Transforming How Texas Government Serves Texans

Classifying Data with Sensitivity Labels

Texas Department of Information Resources

June 2023

Authors

Cristina Blanton, University of Texas System

Pooja Bahadur, Texas Health and Human Services Commission

Patrick McSorley, Texas Commission on Environmental Quality

Grant Cardwell, Texas Commission on Environmental Quality

Suresh Sundararajan, Texas Department of Transportation

Monica Smoot, Texas Department of Information Resources

Table of Contents

Introduction 1
Guiding Principles for Data Sensitivity Labeling 3
Data Classification Types and Terms 4
Regulated Versus Unregulated Sensitive Data 4
Determining the Sensitivity of Unregulated Data 4
Data Classification Terms 5
Data Classification Framework and Policies 5
Data Discovery 6
Data Classification and Labeling – Getting Started 7
Implementing a Sensitivity Labeling Policy 8
What Can Sensitivity Labels Help With? 9
Key Takeaways 10
References 10

Introduction

Designated Data Management Officers, as described in Texas Government Code Section 2054.137, shall coordinate with the agency's information security officer to implement best practices for managing and securing data in accordance with state privacy laws and data privacy classifications.

All day-to-day activities in an organization deal with various aspects of business functions that engage with data. There are many occasions where business needs require handling data of varied importance, from open or public data to highly classified data.

Data classification ensures that organizations comply with regulations and develop data-centric security that can be applied across all levels of an organization. It helps an organization prioritize data protection efforts, thereby improving data security. Classification and proper labeling of data helps reduce costs, increase productivity, and retire data that is no longer needed.

Without the right labeling, there is a higher probability that sensitive data will remain in networks and databases without proper knowledge of them. Implementing data sensitivity labels and data security policies can help organizations identify the level of security and privacy protection that should be applied to enforce the appropriate access controls. Implementing sensitivity labeling in an organization will:

Increase awareness – Data classification and sensitivity labeling enable users to be aware of the types of information they handle and the value of the data. It helps users recognize their obligations in protecting data to prevent data loss or sensitive information leaks, which could lead to regulatory action or reducing the trust of stakeholders, including the public.

Improve compliance – Texas Administrative Code requires state agencies and institutions of higher education to protect specific types of data, such as the personal data of individuals, medical records, financial information, and credit card information, just to name a few. Data classification through sensitivity labeling allows us to identify data subject to specific regulations, so that required controls can be applied to pass audits. While not an exhaustive list, some regulations and guidelines that Texas organizations must comply with are:

- The Health Insurance Portability and Accountability Act of 1996 (HIPAA) is a federal law that required the creation of national standards to protect sensitive patient health information from being disclosed without the patient's consent or knowledge.
- The Family Educational Rights and Privacy Act (FERPA) is a federal law that protects the privacy of student education records and applies to all schools that receive funds under any program administered by the U.S. Department of Education.

- The Payment Card Industry Data Security Standard (PCI DSS) is a global standard that provides a baseline of technical and operational requirements designated to protect payment data.
- The National Institution for Standardization and Technology (NIST) Special Publication 800 series provides guidelines, recommendations, and technical specifications to address and support security and privacy needs to meet statutory responsibilities defined under the Federal Information Security Modernization Act (FISMA). For example, NIST SP 800-53 provides a catalog of security and privacy controls for information systems and organizations.

Data classification must comply with relevant regulatory and industry-specific mandates, which may require classification of different data attributes. For example, the Cloud Security Alliance (CSA) requires that data and data objects include the data type, jurisdiction of origin and domicile, context, legal constraints, and sensitivity.

There are other laws and regulations that may impact an organization's data and should be considered. In the United States, certain classes of information are deemed as sensitive, where laws and regulations protect data based on its impact to organizations and the public. Legislative definitions of personal information have broadened over time, led primarily by the state of California. In other countries and political and economic unions such as the European Union, data protection laws tend to be more comprehensive, such as the General Data Protection Regulation that was passed into law and went into effect in 2018.

The General Data Protection Regulation (GDPR) and, most recently, the California Privacy Rights Act (CPRA) demonstrate the recent increase in regulatory commitment to the protection, use, and transmission of confidential and sensitive data. These regulations are considered the gold standard as states work to implement controls to protect the rights of their citizens. In Texas, the 87th Legislature passed Senate Bill (SB) 475 with specific mandates that increased awareness and focus on the state's data management practices and structures. These laws strengthen organizations' and public entities' expectations in the protection of customer data, including how data is used and who the data is shared with, and often require entities to notify consumers when sensitive personal information is accessed by an unauthorized person. This unauthorized access of sensitive personal information can create negative publicity, resulting in loss of general goodwill, trust and, in more severe cases, may lead to class action lawsuits. Some laws further grant the involved data subject rights to request an organization to erase or stop sharing their personal information.

