Forum Guide to Metadata

National Forum on Education Statistics

Forum Guide to Metadata

National Forum on Education Statistics

National Cooperative Education Statistics System

The National Center for Education Statistics (NCES) established the National Cooperative Education Statistics System (Cooperative System) to assist in producing and maintaining comparable and uniform information and data on early childhood, elementary, and secondary education. These data are intended to be useful for policymaking at the federal, state, and local levels.

The National Forum on Education Statistics (Forum) is an entity of the Cooperative System and, among its other activities, proposes principles of good practice to assist state and local education agencies in meeting this purpose. The Cooperative System and the Forum are supported in these endeavors by resources from NCES.

Publications of the Forum do not undergo the same formal review required for products of NCES. The information and opinions published here are those of the Forum and do not necessarily represent the policy or views of NCES, the Institute of Education Sciences (IES), or the U.S. Department of Education (ED).

November 2021

This publication and other publications of the National Forum on Education Statistics may be found at the websites listed below.

The NCES Home Page address is http://nces.ed.gov

The NCES Publications and Products address is http://nces.ed.gov/pubsearch

The Forum Home Page address is http://nces.ed.gov/forum

This publication was prepared in part under Contract No. ED-IES-16-Q-0009 with Quality Information Partners, Inc. Mention of trade names, commercial products, or organizations does not imply endorsement by the U.S. government.

Suggested Citation

National Forum on Education Statistics. (2021). *Forum Guide to Metadata* (NFES 2021110). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Technical Contact:

Ghedam Bairu

(202) 245-6644

Ghedam.Bairu@ed.gov

Foreword

The Forum is pleased to present the *Forum Guide to Metadata*. The purpose of this document is to provide timely and useful best practice information on metadata, including information on how metadata can help manage data complications and improve data quality. This information is intended to help agencies use metadata to document operational changes that impact data.

Publication Objectives

In 2009, the Forum sought to further one of its chief goals, to improve the quality of education data gathered for use by policymakers and program decisionmakers, by developing a "best practice" guide to address the appropriate and effective use of metadata. To this end, the Forum produced the *Forum Guide to Metadata: The Meaning Behind Education Data* (https://nces.ed.gov/pubs2009/2009805.pdf). This new publication advances the 2009 version to focus on the use of metadata by education data specialists and the public at large. Information in the guide has been reorganized and updated and features current metadata-related case studies provided by members of the Forum.

Note: Work on this update began during the coronavirus disease (COVID-19) pandemic in 2020. Although the updated guide does not focus specifically on the pandemic, it includes content highlighting the importance of quality education data and metadata in the context of a widespread health emergency.

Intended Audience

The 2009 version of this guide targeted an audience of local and state education agency (LEA and SEA) staff members. This updated version addresses a broader readership in the education data world, including teachers, data stewards, data managers, and federal staff at ED.

Organization of This Resource

This resource includes the following chapters:

- **Chapter 1** introduces the concept of metadata, or data about data, especially as related to education agencies and education data systems, and discusses metadata as a critical element of sound data management. Chapter 1 continues with a discussion of the benefits of metadata and an examination of an education metadata system and its common components.
- **Chapter 2** focuses on the varied uses of metadata from perspectives including technical metadata, data management metadata, data reporting and use metadata, privacy metadata, and business rules.
- **Chapter 3** discusses planning processes that contribute to the successful implementation of a metadata system in an education setting.
- **Chapter 4** is composed of metadata-related case studies highlighting the challenges, complexities, and lessons learned from metadata management experiences at the SEA and LEA levels.

National Forum on Education Statistics

The work of the Forum is a key aspect of the Cooperative System. The Cooperative System was established to produce and maintain, with the cooperation of the states, comparable and uniform education information and data that are useful for policymaking at the federal, state, and local levels. To assist in meeting this goal, the NCES within IES—a part of ED—established the Forum to improve the collection, reporting, and use of elementary and secondary education statistics. The Forum includes approximately 120 representatives from SEAs and LEAs, the federal government, and other organizations with an interest in education data. The Forum deals with issues in education data policy, sponsors innovations in data collection and reporting, and provides technical assistance to improve state and local data systems.

Development of Forum Products

Members of the Forum establish working groups to develop guides in data-related areas of interest to federal, state, and local education agencies. They are assisted in this work by NCES, but the content comes from the collective experience of working group members who review all products iteratively throughout the development process. After the working group completes the content and reviews a document a final time, publications are subject to examination by members of the Forum standing committee that sponsors the project. Finally, Forum members review and formally vote to approve all documents before publication. NCES provides final review and approval before online publication. The information and opinions published in Forum products do not necessarily represent the policies or views of ED, IES, or NCES. Readers may modify, customize, or reproduce any or all parts of this document.

Working Group Members

This online publication was developed through the National Cooperative Education Statistics System and funded by NCES within IES—a part of ED. The Metadata Working Group of the National Forum on Education Statistics is responsible for the content.

Chair

Georgia Hughes-Webb,* West Virginia Department of Education

Members

Laura Boudreaux, Louisiana Department of Education

Matthew Danzuso, Ohio Department of Education

Brenda Dixon, Illinois State Board of Education

Stephen Gervais,* San Bernardino City Unified School District (CA)

Dawn Gessel, Putnam County Schools (WV)

Ryan Kuykendall, DeSoto County Schools (MS)

Rose LeRoy, New York State Education Department

Rich Nye, Granite School District (UT)

Melanie Stewart,* Milwaukee Public Schools (WI)

Debbie Yedlin, Arizona Department of Education

Consultants

Kristina Dunman and Andrew Scott Pyle, Quality Information Partners

Project Officer

Ghedam Bairu, National Center for Education Statistics

Acknowledgments

Members of the Metadata Working Group would like to thank everyone who reviewed or otherwise contributed to the development of the *Forum Guide to Metadata*, including the following case study and real-world example contributors.

Case Study, Real-World Example, and Content Contributors

Laura Hansen and Jennifer Lee, Metro Nashville Public Schools (TN)

Sandee Hawkins, Mike Mendez, and Jonathan Wiens, Oregon Department of Education

Raymond Martin, Connecticut State Department of Education

* Working group members marked with an asterisk also contributed case studies or real-world examples to this guide.

Contents

Glossary of Common Metadata Items[1]

The following list provides frequently used names for an array of common metadata items and provides a short description of each.

Business rule. A rule under which an organization operates, and the expression of that rule as a mathematical or logical assertion governing how data can be entered or used within a data system. For example, a business rule may state that values for the data element Age of Student must fall within the range of 5 to 21 (that is, 5 ≤ Age of Student ≤ 21) if the agency serves only students of that age.

Calculations or formulas. The actual mathematical formula for computing a value. All components needed for the calculation should be included as related data elements.

Code set. A list of choices that serves as a response for a data element.

Data source. The collection instrument, data file, or formula from which data originated.

Data target. Any reporting instrument (reports, report cards, publications, and other products), data file, or formulas that use or publish the data.[2]

Data treatment. A description of how the format or presentation of data was modified or otherwise changed after collection.

Definition. A description of the meaning of a data element.

Effective dates. The date a data element is introduced or modified, and the date its use ends in favor of a modification or retirement. All past start and end dates are retained as a part of a data element history.

Element name. The unique word or set of words that identifies the name of a metadata item.

Element type. A description of the form or qualities (that is, the "type") of the data that constitute the element.

Field length. The recommended maximum number of places that the value of a data element would require in an electronic record system. For example, a descriptive alphanumeric (AN) element might require 60 letters or numbers for a response, whereas a date (DT) would require eight digits (MMDDYYYY). Both minimum and maximum lengths generally are specified.

Keywords. Any terms or phrases that relate to or are cross-referenced with an item (for searching functions, as an example).

Metadata. Defined most simply as "data about data," metadata are structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information source.[3] In other words, metadata provide the context in which to interpret data.

Ownership and stewardship. The individual or office that authorizes collection of the data and is responsible for the attributes of a data element. Only this individual or office can change an attribute for the element, and all subsequent use of the element should reflect authorized modifications.

Permitted values. The range of possible acceptable values for a data field. For example, an elementary school may limit the permitted values for the Birthdate data element to a range that reflects the allowable age of elementary school students.

1 Note: This is only a partial glossary and should not be understood as a comprehensive list of metadata items.

2 Only applicable where used; this is not a universal metadata item.

3 Riley, J. (2017). *Understanding Metadata: What is Metadata, and What is it For?: A Primer.* The National Information Standards Organization. Retrieved July 1, 2021, from http://www.niso.org/publications/understanding-metadata-2017.

Purpose or mandate. The reason a data item is collected (for instance, state law, school board requirement, component of a report card indicator formula).

Quality metrics. Measures intended to provide information about the relative quality of a piece or set of data. Quality metrics might include completeness, continuity, contiguity, currency, reliability, accuracy, and coherence of a dataset.

Related data elements or components. Other data elements or indicators commonly used with the data element to enhance understanding or provide additional information. For example, all components needed to calculate a data element should be included in this metadata item.

Restrictions. Any factors that limit the value, use, or interpretation of a data element. For example, data about a student's health conditions often are considered confidential and require appropriate access.

Retention period. The amount of time a piece of data should be retained in active or archived form. A "disposal date" may be appropriate for data that will be destroyed.

Routine use. A description of the most common ways a data item is used appropriately. This metadata item also may warn users about common ways that the data are misused.

Security and confidentiality. The classification for a piece of data that conveys the level of access and security to be applied to that data. In addition to the use of standard passwords, encryption techniques, and user authentication methods, security requirements sometimes specify how to dispose of the data appropriately. For example, a list of staff members' Social Security numbers cannot simply be thrown in a trash can or deleted from electronic disk storage. Instead, it might require random binary overwriting for electronic files or shredding of paper files.

Storage or archival destination. The location (physical or electronic) where a piece of data is stored to maintain an archive of data records. This location includes backup storage and should, as appropriate, be as specific as possible (for example, "the Blue Ridge Backup Facility, eastern wing, section 8, box 4, tape 2").

Translations. The transformation of a data value from one format, language, or presentation to another. For example, a date originally collected as 050819 (August 5, 2019 in the DDMMYY format) might be translated to 08052019 in the MMDDYYYY format in the target or destination system.

Chapter One:
Metadata and Metadata Systems

Metadata[4] are defined most simply as "data about data"—structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information source.[5] In other words, metadata provide the context in which to interpret data.

Product labels are a form of metadata that many people use every day, as shown in figure 1. The label on a bottle of juice details important facts about the product, which help the consumer understand what is in the bottle, where it was produced, and how long the product will remain fresh. Similarly, information about an agency's data, such as where the data are stored, when they were collected, which source provided the data, and the data's verification status, provide information to help data users collect, manage, and use the information they need.

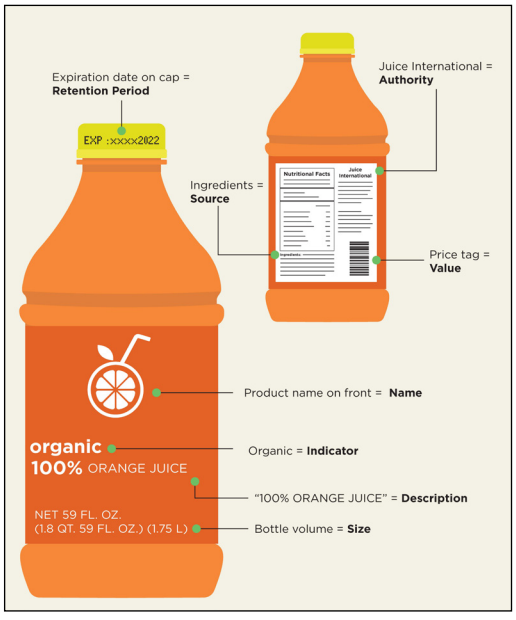

Figure 1: The label on a bottle of juice contains a lot of metadata, or context that conveys important information about the contents.[6]

4 The term "metadata" was coined in 1969 by Jack E. Myers and trademarked in 1986 by his company, The Metadata Company (http://www.metadata.com). The trademarked version is written with a capital "M" and is distinguishable from public use of the term as "metadata" and "meta-data."

5 Riley, J. (2017). *Understanding Metadata: What is Metadata, and What is it For?: A Primer.* The National Information Standards Organization. Retrieved July 1, 2021, from http://www.niso.org/publications/understanding-metadata-2017.

6 This infographic was adapted from a similar image in *What Is Metadata?*, a presentation available from the Federal Geographic Data Committee at https://www.fgdc.gov/metadata/documents/WhatIsMetaFiles/WhatIsMetadataPPT/view.

Education agencies rely on data for decision-making. Education organizations and their stakeholders place significant value on using data to inform instructional, management, and policymaking practices. Agencies are aware that a thorough understanding of data is key to fostering and guiding teachers' careers and the scope of professional development they need, in addition to state and federal policy initiatives, school budgets, and, most importantly, children's education. The volume of information collected can complicate the use of data. Metadata help reduce complexity and promote a better understanding of data by providing contextual information about the data to ensure that data can be managed efficiently and used effectively.

