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

by SCOPE Metadata team[1]
10 July 2020

Abstract

This paper explains the benefits and challenges of metadata systems for U.S. statistical agencies. Metadata systems facilitate the discovery, accessibility, understanding, and use of statistical data. These systems can also increase interoperability and comparability across agency surveys and other data sources, domestically and internationally. New processing and dissemination tools and new policies such as the Evidence-Based Policymaking Act and the Federal Data Strategy represent opportunities for improving metadata practice by the U.S. statistical agencies. This paper discusses basic metadata background, including:

- definitions and descriptions of metadata;
- examples of metadata systems from statistical offices, libraries, and museums;
- an overview of relevant laws, Executive Orders, regulations, and standards;
- how metadata systems are put into practice; and
- recommendations to the statistical agencies.

[1] The SCOPE Metadata team is an informal, longstanding interagency group of U.S. Federal statistical agency representatives. Authors here include Daniel Gillman (BLS), Kathryn McNamara (Census), Peter B. Meyer (BLS), Francisco Moris (NSF/NCSES), William Savino (Census), and Bruce Taylor (IES/NCES). We thank Leighton Christiansen of the Bureau of Transportation Statistics, Marilyn Seastrom of IES/NCES, and Sara Snyder and Diane Shaw of the Smithsonian Institution, and other agency colleagues for helpful comments. Any remaining errors are the responsibility of the authors:

1) BLS – The opinions, analysis, and conclusions put forth in this paper are solely the authors' and do not necessarily reflect those of U.S. Bureau of Labor Statistics or the Department of Labor.
2) Census – All views expressed in this paper are those of the author(s) and not necessarily those of the U.S. Census Bureau.
3) NCES – This paper is intended to promote the exchange of information. The views expressed do not necessarily reflect the position of the U.S. Department of Education.
4) NCSES – The opinions expressed and arguments employed herein are solely those of the authors and do not necessarily reflect the official views of the National Center for Science and Engineering Statistics (NCSES) within the U.S. National Science Foundation.

Table of contents

Introduction

This paper describes metadata systems and possible benefits from more widespread adoption of these systems and related practices by statistical agencies and their survey programs.[2] Metadata systems can facilitate user discovery, access, understanding, and usage of final statistical outputs and products. Users of these systems can benefit from greater access and interoperability across multiple data sources. For agencies producing statistics, metadata systems can support production, processing, and administrative activities. The paper also covers metadata standards and regulations, and provides examples of metadata projects, to encourage the development of new ways to take advantage of metadata tools.

As discussed more fully below, metadata are data used to describe statistical objects, and a metadata system is a set of tools and procedures, usually implemented in software, to make this information usable and shareable.

Metadata have always been an integral part of statistical data. For others to use a data set, they have to be able to find and understand it. Metadata provide the necessary information, whether they are printed on a page, or contained in formatted text, a document file, or a machine readable data set.

Over time, the medium of choice for the delivery of electronic data has changed. This evolution has involved tapes, disks, CDs, and the Internet. As a result, the complexity of systems managing metadata has increased, too. Codebooks and documents provide metadata in a form for humans to read. More sophisticated electronic metadata systems provide both human and machine-readable metadata. As data are distributed online, users need to integrate data and metadata from multiple sources, while protecting privacy and confidentiality (NASEM 2017a, 2017b). Machine-readable metadata further require interoperability across underlying IT platforms.

As we will discuss below, metadata systems help meet a number of external criteria given by the Evidence-Based Policymaking Act, the Federal Data Strategy, the FAIR principles, and widely used metadata standards. Agencies can comply with a variety of standards if they have an effective metadata system.

The next sections cover 1) definitions and goals for metadata systems; 2) examples of metadata systems; 3) relevant US federal laws, orders, and regulations; 4) how metadata systems are put into practice; and 5) recommendations. Appendix 1 contains a glossary of relevant terms and Appendix 2 lists some relevant software companies and systems.

[2] This paper does not cover paradata in a narrow sense of data about the process of data collection (Kreuter 2018a). At the same time, to the extent that "boundaries between data, metadata, and paradata are sometimes blurred" (Couper 2017), some of the information and tools discussed in this paper may be of interest to those interested in documenting or analyzing the survey process.

Basics of metadata

Metadata was originally defined as *data about data.*[3] This definition held until sometime in the 1990s when online digital libraries came into existence. In online digital libraries, metadata has come to mean *data describing a set of objects or resources,*[4] because digital libraries describe things in addition to data. Many museums and libraries have online catalogs describing the artifacts and books they curate and manage.

In this paper, metadata are data used to describe *statistical objects*. Information used in this *role* are metadata. In a traditional statistical survey, for instance, data are collected from persons in households or business establishments. Statistical agencies develop metadata systems to describe the survey questions, units of measure in the responses, steps used to process and analyze the data, sampling plans for ease of collection and control for error, and the observations and time series that result. Data users may also construct metadata from data sets published by an agency.

Metadata can serve several roles. Metadata can make clear who has access to data across the production cycle and where data are stored; these are *administrative* purposes. In *conceptual* or *semantic* roles, metadata describe what the variables and categories mean and how they can be used. Metadata relate directly to *business* processes by framing who the creators and users of the data are. Metadata can address *publication* or *dissemination* issues by specifying display attributes of variables and when data were released.

Metadata helps to relate the concepts in data sets to one another or to administer data sets. Metadata can include variable names, data types, descriptions, display attributes of variables, keywords, and the category systems used in the data set (e.g. NAICS, the North American Industry Classification System). The roles for this kind of metadata lead to it being called *descriptive*, *semantic*, or *detail metadata*. If the data came from a questionnaire, then relevant metadata may include skip patterns, dependencies, and ordering of questions. These help frame how compound objects in a system fit together and here the role is often called *structural metadata.* Some information is associated with how to manage data sets – file locations, file permissions, dates of creation and modification, and whether files are public or exposable by Freedom of Information Act (FOIA) requests. This role is usually identified as *administrative metadata*. All of these roles overlap. For additional information see the glossary and Wilson (2018).

There are many resources for an office to consult when building a metadata system. A worldwide community develops standards and models for the kind of metadata that statistical offices use. Metadata management is more efficient if descriptions of fields can be reused, not reinvented. The UN's Economic Commission for Europe (UNECE)'s High-Level Group for the Modernisation of Official Statistics coordinates development of the main standards for metadata and for the business activities of statistics-producing offices.

[3] Metadata was originally defined by Bagley in 1968 and independently by Sundgren in 1973 at Statistics Sweden as "data about data."
[4] See the Dublin Core Metadata Initiative, DCMI.

Technical committees under the International Organization for Standardization (ISO) have developed some metadata standards used by statistical agencies.

Typical statistical objects

Since metadata are data that describe objects, we need to consider the kinds of objects in scope for statistical metadata. The objects a metadata system is meant to describe vary from system to system; the selection of the objects depends on the purposes of the system. In general, the statistical community tracks objects such as:

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

For instance, metadata about a data set includes descriptions of the data set itself and the underlying data. For a machine readable data set, describing the variables is just as important as describing the set overall.

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 (in the case of marital status, nominal datatype)
- Universe (say, adults in the US).

Once a list of attributes is developed, they may be reused for similar data sets elsewhere. Reuse of metadata terms is a desirable practice to reduce costs and facilitate inter-operability across datasets. In addition to describing objects one by one, it is possible to use metadata and design systems to manage them by identifying similarities among objects. For instance, in our example above, we apply the concept of 'marital status' to a variable. There might be many such variables, especially if a single survey produces a similar data set on a regular ongoing basis. A statistical office might conduct several

surveys that collect a marital status variable. If the same concept applies to each, write that definition once and apply it each time it is needed. We call this the *reusability principle*: write once, use many times.

Metadata systems

A metadata system is a system built to organize metadata. In later sections, we briefly describe many such systems with various levels of complexity. However, there are some common elements and attributes that each system should have.

Repositories, databases of metadata, are a basic component of metadata systems. The organization of the repository is based upon a schema (e.g. column headings and data types) to help formalize and organize the attributes, or metadata elements, of the system.

Another component is the user interface. Users can be humans or other systems. The interfaces will be built based on which kind of user they address. In each case, an API (Application Programming Interface) is built to communicate with the repository, and the API gets its guidance, i.e., what commands to send to the repository, from either a human or another system. The API sends a query desired by the user to the repository and returns metadata to the user.

An increasingly important attribute of metadata systems is interoperability. A metadata system may be interoperable, or not, at different levels of generality, for example, across surveys of a given program or agency, or across a statistical domain within or across agencies (e.g. employment statistics; trade statistics).

A system is interoperable in two main ways: a user is able to operate with the system without help (system interoperability); and the output of the system is understandable to the user without help (semantic interoperability). Both types of interoperability are achievable through the adoption of standards, which provide an open and sharable way to express a specification. Some specifications provide schemas (for the design of repositories, such as the Data Documentation Initiative's DDI Lifecycle), others contain definitions and specifics for conveying shared semantics, such as the categories and organization of NAICS and SOC (Standard Occupation Classification).

International standards and guidance

Adoption of standards

Standards built through a consensus process that includes participants from all the stakeholder communities stand the best chance of being successfully adopted by each potential user. This is because a specification built on consensus by a group including more stakeholder communities stands a better chance of being accepted by users and satisfying the needs of those communities. Further, international standards facilitate cross-country comparisons and reporting to international organizations where the U.S. has been a leading participant in developing statistical data and metadata guidelines.

In a setting characterized by multiple Federal agencies, levels of governments, and private data stakeholders, building metadata systems through the adoption of consensus metadata standards has several advantages:

- Developers do not have to spend time specifying the types and organization of the metadata required, as this has already been thought through.
- Organizations that conform to the same standards gain interoperability. This simplifies the laborious task of finding a mapping between systems.
- Conformance happens at systems interfaces (either through a human user interface or an API), which means that each adopting organization can build its system in its own way, to optimize performance for its own particular needs.
- Transparency of data and systems depends on sets of metadata attributes being available, and this can be achieved through adopting standards that specify those attributes. For example, the MARC (Machine Readable Catalog) library standard allows a person or machine to browse the catalog of the books in the library inventory online. This leads to transparency, supported by the MARC standard.

The UNECE family of metadata standards

The United Nations Economic Commission for Europe (UNECE), headquartered in Geneva, manages a number of cooperative international projects of interest to statistical offices within the UNECE and sometimes around the world. These efforts include demographic, economic, methodological, and computing related activities.

Among the results of the computer-related activities is a family of standards designed to describe the work of statistical offices and modernize the development and management of statistical processing throughout each office. These standards include a broad and deep view of the metadata needs for a statistical office. They are listed here and described further in Appendix 1:

- GSBPM − Generic Statistical Business Process Model. This standard includes an outline of the processes statistical offices use to plan, design, and conduct surveys and other statistical programs.

