TECH**BRIEF**

FHWA InfoMaterials™ Dataset Management

FHWA Publication No.: FHWA-HRT-21-065

FHWA Contact: Y. Jane Jiang, FHWA Long-Term Infrastructure Team (LTIP), (202) 493-3149, jane.jiang@dot.gov.

INTRODUCTION

The Federal Highway Administration (FHWA) developed a Web-based portal, InfoMaterials, to provide easy access to the agency's infrastructure research and materials testing datasets (Federal Highway Administration 2021a). FHWA collects a tremendous amount of valuable data through cutting-edge testing technology, sponsored research, and other collaborative efforts. Data are essential for research and innovation, so it is crucial that the FHWA preserve, document, and provide accessibility to stakeholders and researchers to reduce data collection duplication and maximize the benefits of the collected data. FHWA's InfoMaterials was first released in January 2020. It hosts datasets containing characterization data of asphalt and concrete materials, pavement performance testing and analysis results, and many other types of structured and unstructured highway infrastructure research data. The datasets are supplemented by extensive metadata, descriptions, and references to assist users in their understanding of the datasets.

This document was originally developed to provide general guidance for the FHWA Office of Infrastructure Research and Development (R&D) at the Turner-Fairbank Highway Research Center (TFHRC) for the InfoMaterials Web portal. However, these guidelines can potentially be used for data and metadata management of all research data collected or generated by FHWA programs/studies.

In the many research programs and studies that FHWA conducts, measurements are recorded and data are collected and analyzed by a research team. A well-structured data management approach increases the visibility and utilization of the data. It not only preserves the integrity of the data, but also helps save time by simplifying data access and retrieval processes. This approach facilitates further utilization and new discoveries of data more efficiently.

For research or materials testing projects, the following are key elements to successfully execute a data-management plan:

- Identify data needs and outcomes of the project.
- Map the data collection and testing process.
- Assign responsibilities for each stage of data collection, data entry, data organization, quality control (QC)/quality assurance (QA), and documentation.
- Capture all data and metadata during the process.
- Define nomenclature and a common labeling system.

U.S. Department of Transportation
Federal Highway Administration

Research, Development, and Technology
Turner-Fairbank Highway Research Center
6300 Georgetown Pike
McLean, VA 22101-2296

https://highways.dot.gov/research

- Conduct QC/QA of the collected data.
- Centralize data storage in a secure location and back up regularly to prevent data loss.

This document provides information for data organization and storage in terms of data types, metadata, formats, and organization. Examples on how to organize different types of data are included for illustrative purposes.

DATA TYPES

Effective and efficient data management is critical for making research data accessible and available for the intended audience. The data collected and/or generated during research studies typically fit into one of three major data types: structured; unstructured; or semistructured data (figure 1). The structured data are preferred because they enable easy filtering and data visualization.

The following sections provide information on these three data types.

Structured Data

Structured data conform to a predefined data model that organizes the data in a way that is simple to analyze and comprehend. Structured data typically consist of a tabular format, where rows and columns have a logical relationship. Common examples of structured data are spreadsheets or relational databases such as Microsoft® Access® and SQL Server databases.

Structured data are considered the most traditional form of data storage, since the earliest versions of database management systems could store, process, and access this data type.

The data model for structured data defines how the data are stored, processed, and accessed. Each data attribute is uniquely defined in the data model and can be retrieved individually or along with other data attributes. This strategy makes it very efficient to filter and query the data from multiple tables.

Unstructured Data

Unstructured data consist of information that either does not have a predefined data model or is not organized in a predefined manner. Unstructured data are typically text-heavy content but may contain data attributes such as dates, numbers, and facts. This format causes ambiguities and irregularities, which make the data difficult to analyze. Common examples of unstructured data include images, documents, and audio/video files.

Semistructured Data

Semistructured data are a form of structured data that do not conform with the formal structure of data models associated with relational databases or other forms of data tables but contain tags or other markers to separate semantic elements and enforce hierarchies of records and fields within the data. Therefore, semistructured data are also known as having a self-describing structure. Examples of semistructured data include JavaScript Object Notation (JSON) and Extensible Markup Language (XML) files. The reason this third category exists (between structured and unstructured data) is because semistructured data are considerably simpler to analyze than unstructured data.