SB 475 also added obligations to the Texas Department of Information Resources (DIR) to coordinate data security for state agencies and institutions of higher education. The DIR Security Controls Standards Catalog (based on standards from NIST SP 800-53) provides state agencies with a baseline of data security frameworks and parameters to be incorporated into operating policies and procedures. When applying data classification labels to data, it is important to understand the data security controls involved. Labeling data will help an organization understand its data and how to best use and protect it. However, if labeled data is not properly

secured, it is easier for bad actors to find sensitive information. Data classification and labeling must therefore go hand-in-hand with applied security requirements appropriate to the data.

It is advisable to consult with your organization's general counsel to become more familiar with data protection laws in your country, state, and industry — especially as they relate to cloud computing and the storage of sensitive information. Each regulation has varying levels of compliance requirements which can be used as classification levels within your schema. For example, if an organization stores files classified under the Payment Card Industry Data Security Standard (PCI DSS) and the classified files are leaked outside of its Cardholder Data Environment (CDE), the organization can act immediately to remediate the data through the appropriate best practices before investigating the leak source and prevention methods.

Guiding Principles for Data Sensitivity Labeling

Data classification describes the process of identifying, categorizing, and protecting data according to its sensitivity or impact level. Classifying data according to its type, sensitivity, and impact to an organization if altered, stolen, or destroyed helps stakeholders understand the value of data, determine risk, and implement appropriate controls to mitigate those risks.

Data classification is not a new concept. Many organizations have existing policies on labeling or marking physical data such as documents or files. Organizations may use specific folders for routing, watermarks for internally generated documents, stamps to mark documents received, or handwritten notes to apply labels to their physical data. When applying sensitivity labels to digital data, it is important to know how an organization handles its physical data. Organizational nomenclature, on-boarding, and training should be reviewed to ensure consistency between the labeling of physical and digital data.

Sensitivity labels that distinguish different data classifications is metadata (contextual information about the data) applied to physical or digital data records. These labels must be visible for physical data and searchable for digital data. For physical data records, sensitivity labels may be stamped and noted, folders may be labeled, or storage facilities segregated for different classifications or sensitivities. Digital file sensitivity labels may be applied directly to the digital file or folder, and storage locations may be segregated to ensure security of sensitive records. Additionally, organizational databases and applications should have sensitivity labels applied through a centralized Configuration Management Database (CMDB) or in the database metadata on the record, table, or schema levels.

Sensitivity labels are central to applying controls to different types of data classification. They may be used by third party tools to provide a custom sets of access rules that can be applied across organizational networks and applied to documents, collaboration tools (such as Microsoft Teams), and SharePoint sites. The labels help administer controls over who can access the information in a straightforward and intuitive way, based on laws or regulations or organization policy. Data classification and labeling guidelines help data producers and consumers to correctly apply rules in practice. These may include a series of best practices when sharing documents, sending emails, or collaborating across different platforms and organizations.

Sensitivity labels provide an easy way to universally communicate across an organization the level of sensitivity a specific data component should receive as it moves throughout the organization to serve various uses and purposes.

Data Classification Types and Terms

Regulated Versus Unregulated Sensitive Data

Regulated data is defined by statutes, regulations, or industry standards with defined controls standards to protect and secure the information. **Unregulated data** includes all publicly known information and may not have predefined controls. Therefore, it is not always easy to identify unregulated data which may be considered "sensitive." Viewing data in terms only of regulated v. unregulated promotes an incorrect line of thinking. Almost all data within an organization is valuable, but determining risk associated to misuse depends on the amount available, the impact to the organization if compromised, or the context it's used in. For example, a first and last name in a public record by itself is relatively harmless, but a first and last name in an organization's database with a user identifier should be treated as more sensitive as the identifier could be tied to payment card data, home addresses, social security numbers, and more. So, although unregulated data may contain publicly available information, it should never be overlooked by an organization, as its context can enhance its level of sensitivity.

Some examples of unregulated sensitive data may include customer surveys, job applications or employee contracts. These types of data may not always contain confidential information, but they often can. Examples such as these are why it is critical to apply policies and procedures to all your data, regardless of whether it's regulated or unregulated data.

Determining the Sensitivity of Unregulated Data

At first glance, many cases of unregulated data may not appear to be sensitive. However, in certain contexts, seemingly unimportant data may contain information requiring its classification as sensitive, protected data.