- GSIM − Generic Statistical Information Model. This standard includes a model defining and linking the main information objects needed to describe statistical programs including surveys and the terminology for managing classification systems that change over time such as for industries, occupations, and products.
- GAMSO − Generic Activity Model for Statistical Organizations. This standard lays out the activities (some non-statistical) a statistical office needs to manage, including those defined in GSBPM.
- CSPA − Common Statistical Production Architecture. This lays out a structure to organize software components needed to conduct statistical programs.
- CSDA − Common Statistical Data Architecture. This document, under development, will offer an architecture for managing data within a statistical office.

FAIR Guidelines

The FAIR guidelines are a generic set of principles for direction on how to make scientific data Findable, Accessible, Interoperable, and Reusable. Each of these four goals is subdivided into 3 or 4 specific principles (15 in all) that can be applied to both data and metadata. Originally developed for scientific data, these principles are gaining visibility, are in wide use, and adopted within an increasing number of domains. See https://www.go-fair.org/fair-principles/ for more details. These and other international principles should be used in conjunction with the Federal Data Strategy and related U.S. regulations, discussed below.

Other international metadata standards

International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC) develop and manage standardization needs across many fields, sometimes jointly. National bodies make up the technical committees under ISO and ISO/IEC, which develop standards in various fields. Information technology standards fall under the purview of ISO/IEC, whereas geographic information standards are under ISO. Metadata standards fall under several subject matter areas in ISO.

An example is the ISO/IEC 11179 *Information technology – Metadata registries* standard.[5] In particular, ISO/IEC 11179 "is an international standard that documents the standardization and registration of metadata to make data understandable and shareable. This standardization and registration allows for easier locating, retrieving, and transmitting data from disparate databases."[6] The standard consists of six parts, including a framework, a meta-model for building a registry, and guidelines for forming data definitions. Also, ISO 19115 *Geographic information systems − metadata* is devoted to

[5] See https://www.iso.org/about-us.html & https://www.iso.org/obp/ui/#iso:std:iso-iec:11179:-1:ed-3:v1:en

[6] Pon R.K., Buttler D.J. (2009) Metadata Registry, ISO/IEC 11179. In: LIU L., ÖZSU M.T. (eds.) Encyclopedia of Database Systems. Springer, Boston, MA. pp. 1724–1727.

describing geographic areas and data sets, which are vital to describing statistical data. See Glossary for entries under ISO.

Other international standard-setting organizations include the DDI Alliance, the National Information Standards Organization (NISO), the Object Management Group (OMG), SDMX, the World Wide Web Consortium (W3C), and many others. Standards under the DDI Alliance and SDMX are described in the Glossary.

NISO develops library, information retrieval, and cataloging standards. For example, the Dublin Core, MARC, and thesaurus standards are part of their responsibilities. See the Glossary for descriptions of MARC and Dublin Core.

OMG oversees standards used in modeling systems. The best known and most used example is the Unified Modeling Language (UML). This modeling language is used to express many standards, designs, and systems throughout the world. UML is built into many modeling tools for visually presenting how data and metadata are organized in a system.

The W3C oversees standards for the World Wide Web. Increasingly, these efforts are devoted to machine-readable descriptions. Several important standards in this area are listed in the Glossary under W3C. Two important examples are Resource Description Framework (RDF), which is used to create a graphical organization of linked information. It is well-suited for developing a metadata system on the Web. Another standard is the RDF Data Cube Vocabulary. This standard recasts the SDMX data model into RDF, so it is machine-readable and designed to be used on the Web. SDMX is both the standard and the name of the sponsoring organization.

Metadata in U.S. laws and regulations

Here we highlight selected laws and regulations related to U.S. data standards.[7] Since the founding of our country, there has been the need to define, collect, classify, and report data. Statutes governing the decennial Census since 1790 and the 1800s stipulated variables and their descriptions (metadata) and how to collect the data. From 1820 to 1920 the government developed classification systems for industries, occupations, and patents but they were not easy for users to apply. The National Bureau of Standards, which became the current-day National Institute of Standards and Technology, was founded in 1901 to standardize variables and measurement procedures in the United States. It now sets some standards such as FIPS, the Federal Information Processing Standards.

Federal-wide standards appeared with the *National Archives and Records Act* of 1934 which established the National Archives to preserve and care for the records of the U.S.

[7] For an example from the EU perspective see the Glossary entry for the EU's General Data Protection Regulation.

Government. The National Archives sponsored advances in the usage and modern understanding of schemas, taxonomies, and standards for government data. The National Archives published the *Federal Register* and in recent amendments the federalregister.gov was established as the home for federal regulations and official requests for information and comment.

The *Freedom of Information Act* of 1967 requires the full or partial disclosure of previously unreleased information and documents controlled by the United States government upon request. Classifications were created for disclosure procedures and exemptions. In modern amendments FOIA.gov, and, each agency's [efoia.*Agency*.gov] or [*Agency*.gov/FOIA] were established.

The *Privacy Act* of 1974 established a code of fair information practices that governs the collection, maintenance, use, and dissemination of information (data) about individuals that is maintained in systems of records by federal agencies. The law requires that agencies notify the public of their databases about individuals by publication of a SORN (systems of records notice) in the *Federal Register*, now federalregister.gov. It prohibits the disclosure of a record about an individual from a system of records absent the written consent of the individual, unless the disclosure is pursuant to one of twelve statutory exceptions. The Privacy Act provides individuals with a means by which to seek access to and amendment of their records.

The *Paperwork Reduction Act* (PRA) of 1995 (44 U.S.C. 3501 et seq.) requires that agencies obtain Office of Management and Budget (OMB) approval before requesting most types of information from the public. "Information collections" include forms, interviews, and record keeping, to name a few categories. OMB is thus involved in the workflow of developing survey questions and enforces some standardization and attempts to limit duplication. OMB exercises control of some metadata.

The *Confidential Information Protection and Statistical Efficiency Act* (CIPSEA) of 2002, as amended (Title 44, Chapter 35—Coordination of Federal Information Policy), established uniform confidentiality protections for information collected for statistical purposes by U.S. statistical agencies, and allows data sharing among the Bureau of Labor Statistics, the Bureau of Economic Analysis, and the Census Bureau. Title III of the Foundation for Evidence-Based Policymaking Act of 2018 (PL 115-435) codified CIPSEA and enhanced provisions to use administrative data for statistical purposes by strengthening privacy and confidentiality protections for integrated and inter-operable data.

The *Open Data Policy – Managing Information as an Asset* M-13-13 (2013) followed an earlier Executive Order[8] by institutionalizing the principles of effective information management at each stage of the information's life cycle to promote interoperability and openness. This OMB memo requires the collection or creation of information in a way that

[8] *Making Open and Machine Readable the New Default for Government Information,* Executive Order of May 9, 2013.

supports downstream information processing and dissemination activities, including using machine-readable and open formats, data standards, and common core and extensible metadata for all information creation and collection efforts. The memo encourages designs which "define once, use many" and "collect once, use many." M-13-13 also mandated that Federal agency data inventories be made available in human-and machine-readable format to enable automatic aggregation at Data.gov, which had existed since 2009.

The *Digital Accountability and Transparency Act* (DATA) of 2014 established Government-wide data standards for financial data and provided consistent, reliable, and searchable spending data that is displayed accurately for tax-payers and policy makers on USASpending.gov. This act holds Federal agencies accountable for the completeness and accuracy of the data submitted to USASpending.gov. Subsequent guidance from OMB and the Department of the Treasury established financial data standards for Federal funds and standardized financial reporting elements (variables) such as the use of XBRL standard. This appears to be the first law specifically encouraging the use of a machine-readable interface for audit reports (an API).

The OMB Memo M-16-21, *Federal Source Code Policy* (2016), defines an infrastructure that new custom-developed Federal source code be made broadly available for reuse across the Federal Government. This is consistent with the Digital Government Strategy's "Shared Platform" approach, which enables Federal employees to work together—both within and across agencies—to reduce costs, streamline development, apply uniform standards, and ensure consistency in creating and delivering information. This memo established code.gov (hosted by GSA), which updates its metadata schema, API, compliance metrics, and engagement metrics based on agency inventories.[9]

The OMB Memo M-17-05, *Guidance on Federal Information Security and Privacy* (2016), establishes consistent government-wide performance and best practices to protect national security, privacy, and civil liberties while limiting economic and mission impact of security incidents. Agencies will move toward automated data collection and report to Performance.gov.

The *Geospatial Data Act* of 2018 (GDA) codified into law the existing Federal Geographic Data Committee (FGDC), as the primary entity for developing, implementing, and reviewing the policies, practices, and standards relating to geospatial data according to the guidelines and requirements issued by OMB. The Act codifies the National Geospatial Advisory Committee (NGAC), which will continue providing advice and recommendations to the chairperson of the FGDC relating to the management of federal and national geospatial programs, the development of the National Spatial Data Infrastructure (NSDI). The GSA defines the NSDI as "the technology, policies, criteria, standards, and employees necessary to promote geospatial data sharing throughout the Federal Government, State, tribal, and local governments, and the private sector." The Act directs the FGDC to set metadata standards and delegate the management of data themes (topic areas) to various federal agencies. In accordance with the law, the FGDC operates an electronic service

[9] See also https://resources.data.gov/schemas/dcat-us/v1.1/ and https://resources.data.gov/standards/

providing public access to geospatial data and metadata, known as the GeoPlatform and referred to also as the geospatial data "clearinghouse."

The *Foundations for Evidence-Based Policymaking Act* of 2018 contains Title I on evidence-based activities and Title II on using open data to further support the creation, access, use, and management of Federal data. Title I provides guidance for program evaluation and the organization of the evaluation function in Federal agencies. Title II updates data and metadata policies to, among other purposes, increase data access and transparency. Title III codifies CIPSEA (discussed above). The Act builds on the 2017 Commission report "The Promise of Evidence-Based Policymaking" which made recommendations about metadata. The law prescribed several practices including data inventories or online catalogs of data sets and requires Federal departments to designate statistical officials, chief evaluation officers and chief data officers. This paper identifies specific steps the agencies can take and explains how they aid those using statistical data as evidence for policy or research. The same steps help the data meet the FAIR criteria, which are discussed below.

The OMB Memorandum M-19-15 (2019) implements the regulations required by the *Improving Implementation of the Information Quality Act.* The new Guidelines reflect recent innovations in information generation, access, management, and use; and guide agencies in addressing common problems with maintaining information quality in implementing the Evidence Act and the Federal Data Strategy.