METADATA

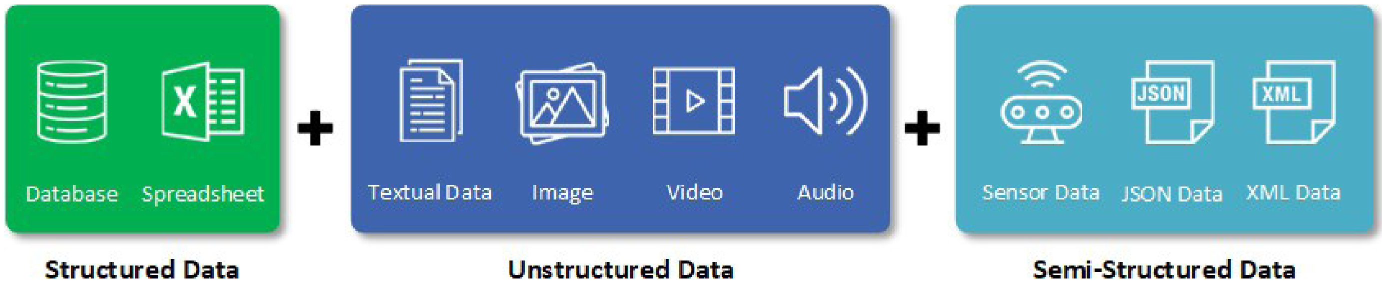
A key to effective data exchange is providing adequate metadata for users. Metadata are detailed data about other data. They provide additional information about a specific set of data to describe, explain, locate, or otherwise facilitate the retrieval of desired information. Well-documented metadata help in managing information resources through elements such as version number, archival date, and other technical information for purposes of file management, rights management, and data preservation. For research data, the metadata are needed both at the project level and the data level, as explained in the following sections.

Project-Level Metadata

Project-level metadata describe the "who, what, where, when, why, and how" of the dataset, which provides context for understanding why the data were collected and how they are used. The following are some examples of project-level metadata:

- Project/Program: Project/program description, such as project name, goal, location, schedule, and other high-level information.
- Protocols/Standards: Information describing the testing protocols and devices used, such as brand, year/version, and calibration information. It could include the standard test specification used, such as American Association of State Highway and Transportation Officials (AASHTO) or ASTM (formerly American Society for Testing and Materials) designation and the version.
- Computation: Information explaining the calculation methodology, algorithms, and formulas used to generate the data.

Figure 1. Graphic. Data types.



Source: FHWA.

- References: Information regarding the reference material, such as reports and documents used in the project/program.

Data-Level Metadata (Data Dictionary)

Data-level metadata, also known as data dictionary, are a collection of the names, definitions, and attributes for data elements and models. The data in a data dictionary are the metadata about the database. For data stored in either a database or a spreadsheet format, a data dictionary is required to help data users understand the structure of the provided tables or worksheets. The key features of a data dictionary include table descriptions, field descriptions, file descriptions, codes, and other reference information.

- Tables: In the case of structured data, the meta data include a table dictionary that contains a description of the table contents and its relationship to other tables.
- Fields: In addition to the table dictionary, a data dictionary is also required, which contains information on the data fields, such as data types, data units, and precision. It also includes information regarding the computed fields (i.e., calculated from other numerical fields), computation algorithms/formulas, and unit-conversion factors.
- Files: The metadata for unstructured and semistructured data, such as files and/or documents, could be the attributes associated with the corresponding files, such as file type, description of the file, or date created. In a set of photographs, for example, metadata could describe when and where the photos were taken. In the case of semistructured data, the metadata could be a mix of both the data dictionary elements as well as the attributes associated with the file's data. In some cases, the semistructured data are machine-generated data files. Metadata that help interpret these files should be documented.
- Codes: Codes are used in databases to reduce data storage requirements, enforce data uniqueness, and provide an easy means to sort data. When codes are used within a database, a two-level table data structure is recommended for users to decode data. The first-level table is a code type table, and the second level is the codes table. The purpose of the code type table is to assign a unique index, or name, to each of the different code types that are specified in the data dictionary by unique combinations of the table name and field name.
- Classification: Data classification is a hierarchical grouping of data that are logically associated with each other or have something in common. Classification is done at the table and/or file level. There may be multiple tables and/or files within a given category, and combining them in a classification allows easy retrieval of related data. For example, materials and performance data are two classifications of data that can contain multiple data elements under each classification. A given data table or file can be associated with multiple data classes.

