FORTRA™

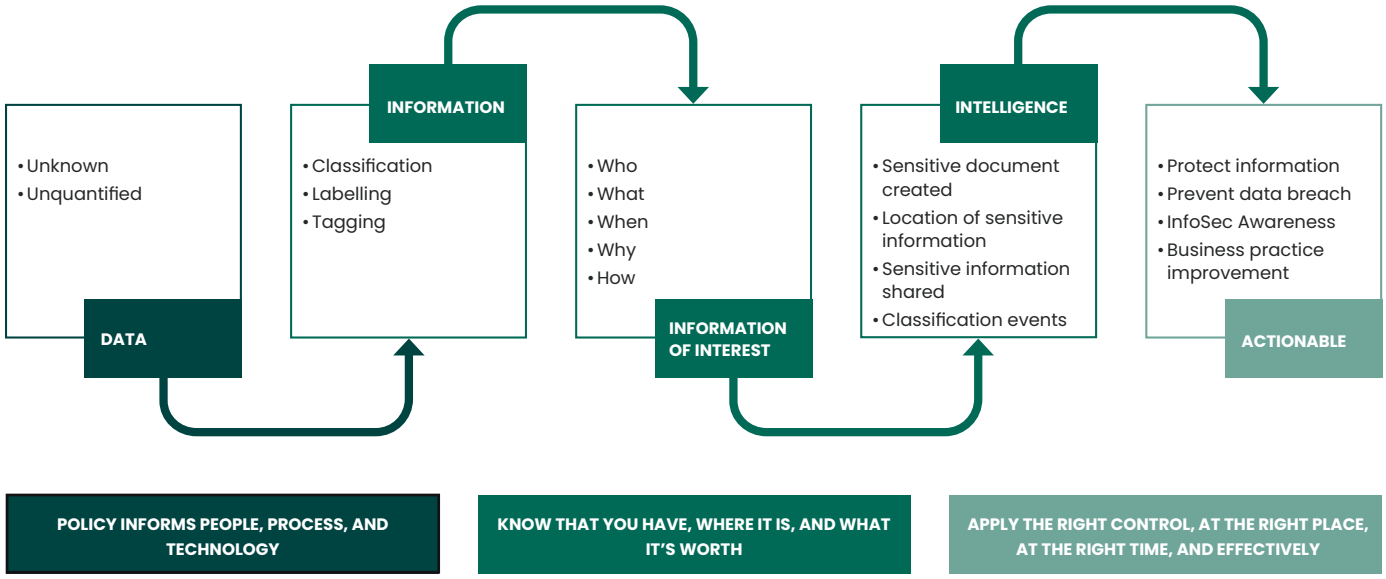
WHITE PAPER (DATA CLASSIFICATION SUITE)

# How Metadata Will Accelerate Your Data Protection Strategy

## Introduction

Unknown data and unquantified exposure risk are key causes for concern among a growing list of organizations around the world. Moving to a position of enhanced knowledge and actionable intelligence requires organizations to understand and control data better. This white paper will discuss the use of rich and persistent metadata as a key component of a data protection strategy.

## Data Protection Transformation




Fortra's Data Classification Suite (DCS) leverages metadata, and its inherent structure and extensibility, as a foundational technology for the products and solutions that are used by Fortune 1000 companies around the world to enable evolving data governance strategies.

## Metadata

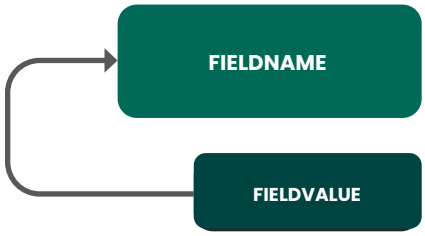
Metadata is data about data. When we use a smartphone to take a photo, the resulting image is "the data" and information embedded in the image file is the metadata, e.g., file name, date, time, location, aperture, shutter speed, and more. We all love to look at the photo, but we gather more valuable information when we also look at the associated metadata.

Fortra's Data Classification Suite uses metadata as a fundamental part of the schema, and the schema, in turn, describes the structure of how the business should interpret the metadata.

Our fictional company **RANOGA** wishes to apply metadata to documents to implement a classification schema. The schema represents a risk to the business should these documents be disclosed. RANOGA wants to set data risk levels to **PUBLIC**, **INTERNAL**, **CONFIDENTIAL** and **SECRET**. To express this schema adequately, we must employ Value Name Pairs.

## Value Name Pairs

Value Name Pairs are used in file formats that hold metadata and constitute a fundamental data representation in computing systems and applications. They provide the capability for open-ended data structures that allow for future extension, without modifying existing code or data.




**Fieldname** provides the container within the schema that will constitute metadata for the first half of a named value pair.

**Fieldvalue** provides the content associated with Fieldname and will constitute metadata for the second half of a named value pair. Each Fieldname can have one or more Fieldvalue(s).

