NIST Internal Report
NIST IR 8496 ipd

Data Classification Concepts and Considerations for Improving Data Protection

Initial Public Draft

William Newhouse
Murugiah Souppaya
John Kent
Ken Sandlin
Karen Scarfone

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8496.ipd

NIST
NATIONAL INSTITUTE OF
STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

NIST Internal Report
NIST IR 8496 ipd

Data Classification Concepts and Considerations for Improving Data Protection

Initial Public Draft

William Newhouse
Murugiah Souppaya
National Cybersecurity Center of Excellence
Information Technology Laboratory

John Kent
Ken Sandlin
The MITRE Corporation

Karen Scarfone
Scarfone Cybersecurity

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8496.ipd

November 2023

U.S. Department of Commerce
Gina M. Raimondo, Secretary

National Institute of Standards and Technology
Laurie E. Locascio, NIST Director and Under Secretary of Commerce for Standards and Technology

Certain commercial equipment, instruments, software, or materials, commercial or non-commercial, are identified in this paper in order to specify the experimental procedure adequately. Such identification does not imply recommendation or endorsement of any product or service by NIST, nor does it imply that the materials or equipment identified are necessarily the best available for the purpose.

There may be references in this publication to other publications currently under development by NIST in accordance with its assigned statutory responsibilities. The information in this publication, including concepts and methodologies, may be used by federal agencies even before the completion of such companion publications. Thus, until each publication is completed, current requirements, guidelines, and procedures, where they exist, remain operative. For planning and transition purposes, federal agencies may wish to closely follow the development of these new publications by NIST.

Organizations are encouraged to review all draft publications during public comment periods and provide feedback to NIST. Many NIST cybersecurity publications, other than the ones noted above, are available at https://csrc.nist.gov/publications.

NIST Technical Series Policies

How to Cite this NIST Technical Series Publication:

Newhouse W, Souppaya M, Kent J, Sandlin K, Scarfone K (2023) Data Classification Concepts and Considerations for Improving Data Protection. (National Institute of Standards and Technology, Gaithersburg, MD), NIST Interagency Report (IR) 8496. https://doi.org/10.6028/NIST.IR.8496.ipd

Author ORCID iDs

William Newhouse: 0000-0002-4873-7648
Murugiah Souppaya: 0000-0002-8055-8527
John Kent: 0009-0001-9989-2277
Ken Sandlin: 0009-0000-3757-4858
Karen Scarfone: 0000-0001-6334-9486

Public Comment Period

November 15, 2023 – January 9, 2024

Submit Comments

data-nccoe@nist.gov

National Institute of Standards and Technology
Attn: Applied Cybersecurity Division, Information Technology Laboratory
100 Bureau Drive (Mail Stop 2000) Gaithersburg, MD 20899-2000

All comments are subject to release under the Freedom of Information Act (FOIA).

Abstract

Data classification is the process an organization uses to characterize its data assets using persistent labels so those assets can be managed properly. Data classification is vital for protecting an organization's data at scale because it enables the application of cybersecurity and privacy protection requirements to the organization's data assets. This publication defines basic terminology and explains fundamental concepts in data classification so there is a common language for all to use. It can also help organizations improve the quality and efficiency of their data protection approaches by becoming more aware of data classification considerations and taking them into account in business and mission use cases, such as secure data sharing, compliance reporting and monitoring, zero-trust architecture, and large language models.

Keywords

data classification; data governance; data labeling; data management; data privacy; data protection; data security.

Reports on Computer Systems Technology

The Information Technology Laboratory (ITL) at the National Institute of Standards and Technology (NIST) promotes the U.S. economy and public welfare by providing technical leadership for the Nation's measurement and standards infrastructure. ITL develops tests, test methods, reference data, proof of concept implementations, and technical analyses to advance the development and productive use of information technology. ITL's responsibilities include the development of management, administrative, technical, and physical standards and guidelines for the cost-effective security and privacy of other than national security-related information in federal information systems.

Audience

The audiences for this publication include managers and executives with responsibilities related to data classification or data protection, organization policy makers, cybersecurity and privacy product and service vendors, cybersecurity and privacy professionals responsible for managing data across the organization, compliance professionals, and legal professionals.