FORMATS

Research data come in many formats: text, numeric, multimedia, models, software languages, discipline specific, and/or instrument specific. Table 1 provides guidance on selecting the appropriate file format and data type based on the nature of the data.

ORGANIZATION

One of the simplest and most overlooked aspects of data management is keeping the research data organized from the beginning. Once the research begins, there may be multiple files in various formats, multiple versions, and methodologies all relating to the same data. Therefore,

Table 1. Data formats

Nature of Data	Recommended Format	Data Type
Quantitative tabular data with extensive metadata: A dataset with variable labels, code labels, and defined missing values, in addition to the matrix of data.	• Microsoft Excel® (.xlsx) • Microsoft Access (.accdb) • Microsoft SQL Server (.bak)	Structured
Quantitative tabular data with minimal metadata: A matrix of data with or without column headings or variable names, but no other metadata or labeling.	• Comma-separated values (.csv)	Structured
Geospatial data: Vector and raster data.	• ESRI shapefile (.shp) • Keyhole Markup Language (.kml) • Keyhole Markup Language Zipped (.kmz) • Georeferenced TIFF (.tif) • CAD Files (.dwg)	Semistructured
Qualitative data: Textual data.	• Widely used proprietary formats, e.g., Microsoft Word® (.doc, .docx) • Extensible Mark-up Language (xml) • JavaScript Object Notation (json)	Semistructured
Digital media: Image/audio/video data.	• Image formats (.jpg, tif, .png) • Audio formats (.mp3, .aif) • Video formats (.mp4, .mj2)	Unstructured
Document data: Files, reports, and documents.	• Rich Text Format (.rtf) • Portable Document Format (.pdf) • Open Document Text (.odt) • HTML (.htm, .html) • Microsoft Word (.docx)	Unstructured
Binary data: Spectroscopy data and other plots that require the capability of representing contours as well as peak position and intensity.	• Proprietary file formats determined by the software program used for generating the binary data	Unstructured or semistructured

it is wise to plan the file organization at the beginning of the research study.

While defining the file organization, it is important to consider how easy and intuitive it will be for someone else to locate required files from the stored data. Although there is no single, correct method to organize data, there are best practices and guidelines to follow that can improve data management. The following sections provide information on these best practices and guidelines regarding folder structure and naming conventions.

Folder Structure

Defining an optimized folder structure for data storage is important for improving the efficiency of data organization and retrieval processes.
The following guidelines should be considered when creating an effective folder structure for research data:

- Organize data in folders and subfolders based on category and/or type of data.
- Store similar data together.
- Avoid overlapping categories.

- Keep folder sizes manageable.
- Keep folder hierarchy manageable.

Naming Conventions

File names should provide context for the files and distinguish them from other similar files. Many files are used independently of their file or folder structure, so the data organizer must provide a sufficient description in the file name. The following sections provide guidelines for folders and file-naming conventions:

- Be consistent: Keep the naming convention consistent throughout the dataset. Avoid cases where one instance of the same information is stored using two different naming conventions.
- Be concise: Keep file and folder names short and concise.
- Retain the order of information: If the data stored in a folder are associated with a certain period (e.g., date and time), include this information in the file name (e.g., YYYYMMDD, not MMDDYY).
- Be descriptive: Make the name relevant and descriptive so that others can understand the type and nature of the data stored in the corresponding file/folder. Use the following tips for making the names more relevant to the content:
  - Use unique identifiers (i.e., project name or grant number in folder name).
  - Create a project or research data name.
  - Note conditions (e.g., laboratory instrument, solvent, temperature).
  - Insert experiment number (sequential).
  - Note date (in file properties, too).
  - Use application-specific file extension in lowercase (e.g., .txt, .csv, or .xlsx).
  - Using sequential numbering, make sure to use leading zeros to allow for multidigit versions. For example, a sequence of 1–10 should be numbered 01–10; a sequence of 1–100 should be numbered 001–100.
  - Do not use any special characters, such as “&*%#;()!@$^~'{}[]?<>-”.
  - Use only one period before the file extension.

EXAMPLES

This section provides examples to help develop data and metadata for a project. Please visit the FHWA InfoMaterials (2021a) Web portal for organized sample datasets as a reference. In this portal, one can see how different projects' datasets are organized and can export datasets to review example data and metadata organization.

Data

The organization of data primarily depends on the nature of projects. Each project may consist of structured, unstructured, or semistructured data. Based on the project data requirements, the first decision to make is about the database file formats for each of the data types.

Typically, all projects contain data that can be stored in a database such as Microsoft Access or a spreadsheet such as Microsoft Excel. The following sections provide examples of using these database formats and considerations for data storage structure.

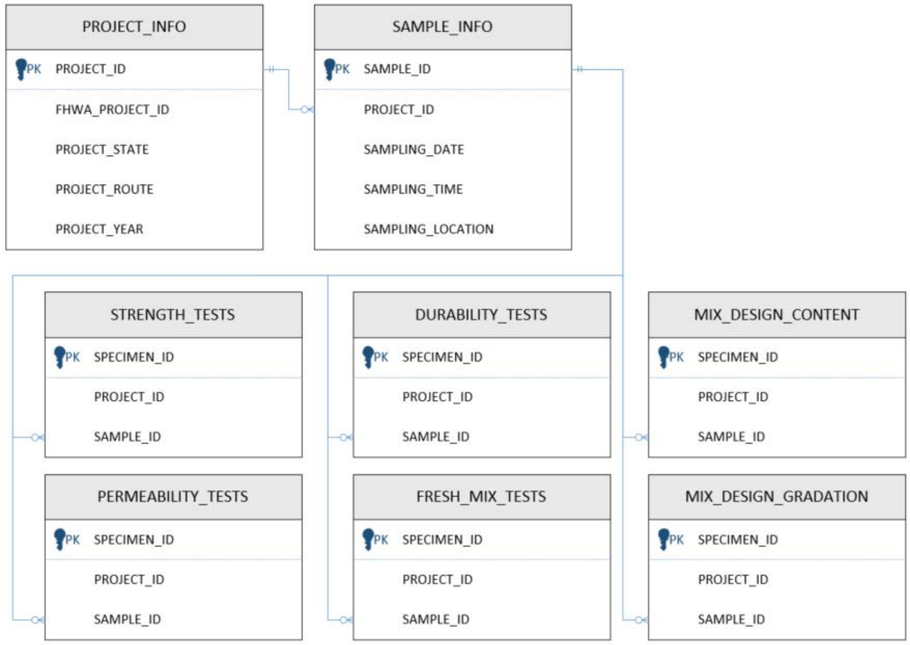
Microsoft Access

A database file in Microsoft Access format contains structured data organized in data tables (also known as entities). Each table consists of multiple data fields that represent the corresponding attributes of the data. Typically, each table contains at least one key field that is used to identify unique records in the table. The data tables are often linked to each other through a relationship. Figure 2 shows an entity relationship diagram for a sample database in Microsoft Access format.

Developing an efficient database design requires some basic skills. For example, Microsoft Office Support (2021) recommends the following high-level steps for getting started with database design:

1. Determine the purpose of the database: Prepare for the remaining steps by solidifying a purpose. This approach will also help with selection of appropriate data types for the data.
2. Gather and organize the required information: Collect all the needed information to store in the database, such as product name, material name, test type, and so on.
3. Distribute the information into tables: Distribute the information items into major entities or subjects, such as products or mixture. Each entity then becomes a table.