Consider an example from Verizon's 10 Things You Should Never Post on Facebook where commonplace information should be regarded as sensitive: an individual is planning a family vacation. The details of the vacation alone may not be sensitive and seem harmless to share on social media. However, posting the dates and timeframe of the family vacation would create a home security risk, considering an individual's address could also be found with relative ease. Publicly announcing on social media that your house will be unattended during a family trip opens opportunities for burglary. The family's vacation times should be considered as *private* or *friends-only* information in this case, given the context and other data available. Thus, when determining the sensitivity of otherwise unregulated, unsecured data, it is important to consider how bad actors might use the data *and* what correlated data they may already have access to, how they could use it to infer sensitive data, and how that could alter the sensitivity classification of the data in question.

Data Classification Terms

Data Classification terms can look unique to each organization, but generally, organizations will categorize data by four types:

Public – Information that is freely and without reservation made available to the public. Data with a public classification typically poses little-to-no risk if disclosed, since public data is freely accessible by anyone. Some examples of public data include agency publications, press releases, or public web postings.

Sensitive (Internal/Controlled) – Information that could be subject to release under an open records request but should be controlled to protect third parties. This is data that isn't meant for public exposure and while there may be some level of harm if exposed, the potential harm is relatively minimal. Examples may include an organization's budgetary plan, or data in transit before it has been published for public use.

Confidential – Information that is typically excepted from the Public Information Act. If this data is exposed, the organization responsible can see negative ramifications, including penalties or fines. Some examples of confidential data include attorney-client communications, computer vulnerability reports, protected personnel information.

Regulated – Information that is subject to specific laws and regulations by a state or federal regulation or other third-party agreement, governing its collection, storage, use, and disposal. This is data mandated by statute or regulation that must be protected from unauthorized access or disclosure. If this data is exposed, the organization responsible for the data can see negative ramifications, including penalties or fines. Some examples of regulated data include data that meets the definition of personal identifying information (PII) and sensitive personal information (SPI) under the Texas Business and Commerce Code 521.002(a)(1) and 521.002(a)(2), HIPAA Security (45 CFR Parts 164), PCI DSS, and Federal Tax Information (FTI).

In some organizations, *confidential* and *regulated* data may be treated the same and combined into a single data classification. An organization may also define additional terms to meet business needs. An example of this could be a *redacted* or *sanitized* classification to denote sensitive or confidential data prepared (but not yet published) for public consumption.

Once an organization's classification levels are defined and a process is established for applying those classifications to data based on specific criteria, it is on the right path for strong data lifecycle management.

Data Classification Framework and Policies

In a codified, formal enterprise-wide policy, a data classification framework is comprised of three to five classification levels, each of which includes a minimum of the classification's name, definition, justification, example(s), and consequences of public disclosure of the classified data.

Another key component of a data classification framework is the set of security controls associated with each classification level. Note that data classification levels by themselves are

simply labels (or tags) that indicate the value or sensitivity of the content. To protect that content, data classification frameworks will need to define controls for each of an organization's data classification levels. These controls may include requirements related to:

- Roles and Responsibilities
- Data Controls
- Access Control
- Transmission Controls
- Audit Controls
- Notification Requirement

Controls will vary by data classification and labeling level, such that the protective measures defined in the framework increase in proportion with the sensitivity of the content. To assist agencies in developing a framework, DIR has published a Data Classification Template that includes suggested controls for each control level.

Data Discovery

Classifying and labeling data requires knowledge of the location, volume, and context of the data. This information is obtained through data discovery, the process of collecting, evaluating, and building insights on data from various sources. Many organizations store and process large volumes of data that is distributed across multiple repositories such as:

- Physical data stores, file rooms, film, and tape
- Big Data platforms
- Heterogeneous databases that are managed on-premises or in the cloud
- Collaborative application platforms such as Confluence, SharePoint, or Office 365
- Cloud services such as Google Docs, DropBox, Amazon S3, and data market portals
- File system objects, such as spreadsheets, PDFs and text files
- Email systems

To perform effective data classification and labeling of these repositories, organizations should undergo an accurate and comprehensive data discovery process. Knowing where the data lives and who has access to it is fundamental to evaluate its impact on the organization.

Data discovery aids the creation of data catalogs — inventories of metadata that give users the necessary information to evaluate the data's accessibility, location, and condition. Data catalogs are a critical tool for effective data management and governance.

At a minimum, adequate data discovery should be able to answer the following:

- Where does the data live?
- Who is responsible for the data?
- What does the data represent?
- How can the data be used?
- How is the data being sourced?