"How Many 8th-Grade English Teachers Are in Your Schools?"

Consider how this apparently simple question actually relies on a clear understanding of what each word represents. Can you be sure that everyone answering this question is on the same page? Metadata define the parts that create the whole.

How many: Does "how many" refer to a head count or full-time equivalent count?

8th grade: Does "8th-grade" include classes with 7th-, 8th-, and 9th-grade students or just classes with only 8th-graders?

English: Does "English" include reading and writing classes? Special education English language classes? Other language arts classes? Other language classes?

Teachers: Do "teachers" include only certified teachers? Only certified English teachers? Certified teaching assistants? Only teachers assigned to teach classes or students this grading period?

Are: At what point in time should the answer be valid? At the beginning or end of the current or previous school year?

In: Does the question include teachers of students cross-enrolled in virtual settings? What if someone teaches English in more than one school—are they counted more than once? Does "in" mean physically present in the school, or does it include remote or virtual teachers?

Your: Does the question include only schools under the authority of the state or local education agency, or does it include all schools within the boundaries of the state or locality?

Schools: Are special education schools included? Correctional institutions that grant education degrees? Other residential facilities? Cross-enrolled virtual settings?

Each element of this question leads to a series of further questions. The responses to these questions may lead different people to different answers about the number of 8th-grade English teachers.

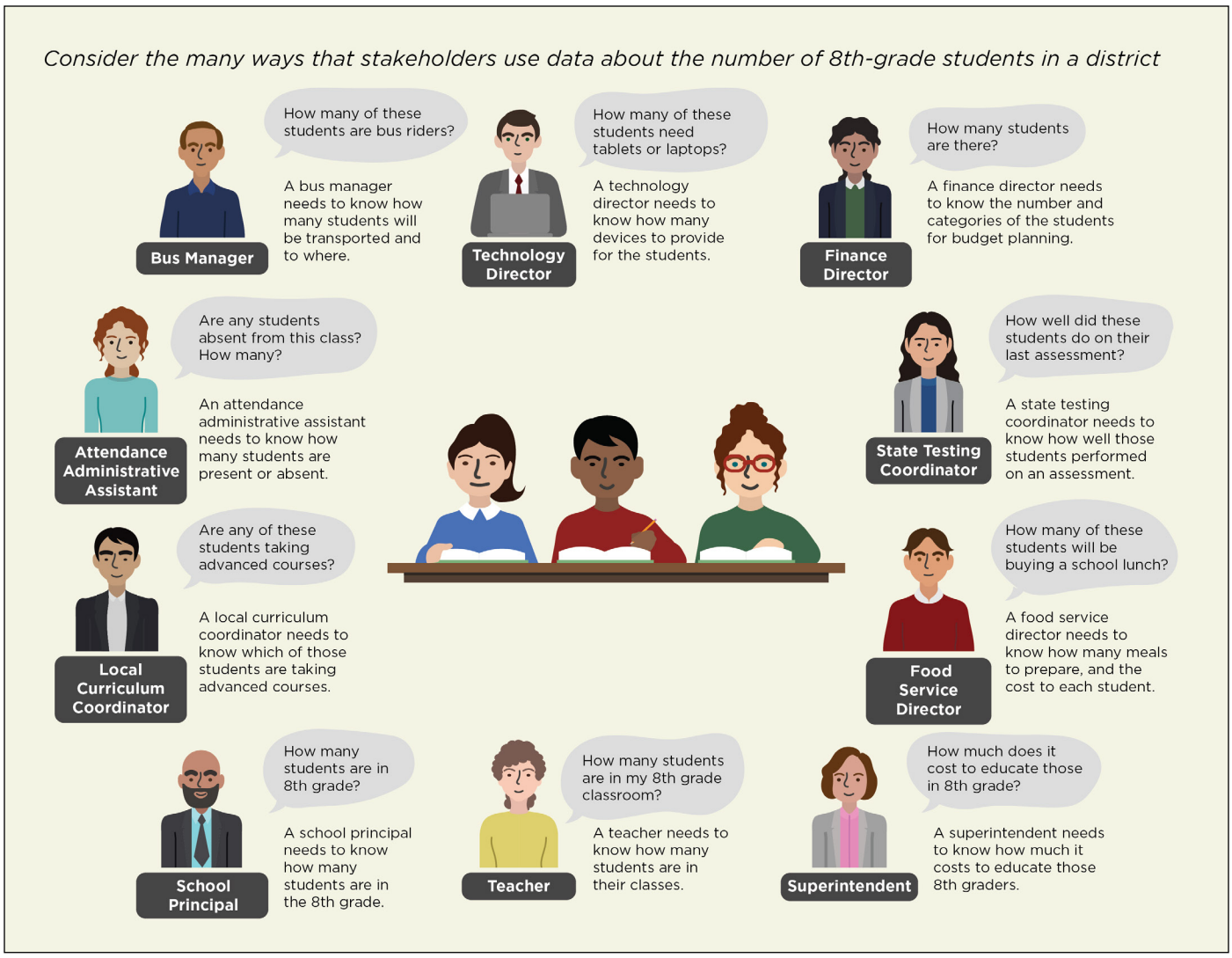

Figure 2: Even if they do not work with data and metadata directly, all educators and staff members whose work involves this 8th-grade class need and use data to perform their jobs and care for students.

Which Metadata Does Your Organization Need?

Many different types of education stakeholders need information to do their jobs. Consider figure 2, which illustrates various ways that LEA staff members use data about a district's 8th-grade students to inform planning, track student outcomes, and support teaching and learning. Each of the stakeholders depicted in the image benefits from metadata. Metadata specify which students were counted as 8th-graders, when and how the students were counted, which system contains the authoritative number of students, how long the number will remain valid, and whether the definition of "number of students" has changed over time.

Although data stewards,[7] information technology (IT) staff members, and others who frequently use metadata may be familiar with its complexities, many other data users may think of metadata as background information about the data. For example, bus managers are likely to check that the data they are reviewing are from the current school year, which is a form of metadata. Similarly, finance directors might check that the data they are using cover all 8th-graders, including those participating in virtual, hybrid, and in-person learning.

> Data stewardship and ownership fall under the broader practice of data governance, which is crucial to the safe management, use, analysis, and communication of education data and metadata. Consult the *Forum Guide to Data Governance* for more information: https://nces.ed.gov/forum/pub_2020083.asp

Accessing and interpreting these data require a host of information management and technology metadata. Technical staff members need to know where each piece of data is physically stored and in what format. Other users, including program staff members and data stewards, need to understand who is responsible for each dataset in the organization, as well as when the data were collected, what time period they represent, why they were collected, and how they are defined. Staff involvement in these processes has grown increasingly complex as practices around metadata have evolved over time. More departments and roles now are engaged in data and metadata management, and organizations need to have systems in place to handle them. Common definitions and understandings about what data mean and how they can and cannot be used are essential tools for the entire organization, no matter its size. With data no longer separated among divisions, it is not sufficient for a select few team members to be data experts. The whole team must understand data and be able to handle data effectively.

The use of metadata has expanded over time to meet the needs of education stakeholders. For example, as state and federal reporting requirements have changed, many SEAs and LEAs have found that metadata help them answer questions more precisely. Moreover, an increased focus on accountability measures in reporting, along with the need for quality data to provide these measures, has increased the number of departments and staff members involved in using and reporting data. Quality metadata explain the meaning and structure of the data, as well as the data source and any collection issues. They also help ensure that users can explain important details about the data they are reporting.

The utility of metadata reaches beyond the reporting requirements that link schools, LEAs, SEAs, and the federal government. Numerous individuals and groups, including parents, guardians, members of the public, researchers, and the media, need to interact with and interpret education data at various times. High-quality, well-managed metadata can help users with varying levels of experience better understand education data.

7 Data stewards are responsible for ensuring the quality of statistical information generated by an organization. Data stewards also generally assume responsibility for enhancing the information reporting process through staff development and by sharing data expertise with the various offices and programs that produce data and information in an organization. National Forum on Education Statistics. (2020). *Forum Guide to Data Governance* (NFES 2020-083). U.S. Department of Education. Washington, DC: National Center for Education Statistics, p. 49

The widespread shift to virtual learning during the COVID-19 pandemic highlighted the usefulness of metadata. Schools and teachers began using education apps and tools to provide instruction and gather key information, such as attendance and measures of engagement. Multiple virtual learning platforms were available for teachers to use. The more robust of these platforms log data such as when and where students log on, how much time they spend on the platform, what activities they participate in, and which web pages they use. Some of these platforms also store assignments, grades, and feedback provided by teachers. Metadata provide information that helps educators, administrators, parents, students, and others understand these data. For example, a virtual platform may provide a screen labeled "progress" that shows what percentage of a course a student has completed. Metadata for the platform explain that the "progress" measure reflects only the percentage of pages within the course that the student clicked on and viewed; it does not reflect how much work the student has done. If a teacher or parent does not have access to the metadata explaining the "progress" measure, they may misinterpret the data. They may see that a student has completed 75 percent of the course and not realize that this means the student has viewed 75 percent of the assignments but has completed only 10 percent of the work.

This type of information is fundamental to a data system's most basic operation. It also helps agencies address some of the deeper—and often more important—characteristics of data, such as the following:

- Are the data private or otherwise sensitive?
- How are the data being used?
- Under what conditions are the data valid for policymaking and reporting?
- How will pending changes in legislation affect current items, definitions, and code lists?

How do metadata benefit different data users?

Anyone who handles data or uses data for decision-making will benefit from metadata, but a few categories of users have the most to gain.

For policymaking and administrative staff members, metadata help

- improve data analysis and use by providing access to instructions, data definitions, and interpretation guidance;
- improve communication with the media and other data users by improving access to supporting or clarifying information about data that are reported publicly;
- improve the accessibility and presentation of data for informing instructional and administrative decisions;
- improve the likelihood that data about schools accurately reflect the state of affairs at the time of collection;
- improve understanding of why individual data elements (such as information about mandates and use) are collected; and
- improve the understanding of connections between data and policymaking.

For data and IT staff members, metadata help

- provide a clear list of technical attributes—such as data type and field length—that can be applied without having to reconsider management parameters each time an item is collected and stored;
- improve the understanding of the business processes driving the collection and use of data that technical staff maintain;

- enable logic-based data quality checks by specifying acceptable values or parameters for each field;
- identify sensitive and confidential data, thereby improving system security;
- simplify and expedite data access and retrieval;
- reduce user inquiries through improved system navigation and data accessibility; and
- simplify the exchange of data between systems, both within and outside the organization.

For program staff members, metadata help

- provide information on the sensitivity of data and all applicable privacy laws and policies;
- reduce the likelihood of incorrect or inconsistent reporting;
- reduce collection demands by identifying redundant data elements;
- minimize questions from technical staff about data maintenance instructions;
- reduce questions from policymaking staff about data use instructions;
- improve data comparability and continuity over time within a program area and across the organization; and
- improve data auditing, thereby increasing overall data quality.

Why should instructional staff members such as teachers and principals care about metadata?

Data are used to evaluate school, student, and teacher performance. Accurate and fair evaluation is supported by accurate and transparent data collection, maintenance, reporting, and use—all of which can be derived from a robust metadata system. When instructional staff and students need information to guide instruction, metadata can help make that information available promptly and in a useful format. Even educators and staff members who do not engage directly with metadata benefit from the high-quality data and reporting made possible by metadata.

Teacher evaluations are a specific example of why teachers should take an interest in metadata—they benefit from knowing what data are included in teacher evaluations, under what conditions they were collected, and how they may have been translated or calculated. Teachers then can understand how those data relate to teacher performance. Metadata also help educators and administrators ensure that their local stakeholders—including board members, parents, and community partners—understand information being shared in school report cards or other public reports about school or district characteristics and performance.

Metadata as a Component of Data Management

> Metadata provide context for a single data item; serve as the backbone for efficient data management; and improve the use, analysis, and management of a body of data.

When datasets were relatively small and simply organized, data typically were used by a handful of people who were intimately familiar with each data element's definition, collection source, uses and limitations, and technical characteristics. The metadata that did exist often were stored in a data steward's memory or a program manager's paper files and could be passed easily from one person to another as a part of the organization's oral and written history. As the education sphere has grown in complexity over the past decades, the field has seen exponential growth of information collected, stored, managed, used, and reported. As in other industries, education metadata have become a necessary component of robust data systems. Without a formal and systematic method for conveying these "data about data," it can be difficult for data, technical, and program staff members to confirm that the information needed to understand the data will be available promptly and in an appropriate format.

A well-managed metadata system keeps metadata organized, defined, and available for all data users, while minimizing disruptions to data management and use. A metadata system ensures that the descriptions, definitions, parameters, usage instructions, and history of each element are accurate and up to date. Metadata are essential for bridging programs and databases because they provide the framework for data exchange and communication within and between organizations. Metadata also inform data policies—such as data retention procedures—and technology planning—such as load time demands—throughout an organization.