The OMB Memorandum M-19-18, *Federal Data Strategy -- A Framework for Consistency* is based on ongoing work under the Federal Data Strategy (FDS) effort. The FDS discussion has guiding principles that encourage harmonizing federal data in ways that increase their value. This could well lead to new guidance on metadata practices for the agencies, particularly geospatial data which has distinctive rules. It is complex and shared in large volumes between governments and private entities globally.

The Executive Order 13859 on *Maintaining American Leadership in Artificial Intelligence* (2019) directs OMB and other interagency councils[10] to update implementation guidance for Enterprise Data Inventories and Source Code Inventories so that their metadata supports discovery and usability in Artificial Intelligence (AI) research & development (R&D).

The *Financial Transparency Act* of 2019 requires federal financial regulatory agencies to adopt specified data standards with respect to format, searchability, and transparency. In this law, Congress is using more of the commonly accepted data management words such as data standards, schemas, taxonomy, metadata, "open" concepts, etc.

The *Grant Reporting Efficiency and Agreements Transparency* (GREAT) Act of 2019 (P.L. 116-103), an amendment to the DATA Act, requires the establishment and use of data

[10] One coordinating committee is the National Science and Technology Council (NSTC) Select Committee on Artificial Intelligence.

standards for information reported by recipients of federal grants. The Act calls for development of a "comprehensive taxonomy of standard definitions for core data elements required for managing Federal financial assistance awards." This law, too, uses established terminology for metadata and data management concepts.

Return on investment

The effort to design and adopt a metadata system has costs and benefits.[11] We refer to *net* benefits as "return on investment." Benefits or advantages of metadata systems can include:

- *Retention of organizational knowledge* – Documentation is metadata, and an organization's knowledge is often stored in document form. Normally, documentation is written as prose and stored in files meant for humans to read. Metadata systems are meant to store a shorthand version of documents by paring down the ideas conveyed in the documents into essential categories and numbers. These categories and numbers are metadata. When stored in a repository, they remain available as long as the system implementing the repository is maintained. Thus, the metadata system supports knowledge management and retention.

- *Enhancing data user experiences* – Metadata support finding, understanding, and using data. By providing relevant information, metadata enhance the public perception of the agency disseminating the data, fulfilling the critical mission of informing the public and facilitating the use of data. If multiple data users are combining data sets and need to construct similar metadata, the agency can save time for future users by publishing that common metadata.[12]

- *Reuse versus rewrite* – Metadata systems store information in a way that makes it easier to reuse and describe the same survey, data collection, or statistical program over many iterations. The Federal Data Strategy uses the phrase "Define once; use many times." For example, the prose documentation of a typical public use microdata file (PUMD) might be 100 pages long, and if this PUMD is released regularly, the documentation is rewritten each time. This is a large undertaking, takes time, ends up repeating much information from one cycle to the next, makes transcription errors possible, and inhibits comparisons of the data over time.

- *Enhancing data producer experiences* – Metadata support the production of statistics by documenting designs and processing steps, often in a machine readable way. Metadata can serve as the input parameters to production processing

[11] Schymik et al (2015) discuss costs and benefits in the context of enterprise search and document management. Studies found that metadata systems improved the users' capability to find what they wanted, but most users were unsatisfied with the system available to them.
[12] Manufacturers often improve products by adopting user innovations. (von Hippel, 2005)

systems, to trace or audit the development from data sources to final microdata and statistics, and to find inconsistencies and avenues for streamlining the production process.

- *System development* – Collection instruments can be developed from specifications (the underlying questionnaire). By laying out the individual questions, response choices, and flow logic as metadata, it is possible to produce a collection instrument by the push of a button. (One example is Q-Bank, discussed in the Glossary.) Metadata systems can also help data sets be more automatically interoperable on release, without planning every detail because the systems are designed to achieve that end through established standards; the data-producing system "knows" how the data is to be organized by users. The system is structured to work using the relevant categories; it doesn't just document them.

Metadata systems – practice and challenges

Metadata projects are often framed by (a) the intended use of the information to be managed in the project; (b) the contractors or other workforce whose tools, software, capabilities, and preferences are relevant; and (c) the standards adopted. These projects are multidisciplinary efforts. Staff working on these projects may include survey program staff, statisticians, methodologists, computer and information science experts, subject matter experts, and statistical officials or the Chief Data Officer.

Systems are also subject to constraints. Cost, time, functionality, and scope are all major factors in designing a metadata system, just as they are for any system. Metadata systems are usually new for any given survey or program, especially in smaller agencies, and the benefits are unknown. All this adds to an inertia that individual programs may find difficult to overcome without cross-agency support or guidance.

Metadata systems are best built iteratively, sometimes with narrow scope at each stage, with management and technical support at all levels of the hierarchy and at each stage of development of the system, with adequate funding for the purposes proposed, with support from the user community, and significant planning at each stage. Because metadata systems remain a relatively new idea within statistical offices, significant risks to management attend to each development project. These risks are attenuated by an iterative approach.

A metadata project can call for deep rethinking of how to reorganize past project information. It can expose past inconsistencies, ambiguities, mistakes, internal agency debates, and critiques. These can be sensitive areas. We have observed that this can be difficult and causes managers to delay or narrow the scope of metadata projects.

Metadata management is often not considered much in the early stages of the statistical or survey lifecycle. This contributes to the justifiable perception that going back to organize the metadata after-the-fact is too expensive and time-consuming. This problem

keeps repeating itself, so metadata systems rarely get built. Managers of surveys and other statistical activities often don't plan a metadata management function in their designs, so metadata management is usually not part of a system when it is instantiated. And as we just discussed, going back to create the metadata is perceived as too expensive. This problem could be remediated by formally building metadata into the documentation of the data collection.

Success is usually gained when all layers of management and technical staff are on board to build a metadata system. Top management, especially, need to be supportive, and this is true for two simple reasons: 1) they control the budget; and 2) they can direct those below them to perform certain tasks. Without this support, systems will not be built.

Another roadblock is scope. Often, after people see the potential in managing metadata, they want to build an all-encompassing system. These systems, sometimes called cathedrals, die because they are too ambitious. Therefore, expectations need to be reined in, scope needs to be managed, and an incremental approach needs to be taken.

This is especially true with metadata management, as new kinds of systems need to be built slowly. Lessons need to be learned and incorporated in future iterations. Potential advantages of metadata management are not always understood completely in the beginning, though that will probably change as more systems are built and experience gained. Finally, the scope of a planned system might need to be altered as resources can change mid-stream. All of these considerations are part of careful system development, but they bear repeating because they are so easily ignored, and they are most important when development of new functionality is being considered.

Examples of metadata systems

This section lists metadata systems from non-statistical applications to draw an inductive understanding of what statistical metadata can offer.

Catalog systems for libraries and museums

Libraries have longstanding metadata efforts. Library catalog records are built using bibliographic elements including MARC21 and Dublin Core. The most common type of record is MARC (MAchine Readable Catalog record). MARC records include a description of the item, subject headings, and a classification (such as a library call number). Each MARC record contains variable fields such as author, title, edition, etc., and each may be divided into subfields. The fields are associated with a 3 digit tag, such as 245 for "title." Since 2002, MARC records have followed rules laid out in the Anglo American Cataloging Rules second

edition (AACR2). Recently, the Resource Description and Access (RDA) standards have begun to supplement and in some cases replace the AACR2.[13]

In recent years, the Library of Congress has led efforts to make library cataloging more agile for a variety of formats. New developments include a pilot of the Bibframe standard, a linked data alternative to MARC.[14] Bibframe uses a linked data model to integrate bibliographic records.[15] The model relies on triples, three high-level "entities" with properties to describe a resource:

- Work: includes author, language, and subject
- Instance: includes publisher, date of publication, format
- Item: refers to a specific holding and includes location in the library and barcode

This page compares a MARC catalog record to a Bibframe one: http://id.loc.gov/tools/bibframe/compare-id/full-ttl.

Bibframe also aims to incorporate Universal Resource Identifiers (URI's) from outside sources into the catalog. Library metadata cataloging is often collaborative, and Bibframe will allow libraries to link to other resources. Many libraries utilize the OCLC consortium, which allows them to share bibliographic records. On the public side, a user can search OCLC WorldCat for which libraries around the world have a certain item.[16] OCLC has also been working on incorporating Bibframe linked data.

The Biodiversity Heritage Library (https://www.biodiversitylibrary.org/) provides an example of metadata implementation in an online catalog. The library is a collaborative international effort to make literature on biodiversity widely available, associated with the *Encyclopedia of Life*, hosted by the Smithsonian Institution's National Museum of Natural History. In addition to a robust catalog search, the online library offers curated thematic collections. The system is diverse and rich, with full metadata readily available in the catalog. The downloadable metadata allows other libraries to link to items in their catalog.[17]

Museums, such as those which are part of the Smithsonian Institution, track metadata associated with objects in their collections. The Smithsonian Libraries has a Metadata Department which provides descriptive information to promote "computer-to-computer discovery" and the "sharing, reusing, and repurposing of data that is used in the processing of digital projects, traditional library materials, and unique hidden collections of the

[13] For examples of RDA metadata records, see https://www.rdatoolkit.org/examples
[14] For more, see: http://www.loc.gov/bibframe/mtbf/, http://id.loc.gov/ontologies/bibframe-category.html, http://www.loc.gov/bibframe/docs/bibframe2-model.html, and https://www.loc.gov/bibframe/faqs/.
[15] Bibframe is expressed in RDF triples, meaning there are URI's associated with three-part statements of the type <subject> <predicate> <object>, e.g. <this record> <refers to an instance of> <a magazine>. For more on RDF, see the glossary in the appendices.
[16] For more information, see https://www.oclc.org/en/worldcat.html
[17] Sample record: https://www.biodiversitylibrary.org/bibliography/42740#/summary

Libraries." These differ from survey and statistical metadata but the comparison may be useful. For more, see https://library.si.edu/departments/metadata.

The Smithsonian American Art Museum (SAAM) offers a search/browse function for objects in its collections online at https://americanart.si.edu/art/browse.[18] For instance, the collection includes the series called the "Brown Sisters", a set of photographs taken of the same four sisters in the same order each year for over 40 years running. The museum collections use the categories of metadata described below. SAAM has about 55,000 unique museum object records in its collections management system, following a standard model.[19] Core fields include:

- A unique identifying number for the object, usually called the object's accession number for this museum
- Object title or name
- Artist or maker of the object
- Cultural affiliation of artist or maker
- Classification of object type (e.g. painting, photograph, sculpture, etc.)
- Geographic place names for site manufacture
- Object provenance
- Object dimensions
- Materials/mediums
- Condition (e.g. conservation treatments performed or needed)
- Location of the object (e.g. on view, in storage, or on loan to another venue)
- Photographs of the object
- Any restrictions on the object, based on copyrights or gift agreements

Some terms have definitions that are formalized and shared with other organizations to make interoperable measurement possible. Several "controlled vocabularies" from the Getty Research Institute are used, along with some local terminology: Art and Architecture Thesaurus (AAT), the Union List of Artist Names (ULAN), Thesaurus of Geographic Names (TGN), and the Cultural Objects Name Authority (CONA), Getty Research Institute. A further effort to create structured linked data on the provenance of museum items into structured linked data is discussed under "Art Tracks" at http://www.museumprovenance.org/ and http://www.museumprovenance.org/reference/standard/. These controlled vocabularies are used to make information available to other institutions by a Web API in a standard way that the receiving institution can interpret.