Fields may have **many** Values. Example:

| Fieldname | Fieldname(s) |
|---|---|
| Classification | PUBLIC |
| | INTERNAL |
| | CONFIDENTIAL |
| | SECRET |

A schema may have **many** Fields. Here we have defined a schema with a single Fieldname with four possible values. If applied to a document created in Microsoft Word, within the custom properties, assuming we restrict each Fieldname to a unique value, it would be represented as:

| Name | Value | Type |
|---|---|---|
| CLASSIFICATION | PUBLIC | TEXT |

## Schema

Data Classification Suite uses schema as a model for describing a structure of information that can be applied to unstructured data such as, but not limited to, documents.

Within DCS, the schema is a set of fields, plus their acceptable values and the intended hierarchy.For example, a simple schema may have a primary structure of **PUBLIC, INTERNAL, CONFIDENTIAL** and **SECRET**, plus a secondary structure with values "HR", "Executives", "Research" which is used when the Level is **SECRET**.

Fortra recommends1 that your schema is:

- Clear
- Meaningful
- Impacting

Many rules state that a document must understand if the content is **confidential**, but does not specify if the actual word "**CONFIDENTIAL**" must be used to mark the material in any way, visually or in the metadata.

One of the reasons there is no universally agreed standard can be found in the very nature of the words used; what is SECRET & what does **SECRET** mean?

- Does it apply to "me"?
- Does it apply to members of a project?
- Does it apply to the Executive team?
- Does it apply to Partners only?
- How do I know whether something should be secret or not?

When we classify a document as **PUBLIC**, does this mean that it poses no risk to the organization, or classifying it as "**Public**" also means the material becomes an official statement from the organization that could be used in a legal case?

We often see schemas with values of **PUBLIC, INTERNAL, CONFIDENTIAL and SECRET**, but do these values meet the requirements of **RANOGA**?

If **RANOGA** were principally involved with Fauna, the company might want to build a schema using the values **BIRD, MAMMAL, INSECT** and **FISH**. The schema does not intrinsically have to reflect risk or value. The purpose of metadata is to provide additional context to the primary data; as such, the naming and value will likely be unique for every organization. Ultimately, each company must assess its data governance, InfoSec, and business intelligence requirements when building and maintaining its schema.

Within Fortra's Data Classification Suite, a schema contains Metadata Fields, Values, Display Text, Tooltips, Descriptions and Field Conditionality. The metadata DCS applies to the content can be used in a conditional expression to invoke Actions that help users handle sensitive information appropriately.

## NamespaceURI

Under the hood, DCS uses XML as its fundamental base for data classification and categorisation. XML documents use NamespaceURI, or Namespace, to provide named elements and attributes. All schema(s) have an identity, and it is within Namespace URI where this capability is realised. There can be only one NamespaceURI.

In lay terms, a Namespace can be described as a unique identifier for the contents of a schema and is used to enable the proper interchange between different schemas[2].

[1] At the time of writing, there is no known international standard for the naming of fields, their values, or ordinal position

[2] Namespaces may be stored with the metadata in some cases

The Namespace distinguishes your organization's metadata from metadata stored by another organization, group, or department.

To accomplish this distinction, and to assist in interpreting metadata from other sources, the Namespace should be different for each schema. It is advised to consider changing the Namespace in the future (adding a version number, for example) when making drastic and modifications to the schema that breaks backwards compatibility. Renaming the Namespace will force the old metadata to pass through Schema Mapping, enabling you to make the necessary transitions.

Fortra recommends using a unique NamespaceURI to avoid the risk of a Namespace clash.

> Namespace clash example: Should Org1 and Org2 both use NamespaceURI: CorpInc, there 's no way to distinguish which organization committed the metadata, and it is possible that metadata values may be overwritten erroneously.

To help prevent any Namespace clashes, DCS provides a suggested unique Namespace URI based on the license information submitted upon purchase. Below is an example of a Namespace autogenerated by DCS[3]. Any organization that implements Fortra's Data Classification Suite can use a namespace of their choosing; best practice is to use a short namespace, unique to your organization, and unique to each schema.

For this document, we will use the fictional company **RANOGA** to help describe the use of metadata within a company environment.

> Suggested Namespace Example: http://www.fortra.com/RANOGA

In some use cases, an organization may need to make drastic changes to the schema that breaks backwards compatibility, but maintain the old schema for legacy purposes. In this event consider changing the Namespace (adding a version number, for example) as this will force the old metadata to pass through DCS Schema Mapping, enabling it to make the necessary transitions.

> Note: Not all data protection solutions that store metadata are using NamespaceURI conventions correctly. For this reason, Fortra also recommends using a unique Metadata Fieldname to prevent possible field clash and enable a successful DCS Schema Mapping

[3] The URL within the namespace is not designed to be resolved to Fortra, but merely is a technique used to generate a namespace that is likely to be unique

# Diving Deeper

## Fieldname

It is entirely possible to use any character to construct the Fieldname and Fieldvalue for metadata. The question to be considered, however, is whether downstream tool(s) will be able to interpret that metadata. Today we expect all solutions to be smart, and to understand information the way the human eye and brain do; it is essential to remember that this may not be entirely true of all technical solutions. For example, let us create the Fieldname **SECURITY MARKING**.