Figure 2. Graphic. Sample entity relationship diagram in a Microsoft Access database.



Source: FHWA.

4. Turn information attributes into columns: Decide the information needed to store in each table. Each attribute becomes a field and is displayed as a column in the table. For example, a project table might include fields such as project name, project description, project location, and project date.
5. Specify primary keys: Decide on a primary key for each table. The primary key is a column that is used to uniquely identify each row in a table. An example might be product ID or mixture ID.
6. Specify the table relationships: Review each entity and decide how the data for one entity is related to the data for other entities. Add fields to tables or create new intermediate tables to set up the necessary relationships between different entities. This relationship can be better visualized using an entity relationship diagram.
7. Refine database design: Review the design for any errors or inconsistencies. Create the tables and populate them with sample data to confirm the expected results are drawn from the database. This critical step and the database design should be updated, as needed.
8. Normalize database structure: Database normalization is the process of structuring a database in accordance with a series of normal forms to reduce data redundancy and improve data integrity. Apply the database normalization principles to see if the tables are structured correctly. Make required adjustments as needed. Normalization is conducted to minimize duplicate data, minimize or avoid data modification issues, and simplify queries.

Microsoft Excel

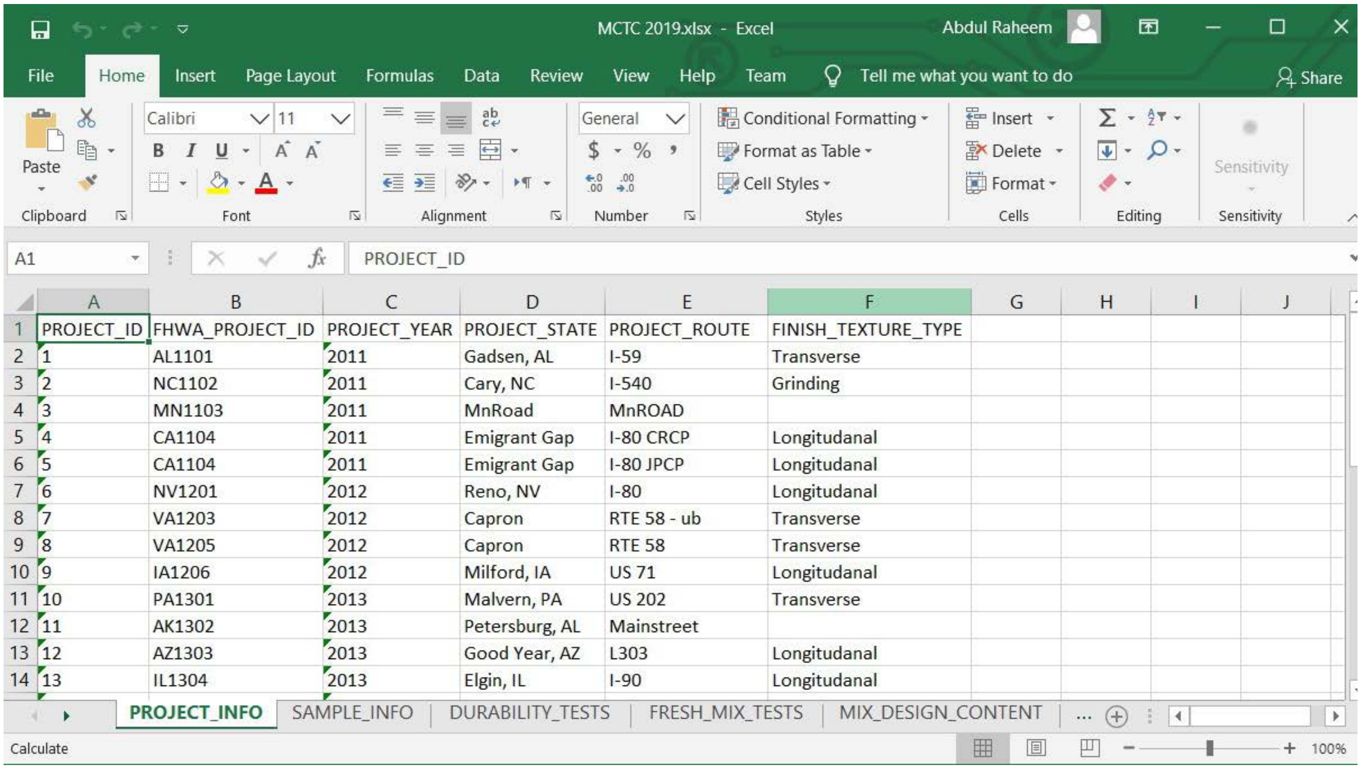
In addition to databases, data can also be stored in Microsoft Excel spreadsheets. The spreadsheets are easy to manage and help simplify the data-entry process during the datacollection phase. Therefore, people often prefer using spreadsheets for collecting data and then migrating the data to the proper database format.