[18] We thank Sara Snyder of SAAM for this detailed information.
[19] They use software TMS for their museum object database. An overview of typical museum fields is shown at https://www.gallerysystems.com/best-practices-for-collections-documentation-and-object-cataloguing/.

Catalog systems for survey questions and data sets

This section lists metadata systems used for statistical applications.

- **Data.gov** is a web site that hosts a catalog of federal government data sets. Each data set is described by a set of elements specified in the Open Data Metadata Schema,[20] for example, the names and description of the datasets, contact information, and when they were last updated. Agencies are asked to post JSON-formatted files describing their data set inventory, conforming to the metadata schema. These files, usually named data.json, are read and offered to the public through Data.gov. The files form an "Enterprise Data Inventory" (EDI). The result is a metadata system providing a catalog of data sets. The catalog is managed through a CKAN implementation.[21] A framework exists for a data set to point to a data dictionary, but there is no standard way to represent one. Data.gov gets updated data sets as the agencies upload the associated data.json files. These files are harvested electronically by GSA's data.gov office and presented on data.gov. OMB memo M-13-13 defined and launched this cross-agency metadata effort with a mandate on many executive branch agencies. [22]

- NCHS runs Q-Bank, and this system provides access to question evaluation research. Reports in Q-Bank provide information on question design and performance. This information can be used to improve surveys and better understand survey data estimates. For more, see https://wwwn.cdc.gov/QBANK/home.aspx. Each record in Q-Bank consists of the following fields:
    - question text
    - response choices
    - survey(s) in which the question appears
    - cognitive research describing the effectiveness of the question for data quality and other issues.

Statistical data dictionaries and projects

Below we list metadata projects to illustrate their purposes and capabilities.

- Eurostat manages the ESQRS, a standard for the production and dissemination of quality reports within the European Statistical System (ESS). ESQRS files provide users with detailed information for assessing the quality of the data sets released by Eurostat. The broad concepts used are compatible with the SDMX cross-domain concepts and with the common terminology as published within the SDMX Glossary (2016). This metadata structure is also increasingly used for national statistical quality reports, and these reports promote FAIR principles.

[20] Open Data Metadata Schema v1.1: https://project-open-data.cio.gov/v1.1/schema/.
[21] CKAN - https://ckan.org/
[22] The memo is available here: https://project-open-data.cio.gov/policy-memo/

- ICPSR, the Inter-university Consortium for Political and Social Research, is an archive for social science data at the University of Michigan. ICPSR offers data sets with multiple levels of metadata. At the bottom level, one sees what the original data issuer offered. Then they add machine-readable codebooks in the DDI 2.x format. ICPSR has a partnership with the Census Bureau to organize metadata for confidential information (Gardner, 2018).

- Continuous Capture of Metadata (C^2 Metadata) – The C^2Metadata Project will create a continuous work-flow for metadata creation by automating the capture of data transformations performed by statistical analysis software. The ICPSR organization is conducting this project and it is intended to help social scientists in academia. It has NSF support.

- The Consumer Expenditure Survey program at BLS is working to document their annual public use microdata files which cover detailed categories of U.S. household expenditures gathered in two surveys: the quarterly Interview, and a biweekly Diary of expenditures. Variables in the microdata are linked to source questions in the surveys, giving provenance information on the data. Changes to questions, variables, and code list over the years are illustrated, easing year by year comparisons. Each data set is linked to descriptions of its elements, using the DDI 3.x Lifecycle standard. The ability to group variables by some common factor – concept, expenditure, source question(s), survey, year, and data set – greatly expands the information potential. For more, see Wilson (2018).

- ED Data Inventory. The goal of the ED Data Inventory is to describe all data reported to the U.S. Department of Education, with the exception of personnel and administrative data. It includes data collected as part of grant and program activities, along with statistical data collected to allow publication of valuable statistics about the state of education in this country. The ED Data Inventory is searchable and includes descriptive information about each data collection, along with information on the specific data elements in individual collections. Variable descriptions include name, label, extended definition, question wording, value labels, and file location. Some of the variables were standardized in NCES (2009). As with other federal data sets, this metadata is shared with data.gov in JSON files. For more information, see: https://datainventory.ed.gov/.

- NIEM – National Information Exchange Model is used at DHS, DOJ, and many other agencies. It is an XML-based information exchange framework developed through a collaborative partnership of agencies and organizations across all levels of government and private industry. NIEM is designed to share critical information among the emergency and disaster management, homeland security, intelligence, justice, and public safety areas. NIEM is essentially a taxonomy of terms with links to variables. Similar variables from different subject matter domains are each linked to the same term in NIEM. This provides a mapping between data across domains, greatly increases interoperability, and addresses concerns in the Evidence-Based Policymaking Act about harmonizing data from multiple sources. The taxonomy provides some meaning for

each variable and the ability to distinguish similar variables from each other. NIEM supports comparison, disambiguation, and harmonization across data sources. There is an effort to extend NIEM into the statistics domain, including federal statistical agencies.

- FGDC. Federal Geographic Data Committee Geospatial metadata describes maps, Geographic Information Systems (GIS) files, imagery, and other location-based data resources. The FGDC is tasked by Executive Order 12906 to enable access (see GeoPlatform.gov) to National Spatial Data Infrastructure (NSDI) resources and by OMB Circular A-16 and the A-16 Supplemental Guidance to support the creation, management, and maintenance of the metadata required to fuel data discovery and access. For details see https://www.fgdc.gov/metadata

- NCSES Metadata Explorer – The National Center for Science and Engineering Statistics (NCSES), within the National Science Foundation, is developing a Metadata Explorer to make published NCSES data and information searchable and discoverable in a new, easy-to-use interface. The smart search functionality will allow users to find or browse content using several terms for a given concept. The metadata schema will include variable descriptions and values, images of the relevant section of the questionnaire, links to pre-populated tables containing the variable, notes on historical changes to how the variable has been collected over time, and information on data quality.

- NCSES-funded CNSTAT Panel 'Transparency and Reproducibility of Federal Statistics' – In late 2018 the NSF's National Center for Science and Engineering Statistics tasked the Committee on National Statistics to undertake a consensus panel study to examine the degree of transparency and reproducibility of federal statistics. The principal questions are: what should an agency should do to make available, both internally and externally, archives of the input data sets used to generate sets of official statistics; documentation of the treatments done to the raw data prior to the computation of the final estimates (for treatment of failed edits, nonresponse, etc.); and documentation of what goes into the computation of the final published estimates. For meeting materials, upcoming reports, and other updates see https://www.nationalacademies.org/our-work/transparency-and-reproducibility-of-federal-statistics-for-the-national-center-for-science-and-engineering-statistics.

- The Office of Population Research at Princeton administers the Fragile Families and Child Wellbeing Studies which are supported by NIH and other funders. The office ran two rounds of a data science challenge contest with hundreds of participants. The metadata were thought to be difficult to use in the first round, and were reorganized for the second round. The resulting metadata were made available as a spreadsheet with rows characterizing variables using more standardized names, missing data codes, and data types. This reduced the numbers of requests and complaints from contest participants, and apparently made it easier to apply machine learning techniques in which many variables are tested for correlations and predictive effect without examining their semantics in detail to start with. (Kindel et al., 2018)

- The International Institute of Social History hosts large scale linked data projects and offers online classification descriptions for multiple data sets. Their web site has services to link categories and records and respond to queries with RDF triples. https://iisg.amsterdam/en.

- BLS's DataFinder[23] is a data dissemination tool available on the BLS web site designed to find and download time series data. The DataFinder can download data from more than one series at a time. The user interface employs a taxonomy of terms organizing all subjects covered by BLS time dependent, multi-dimensional, aggregated data, including time series.

- The Census Bureau built a system called GIDS (Generalized Instrument Design System) with a repository of questions and survey forms. It was used for the quinquennial Economic Censuses.[24] GIDS allows for the reuse of questions and formats to ease the development of their many forms. Census Bureau GIDS saved time and money and an increase in standardization through the use of GIDS over the distributed human-driven process used before.

- The Census Survey Information Tool (CSIT), released in June of 2020, is the Census Bureau Economic Directorate's modernized centralized repository of high-level survey and product metadata. This tool allows survey respondents and data users to quickly answer questions such as whether a survey is mandatory or voluntary, its frequency, and the website where the associated data products are released.

- The Census Bureau's data.census.gov site is a data dissemination tool.[25] Data from the Decennial Census, American Community Survey, and over 100 other surveys are available for search and download. In this new system, the data is disseminated through the Census Bureau's API on a single platform. It offers a table display which allows users to dynamically add filters and customize results. This metadata driven single-cell based approach allows for data visualizations, maps, and other ways of looking at data.

- Census xD is an emerging technologies group, and product studio, at the Census Bureau which partners with federal agencies and universities to improve the delivery of open government services to meet the Presidential Order on Artificial Intelligence (AI). The partners use research-driven, practical approaches to deliver experimental AI solutions. Examples of these groups' work include: building open-source tools to identify and mitigate statistical bias in machine learning implementations, supporting misinformation detection in the 2020 Census, and exploring the future of survey data collection through artificial intelligence.

---

[23] BLS DataFinder - https://beta.bls.gov/dataQuery/search

[24] For more information on GIDS see: https://www.fedscoop.com/census-bureau-wants-support-questionnaire-software-ahead-2020/, https://www.westat.com/project/authoring-questionnaires-us-census-bureau/generalized-instrument-design-system-gids.

[25] It replaces and expands on the now decommissioned American FactFinder.

- The DDI Codebook is applied in two networked systems, the International Household Survey Network (IHSN), managed by the World Bank and the NESSTAR network, developed by the Norwegian Social Science Data Service (NSD). IHSN is a network of metadata systems describing household surveys and censuses in developing countries around the world. The World Bank developed the systems to browse and populate the metadata deposited in the IHSN. NESSTAR26 is a tool to view DDI Codebook metadata describing data sets and research studies in a network of registered sites. Any organization using DDI can join the network. The tool has wide use.

