Classification and Categorization[1]

These two words (classification and categorization) are often used to describe several things at once. As verbs, they show how organizations decide what set of information security controls are needed to preserve and protect the various information assets of an organization. As nouns, they are the labels assigned to those assets to concisely reflect the types of security controls required to achieve the right levels of protection.

Unfortunately, the information security profession lacks a consistent and coherent definition of either of these words, even though it does seem to agree on what these processes need to accomplish. Several different schools of thought or traditions of use have inadvertently created some of this confusion:

- Many nation's military and national security communities use classification to describe the assignment of security labels (such as "Secret," "Top Secret," and even "For Official Use Only") to information assets. Within these communities, long-established policies, procedures, and traditions emphasize the confidentiality of information over almost any other consideration; "classified" information almost by definition is information that must first be kept from unauthorized disclosure. Other information security attributes, such as availability, integrity, or non-repudiation, are of course important; but they are achieved as they serve the need for confidentiality.
- Other national and regional government policies use categorization to describe the assignment of security labels (such as "personal health information," "trade secret," or "proprietary"), usually as part of implementing laws and regulations for various government programs.
- Preferences for terminology within different communities also compounds this confusion. Comparing publications by NIST or other US agencies with ISO, ISO/IEC, or other nation's standards bodies reveals many instances of very similar ideas being described in very different terms.

Across the private sector -- at least, that part of it that does not do business with national governments, their militaries, or their intelligence and security communities -- there is little common agreement on what to call this "security labeling" process.

One way to cut through this fog is to focus attention on the CRUD actions of creating, reading, updating, and deleting an information asset or the datasets and data items it contains. Creating, updating, and deleting data items or datasets have direct bearing on the integrity, availability, authenticity, and even the non-repudiability of that data. Confidentiality, of course, requires control over which processes and people can read the information (and thereby have the ability to retain it in internal memory, copy it, or send it elsewhere). Let's look at some examples to illustrate:

- Marketplace bid/ask quotations, current trading volume, and related data is published; it must be publicly available so that traders, investors, and the businesses can make transactions work

in those marketplaces. Yet this information cannot be subject to change by people or processes that do not have proper authorization. More importantly, each update must be auditable -- it must have a trail of evidence behind it, including non-repudiable contracts or other messages for each buy or sell, bid or ask.

- Hurricane forecasting and evacuation alerting: these processes publish the raw data and the forecasts generated from that data. The believability of the forecasts depends upon the weather agencies' transparency regarding their processes, data, the models that they use, and how they make decisions to issue weather alerts and evacuation notices. Nothing is confidential, but everything must be protected so that its availability, integrity, and authenticity is assured. Non-repudiation is also critical to assert and protect: Imagine the effect on public confidence if a national weather service could attempt to deny issuing an evacuation or safe-to-return warning to the public.
- Postal mail contents and envelope data: In most jurisdictions, the address information (for sender and recipient) on the outside of the envelope or parcel is considered public data (or at least non-private), while the contents of the envelope or package itself is considered private. Postal systems take reasonable and prudent means to protect the integrity of the envelope data but not its confidentiality. Confidentiality of the contents is assured to the degree that the postal service does not deliver it to an incorrect addressee. As a result, registered mail services can only attest that an envelope was sent and received, but can say nothing about the contents of that envelope.

Classification and Categorization As Processes

Where this brings us to is making an impact assessment, which is central to risk management. Before any labels can be attached to sets of data that indicate its sensitivity or handling requirements, the impact or loss to the organization needs to be assessed. This is our first definition:

- **Classification** is the process of recognizing the impacts to the organization if its information suffers any security compromise -- that is, any type or kind of degradation, reduction, or breach of its confidentiality-, integrity-, availability-, non-repudiation-, authenticity-, privacy-, or safety-related characteristics. Classifications are derived from the compliance mandates the organization must operate within, whether these be law, regulation, contract-specified standards, or other business expectations.

One classification might indicate "minor, may disrupt some processes" while a more extreme one might be "grave, could lead to loss of life or threaten ongoing existence of the organization." You'll note that a good impact statement links a summary term with a description. These descriptions should reflect the ways in which the organization has chosen (or been mandated) to characterize and manage risks.

Note that a classification is not a label. (Security labels are part of implementing controls to protect classified information.)