The benefits of properly implementing a robust metadata system include

- improving the likelihood that data meet the users' information needs;
- improving the efficiency of data access and integration;
- improving the probability of accurate data interpretation and use;
- identifying what data exist and where throughout an organization;
- identifying redundancy and disparity in datasets;
- increasing the efficiency of data storage and maintenance;
- improving the accuracy of data transfer across systems;
- improving the application of business rules and edit checks;
- capturing changes accurately in data collection, definition, or use over time;
- reducing user expertise required to conduct effective queries;
- advancing data quality;
- ensuring the proper maintenance of information over time; and
- improving the quality of data-driven decision-making in the organization.

Without a robust metadata system, the following types of serious data problems can arise:

- A single data element may be applied inconsistently within an organization. For example, some staff members may code an absence reason as "excused" while others code the same reason as "unexcused."
- Trend studies may not account for changes in definitions or policies that would influence analysis.
- A data item, or even an entire legacy collection, may be maintained when it no longer provides useful information, placing an unnecessary burden on data collectors.
- Policymakers may not thoroughly understand the data they are using. For example, they may not know the difference between the number of teachers expressed as a "headcount" versus a "full-time equivalent" count.
- Without knowing or understanding the data available to them, policymakers might implement a policy requiring data that are not currently collected.
- Data may have been changed or updated over time, leading to reports that show different results based on when they were run without any explanation.
- Data security can be much more difficult without metadata to show who the owners are and who should have access to the data.

Metadata also support data sharing, which is a core element of the work of data researchers and reporters. When multiple users across agencies have access to data, each will bring their own lens to the data. Metadata can be used to reconcile different perspectives, refine understanding of the data, and secure a measure of accuracy, which will benefit all research using those data.

While metadata cannot eliminate every possibility of errors or inconsistencies in data collection, use, or reporting, a sound metadata system minimizes such risks and provides a framework for better understanding.

Metadata Facilitate Data Sharing

The *Forum Guide to Supporting Data Access for Researchers: A State Education Agency Perspective* (https://nces.ed.gov/forum/pub_2012809.asp) and the companion *Forum Guide to Supporting Data Access for Researchers: A Local Education Agency Perspective* (https://nces.ed.gov/forum/pub_2014801.asp) address the role that metadata play in ensuring that external researchers understand the content and context of education agency data and use those data appropriately. For example, the interruptions brought about by the COVID-19 pandemic have left notable breaks in data collection, which data users will need to be aware of when working with datasets that include the 2019-20 and 2020-21 academic years.

These two Forum guides further explain the importance of metadata to researchers and other stakeholders. They also discuss how metadata can be incorporated into the data sharing process, including training data users so that they understand metadata before accessing agency data.

The data life cycle

Figure 3 illustrates a typical data and information life cycle. A piece of information can be generated directly by data collection and input or by deriving it from existing data. The information then stays in use or in storage until the data are retired, archived, or destroyed depending on their sensitivity and ongoing validity. For example, certain health information, disciplinary records, and assessment scores may be destroyed after a student has left school. Federal, state, and local data retention policies also control when, how, and whether data should be destroyed. Metadata can describe the information at each stage in the cycle, and a comprehensive metadata system can track a single piece of data or a dataset as it evolves. These types of life cycle considerations drive the development of metadata systems. They ensure that the individuals who collect, maintain, and use data have the information they need to manage data effectively and efficiently throughout the life cycle.

The Data Life Cycle: Questions That Metadata Can Address

Phase 1: Definition, Planning, and Development

- What do I want to accomplish with the data?
- What questions do I need to answer?
- How are the needed data defined?
- How quickly do I need the data?
- In what format do I need the data?
- What is the best source for the data?

Phase 2: Data Collection

- Where did I get the data?
- Who supplied the data?
- When did I get the data?
- How did I get the data?
- Who owns the data?
- How are the data defined, derived, etc.?
- How do I know that the data are valid?
- Why do I have the data (for example, are the data mandated)?

Phase 3: Verification and Processing

- Where are the data?
- How can I find the data?
- What is the format of the data?
- How did the data get there?
- Have the data been changed?
- Are the data private or otherwise sensitive?
- Does access to the data need to be limited?

Phase 4: Analysis and Use

- What do the data mean?
- How good is the quality of the data?
- What are the limitations of the data?
- How timely are the data?
- How can I use the data appropriately?
- How do others use the data?
- How do these data relate to other data?

Phase 5: Dissemination

- What do the data mean to the reader?
- Who is the audience for the data?
- What was reported in the past?
- Why have the data changed?
- What may or may not be reported at the individual level? In aggregate?
- Which business rules govern report generation and data and privacy protection?

Phase 6: Disposition

- Where will the data be archived?
- When do the data become invalid?
- What are the implications of preserving the data after that date?
- What procedures are required to destroy the data properly (for example, deleting, shredding, degaussing*)?

* Degaussing is a process of wiping data from a hard disk or other storage device by means of magnetism.

Figure 3: The Data Life Cycle

Description of a Metadata System

Metadata systems are driven by the information needs and characteristics of each specific organization, but most have some common features.[8] The following description is based on these commonalities. In general terms, a robust metadata system will have

- system governance arrangements that include policies and procedures for metadata management and use within the organization, as well as related roles and responsibilities for staff members;
- a metadata model that links metadata items to existing data elements and datasets;
- a list or inventory of relevant metadata items, including a lexicon that identifies shared vocabulary for using terms and naming data elements;
- a comprehensive data dictionary; and
- a training program that conveys practical information about how staff members are expected to use and support the metadata system.

Metadata System Governance

To reflect an education agency's long-range vision, goals, and information needs, a metadata system needs support from the highest levels of the organization for system development, use, and maintenance. Managers also must make sure to consider the organization's broader plans and establish a metadata policy that conforms to existing rules, regulations, and laws to which the organization is subject.

> Learn more about principles and best practices for strong data and metadata governance in the *Forum Guide to Data Governance* at https://nces.ed.gov/forum/pub_2020083.asp.

Members of the organization's data governance team should consider metadata management to be as important as any other aspect of the organization's data. As such, data ownership and stewardship responsibilities extend to metadata, as well. Organizational leaders must ensure that all roles and duties for managing and using a metadata system are delineated, assigned, and accepted throughout the organization. In addition to ensuring that staff members fulfill their assigned responsibilities, senior managers should develop and enforce policies and procedures that sustain the metadata system and its use.

Communication and accountability are as critical to metadata governance as they are to most operations in an SEA, district, or school. Universal data governance policies help ensure high communication standards by requiring coordination, consistency, and standard protocols, such as maintaining a unified data dictionary.

> Although metadata can support information management in many ways, resist the temptation to include too much information. Take care not to create unnecessary data about the metadata (meta-metadata) or even data about those data (meta-meta-metadata). Limit metadata items only to what stakeholders need.

8 The Federal Committee on Statistical Methodology (FCSM) offers a helpful presentation on metadata systems, available from the NCES website at https://nces.ed.gov/fcsm/metadata_systems.asp.

Metadata Managed Through a Metadata Model

A metadata model is a formal description of how metadata are structured to support the information needs of an organization. Like any data model, a metadata model can be described at

- a conceptual level, illustrating relationships between metadata items and the larger body of data around which they are generated;
- a logical level, reflecting the technical and operational parameters in which the metadata items exist; or
- from a physical perspective, specifying layout, file structures, and other characteristics.

In more general terms, a metadata model represents how an organization's metadata items relate to one another and to the data that they describe. At a more detailed level, a metadata model maps and illustrates how data elements, metadata items, business rules, subsystems, data repositories, data flows, and information needs relate to one another within an organization's metadata system architecture.

Metadata Item Inventory

Most organizations with metadata systems maintain an inventory of metadata items. The list of potential metadata items is quite long, but most SEAs and LEAs focus on a subset that addresses most issues for most users. The glossary at the beginning of this guide presents items likely to appear in such a list.

Although it can be helpful to review available metadata item inventories from peer organizations, system planners should not expect to meet their stakeholders' needs simply by copying another organization's item inventory without any modifications. The way an organization uses information will drive the design of its metadata system. Different organizations' metadata item inventories will vary, even within the field of education. When planning a metadata system, it is a good idea to complete a needs assessment that gathers information from stakeholders about the data-related activities required for their jobs. A metadata item inventory should be customized to meet those needs.

Some agencies do not need to do metadata planning internally because their metadata are managed within vendor systems that they use for different purposes. For example, many LEAs use a vendor-provided student information system (SIS) that includes data table structures and detailed metadata that align to state and federal reporting requirements. Staff members often complete training to ensure that data collections meet the quality standards necessary to complete state reporting requirements. LEAs still may customize these metadata systems when they create new reports for new programs or legislative requirements. In some cases, the LEA provides feedback that the vendor uses to develop new features. In other instances, the LEA can customize or extend the SIS without involving the vendor unless needed.

Data Dictionaries: A Critical Tool for Data Management

A data dictionary is an agreed-upon set of clearly and consistently defined elements, definitions, and attributes. Creating a data dictionary while building a new data system or adopting one for data systems that are already up and running leads to more consistent data and easier work. In the same way that standard English dictionaries help people use the English language effectively, data dictionaries help organizations maintain consistency in their information systems. Database users and managers can refer to a data dictionary to find out where specific data are located, whether they were reported correctly, how to use them appropriately, and what their values mean. Like an owner's manual, a data dictionary helps the data user understand and work with data.

Although many items in a data dictionary can be classified as metadata, data dictionaries and metadata systems are not interchangeable. Data dictionaries generally contain only some of the metadata necessary for understanding and navigating data elements and databases. Metadata systems, on the other hand, generally include the entire range of metadata items used to manage and analyze a data system, as well as features for sorting, searching, organizing, and connecting data and metadata.

The Common Education Data Standards (CEDS)

The CEDS initiative is a resource for consistency in data and metadata management. CEDS provides a voluntary common vocabulary for education data and models that reflect that vocabulary. CEDS tools help education stakeholders understand and use education data and align their data with CEDS. SEAs and LEAs that have aligned their data practices with CEDS report improvements in data quality and process efficiency. The CEDS community of education stakeholders continues to develop the standards and discuss how they can be maintained.

For more information, see the CEDS website at https://ceds.ed.gov.

Chapter Two:
Using Metadata

Because metadata systems can offer many benefits, an organization should start by identifying its goals for metadata use. For example, organizations can use metadata systems to improve

- technical systems–for example, by quantifying the processing time or resources needed to build custom tables;[9]
- data management–for example, by defining data elements and indicating when the definition may have changed;
- data reporting and use–for example, by ensuring that publicly released data tables explain limitations of the data; and
- data quality–for example, by ensuring that a dataset is complete and includes only permissible values.

Agencies use different types of metadata to meet their goals. An SEA might provide metadata for learning standards that describe curricula by grade level, subject, and topic. A state teacher credentialing agency might provide metadata for teacher licensing that specify when files were created and restrictions on their use. An SIS vendor might provide metadata that describe how different types of data are organized to align with an LEA's technical systems, data quality, and data reporting needs.

Most organizations find many uses for metadata, especially to improve the quality and use of their data. These metadata may be grouped into categories. This chapter discusses the commonly used categories of technical metadata, data management metadata, data reporting and use metadata, privacy metadata, and business rules.

Technical Metadata

The most basic technical metadata items are known collectively as "data attributes," which are technical specifications and parameters that inform how a piece of data is designed within a technical system. Data attributes include a data element's **field length** (for example, up to 12 characters), **element type** (for instance, alphanumeric, date), **permitted values** (such as 0-999 inclusive), **code sets** (such as 0=No and 1=Yes), and **technical translations** (for example, changing date data from DDMMYY to MMDDYYYY format).

9 ED's Elementary/Secondary Information System, or ElSi, at https://nces.ed.gov/ccd/elsi/ is an example of a tool in which users can build custom tables. The web application allows users to quickly view public and private school data and create custom tables and charts using data from the Common Core of Data (https://nces.ed.gov/ccd/) and Private School Survey (https://nces.ed.gov/surveys/pss/).

The metadata item **storage locations** identify the physical or electronic locations where data are stored. Values can include a building site (such as "in office #213" or "at the offsite storage facility at 123 Jones Street"); the machine (such as server serial number 1234); and the database, table, and column (such as staff_db or assignment_tbl, where "db" stands for database and "tbl" stands for table).

Because data do not just appear in a data system and stay there indefinitely, other useful sets of technical metadata are **data source** and **data target**. Data source identifies where data came from, either technically (such as a particular database) or operationally (such as a particular survey). Data target describes the data's predicted destination, such as another database or a report. Programmers use these critical metadata when designing extract, transform, and load (ETL) processes that move data from one system to another.

Load time can be important metadata for some types of datasets and processes. For systems with strong processing capabilities or simple data loads that range from milliseconds to 1 or 2 seconds, load time may not be worth measuring. But a school district that loads 200,000 attendance records each morning needs to know when the system is going to be engaged at full capacity for a longer time.

Now These Are Useful Metadata!