The process of discovering and cataloging data can require the efforts of a dedicated team to maintain it, which can be time-consuming and require constant review and innovation.

Organizations may consider automated solutions as tools and software exist on the open market that minimize the manual effort involved in maintaining and updating data catalogs. Data catalogs work well when controls are in place that keep data inventories up to date. Organizations without these controls — as data pipelines grow increasingly complex and unstructured data is allowed to become the golden standard — will be unable to maintain data inventories that reflect the present reality or understand the data they own (what it does, who uses it, how it's used, etc.).

Data discovery, when implemented correctly, will allow an organization to comprehend field-level lineages of its data, revealing upstream and downstream dependencies between data assets. An understanding of data lineage is key to drawing connections between data assets and obtaining the right information at the right time.

Data Classification and Labeling – Getting Started

Data classification and labeling initiatives are wide reaching by nature, touching nearly every business function within an organization. Because of this broad scope and the complexity of managing content in modern digital environments, organizations often face setbacks in knowing where to begin and how to maintain a successful implementation, as well as how to measure their progress. Common challenges may include:

- Establishing a governance structure that oversees the ongoing progress and maintenance of data classification efforts.
- Identifying specific key performance indicators (KPIs) to monitor and measure progress.
- Increasing awareness and understanding of data classification policies, why they are important, and how to comply with them.
- Designing an easy-to-understand data classification framework, including determining classification levels and associated security controls.
- Developing an implementation plan that includes confirming the appropriate technology solution, aligning the plan to existing business processes, and identifying impact to the workforce.
- Setting up a data classification framework within a technology solution and addressing any gaps between the capabilities of the tool and the framework itself.
- Complying with internal audit reviews that target data loss and cybersecurity controls.
- Training and engaging end users so that they become mindful of the need for correct classification in their daily work and apply the right measures accordingly.

As an organization develops its data classification and labeling framework, it is best to consider the industry best practices:

Getting the right people involved – Having the right stakeholder(s) is essential for success, and it is critical to have senior management backing. To ensure a framework that truly protects your organization, be sure to include privacy and legal stakeholders such as your Chief Privacy Officer and the Office of General Counsel in the development of your policy.

Balancing security against convenience – It is a challenge to draft a very secure and restrictive data classification framework that has been designed with security in mind and is

also very difficult to implement in practice. A good balance of security against convenience alongside easy-to-use tools will lead to wider end-user support.

Taking an iterative approach – Prioritize features critical to the organization and map them against a timeline. Complete the first step, ensure it was successful, and then move on to the next step, applying lessons learned. Remember that the organization may still be exposed to risk while you design your data classification framework, but it's ok to start small with just a few classification levels and expand later as needed.

Maintaining a practical framework – Data classification frameworks typically contain anywhere from three to five data classification levels. Establishing five classification levels may not always be fitting for an organization. The following criteria should be considered when determining the number of classification levels needed:

- Context of the industry and associated regulatory obligations (highly regulated industries tend to require more classification levels).
- Operational overhead required to maintain a more complex framework (more classification levels).
- Users involved and their ability to comply with the increased complexity and nuance associated with more classification levels.

Implementing a Sensitivity Labeling Policy

As staff collaborate with others from inside and outside of their organizations to fulfill their job duties, there is always an inherent risk that the associated data moves beyond the confines of its intended systems and security controls. It is critical to ensure data moves across devices, applications, networks, and services in a secure, protected way that aligns with an organization's business and compliance policies and avoids the risk of compromise. Sensitivity labels should achieve these outcomes all without limiting productivity and users' ability to collaborate. The sensitivity labels should:

- Provide protection settings that include encryption and content markings.
- Protect content across different platforms and devices.
- Protect content in third-party apps and services.
- Classify content without using any protection settings.

Organizations can successfully deploy sensitivity labels through a working team that identifies and manages the business and technical requirements, proof of concept, internal checkpoints and approvals, and final deployment for the production environment. Once sensitivity labels are created that align to the organization's data classification framework, the organization will need to consider how those labels will be applied to data. This is where the plan is put into action, and the organization can start seeing real benefits. There are a few ways labels can be applied:

Data Classification Based on Default Value – When publishing a label policy, you can identify a specific label to be applied by default to all content created by users and groups included in the policy. This label can set a floor of protection, even if no other action is taken by users or system settings.

Data Classification Based on Query – Labels can be applied automatically when content contains specific types of sensitive information, such as Social Security Numbers. Alternatively, the system can detect sensitive information types and prompt the user to optionally apply the label.

Manual Data Classification – Users can apply labels manually to content. While this approach requires less up-front configuration, it also depends on users appropriately classifying content by sensitivity. For that reason, this approach requires a higher level of training and buy-in to be successful.