The immediate benefit of classification is that it can lead to more efficient design and implementation of security processes, if we can treat the protection needs for all similarly classified information with the same controls strategy. This is our second definition:

- **Categorization** is the process of grouping sets of data, information, or knowledge that have comparable sensitivities (impact or loss ratings), and have similar security needs mandated by law, contracts, or other compliance regimes.

These definitions are as directly applicable to small- and medium-sized enterprises or businesses (SMEs or SMBs) as they are to defense contractors, government agencies, and major globe-spanning enterprises in the private sector. While they are consistent with NIST SP 800-60r1, FIPS 199, GDPR, and many ISO/IEC standards, their benefit to the organization does not depend upon using those as compliance frameworks. Instead, the benefit starts from flowing down from the compliance regimes that establish the boundaries within which the organization must operate.

With these definitions in hand, the organization can then go on to create the controls processes, such as policies and security baselines, as part of implementing their security programs.

Data Classification and Categorization Policy

When classifying and categorizing data, data owners should determine the following aspects of the policy:

- **Data classification and categorization:** Define the criteria and processes used to determine the levels of information security protection required. These should include policies and criteria for reviewing and changing the levels of classification and categorization, if required, throughout the lifecycle of that data.
- **Data access**: Define the roles of people who can access the data. Examples include:
  - Accounting clerks who are permitted to see all accounts payable and receivable but cannot add new accounts.
  - Employees are allowed to see the names of other employees (along with managers' names and departments, and the names of vendors and contractors working for the company). However, only HR employees and managers can see the related pay grades, home addresses, and phone numbers of the entire staff. And only HR managers can see and update employee information classified as private, including Social Security numbers (SSNs) and insurance information.
- **Data security:** Determine whether the data is generally available or restricted by default. As an example, many companies set access controls to deny database access to everyone except those who are specifically granted permission to view or update the data.
- **Data retention**: Many industries require that data be retained for a certain length of time. Many financial regulations around the world require specific retention periods. Data owners need to know the regulatory requirements for their data and base the retention period on regulatory and business requirements.
- **Data disposal**: Data classification and categorization directly influence the method in which the data is to be disposed.
  - Printed data disposal may require cross-cut shredding, as defined by data owner.

- Digital data disposal may require employees to use a utility to verify that data has been fully removed from their PCs after they erase files containing sensitive data to address any possible data remanence issues or concerns.

- **Data encryption**: Data owners will have to decide whether their data needs to be encrypted and how. Data encryption will usually be required to meet specific legal, regulatory or contractual requirements for the use of encryption. The Payment Card Industry Data Security Standard (PCI DSS) is one of the more common forms of such contractually based encryption requirements. (While it's natural to assume that security requirements should be used to select, design, and implement a control strategy, contractual or legal requirements can and often do directly dictate the use of a solution such as encryption.)

- **Appropriate use of data**: This aspect of the policy defines whether data is for use within the company, is restricted for use by only selected roles, or can be made public to anyone outside the organization. In addition, some data have associated legal usage definitions. The organization's policy should spell out any such restrictions or refer to the legal definitions as required. Proper data classification also helps the organization comply with pertinent laws and regulations. For example, classifying credit card data as private can help ensure compliance with the PCI DSS. One of the requirements of this standard is to encrypt credit card information. Data owners who correctly defined the encryption aspect of their organization's data classification policy will require that the data be encrypted according to the specifications defined in this standard.

Classification and Categorization: Levels and Labels

It's reasonable to want to have a simple way of assigning a level of sensitivity to a data asset, such that the higher the level, the greater the presumed harm to the organization, and thus the greater security protection the data asset requires. This spectrum of needs is useful, but it should not be taken to mean that clear and precise boundaries exist between the use of "low sensitivity" and a "moderate sensitivity" labeling, for example.

Very few private businesses seem to publish their internal security categorization and classification policies, nor do they publish their labeling or levels (if they use levels at all). That said, borrowing a page from the military and national security communities, and adding it to a legal and compliance perspective, offers some useful ways to classify and categorize data.

Data Sensitivity Levels and Labels

Sensitivity levels (and labels) capture in one or two words a simple, easy-to-recognize warning statement as to the possible harm that could come to the organization if the data is compromised in any way. Note that sensitivity can embrace both classification (disclosure) and categorization (integrity, availability, or authenticity) concerns. Many such sensitivity level examples can be found in various security blogs, which all tend to suggest the use of three or four levels of information sensitivity, from highest to lowest, and these could be:

- **Highly restricted:** Compromise of data with this sensitivity label could possibly put the organization's future existence at risk. Compromise could lead to substantial loss of life, injury, or property damage, and the litigation and claims that would follow.