Each sheet (also known as a worksheet) in a Microsoft Excel file represents a data table, whereas the columns in the sheets each represent a field. The top row of each sheet contains the column headers that determine the field names for the corresponding table. Since the spreadsheets do not contain any built-in feature for managing relationship between tables, it is recommended to make the first column of the sheet the key field. This approach is used to identify unique rows in the sheet. Figure 3 shows a sample spreadsheet containing multiple tabs for the respective data tables:

The following steps are recommendations for creating effective spreadsheets:

1. Consider future requirements: Consider as many future factors as possible that might possibly force changes in a worksheet. By preparing for any

Figure 3. Graphic. Sample data in Microsoft Excel format.



	A	B	C	D	E	F
1	PROJECT_ID	FHWA_PROJECT_ID	PROJECT_YEAR	PROJECT_STATE	PROJECT_ROUTE	FINISH_TEXTURE_TYPE
2	1	AL1101	2011	Gadsen, AL	I-59	Transverse
3	2	NC1102	2011	Cary, NC	I-540	Grinding
4	3	MN1103	2011	MnRoad	MnROAD	
5	4	CA1104	2011	Emigrant Gap	I-80 CRCP	Longitudanal
6	5	CA1104	2011	Emigrant Gap	I-80 JPCP	Longitudanal
7	6	NV1201	2012	Reno, NV	I-80	Longitudanal
8	7	VA1203	2012	Capron	RTE 58 - ub	Transverse
9	8	VA1205	2012	Capron	RTE 58	Transverse
10	9	IA1206	2012	Milford, IA	US 71	Longitudanal
11	10	PA1301	2013	Malvern, PA	US 202	Transverse
12	11	AK1302	2013	Petersburg, AL	Mainstreet	
13	12	AZ1303	2013	Good Year, AZ	L303	Longitudanal
14	13	IL1304	2013	Elgin, IL	I-90	Longitudanal

Source: FHWA, created in Microsoft Excel.

future developments, the lifespan of the worksheet is increased.

2. Order worksheets logically: Store different types of data on different sheets. Try to put the high-level summary information in the starting sheets and detail data in subsequent sheets.

3. Organize the information flow: Organize worksheets in such a way that information always flows from the upper left to lower right. This layout makes it considerably easier for a user to navigate the sheet.

4. Label columns and rows: Label table columns and rows with simple names that consistently follow previously used naming conventions. Columns without clear and consistent names might not be problematic for the creator of the worksheet, but to other users, inconsistencies can be confusing. The same applies to matrix tables where rows are also labeled.

5. Do not merge cells: Merging cells is not a good practice. It is usually done to improve aesthetics, but it eventually leads to an increased risk of problems with calculations and references.

6. Avoid hiding data: Hiding columns, rows, or even entire worksheets from view is almost never a good idea. It increases the chance that a user overlooks something important when working with the spreadsheet, increasing the risk for errors. The only exception to this rule would be when it is absolutely necessary to hide information that somehow cannot be put into a separate worksheet.

7. Keep styling consistent: Apply styling consistently throughout the entire spreadsheet. A simple and consistent style for formatting is critical to help viewers understand the spreadsheet. Always include a legend. Abbreviations and colored-cell definitions can be listed on a separate worksheet if needed.