Acknowledgments

The authors thank everyone who has reviewed preliminary drafts of this document and contributed to its technical content.

Note to Reviewers

NIST welcomes public comments on any aspect of this publication. Existing data classification practitioners are particularly encouraged to share their insights on how closely the definitions and concepts in this publication correspond with their own experience. NIST is also seeking responses to the following questions:

1. The document currently lists data protection as one of the major elements of data management. Do you agree with that, or do you feel that data protection is at a lower level of abstraction? For example, data protection could be considered one component of data usage, which then would be considered one of the data management elements.
2. Should Section 2.3, Data Governance and Data Management, provide more information about data policy? If so, what should be added?

Trademark Information

All registered trademarks or trademarks belong to their respective organizations.

Call for Patent Claims

This public review includes a call for information on essential patent claims (claims whose use would be required for compliance with the guidance or requirements in this Information Technology Laboratory (ITL) draft publication). Such guidance and/or requirements may be directly stated in this ITL Publication or by reference to another publication. This call also includes disclosure, where known, of the existence of pending U.S. or foreign patent applications relating to this ITL draft publication and of any relevant unexpired U.S. or foreign patents.

ITL may require from the patent holder, or a party authorized to make assurances on its behalf, in written or electronic form, either:

a) assurance in the form of a general disclaimer to the effect that such party does not hold and does not currently intend holding any essential patent claim(s); or

b) assurance that a license to such essential patent claim(s) will be made available to applicants desiring to utilize the license for the purpose of complying with the guidance or requirements in this ITL draft publication either:

    i. under reasonable terms and conditions that are demonstrably free of any unfair discrimination; or

    ii. without compensation and under reasonable terms and conditions that are demonstrably free of any unfair discrimination.

Such assurance shall indicate that the patent holder (or third party authorized to make assurances on its behalf) will include in any documents transferring ownership of patents subject to the assurance, provisions sufficient to ensure that the commitments in the assurance are binding on the transferee, and that the transferee will similarly include appropriate provisions in the event of future transfers with the goal of binding each successor-in-interest.

The assurance shall also indicate that it is intended to be binding on successors-in-interest regardless of whether such provisions are included in the relevant transfer documents.

Such statements should be addressed to: data-nccoe@nist.gov

Table of Contents

1. Introduction

Data are “a representation of information, including digital and non-digital formats.” [NISTPF] A *data asset* is “an information-based resource” such as a database, document, webpage, or service. [CNSSI4009] This publication uses the term “data asset” throughout to indicate the relative importance of specific data resources, as opposed to data in general. *Metadata* are information regarding the context of a specific data asset, like who or what created the data asset (i.e., *data provenance*) and when and where the data asset was collected.

Data classification is the process an organization uses to characterize its data assets using persistent labels so those assets can be managed properly. Examples of possible data classifications include “protected health information (PHI),” “personally identifiable information (PII),” and “financial records.” Applying data classification practices can benefit organizations in:

- enabling application of cybersecurity and privacy protection requirements to the organization’s data assets;
- securely sharing data assets with partners, contractors, and other organizations;
- knowing which requirements from laws, regulations, contracts, and other sources apply to a particular data asset;
- maintaining awareness of data assets and the criticality of each asset, which supports implementation of zero-trust architectures and other cybersecurity and privacy technologies;
- enforcing restrictions on access to and transfer of an organization’s intellectual property;
- capturing metadata about the source of data assets consumed by generative artificial intelligence (AI) technologies (e.g., large language models [LLMs]); and
- identifying and recording metadata for data assets when that metadata might not be needed today but will be needed in the future; an example is for post-quantum readiness and migration planning.

1.1. Purpose and Scope

This publication has two purposes. First, it defines basic terminology and explains fundamental concepts in data classification so there is a common language for all to use, thus alleviating existing confusion and ambiguity regarding what particular terms mean. Second, this publication can help organizations improve the quality and efficiency of their data protection approaches by becoming more aware of data classification considerations and taking them into account in business and mission use cases.