Debbie, the chief financial officer at a district office, never understood why she was submitting the year-end financial reports to the SEA. Information always seemed to go one way—to the state—without being useful to the district. Her outlook changed dramatically, however, when a report from the system noted that the custodial costs were 18 percent higher than comparable districts, flagging an error for her to review.

Debbie knew that her financial records were correct, but it did not make sense that the district was paying 18 percent more for custodial services than comparable districts. She reviewed her submission and quickly realized that she had used the wrong code set when querying the district's financial system. The SEA had asked for a cost for supplies and salaries, but Debbie had given them the cost of supplies, salaries, and benefits. "Well, that would explain the difference," Debbie thought.

Unfortunately, Debbie had used the same number in her preliminary budgeting for the coming school year. "Wow, that correction will reduce the custodial costs in my budget! I am glad the state has a system to identify those types of mistakes!"

Data Management Metadata

At their most basic level, metadata are intended to explain what data mean. Management metadata items include a data **element name**, **definition**, **code sets**, and other data dictionary entries necessary to understand the meaning and context of any single piece of data. For example, when determining the number of students counted as having low socioeconomic status (SES), it is important to know how an SEA defines low SES. Some SEAs define low SES as the number of students who receive free or reduced-price meals through the National School Lunch Program. Other states may define low SES as the number of students who are eligible for the program, regardless of whether they choose to participate. Some SEAs have introduced new SES measures, such as household information or school district poverty estimates. In these SEAs, the definition of low SES has changed over time.[10] The relative meaning of these data depends on the definition of low SES, and anyone using the information would benefit from metadata that clearly and accurately define the term.

10 For more information about how SEAs and LEAs measure SES, see the *Forum Guide to Alternative Measures of Socioeconomic Status in Education Data Systems* at https://nces.ed.gov/forum/pub_2015158.asp.

Similarly, different data users will have ideas and preconceptions that shape how they understand data. Data management teams can anticipate how users will use data and create metadata to guide them toward accurate data use. For example, an LEA may count students as "in attendance" if they are present for at least 50 percent of the school day. However, a researcher who is unaware of this definition may try to use the LEA's data in a project that defines "in attendance" as present for any amount of time during the school day. Comprehensive metadata around the time parameters of attendance will ensure that the researcher understands the data correctly.

Data users often are concerned about data availability, which can be presented as a catalog of what and when data are available. Availability may vary for different users. For example, data might be released earlier for internal planning than for public reporting.

Management Metadata in Times of Disaster

Even when data elements are well defined and consistently understood, data managers and researchers must exercise caution when applying these definitions to data gathered in times of disaster, natural or otherwise. For example, the effects of the COVID-19 pandemic on the 2020-21 school year will need to be factored into research covering this period in future years. COVID-19 has not changed the meaning of data items, but it has changed the environment in which those data exist.

See the following Forum guides for more information:

- *Forum Guide to Planning for, Collecting, and Managing Data About Students Displaced by a Crisis* (https://nces.ed.gov/forum/pub_2019163.asp)
- *Forum Guide to Attendance, Participation, and Engagement Data in Virtual and Hybrid Learning Models* (https://nces.ed.gov/forum/pub_2021058.asp)
- *Forum Guide to Virtual Education Data: A Resource for Education Agencies* (https://nces.ed.gov/forum/pub_2021078.asp)

Restrictions and limitations help users identify factors that limit the use, value, or interpretation of a data element. Restrictions might include a privacy or sensitivity label warning users not to share data or a list of data that cannot be released in combination, such as student names and assessment scores. Limitations often address more practical issues, such as a warning not to compare two similar items that use different sampling techniques.

More advanced users might be interested in **related data elements or components** or **calculations or formulas** that describe how a data value was generated. For example, a dropout rate may include metadata showing the data elements and the formula used to generate it. **Purpose or mandate** generally indicates the underlying reason for collecting the data, including public laws or administrative policies that require collection.

Individuals or offices within an organization must work together to properly maintain metadata. Metadata items like **ownership and stewardship** establish responsibilities for this maintenance. Data owners, who typically have high-level authority over specific data elements or datasets, are accountable for the quality of their data and must understand the responsible use and value of those data. Data stewards, who are typically responsible for implementing data governance policies and standards and maintaining data quality and security, do much of the work related to managing data. This work may include working with data owners to review and update metadata for accuracy. Although organizations may use different terms for the roles of "data owner" and "data steward" based on their governance structures, management terminology, and size, they must establish decision-making responsibilities (data ownership) and management responsibilities (data stewardship) to ensure the effective operation of a data and metadata system. Establishing these responsibilities is equally important in smaller data systems where one person holds the roles of both data owner and data steward.

Data owners are responsible for determining domains that define the range of **permitted values** (for instance, 1-999 inclusive). They also are responsible for the data's **effective date**,

which includes information about the date when the data were collected or loaded and the period for which the data are valid.

Data treatment describes how the format or presentation of data was modified or otherwise changed after collection. This metadata item includes information about mapping and transformations; data cleansing and validation; and rules for significant digits, rounding, cell sizes, business rules, aggregating, and other formulas and derivations. Data history often is presented in the form of an audit trail or other record of how, when, and why data were modified, and by whom.

As an extension of data storage, **retention period** metadata indicate how long data should be maintained and when and how they should be destroyed at the end of their life cycle. For example, some enrollment and fiscal data are maintained indefinitely for historical recordkeeping at a school, LEA, or SEA. **Security and confidentiality** metadata often identify sensitive and private data, as well as appropriate destruction methods for disposing of the data.

Data Reporting and Use Metadata

When data are available to the public, external users and researchers may not have the same grasp of the data's context and meaning as the data professionals inside an agency. Data reporting staff members may decide to publish metadata alongside public-facing data to facilitate understanding. These metadata can take the form of supplemental documentation, concise legends or glossaries, links to related resources, or other materials like the following:

- Subtitles on public-facing dashboards that further explain data to users, as with Michigan's Parent Dashboard for School Transparency: https://www.mischooldata.org/parent-dashboard-page?PageUrl=https://legacy.mischooldata.org/ParentDashboard/ParentDashboardSchoolOverview.aspx?LocationId=S,9730,1254,77
- A list of links to web pages explaining the data used in online resources, like the Wisconsin Department of Public Instruction's explanation of various pages in the WISEdash data system, including data sources and changes: https://dpi.wi.gov/wisedash/about-data
- A brief note advising users of the limits and protections applied to student data, as in the Kentucky Department of Education's State Report Card: https://www.kyschoolreportcard.com/organization/20/school_overview/students/enrollment?year=2020. The web page also alerts users that the COVID-19 pandemic affected 2020 data, and it offers links to more detailed information.
- More detailed explanations of the data available in downloadable reports, such as the Texas Education Agency's overview of its discipline-related data products: https://tea.texas.gov/reports-and-data/student-data/discipline-data-products/discipline-data-products-overview

Privacy Metadata

Metadata can help identify the data that are defined locally as "directory information" per the Family Educational Rights and Privacy Act (FERPA).[11] Metadata also can be used to identify which fields are or may be considered personally identifiable information, such as data that directly identify individuals or could identify individuals when combined with other specific fields.

Metadata can identify which datasets need to be private or confidential and which can be reported or used by various stakeholders. For instance, some systems may present a raw data

11 For more information on FERPA, see https://www2.ed.gov/policy/gen/guid/fpco/ferpa/index.html.

table for assessment results and then generate a cleaned and protected version for public reporting and data or research requests. Within a dataset, metadata can specify when data must be redacted. Many agencies have rules regarding the minimum cell sizes within tables. For example, an agency may require that tables showing fewer than 10 individuals in a single category be redacted.

Metadata also can describe flags for special conditions that apply to certain students who must be protected. Examples include a legal name change, a parent without visiting rights who must not be allowed on or near the school premises, or a medication that a student must take on school trips.

For personnel data, metadata can help to distinguish between elements that are considered public record and those that are considered private and therefore not subject to disclosure under the Freedom of Information Act (FOIA) or applicable state and local privacy statutes.

Business Rules

Business rules are defined as both "directive(s) intended to influence or guide business behavior" and "constraints on a business."[12] Business rules are a form of metadata that express an organization's guidelines for collecting, using, or modifying a particular data element or dataset. For example, an LEA may have a business rule stating that all records of students in grades 3–11 must have a valid score on the annual state math assessment.

Good business rules should

- be explicitly expressed, either in formal language or graphic representation;
- follow an adopted standard for expressing all business rules; and
- be declarative, describing a required or prohibited state.[13]

These declarations should be stand-alone statements of truth about how the organization operates. They should not be further divisible into simpler statements. Users should be able to interpret them under any circumstance as either completely true or completely false. For instance, the business rule "student age cannot exceed 24 years as of September 1 of the current year" means that the age recorded for a student must, under all circumstances, be less than or equal to the value of 24 years as of September 1. Any value in an age field is either completely consistent or completely inconsistent with this rule.

Different realms of metadata will have their own business rules pertaining to their own specific needs, and these rules should be just as absolute. For example, if an organization has a privacy-related business rule that "an N size of 10 or fewer will be reported as an asterisk (*) to ensure privacy," an N size reported as 8 rather than an asterisk violates this business rule.

12 Ross, R.G. (2003) *Principles of the Business Rule Approach*. Boston: Addison Wesley Professional.
13 Perkins, A. (2000) Business Rules = Meta-Data. *Technology of Object-Oriented Languages and Systems*, TOOLS 34.

A Real-World Education Business Rule

For school year 2018-19, the California Department of Education (CDE) modified its existing business rules for submitting discipline data to comply with two adjusted federal reporting requirements:

- Discipline data for all students, including students with disabilities, must follow the same rules.
- Reporting requirements must follow the same rules as the Office for Civil Rights data collection.

These adjustments meant that LEAs now were required to report every incident in which a student committed a statutory offense, not just incidents that resulted in suspension or expulsion. CDE's existing discipline data code "No Suspension of Expulsion" was replaced with "Other Means of Correction or No Action" to accommodate the change.

CDE's definition for the modified code became part of the business rule metadata that describe the agency's discipline data:

An individual committed an offense as defined in Education Code 48900 or 48915, was not suspended or expelled, but the matter was addressed with either no disciplinary action at all or other means of correction. Other means of correction includes, but is not limited to:

1. *A conference between school personnel, the pupil's parent or guardian, and the pupil.*
2. *Referrals to the school counselor, psychologist, social worker, child welfare attendance personnel, or other school support service personnel for case management and counseling.*
3. *Study teams, guidance teams, resource panel teams, or other intervention-related teams that assess the behavior and develop and implement individualized plans to address the behavior in partnership with the pupil and his or her parents.*
4. *Referral for a comprehensive psychosocial or psychoeducational assessment, including for purposes of creating an individualized education program, or a Section 504 plan.*
5. *Enrollment in a program for teaching prosocial behavior or anger management.*
6. *Participation in a restorative justice program.*
7. *A positive behavior support approach with tiered interventions that occur during the school day on campus.*
8. *After-school programs that address specific behavioral issues or expose pupils to positive activities and behaviors, including, but not limited to, those operated in collaboration with local parent and community groups.*
9. *Any of the alternatives described in Section 48900.6 [relating to "community service"].*

SOURCE: California Department of Education. (November 5, 2018). CALPADS Update FLASH #145. Retrieved July 26, 2021, from https://www.cde.ca.gov/ds/sp/cl/calpadsupdflash145.asp.

Data Quality

Quality is a complex yet critical theme in data collection and use. Individuals using data for organizational decision-making, program evaluation, or research must understand the quality of the information they rely on. A host of related concepts, including a wide range of quality metrics, often are used as metadata for assessing and tracking the quality of a data element or dataset. These include, but are not limited to, the following:

- ***Identity.*** Identity directly assesses a dataset's quality and can determine whether every "item"–such as a person, place, concept, or event–is uniquely identifiable and distinguishable from all other entities in a dataset.