- The IMF's Dissemination Standards Bulletin Board shows data on almost all countries. Governments submit data on their national indicators in one of several SDMX formats, which are called eGDDS, SDDS, and SDDS-Plus. The U.S. agencies BEA, Census, BLS, and the Treasury Dept. report such data to the IMF for selected time series. For details see https://www.treasury.gov/resource-center/Pages/imf.aspx. For an example of the metadata schema the US agencies follow, see https://dsbb.imf.org/sdds-plus/dqaf-base/country/USA/category/CPI00. For more see Moris (2017) and IMF 2013, https://www.imf.org/external/pubs/ft/sdds/guide/plus/2013/sddsplus13.pdf

- Wikidata is a database of facts that can be used by Wikipedias of any language, and by other Wikimedia Foundation projects. It incorporates choices of which fields ("properties") are appropriate for the records of various objects by relatively informal consensus processes. Wikidata items link to hundreds of external databases with authoritative descriptions of those objects. It is intended for large-scale use and holds 87 million records of items, most of which are formally linked to an article in a Wikipedia of some language. The software that hosts Wikidata, called Wikibase, can be installed by other institutions. OpenStreetMap uses Wikibase and the Smithsonian Institution is considering doing so.

There are other major metadata standardization projects. We list more such efforts in the Glossary. Statistical agencies need to monitor this growing area.

Classification management systems

Classification systems and code lists are integral to the production of statistics. They are needed for several purposes: as the allowed values for variables, response choices for questions, the categories used in sample stratification or post-stratification, the dimensions used for defining multi-dimensional data, and the stubs in tables (representing multi-dimensional data).

Code lists are normally unstructured lists of categories and associated codes, although they can also have levels and relationships as in classification systems. Examples include lists of products bought or sold, health conditions, etc.

[26] NESSTAR – http://nesstar.com

Classification systems are code lists with a formal structure. The structure, a hierarchy, is determined by a relation, which links each category to a more general one (its parent), with one category, the root, having no parent. The root defines the first level and the child categories of the root compose the second level. The children of the children make up the next level, and so on. Many classification systems have just one level, such as the States of the U.S. The categories in any given level are mutually exclusive (with respect to each other) and exhaustive of the universe the classification system represents, say industries as in NAICS.

The management of classification systems can be complex. Classification systems address and describe some subject field, such as industry (NAICS), occupation (SOC), or geography (e.g., State, County, MSA), and these subject fields change as society does. NAICS, for example, undergoes revisions to account for these changes every 5 years. The results of revisions are called *versions* and are often identified by the year of the revision. Each new version (the latest for NAICS was issued in 2017) is mapped to the penultimate one (for NAICS it is 2012) through an artifact known as a concordance. Concordances show how categories in the latest version are mapped to the previous one and vice-versa.

NAICS is one of many industry classification systems. The Standard Industrial Classification (SIC) system was originally developed in the 1930's to classify establishments by the type of activity in which they are primarily engaged. The *North American Free Trade Agreement* (NAFTA) of 1993 required this classification to become international. The North American Industry Classification System (NAICS) was established in 1997 to replace the SIC system. Separately the International Standard Industrial Classification (ISIC), published by the UN Statistics Division (UNSD), is the international reference classification of productive activities across countries. There are crosswalks between ISIC and NAICS.[27]

For each classification, furthermore, a statistical program using it might alter or add some categories to account for the specifics of the data the program produces. For instance, the Current Employment Statistics survey at BLS employs a category called government to classify some industry data. Government is not a NAICS category. This alteration of NAICS (2017) is known as a *variant*. The variant is tied to the version from which it derives. Mapping variants across consecutive versions is another usage of concordances. Other examples of classification systems are products, injury and illness, gender, race, Likert scale, and many others.

Many national and international statistical offices manage classification systems (versions, variants, and concordances) in information systems known as classification servers. Statistics New Zealand uses the Aria software developed by MTNA – see Appendix 2– to manage and maintain their classification systems.

The advantages to building and maintaining a classification server include:

[27] For ISIC and other international classifications maintained by the UNSD see https://unstats.un.org/unsd/classifications/

- Single source for all current classification systems used
- Access to historical record for how classification systems change over time
- Ability to map from one version of a classification system to another through concordances
- Comparability across variants (used in different programs) of the same version of a classification system
- Support adoption of standards within statistical agencies.

Recommendations

The material above suggests activities to enhance metadata management across the U.S. statistical agencies. We note the relationship of these recommendations to existing FAIR principles and the Federal Data Strategy (FDS) propositions. Agencies can meet these criteria and get a good return on investment.

1) Reuse terminology, classifications, and metadata schemes to save development and deployment time, and to increase comparability and interoperability of data and metadata. Classification management systems help agencies with their internal operations and inform public data users. (The FDS principles include harnessing existing data, and anticipating future use. The R in FAIR stands for reuse.)

2) Consider adopting existing standards appropriate to the subject matter, such as the DDI, SDMX, NIEM, GSIM, GSBPM, and the NGDA standards. They provide a shareable framework for describing the work of statistical agencies, and how those descriptions relate to each other. Statistical agencies share the same statistical data life cycle in general, and they benefit from creating metadata that interoperates. (The FDS refers to the practice of leveraging such standards extensively.)

3) Invest in learning metadata tools. Train staff in areas where the Federal statistical agencies have knowledge and skill gaps. We need better literacy in the relevant areas, such as those mentioned in this paper's glossary, to improve our services and efficiency. Some agencies offer metadata training now.[28] (The FDS Principles refer to a learning culture and opportunities.)

4) Identify opportunities and problems stakeholders have that metadata systems will address. This helps obtain the buy-in required to get a system built. For example, agencies can build systems that help users find the data sets they want. Metadata systems can make it easy to know which variables are in a data set. (The FDS and the Evidence Act ask agencies to identify their data needs, notably to evaluate government policies.)

[28] See for example https://www.ncddc.noaa.gov/metadata-standards/metadata-training.html and the Smithsonian Metadata Department's training.

5) In planning a new metadata system, think big, but build small. Envision how the system could scale up usefully. It may be best to build the first system using established, well-understood technology, rather than risk a failure because the implementation technique is new. Iteratively, it is possible to build a system with functionality and design that are new to the agency. The system can be ported to new technology once the system has proved to be useful. This makes it possible to keep costs manageable, expectations reasonable, and success within reach. (The FDS refers to conscious design, anticipating future uses, and continuous refinement.)

6) Plan, consult, and share with other statistical agencies. When agencies build similar or shared systems, interoperability comes early, inexpensively, and has more supporters. Within government, collaborate between metadata experts, subject matter experts, statistical officials, and Chief Data Officers to evaluate opportunities for metadata systems and metadata management. There is little advantage to building a system that is a repeat of what others have already done. Reusing is simpler, if that is possible. (The FDS refers in several places to making agreements and leveraging partnerships.)

7) Engage with professional experts outside government, and across specialties within government. There are professional conferences and publications associated with several institutions and data repositories discussed in the text and the glossary. Statistical agency managers can also benefit from awareness of metadata projects for libraries, museums, and geographic data.

8) Use the tools and guidance in Data.Gov, Resources.data.gov, and the FDS Action Plan to build and maintain dataset catalogs and dissemination tools. The *federated catalog* includes the data sets from all the principal Federal statistical agencies. In the long term, agencies should enhance data.gov's Open Metadata Schema with selected new fields specific for statistical data, such as whether a series is seasonally-adjusted. The schema is available at https://project-open-data.cio.gov/v1.1/schema/ and at https://resources.data.gov/schemas/dcat-us/v1.1/metadata-resources/. The FDS action plan is at https://strategy.data.gov/action-plan/.

9) Create, or encourage the creation of, standardized data dictionaries and classification schemes for use in data.gov catalog entries. Avoid using PDFs for this because they are not machine-readable. Instead use an established scheme, such as DDI, ISO 19115 (Geographic Information Systems - Metadata), or other metadata standards. Advocate for data.gov's entries about data sets to use such standardized data dictionaries with information about the classifications each data set uses, to help both users and machines mix and match data. The Data.gov catalog would be more useful if a data dictionary were defined along with each data set, so that it becomes straightforward for users and machines to figure out whether two data sets use the same classification system, e.g. for industries. (The FDS promotes wide access and machine-readability, and the I in FAIR stands for interoperability.)

10) Reuse: Create or collect once, and use many times. This is a common adage of metadata management. Metadata are more useful when one description can be used many times. This means the items are related by this shared description and can be linked. An example is the universe associated with a variable. The same universe might apply to many variables. (The FDS refers to conscious design, planning for reuse, and leveraging data standards. The R in FAIR stands for reuse.)

11) Partner with external services that add value to Federal data, or implement their capabilities in the agencies. For example, Google, Statistics USA, and IPUMS are services that offer Federal data to users. These services use metadata from statistical agencies to find and understand data. Machine learning capabilities can be offered by agencies or external partners. (The FDS principles encourage public monitoring of Federal data practices and promoting wide access to open machine readable data and transparency. The F and A in FAIR stand for findable and accessible.)

Conclusion

A goal of this paper was to bring to light the needs and requirements of metadata management, how metadata management helps agencies fulfill their missions, and to show that it is possible to achieve the goals.

Information presented in this paper may help U.S. federal statistical agencies to design and manage metadata systems that promote mission-related goals, increase data quality and transparency, and promote comparability and interoperability. But metadata management projects can also be costly and time consuming. The decentralized nature of the U.S. federal statistical system presents challenges for data users interested in a particular topic but faced with different sources. Ultimately, users don't care much which agency has what data. This is where common metadata tools and standards may help building interfaces across data sources. This harmonized approach may be enhanced by robust metadata systems and metadata management tools.

Progress may be facilitated by cross-agency collaboration, as noted in some of the recommendations, consistent with metadata-related requirements of recent statutes and OMB memos. When agencies partner effectively they increase their return on investment. For example, reusing metadata tools reduces the costs of achieving data quality and interoperability goals.

References

Commission on Evidence-based Policymaking. Final Report: The Promise of Evidence-Based Policymaking. 2017. https://www.cep.gov/cep-final-report.html

Couper, Mick P. 2017. Birth and Diffusion of the Concept of Paradata (in Japanese – translated by W. Matsumoto). Advances in Social Research, 18, 14-26.

EEDI prototype at eedi.referata.com. (Experimentally combines information from the data.gov process with Information Collection Requests, and SORNs; include information on which official category systems a data set uses.)

FCSM and WSS Metadata Workshop. September 2018. Agenda: https://nces.ed.gov/FCSM/2018_FCSM_Metadata_Workshop.asp. Video: https://www.youtube.com/playlist?list=PLqsWXJV2UtreXRxtuPSKD9RbHjZ419Gcq

Foster, Ian; Rayid Ghani; Ron Jarmin; Frauke Kreuter; Julia Lane. 2016. *Big Data and Social Science*. Chapman and Hall/CRC Press.