8. Be clear which units are used: Make sure to include unit information wherever applicable.

9. This action can be done by either setting the units in the cell properties within Excel, or identifying them in a column label.

Metadata

A requirement for hosting datasets on InfoMaterials is the provision of metadata. Metadata should be provided at the project and data levels. The metadata contain additional information about tables, fields, codes, and files. The following sections provide examples of metadata.

Projects

The project-level metadata include information about the program/study, protocols/standards, and references used in the data. The basic format of the metadata for a program/study is shown in table 2.

Table 2. Program/study metadata

Item	Description	Required
Agency name	Name of the agency that conducts the project.	Yes
Project/program/ study name	Name of the project, program, or study for which the data are collected.	Yes
Dataset name	Name of the given dataset. The name can be the same as the program/study name. If the dataset is for a specific period, then the dataset name may also include the period or other identifying information.	Yes
Dataset type	A dataset can consist of structured, unstructured, and semistructured data. If the dataset consists of multiple data types, then appropriate data types should be listed.	Yes
Representative name	Name of the individual who is the representative for the given dataset.	Yes
Representative email	Email address of the individual who is the representative for the given dataset.	Yes
Representative phone	Phone number of the individual who is the representative for the given dataset.	Optional

The basic format for the protocols and standards metadata is shown in table 3.

Table 3. Protocols/standards

Item	Description	Required
Designation number	Designation number of the protocol/standard	Yes
Title	Title of the protocol/standard	Yes
Year	Year of the protocol/standardname. If the dataset is for a specific period, then the dataset name may also include the period or other identifying information.	Yes

The basic format for the references metadata is shown in table 4.

Table 4. References

Item	Description	Required
Title	Title of the reference document/report	Yes
Location	Location and/or URL of the reference material	Yes

Tables

Table 5 contains information on the metadata for data tables in InfoMaterials.

Table 5. Table metadata		
Item	**Description**	**Required**
Table name	The exact name of the table as contained in the database.	Yes
Table alias	A user-friendly alternative name for the table. The most common use of this type of field is to provide an intuitive table name when abbreviations are used in the underlying data storage tables.	Yes
Table description	Brief statement of the general contents of the table.	Optional
Class name	Assignment of a classification or module name to a table. Class names are useful in large databases to display related tables in a group of similar tables.	Yes
Table order	Assignment of the display of tables in a predefined sort order. Table order is useful in large databases where data are stored in a hierarchical manner.	Yes
Public access	Yes/no value. Your dataset may contain tables that should not be published publicly. Use this field, to specify if the table will be published publicly or not.	Optional
File name	The name of the file from which the table is imported. This action is applicable if a given table is imported from some existing data file such as Microsoft Excel or .csv file.	Optional
File title	The title of the file from which the table is imported. The title is used to give a user-friendly name for the file.	Optional

Fields

The basic format of the metadata for the data fields is shown in table 6.

Table 6. Field metadata		
Item	**Description**	**Required**
Table name	The name of the table to which the given field belongs.	Yes
Field name	The exact name of the field contained in the table.	Yes
Field alias	A user-friendly alternative name for the field.	Yes
Field order	Assignment of the display of the field in a predefined sort order.	Yes
Field description	Brief statement on the contents of the field.	Optional
Field units	The units of measurement contained within the field. Preferred practice is to use common text fields associated with the units in different standards, as the method to specify the units. Extra attention should be paid to time fields, as these can vary between software platforms. The unit field can be left null for fields that do not represent a measurement with an associated unit.	Yes
Unit system	The unit system (SI or US) used for the measurement. This action is useful in databases where mixed unit systems are used and conversion between the units is required.	Yes
Field code type	This field provides an index for fields populated with a code. Preferred practice for CODETYPE fields is that they be named using text fields that generally describe the type of code. This field is only required if coded fields are contained within the database.	Optional
Precision	Information on the precision value for the number of decimal places.	Yes

Files

The basic format of the metadata for the files is shown in table 7.