- ***Accuracy and reliability.*** Accuracy metrics determine the extent to which data measure what they purport to measure without the presence of bias. They assess whether the data correspond to the process or outcome being measured. Reliability refers to the consistency, reproducibility, and dependability of the data. If the same item were measured multiple times, would the same results be generated? Reliability may reflect uncertainty in a measurement tool or the amount of random error naturally present in the data.
- ***Completeness.*** Completeness measures the degree to which required records and values exist in a given dataset. For example, if individual student records containing 50 items or fields in each record are being transferred to another data system, a record is considered complete when each of the 50 fields has an entry. Sparsity is the inverse measure of completeness; it measures a lack of data when, for example, only four of nine required fields are available. When data are too sparse, assessing what they mean becomes difficult.
- ***Value set testing.*** Value set testing examines the content of data fields to ensure that each data value falls within the domain of allowable values. Allowable values, such as an age range of 5 to 12 years for students in an elementary grade level, often are based on business rules and other guidelines and standards expressed in metadata. Value set integrity is commonly measured by the frequency or rate of domain violations and the percentage of "defective" values that fall outside of the allowable value set. Coherence is a complementary metric that measures value conflicts across related datasets. Coherence looks not only at whether data fall within a range of allowable values but also at whether data that should be identical in different datasets are indeed the same.
- ***Continuity analysis.*** Continuity analysis confirms a consecutive, non-overlapping, and unbroken history of the events represented by the data. For example, continuity analysis might assess whether daily student membership data are available for each school day, with only one value per day, in an academic year before calculating average daily membership for the entire year. If average daily membership is calculated for each grading period, these data must be available consecutively from the first to the last school day of the grading period. Common continuity measures include the ratio of entities with a defective history to those with a defect-free history. More complex measures examine the size of the gap or overlap when defects occur.
- ***Contiguity testing.*** Contiguity testing further assesses the logical progression of data in a dataset. For example, contiguity measures might assess whether the date that a student passes the state's exit exam always occurs before the date of graduation. Contiguity evaluation generally is based on business rules—as well as other guidelines and standards expressed in metadata—to define the logic against which data are assessed. Typical contiguity measures include the ratio of entities with a defective history to entities with a defect-free history. More complex measures examine the frequency with which particular steps in a required sequence are skipped or recorded out of order.
- ***Currency.*** Currency refers to the age or "freshness" of the data. Currency usually represents the time difference between the present date and the date when data were entered into the database. It often is measured in terms of the gap (for example, the number of hours, days, months, or years) between the current date and the date of the most recent data available. This type of information is most important when data values can change significantly over short periods or when data are used routinely

but not collected frequently. Currency provides valuable information for end users. For example, a user should know if the most recent enrollment data were collected 8 months previously.

- ***Frequency of change.*** Data that are subject to regular changes or updates must be revisited and reevaluated with each new collection and recording. Metadata indicating the frequency or rate of these changes will allow data staff members to stay abreast of the data cycle.
- ***Punctuality.*** Punctuality is an extension of currency and measures how quickly users get access to recent data. For example, if student addresses are updated in May, when will the transportation office have them to plan the following school year's bus routes? Punctuality sometimes is referred to as timeliness and also may be used to establish schedules that describe when users can expect new data. Punctuality may vary for the same set of data depending on the audience. For example, a dataset may be available for internal planning purposes more quickly than for external reporting.
- ***Data verification.*** Data verification is the practice of confirming that data are accurate. The related measure data validation refers to the practice of confirming that data agree with expectations of reasonable values and accepted norms. Metadata can document the results of various statistical and procedural techniques used to verify and validate data. These techniques include response and documentation audits, such as an examination of records that substantiate data submitted by a respondent; cross-checks, which examine data from different collections for consistency; and value edits, which, for example, can compare entered data to maximum or minimum expected values.

> The Federal Committee on Statistical Methodology has created *A Framework for Data Quality*, released in 2020, as a guide to help all federal agencies to identify and report data quality. This document provides a foundation upon which federal agencies can make decisions about the management of data products throughout their lifecycle.
>
> *A Framework for Data Quality* may be viewed and downloaded here: https://nces.ed.gov/fcsm/pdf/FCSM.20.04_A_Framework_for_Data_Quality.pdf

Data Profiling

A data profile is a formal summary of distinctive features or characteristics of a dataset, including the data quality items described in the previous section. Data profiling generally starts by examining what an organization expects to find in its data or database and then determines whether the data reflect those expectations. For example, if a data field is mandatory, the organization would expect 100 percent of the fields to contain data. But data profiling may uncover a different reality. Similarly, profiling may examine what and how many codes are found in a field that stores coded values. More advanced data profiling techniques can determine whether a particular information system tends to overcount or undercount some aspect of the dataset—such as the number of students—relative to expected results. Profiling often is used to evaluate data quality, assess whether a collection system supports quality, and determine whether documentation and other available guidance are being used correctly.

Example: How Metadata Concepts Are Applied to a Data Element in a Metadata System

Element name	Birthdate
Description	The year, month, and day on which a person was born.
Field length	10
Element type	DT (date)
Translations	These data are available to authorized viewers in the operational data system but otherwise are encrypted (via master algorithm) and suppressed in all public reporting.
Storage location	Server = svr10079prod; database = Student_Information; table = sat_student_core; field = birth_date
Source	Student Enrollment Collection System
Target	Provided for fiscal auditing internal management and used in data verification audit processes
Restrictions	Only users with access to individual student data are permitted to view this element.
Limitation	This number does not automatically reflect the student's grade.
Components/operations	The element can be compared to the current date to calculate a student's age.
Purpose or mandate	To serve as the district's principal method for determining a student's age. Also used in matching criteria to identify a student.
Owner	District Registrar
Steward	Element is managed by Enrollment Specialist (Mary Jones) and backup is Senior Business Analyst (James Smith)
Time parameters	Student Birthdate is active upon assignment at enrollment and continues until all individual records are removed from the system.
Treatment/layout	Birth dates entered in alternative formats (MM/DD/YY or name of month, day, and 2-digit year) are converted to a YYYY-MM-DD format.
History	Once entered, the element is never changed for an individual student.
Retention	5 years after student has exited the school district
Requirement	Element is required for each individual.
Security/confidentiality	Sensitive and confidential
Identity	Each individual may have only one Birthdate on record.
Accuracy	Audited once after original entry
Reliability	Assessed every 3 years
Completeness/sparsity	94 percent of the 2017–18 records loaded contain values for this field.
Value set	89 percent of the 2017–18 records loaded contain values within the domain of permitted values.

Note: These entries are presented as examples and do not represent metadata from an actual school, district, or state data system. Some types of metadata described in this chapter are more appropriate for describing sets of data rather than individual data elements; these are not included in this exhibit.

Chapter Three:
Implementing a Metadata System

Introducing a metadata system is a complex endeavor that requires planning comparable to any other large organizational initiative. This chapter focuses on steps that are particularly critical for, or unique to, planning and implementing a metadata system in an education setting.

If the organization's staff members do not already have expertise in metadata, the first step in building capacity will be to train prospective team members on metadata and their potential benefits for education organizations.

Metadata systems are built around existing data systems and, ideally, the organization's vision for future data use and management. System development should be driven by the information and business needs of the organization. In other words, what do data users need to know to effectively manage and maximize the quality and utility of the data? A thorough planning process that incorporates data, technical, administrative, and management perspectives improves the likelihood that the system will meet user needs and organizational goals.

Metadata 101: Metadata Do Not Fix Broken Data Systems

Adam, a district data steward, had high expectations for the new metadata system. But the astute data expert noticed something strange happening during the planning process.

As Adam helped the vendor map the district's data elements and datasets to the new system, he identified numerous mistakes in the data's format, structure, and logic.

The first couple times this happened, Adam kept a mental note of what needed to be corrected in the system, assuming he would fix the problems at a later date. When the list grew too long to trust his own memory, he decided to raise the issue.

Adam understood that a metadata system cannot function properly when the main data system is not configured consistently. He knew that without a clear sense of the data in your system, you cannot expect a metadata system to help you better use and manage the data. He also understood that a metadata system is only as good as the main system it is intended to support. He paused the implementation to allow time to clean up the main data system and clearly define the rules that govern the data.

Establishing a Planning Team

Whether the metadata system is developed from scratch or purchased off the shelf, planning requires time, considerable data and technical expertise, a thorough understanding of the organization and its data operations, and extensive project management skills. A planning team should be established to set the course for the project. Team members likely will include the organization's data manager, a technical authority, and a representative from the organization's data governance body. The team also likely will include representatives of other stakeholders who eventually will use the system, such as data entry staff members, data analysts, program staff members, and policymakers responsible for data-driven decision-making. The team should have executive sponsorship and be led by a project manager with sufficient leadership skills and authority to direct the team and make day-to-day decisions without additional permission.

Conducting a Metadata Needs Assessment

> Planners must be able to distinguish between "wants"–those features that stakeholders would like to have–and "needs"–those features that are required to run the organization.

One of the planning team's first challenges is determining how to shape the metadata system to meet the needs of many stakeholders. A comprehensive needs assessment gathers information about how stakeholders will use a metadata system so that planners can ensure that the system will meet those requirements. System users likely will already have some idea of metadata's importance to their work and of how the system can help them. In this case, the needs assessment can focus directly on the known needs of the users and the system itself. When assembled with care and foresight, this needs assessment also can cover future needs and plans to meet them. In other cases, metadata may be a new or poorly understood concept for an organization. In such an environment, the needs assessment must begin by building a basic foundation of support among users. When staff members understand their data needs, the planning team can help them to understand the benefits of good metadata.

> The *Forum Guide to Technology Management in Education* (https://nces.ed.gov/forum/tec_intro.asp) is a helpful resource for educators and LEA and SEA staff members tasked with making decisions about technology. It addresses best practices for implementing a framework and process for making decisions about technology, and it provides additional information on conducting a needs assessment and a build-versus-buy analysis.

The end product of a needs assessment is a needs statement. When writing the needs statement, it can be helpful to imagine that all staff members involved in creating the statement will leave the project and new staff members will implement the next phase of the metadata system. The needs statement is effective if new staff members can understand its findings without additional input from the team that created it. A needs statement should describe both functional needs and technical needs. Functional needs are tasks that the metadata system will accomplish, including

- locating the definitions and other attributes of all metadata items in the system;
- entering metadata into the system;
- searching by keywords and terms;
- customizing and generating metadata reports;
- aligning with the data dictionary;
- linking to external data standards;
- updating metadata items;
- identifying the modification history of metadata items;
- mapping metadata items to individual data elements;

- identifying data element owners and stewards;
- enabling data owners and stewards to modify data and metadata;
- mapping data elements to their physical storage location within a data system;
- assessing data quality; and
- regulating system access.

The technical needs included in the needs statement should not be overly technical or complex. They simply state the capabilities that the technology solution supporting the metadata system will need to have. These capabilities might include

- meeting all relevant technical standards and specifications;
- accomplishing expected performance requirements;
- achieving expectations for the system interface and ease of use;
- safeguarding access and security for sensitive and confidential information;
- handling peak user capacity;
- accommodating connection needs for users based on their location and how often they need to access the system;
- controlling versions of the data dictionary and business rules; and
- automating loading and updating capabilities.

Suggested Outline for a Metadata System Needs Statement

Section 1: Introduction
1.1 Background
1.2 Objectives and scope

Section 2: System contents
2.1 Types of metadata
  2.1.1 Metadata management
  2.1.2 Program administration
  2.1.3 Technical operations
  2.1.4 Usage guidance
2.2 Metadata standards
2.3 Volume of information

Section 3: System functions
3.1 Storage and retrieval capabilities
3.2 Calculation and processing capabilities
3.3 Collection and output capabilities

Section 4: Access and capacity
4.1 Interface requirements
4.2 Hours of operation
4.3 Number of users
4.4 Transmission volume
4.5 Security and access requirements
  4.5.1 User categories
  4.5.2 Permission restrictions
  4.5.3 Remote access
4.6 Reassessment schedule

Section 5: Technical parameters
5.1 Adherence to technical standards
5.2 Requirements for system interfaces

Incorporating Relevant Metadata Standards

Using generally accepted standards in a metadata system can yield many benefits. By using standards, an organization gains greater access to expertise shared by the standards' publishers rather than waiting for staff members to develop comparable levels of expertise through training or trial and error. Staff members can learn existing standards and gain expertise quickly, although not always at the same level as the developer. Using existing standards also decreases the time needed to develop a new system. Rather than starting from scratch, standards can provide a template for a development project. Finally, using accepted standards makes it easier to compare an organization's data with other elementary or secondary education data systems and partners, with other LEA or SEA offices, or with institutions that commonly exchange data with the organization, such as colleges and universities.

Conducting a Cost-Benefit Analysis and Estimating Return on Investment

Regardless of a metadata system's anticipated benefits, two questions will help a planning team decide whether to proceed with its development:

- How much will the metadata system cost?
- Will the benefits outweigh the cost?

To answer these questions, planners use a cost-benefit analysis to ensure that they consider both the positive and negative implications of a metadata system. As an extension of the cost-benefit analysis, return on investment (ROI) is used to express the amount of benefit (return) relative to the number of resources (investment) needed to produce the return. Based on thorough analyses, many organizations find that the potential improvements to data quality and use are worth the costs of developing and implementing a metadata system.

In addition to costs for hardware and software, staff and consultants, and other direct development requirements, planners should expect indirect costs. These expenses often are referred to as "unanticipated costs," although many of them can be anticipated with careful planning. These types of costs include initial and ongoing staff training, user support such as help desks and tutorial development, system maintenance costs, licensing agreements, and ongoing system evaluation.

The absence of a market price for good data presents a challenge to cost-benefit analysis for metadata systems. However, organizations can measure some cost savings from improved data quality in the areas of purchasing, staff allocation, and maintenance and operations. The analysis also can account for cost avoidance, such as not needing to hire consultants or purchase products to revamp aspects of the data system.

> Metadata can reduce reporting burdens, make data more accessible, and improve data quality—all of which can result in more time to teach and otherwise support students.