This publication’s terms and concepts will be used throughout the NCCoE’s Special Publication (SP) 1800-39, *Implementing Data Classification Practices* series of practice guides [SP1800-39], and will also be used by other NIST efforts, including the NCCoE’s Data Security and Zero Trust Architecture projects. This publication also may inform future versions of NIST SP 800-60, *Guide for Mapping Types of Information and Information Systems to Security Categories*

[SP800-60], as well as help organizations with adopting the NIST Cybersecurity Framework [NISTCSF], the NIST Privacy Framework [NISTPF], and other NIST frameworks and guidance.

The scope of this publication is data classification considerations to enable data protection. Details of how technologies enforce data protection requirements are out of scope for this publication.

This publication applies to any data classifications and data classification schemes that organizations may use, not just those used by the U.S. government or military.

1.2. Publication Structure

The rest of this publication is comprised of the following sections and appendices:

- Section 2 provides background information on the data lifecycle, data governance and management, and types of data.
- Section 3 describes the primary practices involved in data classification and discusses considerations that organizations should take into account for their data classification practices.
- The References section lists the references cited throughout the publication.
- Appendix A lists the acronyms used in the publication.
- Appendix B provides a glossary with definitions of selected terms from the publication.

2. Background

This section defines basic terminology and explains fundamental concepts from data governance and data management as background for understanding the data classification practices and considerations explained in Section 3.

2.1. Data Lifecycle

An organization manages its data assets through the data lifecycle. There are many valid data lifecycles that originate from different technical practices. This publication describes a simple lifecycle that focuses on those high-level phases important to data classification: Identify, Use, Maintain, and Dispose. Not all data lifecycle phases occur for every data asset.

- **Identify**: The organization identifies data assets. Section 3.2 contains more information on methods for identifying data assets.
- **Use:** The organization accesses, views, shares, and modifies part or all of a data asset. As part of use, new data assets may be created by the aggregation (multiple assets joined into one) or disaggregation (one data asset broken into multiple assets) of existing data assets. Data assets may also be repurposed (i.e., used for a different reason or in a different way than originally intended).

- **Maintain**: The organization preserves data assets over time. This may include converting a data asset to a different format or representation as technologies change so it will continue to be usable.
- **Dispose**: The organization disposes of data assets at the end of the data lifecycle. Data assets that are no longer needed are destroyed or otherwise disposed of to free resources and to prevent data from being accessed by unauthorized parties—for example, when storage media is disposed of.

2.2. Structured, Unstructured, and Semi-Structured Data

How a data asset is represented can be described in three broad categories: structured, semi-structured, and unstructured. Each of these terms describes the degree to which a data asset conforms to a logical or physical *data model*—a specification for the elements of data contained within a data asset—within the context of a particular business domain.

- *Structured data* follow a physical data model that describes in detail how the data are to be represented and how a representation should be interpreted. Structured data may be found in a database or other mechanism that clearly indicates what type of information each data field contains, like customer ID or part number. Structured data can be validated against the data model to ensure their meaningfulness.
- *Semi-structured data* describe their own data model (self-describing). Semi-structured data are expressed in formats like the Extensible Markup Language (XML) and JavaScript Object Notation (JSON) for sharing proprietary data sets, sensitive configurations parameters, and other information.
- *Unstructured data* do not follow a detailed data model that is meaningful to a business domain. Examples include documents and videos. Unstructured data might be stored in a specific format, such as a proprietary document format or a standards-based video format. For example, a video could show a patient's medical procedure, people entering and exiting a facility, or a training course for new employees. A document with unstructured data not only could contain nearly any type of information, but it may also have other types of data embedded within it, such as graphics, videos, and other documents, each containing one or more other instances of data.

2.3. Data Governance and Data Management

Data governance encompasses the actions an organization needs to perform to ensure that its data assets are managed properly. Aspects of data classification that are particularly important for data governance are defining the organization's data classification policies and related data protection requirements, and determining how those policies should be implemented and enforced, including roles and responsibilities both within the organization and outside the organization.