GAO. 1999. Standards for Internal Control in the Federal Government, GAO/AIMD-00-21.3.1. Washington, D.C.: November 1999.

GAO. 2013. Status of the Department of Education's Inventory of Its Data Collections. GAO-13-596R. https://www.gao.gov/products/GAO-13-596R or https://www.gao.gov/assets/660/655668.pdf

Gillman, Dan. July 2017. Presentation 'Metadata – The Basics'. https://nces.ed.gov/FCSM/pdf/Gillman.pdf

Gillman, Dan. 2018. What You Need to Know Too: Standards and Interoperability. WSS & FCSM Metadata Workshop. http://washstat.org/presentations/20180914/Gillman.pdf

Kindel, Alexander T.; Vineet Bansal; Kristin D. Catena; Thomas H. Hartshorne; Kate Jaeger; Dawn Koffman; Sara McLanahan; Maya Phillips; Shiva Rouhani; Ryan Vinh; Matthew J. Salganik. 2018. Improving metadata infrastructure for complex surveys: Insights from the Fragile Families Challenge. https://osf.io/preprints/socarxiv/u8spj

Kreuter, Frauke. 2013. Improving Surveys with Paradata: Analytic Uses of Process Information. John Wiley & Sons.

Kreuter Frauke. 2018a. Getting the Most Out of Paradata. In: Vannette D., Krosnick J. (eds) *The Palgrave Handbook of Survey Research*. Palgrave Macmillan, Cham.

Kreuter, Frauke. 2018b. Metadata. WSS & FCSM Metadata Workshop. http://washstat.org/presentations/20180914/Kreuter.pdf

Lane, Julia. 2018. Administrative Data Research Facility and Metadata. WSS & FCSM Metadata Workshop. http://washstat.org/presentations/20180914/Lane.pdf

Moris, Francisco. 2017. Statistical Metadata Quality: toward inter-operable and machine readable metadata. OMB/OIRA/SSP internal memo, April 14, 2017

Moris, Francisco. 2019. Thoughts on Inter-operable Metadata for Data Quality, Transparency, and Reproducibility. Presentation prepared for NASEM/CNSTAT Panel on Transparency and Reproducibility, The National Academies, Washington DC, May 21, 2019.

Moulton, Mary; Leighton Christiansen; Xin Wang. 2018. Freight Data Dictionary. WSS & FCSM Metadata Workshop. http://washstat.org/presentations/20180914/Moulton.pdf

National Academies of Sciences, Engineering, and Medicine (NASEM). 2017a. Federal Statistics, Multiple Data Sources, and Privacy Protection: Next Steps. Washington, DC: The National Academies Press. doi: https://doi.org/10.17226/24893.

National Academies of Sciences, Engineering, and Medicine (NASEM). 2017b. Innovations in Federal Statistics: Combining Data Sources While Protecting Privacy. Washington, DC: The National Academies Press. doi: 10.17226/24652.

National Forum on Education Statistics. 2009. *Forum Guide to Metadata: The Meaning Behind Education Data* (NFES2009−805). U.S. Department of Education. Washington, DC: National Center for Education Statistics. https://nces.ed.gov/forum/pub_2009805.asp

Pon, R. K.; Buttler, D.J. 2009. Metadata Registry, ISO/IEC 11179. In *Encyclopedia of Database Systems,* L. Liu, M. T. Özsu, eds. Springer, Boston, MA. https://digital.library.unt.edu/ark:/67531/metadc926399/m2/1/high_res_d/973862.pdf

Schymik, Gregory; Karen Corral; David Schuff; Robert St. Louis. The benefits and costs of using metadata to improve enterprise document search. *Decision Sciences*, 46:6 (December 2015) 1049-1075.

Seastrom, Marilyn. 2018. Thoughts on the Importance of Metadata for Integrated Data. WSS & FCSM Metadata Workshop. http://washstat.org/presentations/20180914/Seastrom.pdf

Teeter, Jared. Censusofsurvey.com. This site has information on Federal surveys which was web-scraped and recombined from Information Collection Requests submitted to OMB/OIRA.

Vale, Steven and colleagues at UNECE. Part A - Statistical Metadata in a Corporate Context. https://statswiki.unece.org/display/VSH/Contents. Last updated 2012.

Von Hippel, Eric. 2005. *The Democratization of Innovation*. MIT Press.

Wilson, Taylor J. 2018. Implementing the Data Documentation Initiative (DDI) for the Consumer Expenditure Surveys. http://washstat.org/presentations/20180914/Wilson.pdf or https://nces.ed.gov/FCSM/pdf/Wilson.pdf

Appendix 1: Glossary

Administrative metadata: The role of metadata when used to manage data sets, such as file locations, file permissions, dates of creation and modification, and whether files are public or exposable by FOIA requests.

API: An **Application Programming Interface** is a set of functions allowing a software programmer to write software code that interacts with some other system, often a database. These days, APIs often rely on the Web, and through URLs allow a user to access data across the Internet. A developer portal to access services and documentation for the Census Bureau's APIs is at http://www.census.gov/developers/.

Business metadata: The role of metadata when used to describe statistical activities, such as post-collection processing, the management of frames, the designs of samples and questionnaires, outreach activities, etc.

Common Education Data Standards (CEDS) is a data management initiative of the National Center for Education Statistics (NCES) which includes a common vocabulary, data models that reflect that vocabulary, and associated tools and stakeholders. CEDS includes elements from the NCES Handbooks which had been developed earlier by NCES to standardize data definitions so that education data could be more accurately aggregated and analyzed. Information about CEDS is available at https://ceds.ed.gov/Default.aspx.

Census Geocoding Services, and Census TIGERweb GeoServices: A Web API for the Census Bureau's TIGER geographic information system. The input is a structured address or latitude/longitude location, and the system sends a response which can include the lat/long and/or census geographies. For more, see http://www.census.gov/data/developers/data-sets/TIGERweb-map-service.html and http://www.census.gov/data/developers/data-sets/Geocoding-services.html.

Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA): Each statistical agency or unit shall (1) produce and disseminate relevant and timely statistical information; (2) conduct credible, accurate, and objective statistical activities; and (3) protect the trust of information providers by ensuring the confidentiality and exclusive statistical use of their responses. An agency shall make data assets available to any statistical agency or unit for purposes of developing evidence. This shall not apply to any data asset that is subject to a statute that prohibits the sharing or intended use of such asset. Each statistical agency or unit shall expand access to data assets acquired or accessed to develop evidence while protecting such assets from inappropriate access and use.

CKAN: **Comprehensive Knowledge Archive Network**, a software standard for implementing a catalog of some resources on the Web. Data.gov runs it as does the UK equivalent to offer a catalog of data sets. Keywords can be listed with each catalog entry. CKAN may not yet have features to offer detailed metadata on variables in each dataset of the kind needed by statistical agencies to enable users to match datasets by the variables which match because they use common classifications, e.g. industry, or other metadata.

Coleridge Initiative: See Lane (2018); builds from the Jupyter software.

Common Metadata Framework: A four part set of recommendations for metadata management in statistical offices developed by the UNECE around 2004. The four parts are:

1. Statistical metadata in a corporate context
2. Metadata concepts, standards, models, and registries
3. Metadata and the statistical business process
4. Implementation

This framework has been superseded by more recent work. However, Part 1 is still relevant, and Part 3 evolved into the GSBPM. Part 2 is a good historical reference, and implementations (formerly listed in Part 4) change all the time.

Conceptual metadata (or semantic metadata): The role of metadata when used to define concepts and categories; describe variables, such as variable names, datatypes, display attributes, keywords, allowed values, etc.; and classification systems and code lists. For example, the code/category pairs <m, male> and <f, female> are the allowed values for a variable on gender. Another example is the NAICS sector categories (agriculture, mining, utilities, construction, manufacturing, … , and public administration) and their definitions.

CSDA: **Common Statistical Data Architecture**, is under development by the UNECE. It will contain recommendations for how data are managed, transferred, and disseminated in a statistical office.

CSPA – Common Statistical Production Architecture, developed and managed under the UNECE, describes a series of recommendations for modernizing the IT infrastructure designed to process statistical data. It does not specify requirements. Instead it takes as a given the idea that each statistical program performs functions that are very similar to those in other programs. Then, it lays out 3 important ideas:

1. These processes can be subdivided into reusable components
2. These components should be used as the building blocks for implementing the processing for any statistical program
3. The metadata held in a companion GSIM metadata store are then used as the parameters to drive each process specifically for each program that uses them

In this way, the component software becomes standard for each office. CSPA incorporates the idea that the component software can be shared across all offices, but implementation of that is many years away.

Cube or n-cube: A conceptual model for multi-dimensional data. The number of cells in the n-cube is the product of the number of categories in each of the dimensions. The value in each cell is determined by some measure restricted to the categories in each dimension (one category from each) defining each cell. N-cubes are often represented as tables. See W3C Data Cube at https://www.w3.org/TR/vocab-data-cube/. Statistical data are unlike library or museum data in that (1) a statistic like "Unemployment rate among Hispanics in June 2019" is a statement about a class of people; (2) the class adds up to larger totals (it's "dimensional"). The multiple dimensions can be thought abstractly of making up a many-dimensional space.

da|ra: The Registration Agency for Social and Economic Data in Germany. This is a joint project under GESIS (the German Leibniz Institute for Social Sciences) and ZBW (the

German Leibniz Information Center for Economics). The main purpose of the project is to provide the infrastructure for persistent identification, secure management, and reliable citations of research data in the social sciences.

DATA Act: The *Digital Accountability and Transparency Act of 2014* or the *DATA Act*. The act requires open reporting of Loan, Grant, and Contract metrics and variables as defined by OMB and the Treasury Department using XBRL metadata standard and taxonomies, eventually reporting electronically through agency APIs (https://www.congress.gov/113/plaws/publ101/PLAW-113publ101.pdf) .

DataFinder: The system under development at BLS to provide a single point of access to all time dependent, multi-dimensional, aggregated data, including time series. Currently, there is a separate tool at BLS for each subject matter area for this kind of data. It is possible to download only one data series at a time. DataFinder is being designed to overcome these limitations by using a taxonomy of terms for BLS data. It will be possible to download more than one data series. The current system is available through the BLS web site on the Beta page.

Data.gov: A web site offering a large catalog of federal government datasets. GSA runs it. The information is based on the Data.json files – the EDI (Enterprise Data Inventory) published by federal government agencies. The catalog is built using CKAN software. Its Open Metadata Schema has some fields that might be applicable to statistical data, available at https://project-open-data.cio.gov/v1.1/schema/.

Dataverse: Harvard Dataverse is a repository for research data. Data and code can be deposited. The result is a shared system for managing and searching for data sets. There is also a project among several universities to develop open source software to support the Dataverse idea.