Some of the benefits of a metadata system are easily quantifiable, but many are not. Even so, organizations can estimate potential financial benefits. For example, a robust metadata system can reduce redundancy in a data system. This reduction can, in turn, decrease the burden of data collection, access, and reporting—each of which has a significant cost. Similarly, metadata systems can make data more accessible, saving staff time. Metadata also improve data quality and use, reduce the need to rerun or correct reports by ensuring that data are reported correctly the first time, and help users better understand the data they are analyzing. These benefits can lead to better decisions about purchasing, staffing, and even academic preferences, such as curriculum selection, teaching assignments, and leadership.

The following example of cost-benefit and ROI analyses covers several frequently recognized categories of costs and benefits, including cost avoidance and ROI for metadata solutions. These categories may vary for different organizations based on a wide range of factors. Monetary values for these costs and benefits can be placed in a spreadsheet for detailed estimates.

Example of Metadata System Cost-Benefit and Return on Investment (ROI) Analysis

Costs

Hardware and software	Purchase of the computers, networking equipment, and software needed to operate the system
Installation	Payment to in-house staff or external contractors to install the system
Consulting	Payment to external contractors for technical or other expertise during system development, installation, implementation, and training
Initial training	Costs associated with providing introductory system training, including staff time and logistical expenses
Ongoing training	Costs associated with providing ongoing training, including staff time and logistical expenses
Opportunity	Unavailability of IT and data staff members to take on other essential tasks while working on the metadata system
Staffing changes	Costs associated with reassigning staff tasks because of system maintenance or use requirements
Support and maintenance	Costs to maintain a system over time, such as upgrades, routine maintenance, and malfunctions
Evaluation	Analysis and reporting costs associated with determining whether the system is meeting user needs and organizational expectations

Benefits*

Reduced IT costs	Savings associated with reduced technical demands because of efficiencies, such as removing redundant data and decreasing storage needs
Interoperability	Savings associated with improved effectiveness and efficiency when sharing data across two or more systems
Productivity gains	Savings associated with increased staff output and efficiency because of improved data access and understanding
Reduced data burden	Savings related to a reduction in the resources (for example, staff time, collection demands, and reporting effort) required to collect, manage, or report data
Reduced redundancy	Savings associated with reducing unnecessary data (for example, data that are no longer used)
Data quality	Savings associated with improving the validity, reliability, utility, and timeliness of data, such as decreased auditing costs
Improved decision-making	Savings associated with making better decisions because of improved data quality and access
System security	Savings associated with decreased risks to an organization's data (for example, improved identification of sensitive or confidential data to support security efforts)

* In addition to readily measurable benefits, less quantifiable benefits—sometimes called "soft" or "intangible"—also occur. Examples include improved data use to keep more students in school, improved staff morale because employees trust the organization to maintain accurate human resources files, and more effective auditing procedures like error checking to confirm calculations. Although assigning monetary values to these "soft" benefits can be hard, they can be estimated and reasonably included in a cost-benefit analysis.

Net cost =	Sum of benefit savings - sum of implementation costs
ROI =	((Total cost savings - total cost of ownership)/total cost of ownership) x 100

Build-Versus-Buy Analysis

Deciding whether to build or buy a metadata system can be a challenge. Starting from scratch without being sure that the human resources needed to handle the job are available can be overwhelming, but commercial products bring their own limitations. For example, most commercial packages are proprietary and cannot be modified without invalidating warranties and, in some cases, preventing upgrades from working properly. The choice to build or buy the metadata system also may dictate whether the system uses centralized, federated, or distributed architecture. Responses to the following questions and considerations can help planners decide whether to build or buy a metadata system:

- **What solutions have similar agencies found?** Have other organizations with comparable needs and budgets found acceptable commercial solutions? If so, those technology solutions might work for your organization, as well. If not, an off-the-shelf product may not work for your organization either.
- **Will you need to modify a purchased solution to meet your needs?** Do commercially available products meet all of your organization's needs, or will they need to be modified? If a product meets most, but not all, of your requirements, you may wish to determine whether it can be modified or reconsider the importance of any unmet needs. A proprietary product's existing functionality sometimes can be altered, but modifications to improve processing speed or other performance aspects may not be feasible. In addition to potentially invalidating warranties, customizing commercial products often makes them incompatible with future releases or updates from the developer. Before proceeding, confirm that support still will be provided for the modified product.
- **Will your purchased solution be adaptable?** Will commercially available products accommodate changes over time? Policies, business rules, and metadata characteristics are not constant. Priorities and procedures occasionally change, and a metadata system must be able to accommodate these changes.
- **Will you have support in the future?** Are commercially available products guaranteed to receive continued support and services from the vendor in the future? Any work with an external vendor must account for that vendor's stability. If a vendor goes out of business or is acquired by another company, you may no longer be able to receive support from them.
- **Can your staff build the system you need?** Do you have access to staff members or consultants with the necessary expertise to build your system? If so, does your project have the resources to cover the staff time or the cost to hire outside expertise? If you must hire external consultants, have you determined how your staff members will support a system that they did not develop?
- **Can you provide consistent system support?** Do you have resources to support the system on an ongoing basis? Have you planned for ongoing costs such as new staff member training, system upgrades, and licensing? Whether you build or buy the metadata system, its initial development costs—though substantial—are not the only resources needed to maintain it over time. A system developed in house needs staff for system maintenance, regular updates, and new development.
- **How soon do you need the system working?** What is the time frame for implementing the new metadata system? If the system has to be up and running urgently, the time needed to build a system in house may rule out that option. A vendor

team with an available solution may be able to supply and implement the new system quickly. In this case, it is vital to have a clear, thorough picture of what you need from the system and a list of questions to help you choose the right solution.

Metadata System Architecture

Metadata system architecture often is driven by the results of a build-versus-buy analysis that, in turn, depends on the organization's existing management, governance, and technology considerations. Metadata system architecture can be divided into three main designs: centralized, federated, and distributed.

With **centralized architecture**, all metadata exist in a single database that stores nothing but metadata. The greatest challenge to implementing centralized architecture is finding a single model that meets the needs of all data systems and users. If a single metadata model has been designed for the entire organization, implementing a centralized metadata system generally is fairly straightforward. Centralized systems are governed, managed, and operated as a single entity. In other words, decision-making also is largely centralized, which helps ensure that metadata are consistent across subsystems throughout the entire organization. For example, the definition and attributes of the "class" data element would be the same in the finance system as in the student record system. Data stewards and data users generally access a centralized metadata system via a single interface, although the core interface may be modified to accommodate differences in access privileges or other user rights.

In **federated architecture** designs, each stand-alone data system in the organization maintains its own metadata system within the constraints of a centralized technical framework and governance structure. This design allows metadata to reflect the specific information needs of each independent data system while still communicating with other systems. Users who access multiple data systems may do so through separate interfaces, and data stewards likely manage each system independently. Metadata items that affect more than one system can be coordinated through automated translation and update processes or by manual modification. Federated designs require central planning and rulemaking within a distributed architecture, as well as a fairly sophisticated technical infrastructure and strong system governance.

In a **distributed architecture** design, each stand-alone data system has a corresponding metadata system. The major benefit of a distributed system is that metadata can be modified and updated without needing to coordinate with other systems. Metadata items also directly reflect the operational data. Despite these benefits, distributed architecture generally lacks cohesiveness and integration. Stand-alone components tend to evolve without adhering to rules and conventions that would synchronize them with the rest of the system. Moreover, vocabularies and definitions often "drift," or start to deviate from those in other systems. This drift can lead to multiple terms for one item and, conversely, multiple items for the same term. Both situations result in duplication and affect data quality. These stand-alone components, sometimes called "silos," can become autonomous over time and eventually unable to exchange data or otherwise work with the rest of the system.

Metadata in the Cloud

Cloud-based data storage services have brought many positive changes to metadata management, including making metadata easier to access. But by moving away from in-house networks and working online, agencies that use cloud architecture encounter challenges on top of those for traditional centralized, federated, and distributed architecture. The most obvious challenge is that users cannot access the data and tools without an internet connection or sufficient bandwidth. Data teams accustomed to working in an office with high internet connectivity may be unable to work with their usual efficiency if required to work from home with less stable internet connections.

Browser compatibility presents another challenge for cloud-based work. Some browsers may be unable to meet a data system's specifications for access. Users may not be able to work at all if they have to use an incompatible browser. Teams using cloud-based data systems must take these issues into account from the outset and ensure that all staff members are properly outfitted and trained.

Establishing a Project Implementation Plan

A thorough and realistic project plan is critical to implementing a metadata system efficiently and effectively. Planners must recognize the iterative nature of developing and implementing a complex technology initiative and budget time for planning, implementing, testing, and refining the system until it meets user requirements. The implementation plan and schedule should address all aspects of the project, from planning through post-implementation training. Good plans often

- start with a basic and understandable feature that stakeholders are likely to care about, rather than a component that may be important but does not address user needs or experiences;
- include time for a "feedback loop" that supports iterative development and implementation; and
- stress extensibility, which allows modules to be expanded or customized with more specialized capabilities after stakeholders have mastered the basics.

A development schedule is only effective if its goals and deadlines are realistic. If they are unattainable and targets are missed, subsequent deadlines lose their credibility.

The project implementation plan should present work in discrete, manageable tasks. For example, mapping a metadata item inventory to all active data elements in a large education data system may be too big a job to accomplish in a single step. Instead, the planning team might identify and prioritize smaller, more manageable tasks such as mapping a smaller set of core metadata items. Alternatively, planners might divide mapping into subtasks based on data categories, such as student personal information, student enrollment, student assessment, staff personal information, and staff assignments. The tasks in the project implementation plan then are assigned, carried out, monitored, and completed in discrete units that can be understood and undertaken by members of the implementation team.

The planning team must give special consideration to coordinating the metadata system with the existing or envisioned data systems. If an organization does not understand what data it has, what format data are in, where they are located, and their quality, a metadata system that depends on those data is unlikely to provide useful information.

Review and Final Assessment

Whether the organization builds or buys its metadata system, the system needs a final, thorough review and assessment before its release. All personnel involved with the planning, procurement, and construction or acquisition of the system should take part in this final assessment. The team must understand what to check, which features to test, and which data elements are most crucial for successful implementation. The metadata needs assessment compiled by the planning team is a useful checklist for the review stage. Once the system is implemented, adding or adjusting functions becomes more complex. The review is vital to verify that the system meets all requirements before it launches.

Tips for Developing an Implementation Schedule

- Reduce large tasks to more manageable subtasks to keep jobs achievable.
- View the first attempt at a task that must be repeated later as a pilot effort. Learn from the experience and modify subsequent efforts and timelines to reflect lessons learned.
- Phase in functionality rather than trying to release every planned feature or capability at once. A phased approach may require more time initially, but it will reduce wasted effort in the long run when lessons learned in early phases improve subsequent decision-making.

New metadata systems are unlikely to be perfect on first use. Some priority features will be part of the initial implementation, while others may have to wait for adjustments or fine-tuning. A final review lets the development team identify features that can be improved after the initial release.

Training Users to Maximize System Utility

Even the best designed metadata system will not work well if the people expected to use it do not understand its purpose or how to operate the system effectively.

Many stakeholders may be unfamiliar with metadata and will need professional development to learn about the concept and its uses. In many fields, including education, readily available data tools are not used to their full potential because ineffective or insufficient training makes using the system more a challenge than a benefit. As with any other effective professional training endeavor, the organization must commit to identifying or developing skilled metadata system trainers, customizing training curricula to reflect specific user needs, and allocating professional development time for stakeholders at the system's initial release and on an ongoing basis. Without comprehensive training, stakeholders are unlikely to appreciate the power and benefits of a metadata system.

The primary purpose of stakeholder training is to teach users to (1) understand the concept, use, and purpose of metadata; (2) operate a metadata system effectively and efficiently; and (3) use metadata to inform their data use. If these major objectives are not accomplished, only technical staff members may have the confidence to use the metadata system, and its potential value will not be realized.

A metadata system training program should accomplish the following:

- **Introduce the concept of metadata.** Different stakeholders will have different understandings of metadata. Training programs should be designed not to overwhelm those unfamiliar with the concept with technical details while also not boring anyone with some familiarity. One strategy for customizing training is to adopt a modular approach, with each module building on content from the previous one. Stakeholders can begin their training at the level most appropriate for their knowledge and

experience. For example, an initial training module might introduce the concept of metadata without delving too deeply into technical details and terminology. A subsequent module might address more formal terms and model relationships between metadata, data, and information needs. A third module then might describe the organization's preferred practices for entering, managing, and using metadata.