Data management is the implementation and enforcement of the policies and practices resulting from data governance. Data management should occur for all data assets throughout the data lifecycle. Metadata are a form of data, so metadata also need to be managed. Although explaining data management in detail is outside the scope of this publication, some basic understanding of the following areas of data management is necessary in order to understand data classification's role as part of data management:

- **Data definition:** In order to manage a data asset, an organization first needs to define it. *Data definition* varies by data asset, but it usually includes identifying the applicable data type and data model, as well as collecting metadata regarding the origin, nature, purpose, and quality of the data asset *(data cataloging)*. Data definition strives to gather sufficient information about a data asset so that the organization can ascertain its data classifications. The formality and rigor of data definition varies greatly among data assets, but it is typically related to whether the data asset is structured, semi-structured, or unstructured.
- **Data classification:** The data classifications for a data asset are selected and assigned based on one or more of the following: its data definition, its catalogued metadata, and review or analysis of its contents. Section 3 discusses this topic in detail.
- **Data protection:** Once data classifications are assigned, the organization needs to enforce the *data protection* requirements associated with each of those classifications. These encompass all of the controls needed to protect each data asset in accordance with its classifications. An example is a data classification associated with requirements to encrypt the data asset when at rest or in transit, use a data integrity mechanism to detect tampering, allow access by members of a particular group only, and retain the data asset for at least two years from the date it was acquired.
- **Data monitoring:** *Data monitoring* is needed to identify any changes to the data definition or the data asset itself that might necessitate changes to data classifications and/or data protection. Data monitoring can also identify lessons learned from real-world data classification and protection experiences that may improve data management.

3. Data Classification Functions

The process of data classification includes the following functions:

1. Define the organization's data classification policy, which is the taxonomy of data asset types and the rules for identifying data assets of each type.
2. Identify the organization's data assets to be classified.
3. Analyze the data assets and determine the appropriate data classifications for each.
4. Associate data classification labels with each data asset. (Once labels are assigned, the applicable cybersecurity and privacy requirements can be enforced for each data asset.)
5. Monitor each data asset for changes that may necessitate updating its data classifications and/or the data classification policy.

This section provides more information on each of the functions, including considerations that organizations may choose to adopt. Taking these considerations into account can help organizations improve the quality and efficiency of their data classification implementations, which can have positive impacts throughout the data lifecycle.

3.1. Defining the Data Classification Policy

A *data classification scheme* is a taxonomy of all of an organization's known data asset types. For example, part of a classification scheme might involve data classifications that characterize high-level business data types of a data asset—for instance, "vendor invoices," "customer invoices," "employee records," etc. Data assets may also be classified based on source information, like "internally created," "licensed data," or "acquired data." Additional data classifications could include geopolitical information about the data asset, e.g., "US person" or "EU entity". With those three independent classifications applied to a data asset, the organization can then protect the data asset according to the requirements corresponding to its business data type, source, and geopolitical origin. When data are shared from one organization to another organization, the two organizations' data classification schemes may need to be mapped to a common, shared taxonomy.

A *data classification policy* is comprised of the data classification scheme and the formal description of the data types within an organization. It is used to enable identification of data types from a data asset. Classification policies can be expressed as digital policies to enable automated classification determinations. Organizations should define their data classification policies in such a way that all affected parties, including external parties who share or receive data assets, have a common understanding of them. Any ambiguity in these policies may cause errors and inconsistency in how data are classified and protected, which could increase the risk of compromises and compliance violations.

The data classification scheme and policy do not directly indicate how the data assets must be protected; instead, each data classification is linked to a set of associated data protection requirements. Generally, a data asset must be protected in accordance with the consolidated requirements of all of its data classifications.

The specificity of a data classification scheme will determine the nuance afforded to developing data protection policies. For instance, classifying a data asset only as "sensitive data" typically does not provide enough information to identify all the data protection requirements for that data asset, since many types of data are considered sensitive. Classifying a data asset as "PHI" instead of "sensitive data" enables more fine-grained protection policies, such as preventing certain types of PHI from being sent to certain business partners. However, more specificity in the data classification scheme can make the process of classifying new or modified data more difficult or costly. An organization should balance the effort and costs of analyzing its data to determine classifications against the versatility it requires for protecting various types of data assets.