Government agencies face a variety of regulations governing their data, which vary based on territory and the nature of their work. While the basic principles for developing a strong data classification framework are universal, the details of an organization's framework will depend on the nature of that organization, its requirements for the use of data, and the unique compliance .and security factors that data demands.

What Can Sensitivity Labels Help With?

Provide protection settings that include encryption and content markings. For example, apply a "Confidential" label to a document or email, and that label encrypts the content and applies a "Confidential" watermark. Content markings include headers and footers as well as watermarks, and encryption can also restrict what actions authorized people can take on the content.

Protect content in Office apps across different platforms and devices supported by Word, Excel, PowerPoint, and Outlook on the Office desktop apps and Office on the web, on Windows, macOS, iOS, and Android.

Protect content in third-party apps and services by using Microsoft Defender for Cloud Apps. With Defender for Cloud Apps, you can detect, classify, label, and protect content in third-party apps and services, such as SalesForce, Box, or DropBox, even if the third-party app or service does not read or support sensitivity labels.

Protect containers that include Teams, Microsoft 365 Groups, and SharePoint sites. For example, set privacy settings, external user access and external sharing, and access from unmanaged devices.

Extend sensitivity labels to Power BI. When you turn on this capability, you can apply and view labels in Power BI, and protect data when it's saved outside the service.

Extend sensitivity labels to assets in Microsoft Purview Data Map. When you turn on this capability, currently in preview, you can apply your sensitivity labels to files and schematized data assets in Microsoft Purview Data Map. The schematized data assets include SQL, Azure SQL, Azure Synapse, Azure Cosmos, and AWS RDS.

Extend sensitivity labels to third-party apps and services. Using the Microsoft Information Protection SDK, third-party apps can read sensitivity labels and apply protection settings.

Classify content by applying a label without applying any protection settings. This gives users a clear understanding of how the content is classified within your organization by linking it to the label name. The labels can then be used to generate usage reports and track activity for sensitive content. With this information, protection settings can be applied at a later stage if deemed necessary.

Key Takeaways

- Gather allies within your organization, form a stakeholder group, and obtain executive sponsorship.
- Assess all organizational needs, risks, and regulatory requirements associated with data classification and labeling.
- Decide how many data classification levels are appropriate for your organization.
- Define a Glossary for terms to be used in guidance, including clear data classification definitions and plain language examples.
- Establish organizational policies for data handling based on each data classification level, while ensuring each classification level is mapped and/or differentiated from sensitivity labels.
- Create broad and lasting organizational change through a Crawl-Walk-Run approach — start small with a data classification and labeling pilot program, measure the business impact and feedback, and iterate an improved strategy at the organization level.
- Incorporate data handling policy and best practices into a sustainable training plan for all users and stakeholders involved at your organization.

References

https://texreg.sos.state.tx.us/public/readtac$ext.TacPage?sl=R&app=9&p_dir=&p_rloc=&p_tloc=&p_ploc=&pg=1&p_tac=&ti=1&pt=10&ch=202&rl=22

https://texreg.sos.state.tx.us/public/readtac$ext.TacPage?sl=R&app=9&p_dir=&p_rloc=&p_tloc=&p_ploc=&pg=1&p_tac=&ti=1&pt=10&ch=202&rl=72

https://www.cdc.gov/phlp/publications/topic/hipaa.html#:~:text=The%20Health%20Insurance%20Portability%20and,the%20patient's%20consent%20or%20knowledge.

https://www2.ed.gov/policy/gen/guid/fpco/ferpa/index.html

https://www.pcisecuritystandards.org/document_library/?category=pcidss&document=dss4aag

https://www.nist.gov/itl/publications-0/nist-special-publication-800-series-general-information

https://csrc.nist.gov/publications/detail/sp/800-53/rev-5/final

https://gdpr.eu/

https://oag.ca.gov/privacy/ccpa

https://capitol.texas.gov/tlodocs/87R/billtext/pdf/SB00475F.pdf#navpanes=0

https://dir.texas.gov/resource-library-item/security-controls-standards-catalog

https://www.verizon.com/articles/device-protection/10-things-you-should-never-post-on-facebook/

https://statutes.capitol.texas.gov/Docs/BC/htm/BC.521.htm

https://www.ecfr.gov/current/title-45/subtitle-A/subchapter-C/part-164

https://www.irs.gov/privacy-disclosure/protecting-federal-tax-information-fti-in-databases-through-labeling

https://dir.texas.gov/sites/default/files/Data%20Classification%20Template.xls