- **Present meaningful, real-world examples to illustrate training points.** Trainees often appreciate lessons that they can apply readily to their everyday responsibilities. Good trainers illustrate points with realistic examples that relate directly to participants' duties. In addition to explaining concepts in understandable terms, examples demonstrate how to use metadata on the job and illustrate metadata's power to improve data use.
- **Communicate how metadata benefit the user.** Trainees will be more receptive to training when they understand how it will support their work and processes. Include content in the training program that depicts how good use of metadata can lighten a data professional's workload, make certain tasks less burdensome for them, guarantee higher data quality, and other benefits.
- **Customize training to match audience needs.** Not all stakeholders will use metadata the same way. For example, data stewards generally will be responsible for entering and updating most nontechnical metadata, whereas database administrators often are in charge of technical metadata. Program staff members and other data users need to focus on accessing metadata to improve their analysis and use of program data. Because each stakeholder group uses a metadata system differently, it often makes sense to develop separate, modular training resources that can be combined to meet the needs of each group. Customizing content to meet functional needs and minimize less relevant information generally makes training efforts more efficient and effective.

Teaching Metadata in a Training Program

Effective training sessions often begin with ideas that stakeholders understand and then proceed to more advanced topics. The sequence of topics covered in a metadata system training program might look like the following:

- What are metadata?
- How do metadata affect you and your data use?
- What do metadata do for your organization?
- Metadata system overview
    - Access rights and tools
    - Governance
    - Policies and procedures
- What are the basic (or advanced) system components, and how can you access them?
- How will metadata affect your understanding of data?
    - Data element definitions
    - Permitted values
    - Usage guidance
    - Restrictions
- Use examples (related to audience)
- How are you expected to maintain system security?
- How can you learn more about the metadata system?

Metadata will be a new concept to many participants. Training stakeholders to use a metadata system does not necessarily ensure that they understand when or why to use the system. In addition to describing the concept of metadata, trainers need to explain why metadata are relevant to each stakeholder group's roles and responsibilities:

> Do not assume that stakeholders understand the power and possibilities of metadata. Teaching them how and, sometimes more importantly, why to use a metadata system are critical aspects of any implementation effort.

- Policymaking staff members might learn how metadata can show them how to use data, define terms, and guide interpretations to ensure that their policy decisions are based on an accurate understanding of the data. They also might learn how the data are commonly used and the implications of mistakes in data collection and processing.
- Data and IT staff members might learn that metadata provide a clear list of technical attributes (such as data element type and field length) that do not need to be reconsidered each time an element is collected. They also might learn how metadata can identify sensitive or confidential data and improve system security, and that metadata will make exchanging data between systems easier, both within and outside the organization.
- Program staff members might learn how metadata can help identify redundant data elements and collections, potentially reducing collection demands and improving data comparability and continuity over time. They also might learn that metadata can improve data checking and auditing to increase the overall quality of the data.

Regardless of the examples used, stakeholders should leave a training session with a clear sense of what metadata are and why using metadata is worth their time and effort. Metadata training should be tailored as much as possible to the specific needs of the user. For example, a trainer working with educators and school administrators should understand that these users likely will access the system solely to look up data element definitions or to project different data results when planning classroom instruction or working on institutional improvement. Administrators are likely to be more focused on improving the quality of the data rather than putting the data to a specific use. Stakeholders will better retain their training, and better employ what they learn, when the training is targeted to their practices.

Chapter Four:
Case Studies

Milwaukee Public Schools (WI): Clear, Collegial Communication

As the largest LEA in Wisconsin, Milwaukee Public Schools (MPS) experiences the evolutions and challenges of metadata in a way that similar LEAs in other states can relate to and learn from. Regardless of their size, all LEAs face the challenge of bringing an array of people, agencies, and committees—all with their own concerns—to a common understanding of education data. The most thorough data collection effort only will be effective if the metadata are consistent and everyone involved can understand the data in the same way.

Data Systems Within Data Systems

MPS long has taken a proactive approach to data and metadata, growing and refining its databases to effectively meet the data needs of regular users, such as educators and administrators, as well as researchers. To this end, MPS takes part in the DataShare collaborative, which also involves city and county agencies outside of education, such as the Medical College of Wisconsin, the Milwaukee Community Justice Council, and the City of Milwaukee Health Department. DataShare partners contribute to a de-identified database for research, which needs commonly accepted definitions for any shared data elements.

MPS's commitment to ensuring the usefulness of data also is a key aspect of other district data efforts. The district found that data system technology helps make data definitions and metrics more consistent but is not sufficient on its own. Technology relies on users for best performance. The more users there are, the greater the chance of complications. The Wisconsin Information System for Education (WISE) stores both public and district-level data, and automatically updates data every night. LEAs use smaller data systems to network effectively with WISE. The number of staff members who need to work together across all of these data systems makes having common data definitions and common metrics essential.

Effects of a Pandemic

The COVID-19 pandemic highlighted the importance of metadata for the collection and management of MPS data. Attendance data and graduation data illustrate two areas where metadata are necessary to understand how data collection, definitions, and use changed during the pandemic:

- **Attendance.** With fewer students and faculty on school campuses, a district-wide move to remote learning for many, and a variety of learning methods (concurrent, in-person, fully virtual), MPS had to adjust how it measures and records student attendance. Traditional attendance often cannot be taken in the same way when

teachers and students are in different locations. A variety of data points become essential for accounting for student presence or absence; time or frequency of logins to an online workspace may be used as a metric, for instance. MPS adjusted metadata items associated with these data points, including time parameters, permissions for login IP addresses, and schoolwide collaboration of attendance documentation through instructional communication and submitted coursework. Adapting its data collection processes to changed circumstances let MPS reduce the frequency of data errors and improve data quality.

- **Graduation.** MPS received a waiver to adjust its graduation requirements in 2020 to let students graduate based on the state's requirements rather than more specific district requirements. This flexibility allowed students who were unable to take district-specific courses during the pandemic to graduate, leading to higher rates of 5-, 6-, and 7-year high school graduations.

MPS has documented the reasons and methods for these adjustments. An academic year like 2020-21 will appear as an anomaly to future data researchers, who will need to understand why and how the data for that period diverge from other years.

Communication, Contact, and Flexibility

The complexities of data and the differences between local, state, and federal education agencies' data needs can hinder efforts to meet metadata challenges, as can the human tendency to resist changes to existing practices. A new, more efficient data management platform will be of little benefit to a user who persists in using an older, more familiar model. Every data officer and researcher is an individual, and getting all of them on the same page requires more than just a data dictionary.

For MPS, the solution has been person-to-person communication. The LEA has two main data teams: the Department of Research, Assessment, and Data, and the Department of Student Services. These teams meet once a week to discuss issues and to clear up any areas of confusion between departments. Collegiality and open dialogue can go a long way to solve problems, discover areas for improvement, and introduce staff members to new metrics and definitions. Together, the two teams are better able to take care of issues on their own and to advocate to the SEA when needed. When LEA teams work well together, appeals to the SEA are less frequent and are understood to be essential when they occur.

Collegial communications also help everyone on the team feel validated and valued, leading to stronger relationships. Positive interactions allow all team members to contribute with confidence, which improves the data team's synchronicity and, ultimately, the quality of the data. An open, convivial team environment where everyone communicates has tangible benefits for data collection and reporting, including fewer errors, higher data quality, and a clearer path forward for future collection efforts.

West Virginia: The Importance of a Metadata Plan

The West Virginia Department of Education (WVDE) stores its data in a statewide centralized data system known as the West Virginia Education Information System (WVEIS). WVEIS was developed in the late 1980s at the behest of the state legislature to improve the consistency and timeliness of the data coming from LEAs. Mandated by the state, WVEIS proved to be a useful and resilient data management system, serving educators and researchers for three decades and counting.

Updating a Deep-Rooted Data System

During the 2018-2019 school year, WVDE began the multi-year project of updating WVEIS with a more modern architecture and interface. The SEA has a dedicated data team moving all data from the old system and creating consistent definitions aligned with CEDS. This work includes developing a data inventory and adding new metadata, such as information on when particular codes were added to the system and how long those codes remain valid. Much of the data inventory work builds on progress made with the agency's 2012 Statewide Longitudinal Data Systems (SLDS) grant project, which included building a data warehouse to keep all data accessible for use. Where possible, WVDE staff members are capitalizing on the metadata within the warehouse, as well as the work done while developing the warehouse, to ensure that data are defined, constrained, and used correctly regardless of where they are accessed within the system. The detailed, easily exportable metadata in both WVEIS and the data warehouse help ensure that anyone using WVEIS can readily understand what the data elements mean and how they have changed over time. Although WVEIS and the data warehouse reside in different locations and serve slightly different purposes, WVDE leaders and staff members consider the two systems as parts of a comprehensive whole. The full team works together to create an open data environment where data are owned by everyone rather than one or a few select staff members who "know the system."

Data Collection in a Closed School System

The COVID-19 pandemic in 2020 had the same destabilizing effect on West Virginia's data collection efforts as it had in SEAs and LEAs nationwide. Data still can be collected and reported, but data quality and validity are much harder to verify in such an unsettled climate.

For example, the complete and sudden closure of the school system in mid-March 2020 required education agencies to review and quickly change their methods of measuring student attendance. Following a gubernatorial order on March 13 that closed schools statewide, schools and districts had to work quickly to ensure that students still could learn and were safe and fed. Some schools offered virtual learning opportunities, while many relied on paper-based work packets given the limited availability of broadband internet access in many areas of the state. Teachers called and emailed students in their classes to check on their welfare. Because the state emphasized student safety and wellbeing after schools closed, most LEAs opted to count all student work assigned after the closure as bonus work that only could help, not hurt, a student's final grade in a course. WVDE determined that reporting accurate attendance in spring 2020 would be extremely difficult given the understandable inconsistencies in how LEAs offered learning opportunities and checked on students, many students' limited ability to participate in virtual classes and school meetings, and the decisions to treat post-closure schoolwork as supplemental rather than required. WVDE decided to stop collecting school year 2019-20 attendance data after March 13 because attendance data after that date could not be validated and verified.

Planning Is Essential

Maintaining accurate attendance data is essential. After the pandemic, teachers will need to know exactly how much learning opportunity their students may have missed during the months of remote or virtual instruction. Students who learn best in a classroom environment may require remediation to bring them up to grade level. Conversely, a remote learning environment may have enabled some students to excel beyond their classroom performance. These students could benefit from accelerated learning. Student attendance also is a key metric for other support programs in West Virginia.

Before the school year started in fall 2020, WVDE exhorted all LEAs to find effective ways to track attendance. WVDE added new codes to WVEIS to assist LEAs, including two new attendance codes to account for virtual presence/engagement and virtual absence/non-engagement. WVDE normally gives LEAs time to implement new codes before including them in reports, but that was not the case for the 2020-21 school year. Although LEAs generally have the flexibility to use attendance or absence codes that work for them and their students' situations, they must use and report codes that reflect virtual attendance/engagement appropriately during the pandemic (and beyond). Throughout the pandemic, LEAs and WVDE have included notes in all reports explaining the changes affecting the data. These metadata will help ensure that data users easily can identify and understand data anomalies resulting from the pandemic.

Going Forward

West Virginia's experience illustrates the fundamental importance of having a detailed plan, even for work that already is underway. Although WVDE's WVEIS has a long history of effectiveness, improvements such as the new data inventory will help strengthen metadata and ensure that the data are easy to understand. Moreover, the COVID-19 pandemic highlighted the fact that high-quality data become more important in times of crisis. It is essential to have a documented and carefully thought-out plan for data management during a crisis that can be easily updated with changes in definitions and circumstances.

Oregon: Consistency Through Collaboration

The Oregon Department of Education (ODE) has a longstanding commitment to using metadata to improve the consistency and quality of data while also reducing data reporting, management, and collection burdens. The systems and technology underlying ODE applications are highly driven by metadata, and staff members involved in data collection and reporting are well practiced in using metadata in their work. To further improve its use of metadata, ODE launched a new large-scale metadata project in 2018.

Collection and Reporting

ODE's business analysts use and update metadata the most, and the principal use for metadata is to make data consistent across multiple applications. Metadata support the development of business rules, data validation, and maintaining data consistency throughout ODE's schools and districts. Metadata also help standardized data entry processes whenever possible. For example, school and district staff members can use standardized dropdown menus to enter or search for data.

Business analysts and system developers work together to manage metadata, thereby ensuring that any new parameters put on data are validated through use. Thorough data reporting helps validate metadata further; by checking school and district data submissions for accuracy and consistency, ODE staff members quickly can identify whether the metadata meet the needs of data reporters. This concerted effort at consistency helps prevent discrepancies that otherwise might need to be resolved by changing the data system code.

Flexibility Through Metadata

ODE's metadata project has evolved over the years. It began in part to address a report-generation issue in which staff members had to modify the code for each report in order to change the reported data. Over time, ODE staff members began to use metadata when writing code. Developers subsequently found that building projects with an eye to metadata sourcing from the beginning had a number of benefits, including the following:

- Data pulls and filters could be embedded in metadata business rules, removing the need to work directly with code.
- Business rules could be adapted quickly just by changing the relevant metadata. For example, the dates when a particular data collection opens and closes can be modified by making changes to metadata.
- Data searches can incorporate points in time, and codes can be reassociated with varying definitions without changing them. For example, the term "Limited English Proficient (LEP)" has been replaced by "English learner" over time, but both definitions use the same code. A point-in-time search always will turn up the appropriate definition for the specified time.