In most situations, three groups of people need to work together to ensure the data assets are properly protected:

- The data asset's business owner understands the origin, nature, and purpose of the data asset and its importance to the organization's mission. The business owner is key for determining the data classifications.
- The compliance staff understands the legal and regulatory requirements for protecting data assets associated with each of the organization's data classifications. Compliance staff also perform auditing and reporting to ensure and document adherence to those requirements.

- The technology owners understand the technology that houses, interacts with, and safeguards the data asset throughout the data lifecycle. Cybersecurity and privacy professionals, system administrators, and others acting on behalf of technology owners are responsible for implementation and enforcement of the requirements for protecting data assets based on the assets' data classifications.

Cybersecurity, privacy, compliance, and business requirements should all be addressed holistically in the data classification definitions and policies. Personnel from each of these areas should be involved in developing, reviewing, and updating the definitions and policies.

Generally, data classifications and classification schemes should be defined separately from data protection requirements. The protection requirements for any particular data asset are highly likely to change over time, while the data classifications themselves tend to be static. For example, the text of laws defining what PHI is does not change, but the technologies that house PHI and the cybersecurity and privacy controls that protect PHI may change over time.

Data classification policies should be monitored and auditable, and changes to the policies should be controlled to prevent unauthorized changes to the data classification definitions or assignments. Access, especially modifications, to policy stores should be logged so organizations can verify and validate the effective state of their data classification processes at any time. Also, the data classification policies and protection requirements should each be versioned. Over time, version information will allow individuals and automated systems to quickly and reliably identify stale or obsolete classification information and take appropriate actions such as flagging the discrepancy or requesting updated information.

3.2. Identifying Data Assets to Classify

A data asset is identified as needing classification when activities such as the following take place:

- **Creating:** Data assets are identified as part of their creation process. Examples include an employee entering a customer's personal information into an application, a process automatically producing new data by analyzing existing data, or a sensor capturing measurements of environmental characteristics (e.g., temperature).
- **Discovering:** Existing data assets within an organization that have not been classified are located. Discovery searches an organization's technology assets such as desktop workstations, servers, and cloud services for data. An example is an employee having written a new ad hoc document.
- **Importing:** An external organization's data assets are identified within the organization. It is responsible for ensuring an organization's commitments for managing and protecting data assets belonging to external organizations are met. An example is a business partner providing a copy of one of its databases for the organization to use.

An organization's business processes should take all these means into account so that all data assets are classified promptly and appropriately.

Data assets should be classified as close to the time of their creation, discovery, or importation as possible. One reason for this is to support properly protecting the data as soon as possible. Another reason is that capturing the original metadata for a data asset may be particularly helpful in providing context and transparency vital for assigning data classifications. The later the metadata are collected, the less helpful they will generally be for data classification purposes, both now and in the future. For example, a new classification need, like a new regulation or a change to an existing regulation, may require analyzing existing data assets to determine if the new data classification applies to them. Having more metadata on hand may make this analysis easier and more accurate.

When data assets are identified, an organization may need to revise its data classification policy to fully address the assets. Even information of the same type that is found may be structured differently in newly found data sets. The tools used to analyze and label data assets may also need to be updated to properly classify these data assets in the future.

Data assets imported from another organization should usually be re-classified, even if that organization provided their classification information. The data may have been misclassified by that organization, or your organization may be subject to additional requirements. The act of sharing the data may itself introduce additional requirements. At this time, many industries lack standards for classifying data cross-organization or cross-sector. Moreover, there is limited interoperability among technologies for data classifications. These limitations alone are likely to necessitate the re-classification of imported data so that the organization can ensure the appropriate protection of received data.

When possible, the original classification information from the originating organization should be preserved. To disambiguate external data classifications, their identifiers and labels should be prefixed with a scope that identifies the origin of the classification. This could simply be the name of the organization providing the data asset, or it could refer to an external standards organization if or when such standards exist. For data imported from other organizations, this allows maintenance of the original classification information in addition to labeling the data with the importing organization's classifications.