Table 7. Files metadata

Item	Description	Required
File name	Name assigned to the file as it exists on a storage medium.	Yes
File title	A user-friendly title given to the file for display on the user interface.	Yes
Description	Brief information on the purpose and contents of the file.	Yes
Class name	Assignment of a classification or module name to a file. Class names are useful in large datasets to display related files in a group of similar files.	Yes
File order	Assignment of the display of files in a predefined sort order. File order is useful in large datasets where files are stored in a hierarchical manner.	Optional

Codes

The basic format and contents of the code type table is shown in table 8.

Table 8. Code type metadata

Item	Description	Required
Code type	Name assigned to a common set of codes	Yes
Code description	Description of the type of codes associated with the specified code type	Optional
Source	Text reference to other documents where more information on where the codes used for this code type are further defined	No

The codes table provides a translation of the meaning of the different code values based on code type, and in some cases, other information. For example, the use of old Federal Information Processing Standard (FIPS) codes for counties requires additional knowledge of the State since all States have a county code of 1. In these cases, there is a requirement for the codes table to have additional fields to properly relate back to data tables. The basic format and contents of the codes table include the following (table 9):

Table 9. Codes metadata

Item	Description	Required
Code type	Name assigned to a common set of codes.	Yes
Code	Index-associated description of the type of codes associated with the specified code type.	Yes
Code title	Title for the code.	Yes
Detail	Description that defines the code.	Optional
Additional code	This code value is needed to properly interpret the relationship between multiple code types, such as State and county codes using the FIPS system.	Optional
Additional code type	Code type of the field entered in the additional code field.	Optional

RESOURCES

The following resource documents can be useful in providing data for public dissemination:

- *Department of Transportation Information Dissemination Quality Guidelines* (U.S. Department of Transportation 2019).
- *The FHWA Data Governance Plan*. While this plan is still in development, some of the essential features can be found in this series of documents (Federal Highway Administration 2021b).
- National Information Standards Organization (NISO), *Understanding Metadata*. This document provides a primer on the basics of modern metadata standards (NISO 2017).
- *FHWA InfoMaterials Brochure*. This document provides a brief overview of the FHWA InfoMaterials website (Federal Highway Administration 2021c).

REFERENCES

1. Federal Highway Administration. 2021a. "FHWA InfoMaterials" (website). McLean, VA. Available online: https://infopave.fhwa.dot.gov/InfoMaterials, last accessed April 1, 2021.
2. Federal Highway Administration. 2021b. "FHWA Data Governance" (website). Washington DC. Available online: https://www.fhwa.dot.gov/datagov/, last accessed April 1, 2021.
3. Federal Highway Administration. 2021c. *FHWA InfoMaterials*. Report No. FHWA-HRT-21-018. Washington, DC: Federal Highway Administration.
4. Microsoft Office Support. 2021. "Database Design Basics." (website). Redmond, WA, Available online: https://support.microsoft.com/en-us/office/database-design-basics-eb2159cf-1e30-401a-8084-bd4f9c9ca1f5, last accessed April 7, 2021.
5. NISO, 2017. *Understanding Metadata*. Baltimore, MD: National Information Standards Organization.
6. United States Department of Transportation. 2019. *The Department of Transportation's Information Dissemination Quality Guidelines*. Report No. DOT-OST-2019-0135. Washington DC: U.S. Department of Transportation.

Researchers—This study was conducted under contract DTFH-15-D-00004 by Wood Environment & Infrastructure Solutions, Inc. and iENGINEERING, Inc.

Distribution—This TechBrief is being distributed according to a standard distribution. Direct distribution is being made to the FHWA divisions and Resource Center.

Availability—This TechBrief may be obtained at https://highways.dot.gov/research.

Key Words—data, data governance, data management, metadata, databases, spreadsheets, dataset.

JULY 2021

Recommended citation: Federal Highway Administration, FHWA InfoMaterials™ Datasets Management (Washington, DC: 2021) https://doi.org/10.21949/1521645

FHWA-HRT-21-065
HRDI-30/07-21(WEB)E