- **Moderately restricted:** Compromise of data with this sensitivity label could lead to loss of temporary competitive advantage, loss of revenues, disruption of planned investments or activities, etc.
- **Low sensitivity (sometimes called "internal use only"):** Compromise of data with this sensitivity label could cause minor disruptions, delays, or impacts.
- **Unrestricted public data:** As this data is already published, no harm can come from further dissemination or disclosure.

The Center for Internet Security, for example, offers the three levels of "sensitive," "business confidential," and "public" as a way of simplifying classifying low, medium, and high levels.

A natural question might be whether it is ethical to classify information that could be embarrassing if inappropriately disclosed or made public. Personal choice should be able to dictate what information about a person is to be made public, in consonance with due process of law. Businesses and organizations (as legal persons) are no different; there is no requirement in law to publicly admit to all of one's errors in judgment, only to accept the responsibilities to correct damages those errors may have caused.

Data Categorization Labels

Data categorization levels often reflect the source or nature of the security requirements that dictate their use. These often do not fit neatly onto a hierarchy of most damaging to least damaging, as do sensitivity concerns. For example:

- **Human safety critical** as a category would include information which, if compromised, could cause loss of life or injury.
- **Equipment and property safety critical** as a category would include information which, if compromised, could lead to physical property or systems being damaged.
- **Personally identifiable information (PII) critical** as a category would include information that can be used to uniquely identify a particular person.
- **Private data** as a category would be information that the organization has agreed (internally or with another party) to keep private, or to limit the distribution, dissemination, and use of.
- **Proprietary data** as a category would be information regarding the organization's internal business logic, processes, decisions, or criteria for decision making.
- **Compliance data** as a category (or as a set of categories) could be any data that has special handling and protection requirements defined by law, regulation, or contract.
- **Time-critical data** as a category would include information which, if compromised or delayed, could cause a significant business or organizational activity to fail or be cancelled or postponed.

Note that in some organizations, further categorization labels may be needed to explicitly and clearly separate data relating to individual clients, projects, or to individual compliance requirements. FERPA[2]

[2] In U.S. law, the Family Educational Rights and Privacy Act of 1974, as amended.

and HIPAA,[3] for example, might be appropriate category labels and may be needed to make clear the handling requirements for family educational data or health care information.

Note that any given data or information asset can have multiple labels, such as "human safety critical HIPAA highly restricted."

Benefits of Classification and Categorization

Other than the obvious benefit of protecting assets based on value, there are other potential benefits that can be realized by using asset classification and categorization systems:

- Awareness among employees and customers of the organization's commitment to protect information.
- Identification of critical and sensitive information.
- Identification of vulnerability to modification.
- Enable focus on integrity controls.
- Sensitivity to the need to protect valuable information.
- Understanding the value of information.
- Meeting legal requirements.

Issues Related to Classification and Categorization

Classification and categorization need to be driven by the assets owners because they are in the best position to understand the value and sensitivity of the asset. In some instances, the owner may delegate the **responsibility** for classification or categorization (that is, the duty of due care to accomplish the task) of the asset to someone else. However, it is important to understand that even though the owner has delegated the responsibility, the owner will always remain accountable for protecting the value of the asset. **Accountability**, in many legal frameworks, remains with the person or organization it is assigned to, and cannot be delegated.[4]

Asset classification and categorization may have some other issues that the organization needs to address. These can be sources of errors leading to information security compromises, and may include but are not limited to:

- Human error
- Data owners' limited breadth and depth of knowledge
- Inconsistent classification and categorization methods
- Inconsistent, arbitrary or capricious classification or categorization determinations
- Confusing, unclear, or incomplete labeling of all classified and categorized items

[3] In U.S. law, the Health Insurance Portability and Accountability Act of 1996, as amended.

[4] Maritime law provides the well-known example that while a ship's captain may delegate responsibility for the operation of the vessel to a subordinate, such as a watch officer, the captain cannot hand off the ultimate accountability for the safety of the ship, its crew, passengers, and cargo, nor for any missions assigned to that ship.