3.3. Determining Data Classifications for Data Assets

Classifying data is the process of analyzing a data asset and determining which data classifications to assign to it. Classification is performed by a *classifier,* a person or technology that applies the organization's classification policy to a data asset to determine what data classifications that asset should be assigned. For some types of data, data classification can be solely based on the data definition and thus fully automated, but more often—especially for unstructured data—classifying data involves additional analysis of the metadata and/or the data itself. Responsibilities for data classification decisions are sometimes assigned to end users, like requiring them to manually determine the classifications for the documents they create.

Highly controlled structured data, like a set of databases being created for use within a major enterprise application, normally have well-defined fields and extensively validate data values to ensure they comply with the data model. The field for a person's first name could not contain a driver's license number, birthdate, or other unexpected information. Data classifications would be identified as part of the data model's creation, recorded in the databases, and enforced by the enterprise application and its supporting platforms.

While their flexibility may present some challenges, semi-structured data may provide some of the context necessary for classification through its self-described data model.

Unstructured data, where the data model is informal or nonexistent—such as a new text document—present the greatest challenge to data classification. Most organizations will need to use a combination of approaches such as the following for classifying their unstructured data:

- **Automatically select classifications based on metadata analysis.** Ideally, data classifications can be derived from existing metadata such as filename, file extension, author, creation date, and location. Metadata can act as a proxy for specific characteristics of the data that drive classification, but their accuracy as a proxy will vary. For instance, if existing business processes and systems adequately control where data are stored, and storage is compartmented such that data's inherent attributes dictate its storage location, then location would be an accurate proxy when selecting location-specific data classifications. Conversely, if the location of the data is a shared document folder with few controls and broad access, its location would not reflect its inherent attributes and would not be a valid proxy for data classifications.
- **Automatically select classifications based on content (data) analysis.** Deriving data classifications from the contents of the data may provide the most accurate results when there is no enforced data model. However, especially with unstructured data, it can be difficult to correctly interpret the significance of its contents. Technologies like optical character recognition (OCR) can assist in locating content in files. Examples of content analysis tools for data classification purposes include:
    - *Token-based analytical approaches* scan the data looking for the presence and count of specific tokens (i.e., keywords) within the data. These tools are simple to understand and use, but they are limited in determining how each token is used and may be ineffective for many classification schemes.
    - *Regular expression matching tools* allow for more sophisticated matching of strings within the text compared to token-based analytics, including patterns such as telephone numbers, social security numbers, credit card numbers, physical addresses, and email addresses. These tools can be used to identify more complex patterns in the data that are necessary to support more nuanced classification schemes.
    - *Machine learning (ML) tools* can be used to look for the patterns in the data that indicate the attributes that drive classification. In this approach, a set of example data is classified, and then one or more models are trained to analyze and classify the data. This approach appears to be the most capable means of deriving classifications for data automatically but can be complex to establish and manage. The data sets used for training the model(s) must be a comprehensive corpus of data that provides sufficient information for each classification to be detected.
- **Manually select classifications.** Automatic classification may not be feasible for all instances of data, especially ad hoc instances. In these cases, manual classification performed by a human is necessary. Unfortunately, manual classification is usually difficult to implement consistently at scale, and it relies on the accuracy and understanding of each person performing classification.

3.4. Labeling Data Assets

A *label* is a metadata attribute that represents a data classification. A data asset may have more than one label. *Labeling* is the process by which the labels are associated with a data asset, such as by cryptographic binding or by associating the data asset and its labels in a data catalog.

Note that while some people consider the term “label” to be synonymous with the term “tag,” others do not. Also, “label” is increasingly being used as the primary term for this concept, so this publication only uses “label” for consistency.

Data classification assignments, including labels and metadata used for data classification purposes, need to be safeguarded. Without adequate protection, labels and metadata can be altered or deleted. When data or data classifications change, the data’s labels and metadata may need to be updated in a controlled fashion. This is especially true if data are aggregated.

Making data labels “stick” with data as it moves from place to place, and especially from one organization to another, is one of the largest challenges in data classification for most organizations. There are additional challenges involving *portion marking*, when different portions of a data asset, such as sections of a document or file, each have different classification labels. Numerous technological approaches to labeling are currently in use, but no approach works universally across data assets, technologies, and organizations. Further discussion of labeling technologies is outside the scope of this document.

