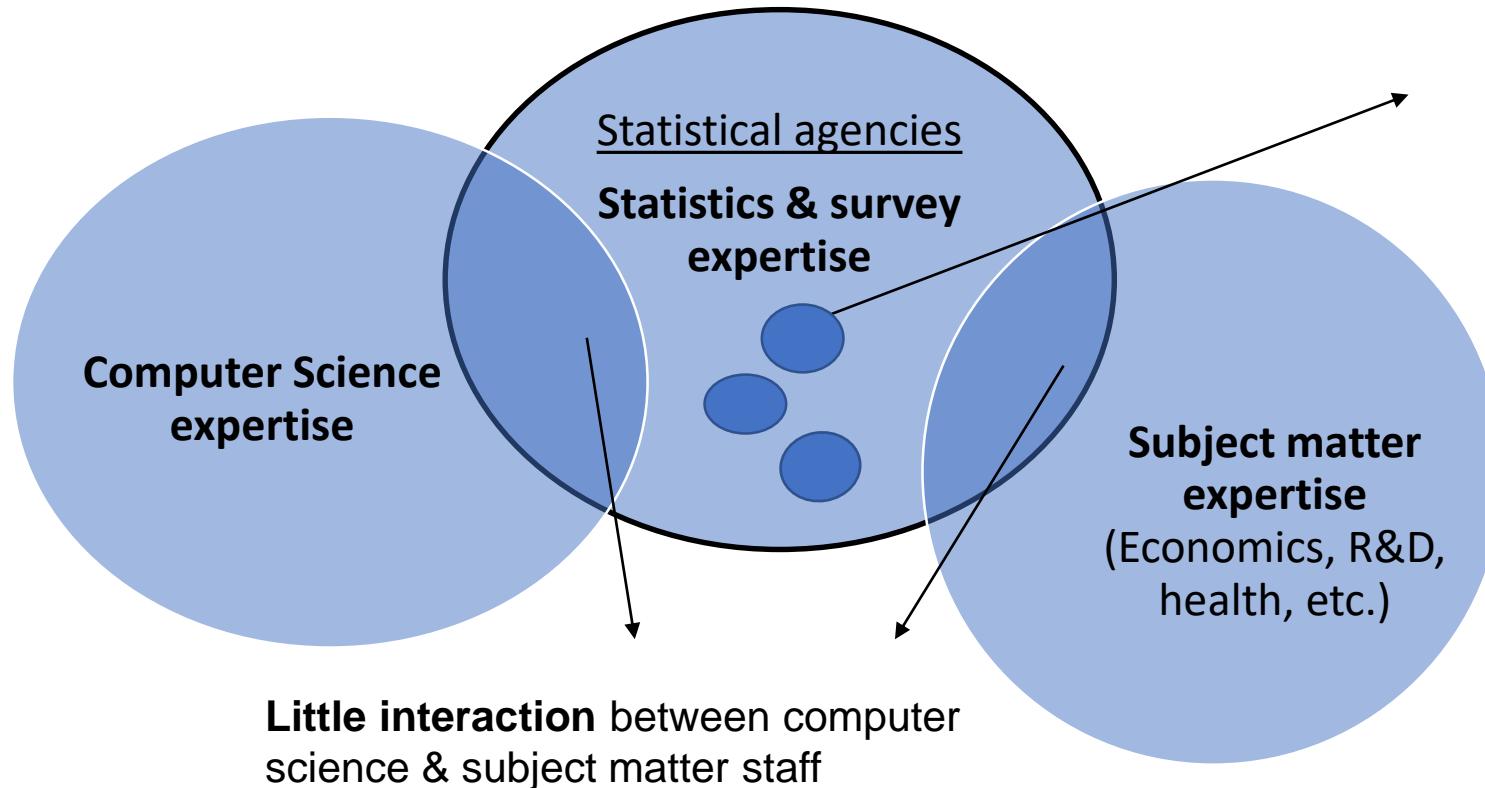
Opportunity: Facilitating inter-disciplinary collaboration (Data Science focus) to develop inter-operable metadata

Computer science agency staff

GSA Github, Open data tools, and '18F'