- Inadequate or incomplete processes for downgrading or destroying sensitive information, including recording media

Human Error

The human factor in security is often viewed as the weakest link; arguably, this is denying the organization the most resilient, flexible, and responsive asset when it may need it the most. When discussing classification and categorization, the human error can also impact the entire process. This is the reason it is vital to make clear distinction between different classification and categorization levels, labels, and their associated handling procedures, so that personal judgment will not impair the integrity of the overall security process. To avoid problems caused by the human factor, all assets must be classified and categorized, and all staff that handle those assets need to understand and apply the same classification and categorization schemes. When staff doesn't understand complex policies, procedures, and supporting elements, or don't have the skills to handle the assets, they may judge the value of those assets subjectively rather than objectively, which may cause inconsistent classification and categorization.

Data Owners' Limited Knowledge and Awareness

The owner needs to have the proper knowledge and ability to classify and categorize properly. This requires both in-domain knowledge and experience of the business logic and its use of the information, a working knowledge of security controls, and knowledge of compliance requirements that might apply. When any of these knowledge and skill sets are incomplete, incorrect, or out of date, data owners may be tempted to err on the side of caution and apply a higher sensitivity category or classification to the data than is actually required. The establishment of an asset classification and categorization board, or committee, with proper membership from key areas of the organization, will have the overall corporate perspective of the value of assets and can alleviate and address this problem.

Classification and categorization should be done based on the value of the asset, but there are elements that owners need to take into consideration to determine the true, correct value of the asset. Awareness and understanding of the laws and regulations, as well as business needs, would allow the data owner to fully assess the asset value.

Inconsistent, Arbitrary or Capricious Classification and Categorization Determinations

When security determinations of any kind are made in inconsistent ways, this can create the perception that all security policies and procedures are arbitrary and capricious. This inadvertently transforms deliberate, thoughtful governance and processes into rule by whim. Classification and categorization determinations in particular can, if made in inconsistent ways, damage the trust and confidence in these processes that all employees, stakeholders, customers, and the organization, depend on for their proper function.

Inconsistent determination or application of security policies (or any set of policies) should be considered a violation of due care and due diligence. Making such determinations in an arbitrary or spur-of-the-moment way puts both the immediate situation and the longer-term needs of the organization at undue risk.

Confusing, Unclear, or Incomplete Labeling of All Classified and Categorized Items

It's important to remember at this point that sensitive information (be it classified, categorized, or both) is what needs protecting; part of that protection is achieved by restricting the ways in which users can create tangible copies of that information.

Traditionally, this required physically labeling, marking, or color-coding of removable storage media or devices, and the marking of printed documents, to show the classification, categorization, and any special handling needs dictated for that information. Users have the responsibility to ensure that data is only written to media or devices that are suitably labeled, which warns the next person who might use that media of the presence of sensitive information being present.

Systems that can simultaneously store, move, and use multiple levels of classified and categorized information need special processes (people and software) to be able to ensure that these security needs are not compromised. These so-called system-high environments usually require users to manually determine the classification and categorization of new data as they create it, apply the required labeling, and then handle that labeled information in correct ways. Systems can be built to provide additional fields in the data structures and logic to support security enforcement.

In many cases, however, it is far too easy to end up with multiple datasets that are incorrectly marked as to their security sensitivities. Email and other online productivity suites are beginning to use keyword scanning, sentiment analysis, and other machine learning techniques to identify information that might have incorrect security labels, but these are not in widespread use yet.

Inadequate or Incomplete Processes for Downgrading or Destroying Sensitive Information, Including Recording Media

As an information asset moves through its useful life, it may need to shift to a higher or lower level of classification or categorization, based on changes in the value of the asset and the external environment. Monitoring the value of the asset as it moves through its lifecycle is necessary for this to work properly. As the value changes, the asset may need to be re-classified or re-categorized and, therefore, protected according to the new value. For example, the financial reports of a traded company are categorized as restricted until they are published, since the data can be used for insider trading. However, once the reports are made public, they have to be re-categorized as public, allowing for further dissemination (but still protected from unauthorized alteration).

At some point, such as the end of its useful life and required retention period, an information asset may either be completely downgraded or need to be destroyed. The destruction procedure, the methods used, and how effective those methods are should reflect the classification and categorization levels, as well as the media on which the copies of the information reside. This usually dictates that removable media be handled (in this case, destroyed) by the techniques required for the highest sensitivity level of data that has ever been placed on that media. This eliminates the opportunity that any of the data can be recovered. Examples include the shredding of hard drives, degaussing technologies, purging methods, overwriting, and sanitizing. At the end of the day, the costs of new or replacement media are far, far less than the potential losses to the organization, its mission, and its stakeholders of an information security compromise.