3.5. Monitoring Data Assets

Each data asset should be monitored after its data classification and labeling to identify any changes that may necessitate updating its data classifications and labels. The appropriate monitoring method will depend primarily on whether the data are structured, semi-structured, or unstructured. For example, changes to the nature of structured and semi-structured data are most likely detectable by monitoring their data models for changes to the data definition. However, changes to the content of unstructured data, especially ad hoc files, may be happening all the time, and many of those changes will not affect data classifications.

Further discussion of technologies and methodologies for monitoring data assets for changes impacting their data classifications is outside the scope of this publication. Please refer to the NCCoE’s SP 1800-39, *Implementing Data Classification Practices* series of practice guides [SP1800-39].

References

[CNSSI4009] Committee on National Security Systems (2022) Committee on National Security Systems (CNSS) Glossary. (National Security Agency, Ft. Meade, MD), CNSS Instruction (CNSSI) No. 4009. Available at https://www.cnss.gov/CNSS/issuances/Instructions.cfm

[NISTCSF] National Institute of Standards and Technology (2018) Framework for Improving Critical Infrastructure Cybersecurity, Version 1.1. (National Institute of Standards and Technology, Gaithersburg, MD), NIST Cybersecurity White Paper (CSWP) NIST CSWP 6. https://doi.org/10.6028/NIST.CSWP.6

[NISTPF] National Institute of Standards and Technology (2020) NIST Privacy Framework: A Tool for Improving Privacy Through Enterprise Risk Management, Version 1.0. (National Institute of Standards and Technology, Gaithersburg, MD), NIST Cybersecurity White Paper (CSWP) NIST CSWP 10. https://doi.org/10.6028/NIST.CSWP.10

[SP1800-39] Newhouse W, Souppaya M, Kent J, Sandlin K, Scarfone K (2023) Implementing Data Classification Practices. (National Institute of Standards and Technology, Gaithersburg, MD), NIST Special Publication (SP) 1800-39A. Available at https://www.nccoe.nist.gov/data-classification

[SP800-60] Stine KM, Kissel RL, Barker WC, Fahlsing J, Gulick J (2008) Guide for Mapping Types of Information and Information Systems to Security Categories. (National Institute of Standards and Technology, Gaithersburg, MD), NIST Special Publication (SP) 800-60, Vol. 1, Rev. 1. https://doi.org/10.6028/NIST.SP.800-60v1r1

Appendix A. List of Symbols, Abbreviations, and Acronyms

AI
artificial intelligence

JSON
JavaScript Object Notation

LLM
large language model

ML
machine learning

NCCoE
National Cybersecurity Center of Excellence

OCR
Optical Character Recognition

PHI
protected health information

PII
personally identifiable information

SP
Special Publication

XML
Extensible Markup Language

Appendix B. Glossary

classifier
A person or technology that applies the organization's data classification policy to a data asset to determine what data classifications that asset should be assigned.

data
A representation of information, including digital and non-digital formats. [NISTPF]

data asset
An information-based resource. [CNSSI4009]

data cataloging
Collecting metadata regarding the origin, nature, purpose, and quality of a data asset.

data classification
The process an organization uses to characterize its data assets using persistent labels so those assets can be managed properly.

data classification policy
An organization's data classification scheme and the formal description of the data types within the organization.

data classification scheme
A taxonomy of all of an organization's known data asset types.

data definition
Identifying a data asset's data type and cataloging the data.

data governance
The actions an organization needs to perform to ensure that its data assets are managed properly.

data management
The implementation and enforcement of the policies and practices resulting from data governance.

data model
A specification for the elements of data contained within a data asset.

data monitoring
Identifying any changes to a data asset's data definition or the data asset itself that might necessitate changes to data classifications and/or data protection.

data protection
The controls needed to protect a data asset in accordance with its data classifications.

data provenance
Who or what created a data asset.

label
A metadata attribute that represents a data classification.

labeling
The process by which labels are associated with a data asset.

metadata
Information regarding the context of a specific data asset.

semi-structured data
Data that describe their own data model.

structured data
Data that follow a physical data model that describes in detail how the data are to be represented and how a representation should be interpreted.

unstructured data
Data that do not follow a detailed data model that is meaningful to a business domain.