DDI: **Data Documentation Initiative**, is a family of statistical metadata standards and other semantic products. The work is organized under a consortium called the DDI Alliance. The secretariat for the DDI Alliance is at ICPSR. The main products under DDI are as follows:

- **DDI2: Codebook**, version 2.5, is a statistical metadata standard written in XML for describing a socio-economic research study, a one-time statistical survey, or the data each might produce in a stand-alone way. DDI2 is in wide use throughout the world. It does not support reuse of metadata, i.e., sharing metadata, across XML codebook instances. However, this feature also makes it easy to implement and is often used as a first development step towards more complex metadata management systems in the statistical domain. See http://www.ddialliance.org/Specification/DDI-Codebook/2.5/
- **DDI3: Lifecycle**, version 3.3 is a statistical metadata standard written in XML for describing the full production cycle for statistical activities, be they censuses, surveys, or some others. Lifecycle supports extensive reuse of metadata and is designed to support a full metadata repository for a statistical office. Many surveys and their data can be described together. Increasingly, national statistical offices around the world are turning to DDI3 for their metadata needs. See http://www.ddialliance.org/Specification/DDI-Lifecycle/3.3/

- **DISCO**, for data discovery, is written in RDF-S, enables discovery of research and survey data and data sets on the Web. It is based on DDI XML formats. http://rdf-vocabulary.ddialliance.org/discovery.html
- **XKOS, eXtended Knowledge Organization System**, an extension of W3C SKOS, written in RDF-S, is designed to describe statistical classifications. XKOS adds elements to describe levels, the ability to attach concepts to levels, and semantics of extended relationship types. Version 1.2 is the latest as of this writing. http://rdf-vocabulary.ddialliance.org/xkos.html
- **DDI4: Cross-Domain Integration** is an emerging metadata standard that includes the ability to describe data from any source (not just statistical data) in a wide variety of formats. Sensor, administrative, streaming, and any multi-dimensional data can be described, as well as traditional statistical microdata. A "datum-centered" approach is employed so any datum can be tracked throughout its lifecycle and for any data structures. A generalized process model is included to provide information on the provenance of data. DDI4 supports the needs of using and creating data from multiple sources. Finally, the entire specification is managed through a UML (Unified Modeling Language) model for easy maintenance, changes, and the generation of several language representations, e.g., XML, RDF, SQL, and others. A first draft release was issued in April 2020. https://ddi-alliance.atlassian.net/wiki/spaces/DDI4/pages/491703/Moving+Forward+Project+DDI4

Dimensional data: Data based on a combination of categories each taken from one of several sets (the dimensions). Each category combination defines a cell, which holds the output of some quantitative variable restricted to the category combination defining the cell. In mathematical terms, each cell is defined by an element in the cross-product of dimensions. The categories defining each cell serve to specialize the universe of objects the data measure. The union of the all the cell universes is the whole universe of objects.

Dublin Core schema: The Dublin Core Schema is a limited set of bibliographic attributes used to describe digital resources in a library such as books, artworks, images, video, audio, and web pages. Library of Congress is drawing from the Dublin Core and from MARC in developing RDA. Many metadata standards and systems are based on the Dublin Core.

Element registry: An element registry is associated with a collection of data sets, and describes standardized data elements (variables) in them. For a survey, the other elements might include the phrasing of the question asked, or skip patterns among the questions. A registry is a catalog with an explicit process for maintaining it.

EU's General Data Protection Regulation (GDPR) of 2018: This is perhaps one of the strictest privacy and security laws for personal data. It imposes obligations onto organizations or institutions, even if they are not in the EU, if they target, collect, or process data related to EU citizens or residents, or offer goods or services to EU citizens. The GDPR authorizes fines against those who violate its privacy and security standards.

Exchange metadata: The role of metadata when used to describe data exchanges within and among organizations. Exchange agreements, standards, partners, inputs, and formats are all involved.

Executive Order 13859 on Artificial Intelligence (February 11, 2019): This order launches and supports a public/private and interagency effort called the American AI Initiative. It includes the directive to use the other Executive Orders, OMB Memos, and Laws. The ideas of machine readable define once – use many metadata, variables, and data sets; and locations where to find government data to use in AI (M13-13, Data.gov, and FEBPM).

FAIR: This is a generic set of principles for direction on how to make data *Findable*, *Accessible, Interoperable*, and *Reusable*, described further in the text. See https://www.go-fair.org/fair-principles/ for more details.

Federal Data Strategy: The Federal Data Strategy is a policy effort to fully leverage the value of federal data for mission, service, and the public good by guiding the Federal Government in practicing ethical governance, conscious design, and a learning culture. For more, see https://strategy.data.gov/

FGDC: The Federal Geographic Data Committee, is a high-level U. S. Federal interagency group that sets standards for geographic data. It is, headquartered at the Interior Department. It started around 1990 as an OMB-authorized interagency committee, and the GDA establishes its existence in law. The FGDC is responsible for the NSDI standards and the GeoPlatform. Source: https://www.fgdc.gov/gda. (OpenGIS, and ISO/TC211 have related scopes. See ISO 19115 – GIS Metadata.)

FIBO – Financial Industry Business Ontology, is a specification written in OWL under the auspices of the Object Management Group (OMG), a standards consortium. FIBO is being built to describe financial and economic indicator data. Special emphases are time series, multi-dimensional, and statistical indicator data.

FIPS – Federal Information Processing Standards, include standard code sets, for example for geographic areas. Use of a FIPS standard enables matching of data between data sets by simply comparing codes. NIST sets these standards.

Foundations for Evidence-Based Policymaking Act of 2017 (FEBPM): This law has 3 titles: I) The Evidence-Building activities, which includes developing evidence, data, and methods; each Agency appoints a Chief Evaluation Officer; strategic plans address this law; and an annual report to congress. II) The OPEN Government activities, which includes publishing a catalog and it's data assets as machine-readable; each Agency appoints a Chief Data Officer, who maintains a comprehensive data inventory; a Chief Data Officer Council at OMB; Reports to Congress. III) Codifies and revises the Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA). https://www.congress.gov/bill/115th-congress/house-bill/4174

GAMSO – Generic Activity Model for Statistical Organizations, developed and managed under the UNECE, is a generalization of GSBPM and addresses all management activities a statistical office might need. GAMSO is divided into 4 main sections: strategy and leadership, corporate support, capability development, and production. The production section incorporates GSBPM. In GAMSO, a capability is defined as a potential, the means to

be able to accomplish some task. Development of these means is necessary to get the work of the office done. Strategy and leadership addresses defining a vision, governance, and strategic partnerships. Corporate support focuses on all the main functions of the office: performance, methodology, quality, information, consumers, suppliers, finance, HR, IT, and physical needs. Similarly to GSBPM, GAMSO sets forth an outline of activities that are needed for the smooth operation of a statistical office.

GDPR: The European Union's recent **General Data Protection Regulation** governing personal data privacy. U.S. agencies sometimes honor or follow it in practice, partly to maintain compatibility with European partners.

Geospatial Data Act: The Geospatial Data Act of 2018 (GDA) establishes the FGDC in law, and the idea of the NSDI and the technical advisory committee more active NGAC, and requires certain agencies to report on geospatial data and its costs to Congress periodically. For more, see Folger, 2018, https://www.fgdc.gov/gda, and https://www.fgdc.gov/gda/geospatial-data-act-of-2018-summary.pdf.

GSBPM: **Generic Statistical Business Process Model**, developed and managed under the UNECE, identifies over 40 processes, organized into 8 phases of the statistical production lifecycle. The phases cover all the major milestones in the statistical lifecycle, from planning and design to dissemination and evaluation. In addition, there are two over-arching processes, quality and metadata, which are relevant to every process in a statistical office. Some of the processes have sub-processes identified for them as well, and the order in which the processes are laid out is not normative. Many offices, including the Bureau of Labor Statistics and the Census Bureau have adapted GSBPM into a local version (or profile) for their own purposes and in their own languages.

GSIM: Generic Statistical Information Model, developed and managed under the UNECE, is meant to be a companion standard to GSBPM. The inputs and outputs of processes defined in GSBPM are describable in GSIM. In this sense, they are dual standards. GSIM is divided into 4 major sections, each dealing with a set of information objects, and with all the sections inter-linked. The 4 sections are Concepts, Business, Structures, and Exchange. The Concepts section addresses the basic concepts behind a survey and its data. Variables, code sets, classifications, and related materials are described. The Business section describes each collection activity. The Exchange section describes how data are moved during the statistical lifecycle, and the Structures section describes how data are organized. In the GSIM framework these are mutually exclusive categories; each attribute is in only one of them, though they have relations to one another. GSIM contains over 100 information objects. These allow for several layers of detail, however the model itself is conceptual in nature. It is not designed to be implemented directly. The DDI and SDMX standards generally fit the GSIM model.

ICSP: Interagency Council on Statistical Policy, the OMB-managed interagency group that authorizes FCSM and other interagency working groups. The Chief Statistician and ICSP coordinate with the Chief Data Officer Council, and are responsible for the annual *Federal Statistical Programs* Report to Congress.

ICPSR: **Inter-university Consortium for Political and Social Research** is an international consortium of academic institutions and research organizations. It maintains a large

archive of data collections and also disseminates data gathered by a variety of Federal statistical agencies. ICPSR sponsors research on data curation, data science, and data stewardship including metadata quality. ICPSR also hosts the DDI Alliance. ICPSR is a unit of the Institute for Social Research at the University of Michigan.

INEXDA's Metadata Database: **INEXDA (**International Network for Exchanging Experience on Statistical Handling of Granular Data**)** is a partnership among European central banks, but other statistical organizations can join. INEXDA encourages data harmonization, which will improve the efficiency of working with granular data. The aim of the network is to help with the use of granular data for analytical, policy, and research purposes. Confidentiality will be maintained. INEXDA is associated with Stefan Bender; cited by Kreuter (2018b). https://www.bundesbank.de/en/bundesbank/research/rdsc/inexda/inexda-international-network-for-exchanging-experience-on-statistical-handling-of-granular-data-617914

ISO/IEC 11179: *Information technology — Metadata registries (MDR)*. International standard for registration of metadata to make data understandable and shareable. This is a multi-part standard addressing the semantics and representation of data and the procedures for maintaining a registry of this information. The current parts (in order) are: Framework; Classification; Meta-model; Formulation of data definitions; Naming and identification; and Registration. See Pon and Buttler (2009), pp. 1724–1727 and https://www.iso.org/obp/ui/#iso:std:iso-iec:11179:-1:ed-3:v1:en for a description of the standard and each of its parts. The latest edition of each part is freely available on the web at https://standards.iso.org/ittf/PubliclyAvailableStandards/

ISO 3166: Country codes and their subdivisions, a widely used standard maintained by the UN Statistics Division. This standard contains code lists for countries of the world and subdivisions of them, such as the US states and counties. Three code lists for countries are included.