Statistical agencies
Statistics/survey expertise

Computer Science/IT tools

Subject matter expertise

Further collaboration leveraging FCSM WGs, GSA digital services, and int'l tools
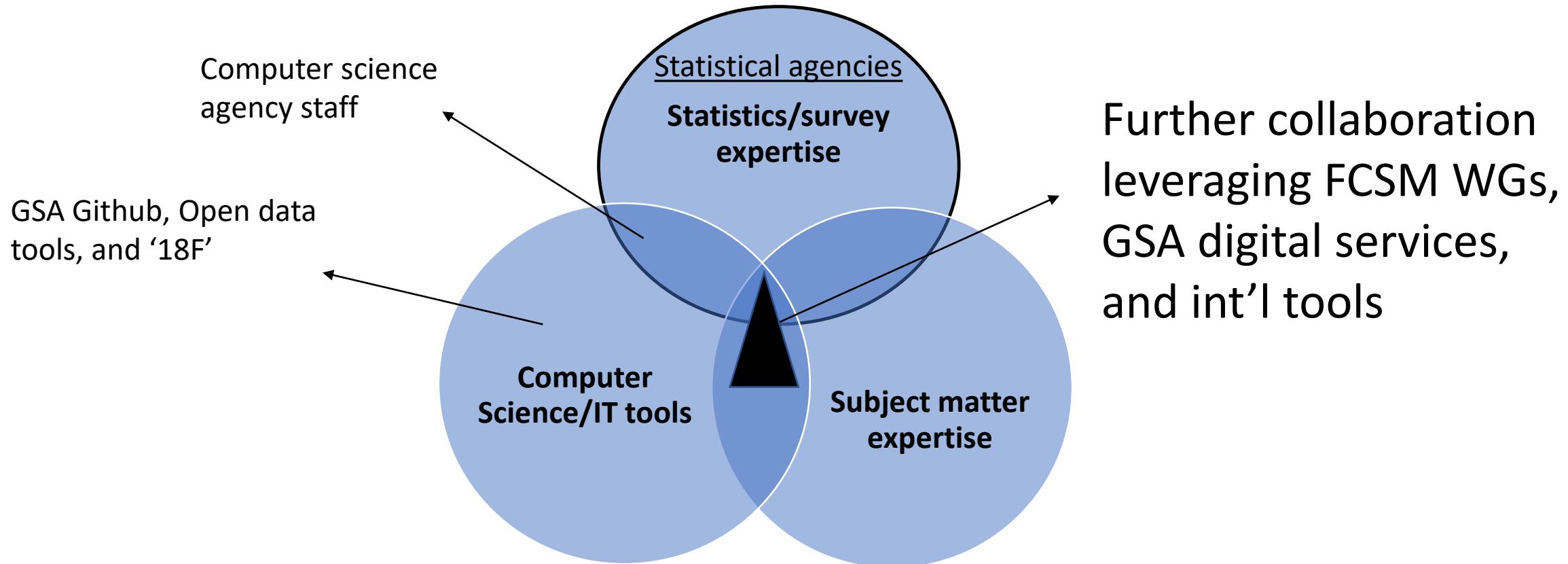
Scenarios with possible outcomes building on this panel (post-panel)

Federal System-wide	NCSES-specific
Near-term	
Data APIs with some common metadata fields across agencies**?** (freq., unit of measurement, geographic labels**)**	Machine-readable data from Federal and other contractors**?**
Open metadata schema (data.gov) with enhanced fields for statistical data**?**	
Long-term	
Cross-agency metadata systems**?**	Metadata mgmt. system for selected domains (business, HR statistics)**?**
Directive on inter-operable metadata and metadata systems**?**	

Questions for the Panel on transparency and metadata quality

1. How agencies could better address *inter-disciplinary needs* to advance metadata quality?

2. Given complexity of metadata across agencies, only selected (cross-domain) metadata can be harmonized.
   Which metadata should be considered as *starting points for metadata inter-operability*?

3. Which *domestic and international metadata tools and platforms* may be leveraged by NCSES and other agencies in the *near and long-term*?