During the COVID-19 pandemic, this enhanced flexibility allowed ODE to manage changes and extensions to COVID-19 data collection efforts and even to generate new pandemic-specific emergency codes quickly without causing difficulties for LEAs.

Toward Centralization and Comprehension

ODE's metadata project spans several related entities that are not integrated. ODE is working to centralize these metadata, helped by a standing catalog populated with metadata about each entity's data collection. This catalog is open to the public at https://www.ode.state.or.us/apps/CollectionCatalog, allowing anyone to search for specific types of collections. The catalog has benefitted many stakeholders, including the Oregon legislature, in part because it allows researchers to find which laws drive data collection efforts.

Centralization poses the challenge of educating potential users and researchers about the metadata they can access and training them to search for data more effectively. ODE is creating an application to let data owners, analysts, and users keep up with yearly changes to codes and to make changes easily through the workflow process. A web-based user interface for this application is forthcoming. ODE also is considering developing a researcher's guide to data use.

Lessons Learned

ODE advises that similar metadata projects will require dedicated investments of time and resources, as well as solid relationships between the data team and the agency's other divisions. In addition, metadata management should be part of the agency's strategic plan if possible. Integrating a metadata project with the agency's formal goals helps secure buy-in from agency leaders, gives the project team tools to overcome anticipated or unexpected challenges, and smooths the project's path to successful completion.

Metro Nashville Public Schools (TN): A User-Focused Approach to Metadata

Metro Nashville Public Schools (TN) (MNPS) has an established data governance structure that informs the district's approach to metadata. Within the district, collaboration between technical and data staff members helps ensure that data users have the metadata they need to report and use district data effectively and accurately. MNPS is attentive to data needs such as storage, maintenance, and cataloging, as well as the needs of those who rely on data for their work.

State-Level Support

The Tennessee Department of Education (TDOE) robustly documents the metadata that govern the outputs from its SIS. MNPS uses this documentation as a standard to keep LEA data and definitions aligned with the SEA's. In addition, the TDOE data manual defines each data field, provides parameters for the format of the data and includes a set of business rules for data collection. TDOE's online Education Information System (EIS) is available at all times to connect LEAs with answers, resources, and support personnel. The combination of SEA data extracts and the business rules that shaped them have helped keep data work manageable for MNPS.

Building Data Literacy

MNPS departments share responsibility for data governance, and the LEA emphasizes the importance of data literacy for all staff members who use data. The LEA created data guides for commonly used reports, which explain the origins and appropriate use of the data in the report and include the definitions that are most needed to understand the data. A data quality dashboard helps schools to identify data that do not conform with business rules. Data-literate department leaders who are familiar with MNPS's business rules, data definitions, and data conventions support the teachers and staff members they supervise, who have their own data collection and reporting responsibilities.

Giving staff members who use data a role in the LEA's data management makes them more open to learning about data and metadata. They learn about the benefits of quality data, and they can identify shortfalls or gaps in the data that need to be addressed. As they learn, they also become able to collaborate on solutions. Staff members receive help in this work from designated data quality managers and specialists who are assigned to schools in an advisory capacity. Learn more about the data quality managers and specialists in the MNPS case study in the *Forum Guide to Data Governance* at https://nces.ed.gov/forum/pub_2020083.asp.

Establishing Data Standards

MNPS fulfills the need for organized, accessible data in part through a data warehouse that brings data from different sources into one place. From the warehouse, MNPS creates analytic reports that meet stakeholder needs and help users visualize data in useful and targeted ways.

MNPS also is aligning its data with established data standards. This work has helped the district identify and document metadata and promote data system interoperability. For example, school names are important for MNPS's reports to the SEA, but schools often are known by nicknames or shortened versions of the official school name. In some cases, these school nicknames or abbreviations are known so widely that they are used in place of the official name on websites or in stand-alone datasets. As part of its effort to establish data standards, MNPS determined that the primary source for all school names should be the SIS. MNPS's data team communicated this change to staff members throughout the district. Information Technology (IT) staff members were informed to pull school name data from the SIS rather than store it in different systems. There is an established change management process to handle changes to school names. Once a change is approved by the SEA, it is implemented at the source system and then communicated across all systems.

Lessons Learned

MNPS offers two key takeaways from its experience with data and metadata management:

- Partnerships and collaboration are key. Whether metadata users are IT professionals, data staff members, specialist advisers, or staff members networking across departments, data and metadata are best managed when everyone works together. By collaborating in this way, staff members can exchange ideas, form solutions, and strengthen the data literacy and confidence of the team as a whole.
- Focus on what is necessary. Begin by evaluating your agency's most urgent data and metadata needs, and use them as the starting place for developing systems and processes. Staff members will be more engaged, receptive, and proactive when their data work serves a known need, and they can see how their use of metadata resolves issues to make their work more effective.

Reference List

Citations and References

Clyde, A. (2002). Metadata. *Teacher Librarian*, 30(2): 45-47.

El-Sherbini, M. and Klim, G. (2004). Metadata and Cataloging Practices. *The Electronic Library*, 22(3) 238-248.

Family Educational Rights and Privacy Act, 20 U.S.C. § 1232g (1974)

Federal Committee on Statistical Methodology. (2020). *A Framework for Data Quality* (FCSM 20-04). Retrieved October 18, 2021, from https://nces.ed.gov/fcsm/pdf/FCSM.20.04_A_Framework_for_Data_Quality.pdf

Federal Geographic Data Committee. (2016). *What is Metadata?* [PowerPoint slides]. https://www.fgdc.gov/metadata/documents/WhatIsMetaFiles/WhatIsMetadataPPT/view.

Gillman, D., McNamara, K., Meyer, P.B., Moris, F., Savino, W., and Taylor, B. (2020). *Metadata Systems for the U.S. Statistical Agencies, in Plain Language*. The Federal Committee on Statistical Methodology. Retrieved July 1, 2021, from https://nces.ed.gov/fcsm/metadata_systems.asp.

Lee, H., Kim, T., and Kim, J. (2001). A Metadata Oriented Architecture for Building a Datawarehouse. *Journal of Database Management*, 12(4): 15-25.

Michener, W.K., Brunt, J.W., Helly, J.J., Kirchner, T.B., and Stafford, S.G. (1997). Nongeospatial Metadata for the Ecological Sciences. *Ecological Applications*, 7(1): 330-342.

Perkins, A. (2000). Business Rules = Meta-Data. *Technology of Object-Oriented Languages and Systems*, TOOLS 34.

Riley, J. (2017). *Understanding Metadata: What is Metadata, and What is it For?: A Primer*. The National Information Standards Organization. Retrieved July 1, 2021, from http://www.niso.org/publications/understanding-metadata-2017.

Ross, R.G. (2003). *Principles of the Business Rule Approach*. Boston: Addison Wesley Professional.

Shankaranarayanan, G. and Even, A. (2006). The Metadata Enigma. *Communications of the ACM (Association for Computing Machinery, Inc.)*, 49(2): 88-94.

Related Resources

National Forum on Education Statistics Resources

The National Forum on Education Statistics has produced a wide range of publications related to data quality and data management. These resources are available at no cost at http://nces.ed.gov/forum/publications.asp.

Forum Guide to Virtual Education Data: A Resource for Education Agencies (2021)

https://nces.ed.gov/forum/pub_2021078.asp

This guide is designed to assist agencies with collecting data in virtual education settings, incorporating the data into governance processes and policies, and using the data to improve virtual education offerings. This resource reflects lessons learned by the education data community during the coronavirus disease (COVID-19) pandemic and provides recommendations that will help agencies collect and use virtual education data.

Forum Guide to Attendance, Participation, and Engagement Data in Virtual and Hybrid Learning Models (2021)

https://nces.ed.gov/forum/pub_2021058.asp

This guide was developed as a companion publication to the 2018 *Forum Guide to Collecting and Using Attendance Data*, drawing upon the information included in that resource and incorporating lessons learned by state and local education agencies (SEAs and LEAs) during the COVID-19 pandemic. The information is intended to assist agencies in responding to the current need for these data, as well as future scenarios, such as courses with blended/hybrid learning models or natural disaster situations in which extended virtual education is required.

Forum Guide to Data Governance (2020)

https://nces.ed.gov/forum/pub_2020083.asp

This resource provides timely and useful best practices, examples, and resources for agencies implementing or updating their data governance programs. It provides an overview of data governance; discusses effective data governance practices, structures, and essential elements; describes how to meet privacy and security requirements while also meeting data accessibility and sharing needs; and includes detailed case studies from education agencies about their data governance efforts.

Forum Guide to Exit Codes (2020)

https://nces.ed.gov/forum/pub_2020132.asp

This guide is an update of the 2006 Forum publication *Accounting for Every Student: A Taxonomy of Standard Student Exit Codes*. The guide defines and presents a model taxonomy of student exit codes, discusses best practices and methods for addressing challenges in exit codes data collection, and provides case studies illuminating how SEAs and LEAs have navigated these challenges.

Forum Guide to Planning for, Collecting, and Managing Data About Students Displaced by a Crisis (2019)

https://nces.ed.gov/forum/pub_2019163.asp

This resource provides timely and useful best practice information for collecting and managing data about students who have enrolled in another school or district because of a crisis. It highlights best practices that education agencies can adopt before, during, and after a crisis and

features contributions from agencies that have either experienced a crisis or received students who were displaced by a crisis.

Forum Guide to Technology Management in Education (2019)

https://nces.ed.gov/forum/tec_intro.asp

This resource is designed to assist education agency staff with understanding and applying best practices for selecting and implementing technology. It addresses the widespread use and integration of technology in modern education systems and focuses on technology governance and planning, technology implementation, integration, maintenance, support, training, privacy, security, and evaluation.

Forum Guide to Education Data Privacy (2016)

https://nces.ed.gov/forum/pub_2016096.asp

This resource provides SEAs and LEAs with best practice information to use in assisting school staff in protecting the confidentiality of student data in instructional and administrative practices. SEAs and LEAs may also find the guide useful in developing privacy programs and related professional development programs.

Forum Guide to Alternative Measures of Socioeconomic Status in Education Data Systems (2015)

https://nces.ed.gov/forum/pub_2015158.asp

This resource provides "encyclopedia-type" entries for eight plausible alternative measures of socioeconomic status (SES) to help readers better understand the implications of collecting and interpreting a range of SES-related data in education agencies. Chapter 1 reviews recent changes in how SES data are collected in many education agencies and presents a call to action to the education community. Chapter 2 reviews practical steps an agency can take to adopt new measures. Chapter 3 describes each of the eight alternative measures, including potential benefits, challenges, and limitations of each option.

Forum Guide to Supporting Data Access for Researchers: A State Education Agency Perspective (2012)

https://nces.ed.gov/forum/pub_2012809.asp

Forum Guide to Supporting Data Access for Researchers: A Local Education Agency Perspective (2013)

https://nces.ed.gov/forum/pub_2014801.asp

These two Forum guides recommend core practices, operations, and templates that can be adopted and adapted by SEAs and LEAs as they consider how to respond to requests for data about education.

Traveling Through Time: The Forum Guide to Longitudinal Data Systems (Series)

Book I: What is an LDS? (2010) http://nces.ed.gov/forum/pub_2010805.asp

Book II: Planning and Developing an LDS (2011) http://nces.ed.gov/forum/pub_2011804.asp

Book III: Effectively Managing LDS Data (2011) http://nces.ed.gov/forum/pub_2011805.asp

Book IV: Advanced LDS Usage (2011) http://nces.ed.gov/forum/pub_2011802.asp

The *Traveling Through Time* series is intended to help SEAs and LEAs meet the many challenges involved in developing robust systems, populating them with quality data, and using this new

information to improve the education system. The series introduces important topics, offers best practices, and directs the reader to additional resources related to longitudinal data system (LDS) planning, development, management, and use.

Other Related Resources

California Longitudinal Pupil Achievement Data System (CALPADS) System Documentation https://www.cde.ca.gov/ds/sp/cl/systemdocs.asp

Common Core of Data https://nces.ed.gov/ccd/

Common Education Data Standards (CEDS) https://ceds.ed.gov

Kentucky Department of Education State Report Card https://www.kyschoolreportcard.com/organization/20/school_overview/students/enrollment?year=2020

The Metadata Company http://www.metadata.com

Michigan School Data Parent Dashboard for School Transparency https://www.mischooldata.org/parent-dashboard-page?PageUrl=https://legacy.mischooldata.org/ParentDashboard/ParentDashboardSchoolOverview.aspx?LocationId=S,9730,1254,77

Oregon Department of Education Collection Catalog https://www.ode.state.or.us/apps/CollectionCatalog

Private School Survey https://nces.ed.gov/surveys/pss/

Texas Education Agency Discipline Data Products Overview https://tea.texas.gov/reports-and-data/student-data/discipline-data-products/discipline-data-products-overview

U.S. Department of Education (ED) Elementary/Secondary Information System (ElSi) https://nces.ed.gov/ccd/elsi/

Wisconsin Department of Public Instruction WISEdash – About the Data https://dpi.wi.gov/wisedash/about-data