ISO 19115: A multi-part standard under ISO for metadata for geographic data and geo-information systems.

JSON: A file format for data sets, used online and to send data to a Web browser. Another such format is XML. JSON digital objects are framed by {} and include []. Here's a simple JSON object: `{ "firstName": "John", "lastName": "Smith" }`

JSON-stat: A JSON file format designed for aggregate statistical data, used in the EU. JSON-stat is intended for aggregate statistics organized along several dimensions, sometimes called a cube model in which observations can be cut, split, or analyzed along several dimensions. The format is not well suited to microdata nor to multi-dimensional aggregates, and not for data with more than 10,000 elements. In general if one would never print an entire table, it is not a table suitable for JSON-stat. JSON-stat is used now by a number of international statistical agencies including the Central Statistics Office of Ireland, Statistics Norway, Statistics Sweden, and the UK Office of National Statistics. The JSON-stat format is text-readable JSON, with certain required elements and standard customary elements. It adds a few required elements to the JSON format. Any program that reads JSON can read a JSON-stat file. It is intended for aggregate statistics that are organized along several dimensions. The format has limitations that are described on the JSON-stat web site, but they are changing as new versions are released.

MARC: A standard or set of standards for bibliographic information, historically developed and used by the Library of Congress and other libraries. The concepts of the Dublin Core and RDA (Resource Description and Access) and METS build upon MARC standards.[29] METS is for items in a digital library. MARC and Dublin Core are widely used. OCLC is the organization behind the Dublin Core.

Metadata registry and repository: A repository is a database for managing metadata. A registry implements a process for keeping records about some specific kind(s) of objects (e.g., cars, births, or marriages). Often, the records managed by a registry are kept in a repository.

M-13-13: Refers to OMB memorandum backing data.gov, titled "Open Data Policy—Managing Information as an Asset." https://project-open-data.cio.gov/policy-memo/ It mandates that agencies share descriptions of datasets with the data.gov people at GSA. It includes metadata requirements. It is now incorporated into the FEBPM Law.

NIEM: The National Information Exchange Model (NIEM) is a reference model and a common vocabulary designed for the exchange of justice, security, and defense data. Plans are underway to expand NIEM for statistical data. NIEM is a taxonomy of terms with links to variables. The taxonomy provides some meaning for each variable and the ability to distinguish similar variables from each other by supporting comparison, disambiguation, interoperability, and harmonization across data sources. NIEM supports cross-domain data comparisons, exchange, and shared semantics.

NISO: The National Information Standards Organization, is an ANSI accredited standards development organization. Its standards address content creation, curation, discovery, interchange, analytics, and business processes that facilitate content exchange for and about information providers such as publishers, libraries, and information aggregators. NISO publishes the Dublin Core standard among others, and support the development of library metadata standards. For more, see https://niso.org

OPEN: The *Open*, *Public, Electronic and Necessary* (*OPEN*) *Government Data Act* (S. 760, H.R. 1770) would require that (1) agencies designate a Chief Evaluation Officer to coordinate evidence-building activities and a Chief Data Officer in charge of related statistical policy, techniques, and procedures, and an official inventory of data "assets"; (2) OMB establish a Chief Data Officer Council and an Advisory Committee on Data for Evidence Building to advise on expanding access to and use of federal data for evidence building, and OMB should promote data sharing agreements among agencies; (3) open government data "assets" be published as machine-readable data. It is now incorporated into the FEBPM Law. https://www.datacoalition.org/open-government-data-act/

OWL: **Web Ontology Language** – similar to RDF-S, except it supports the far more powerful first order logic, the set of logical rules on which mathematics is based. Because the basis for any model built through OWL is logic, consistency checks and inferencing are supported.

[29] For more on RDA and other bibliographic metadata standards, see https://www.oclc.org/en/rda/about.html and https://en.wikipedia.org/wiki/Metadata_Encoding_and_Transmission_Standard).

Paradata: Data about the data collection process on a micro level, usually focused on a survey, e.g. description of the survey phone call or web site; time of day the respondent replied; uncertainty associated with information about the respondent. Paradata is in practice used to adapt during survey periods to attributes of the early respondents, e.g. to improve coverage of groups which are under-covered in the early sample. (Kreuter 2018b)

Q-Bank: A registry managed by NCHS that contains survey question subjects, wording, and the research and applications that support the usage of the question. See https://wwwn.cdc.gov/qbank/home.aspx.

RDF: **Resource Description Framework** – a modelling language for metadata on the Web. RDF is expressed as a series of inter-linked sentences or triples that consist of a subject, predicate, and object. The subject, predicate, and object are each identified through the use of a URI (uniform resource identifier, of which the more common URL is a kind). The simple sentence structure can be combined, e.g., the object of one sentence can be the subject of another: "Rock beats scissors. Scissors beats paper. Paper beats rock." These combined sentences form a graph with subjects and objects as nodes and predicates as arcs. The graph can be queried and simple inferences made based on the semantics. See https://www.w3.org/RDF/. See also RDF-S and OWL. Some of the same goals can be achieved by using other controlled vocabularies, without the RDF standard specifically.

RDF-S: **RDF Schema** – a language for specifying a reusable overall structure of RDF graphs. A schema is used to provide semantics to an RDF graph, increasing the power of using inference when performing queries. Schemas add rules to the way subjects, predicates, and objects are specified and related.

SCOPE/Metadata: A team launched by ICSP/SCOPE some years ago to address metadata issues across agencies in a consistent way. Team members represent statistical agencies. Paul Bugg, then of OMB/OIRA, coordinated the launch of the team. This team authored this report.

School Courses for the Exchange of Data (SCED) is a classification system for prior-to-secondary and secondary school courses so there are common and comparable course codes across schools, districts, postsecondary institutions, and states. The National Forum on Education Statistics maintains SCED. SCED facilitates research on transcripts and other topics. Five-digit code classify course content, and other descriptive information is in elements and attributes. The Forum's SCED working group includes federal, state, and local education agency representatives who receive suggestions and assistance from subject matter experts with different needs. The current SCED File and resources to assist data users is available at https://nces.ed.gov/forum/sced.asp.

Schema.org is an industry-supported collaboration which publishes a library of reusable schemas and vocabularies for data on each of hundreds of topics such as invoices. Systems using the same schema are interoperable. Some of these data layouts are integrated into search engines.

SDTL: The Structured Data Transform Language (SDTL) is a model for describing data transformations. It is being developed as part of the C2Metadata project. (http://c2metadata.gitlab.io/sdtl-docs/)

SDMX: A set of technical standards and tools for transmitting statistical data and metadata together often in an XML format. SDMX can express a cube of data or table of dimensional data in time series of the kind issued by official statistical agencies. SDMX stands for Statistical Data and Metadata eXchange. It was developed by seven international banks and statistical offices: World Bank, IMF, OECD, Eurostat, European Central Bank, Bank of International Settlements, and the UN Statistics Division. These organizations had worked in the UN/CEFACT context together earlier. Eurostat has promoted SDMX as a way to describe microdata as well, and the development of the Validation and Transformation Language was an attempt to incorporate all statistical processing into SDMX. SDMX and DDI overlap but address different areas; for example, DDI is more likely to be used for microdata whereas SDMX has been used extensively for aggregate or macroeconomic time series. SDMX was standardized under ISO as ISO 17369. For more about SDMX, see http://www.sdmx.org, https://en.wikipedia.org/wiki/SDMX, and https://www.iso.org/standard/52500.html.

Semantic metadata: Same as conceptual metadata.

Structural metadata: The role of metadata when used to describe file formats (physical data formats), structures for organizing metadata (logical data formats), and the organization of processing steps in the statistical lifecycle.

UNECE: **United Nations Economic Commission for Europe** includes Canada, the United States, and all European countries as members.

VTL: **Visualization and Transformation Language**, part of SDMX.

W3C: **World Wide Web Consortium**, a standards consortium for specifying how the Web - works, especially, in this case for this report, for describing data.

W3C DCAT: DCAT is the **Data Catalog Vocabulary**, a W3C recommendation written in RDF for organizing descriptions of data sets in a catalog. The data set registry used by Data.Gov is based on DCAT. https://www.w3.org/TR/vocab-dcat/

W3C Model for Tabular Data and Metadata on the Web: A W3C recommendation for describing data in a tabular format. This addresses both the layout of data in a table and some metadata needed to describe that data. The spec is written in RDF-S. In particular, it can be used to represent multi-dimensional cube data or record layouts in microdata. https://www.w3.org/TR/tabular-data-model/

W3C PROV: A W3C recommendation written using OWL used to describe the provenance of resources. As stated on the W3C page for the project, it "provides a set of classes, properties, and restrictions that can be used to represent and interchange provenance information generated in different systems and under different contexts. It can also be specialized to create new classes and properties to model provenance information for different applications and domains." https://www.w3.org/TR/prov-o/

W3C RDF Data Cube: A (W3C) recommendation (standard) that provides a means to publish multidimensional data on the Web. This spec is written in RDF-S. The basic cube model underlying the standard is from SDMX. https://www.w3.org/TR/vocab-data-cube/

W3C SKOS: **Simple Knowledge Organization System**, written in RDF-S, is designed to allow the publication of terminological systems in the Web. It is based on the ISO 25964 standards, which superseded the ISO 2788 and ISO 5964 thesaurus standards (now withdrawn). https://www.w3.org/2004/02/skos/

Wikidata: A platform for reference facts and metadata, supporting Wikipedia projects globally and offering Linked Data. It is at http://wikidata.org.

XBRL is an XML standard for accounting reports and other business information. The name is an abbreviation for eXtensible Business Reporting Language. XBRL came from the accounting firms and addresses the items accountants need for their work auditing businesses. In the DATA Act, OMB and Treasury were given authority to set the standards for how agencies send financial information regarding grants, loans, and contracts to them, and they chose XBRL which has tags for business information. XBRL is also used by some government agencies as the format establishment survey responses. We are not aware of it being used for transactions and it is not usually used for statistical data or publications.

Appendix 2: Private organizations and software tools

This list contains names of some private companies and software supporting metadata management. The mention of a specific company or product on this list does not constitute a recommendation or endorsement by any author or the Federal government or paper co-authors.

Aria: Software product developed by MTNA for managing classification schemes, code lists, concepts, and terminology.

Colectica: A software company offering metadata management software based on the DDI standard.

Colectica Repository/Portal: Software suite for managing metadata based on DDI 3.n (Lifecycle). The Repository is the database for the metadata based on DDI, and the Portal is the web interface to the Repository application.

MTNA: Metadata Management of North America is a software development and consulting company.