Metadata is primarily expressed in a machine-readable format, and we must consider that some downstream solution will interpret the SECURITY MARKING string as **SECURITY%20%MARKING** with the "space" replaced by the ASCII 0x20 character. If downstream technology is making decisions based on the value, we can see how it could lead to an issue.

You can minimize any misinterpretation merely by changing the Fieldname to **SECURITYMARKING**, again remembering that Fieldname and Fieldvalue are not designed for human consumption.

> Note: It is recommended for best practice to avoid the use of special characters, including but not limited to, (!@#$%^<*>&,-.[/

Fortra's Data Classification Suite supports Alphanumeric characters only for Fieldname fields.

## Field CLASH

As discussed, many downstream technologies may not adhere to the NamespaceURI convention, and as such will fall back on the strict confines of Fieldname. Here again, we must carefully consider our Fieldnames to ensure they are unique to avoid any Fieldname clashes.

As a comparison to our fictional company RANOGA, let's look at how a fictional company CorpInc uses a metadata schema:

| RANOGA Fieldname | RANOGA Fieldvalue | CorpInc Fieldname | CorpInc Fieldvalue |
|---|---|---|---|
| Classification | PUBLIC | Classification | OPEN |
| | INTERNAL | | EMPLOYEE |
| | CONFIDENTIAL | | MANAGEMENT |
| | SECRET | | BOARD |

For our example, if **RANOGA** sets a document to INTERNAL, the document properties would resemble:

| Name | Value | Type |
|---|---|---|
| CLASSIFICATION | INTERNAL | TEXT |

Next, we send this document to **CorpInc** who appends some information and saves the file as OPEN. The document properties would then read:

| Name | Value | Type |
|---|---|---|
| CLASSIFICATION | OPEN | TEXT |

This process overwrites the original value and is considered a "clash" because the same Fieldname was used by both organizations but with differing values. If the document were returned to **RANOGA**, it would be treated as a document not classified by the **RANOGA** schema, which may pose a risk to the organization.

Outside of business-to-business clash possibilities, internal solutions such as Microsoft SharePoint Library column name may inject Fieldnames and values outside of an agreed classification schema, thus breaking the original intent of metadata schema.

To avoid the potential of a clash, it is best practice to use a unique field name, possibly prefixed by the org name, as in the following example:

| RANOGA Fieldname | RANOGA Fieldvalue | CorpInc Fieldname | CorpInc Fieldvalue |
|---|---|---|---|
| RANOGACLASSIFICATION | PUBLIC | CorpIncCLASSIFICATION | OPEN |
| | INTERNAL | | EMPLOYEE |
| | CONFIDENTIAL | | MANAGEMENT |
| | SECRET | | BOARD |

In this example, if **RANOGA** sets a document to INTERNAL, the document properties would read:

| Name | Value | Type |
|---|---|---|
| RANOGACLASSIFICATION | INTERNAL | TEXT |

Again, if **CorpInc** receives the document, appends some information, and saves the file as OPEN, the document properties would read:

| Name | Value | Type |
|---|---|---|
| RANOGACLASSIFICATION | INTERNAL | TEXT |
| CorpIncCLASSIFICATION | OPEN | TEXT |

The metadata would be appended, not overwritten. This provides two clear benefits; firstly, each organization understands its risk position with the document clearly, with downstream technologies enabled to make a smarter decision; and, secondly, an ability to determine if the document has been exposed to additional schemas.

Because Fieldname metadata is primarily used in downstream technologies value-add solutions, one may consider creating a short form to make encoding simpler and use fewer bytes. e.g., RANOGACLASSIFICATION can be shortened to RANOGAClass; but avoid reducing it too far to something such as RClass, or RC, as it may not provide a unique enough name and be vulnerable to clash.

## Fieldvalues

Unlike NamespaceURI or Fieldname, there is no apparent technical driver to make the field value "unique" for each schema; however, there is still a need to consider the actual "value" and its downstream use. Example:

| FROM | | TO | |
|---|---|---|---|
| **RANOGA Fieldname** | **RANOGA Fieldvalue** | **RANOGA Fieldname** | **RANOGA Fieldvalue** |
| RANOGACLASSIFICATION | PUBLIC | RANOGAClass | P |
| | INTERNAL | | I |
| | CONFIDENTIAL | | C |
| | SECRET | | S |

If applied to a document created in Microsoft Word, within the custom properties, assuming we restrict each Fieldname to a single value, it would be represented as:

| Name | Value | Type |
|---|---|---|
| RANOGAClass | I | TEXT |

Other downstream technologies such as DLP (Data Loss Prevention) or CASB (Cloud Access Security Broker) could be tuned to seek and evaluate policy based on the value RANOGAClass=I. Having a single character Fieldvalue, though technically feasible, can also lead to high false positives when dealing with downstream technology including, but not limited to, DLP and CASB interoperability. For example, when a DLP tries to look for metadata by looking at proximity to increase accuracy.

In this example, the DLP looks for: **'RANOGAClass' "within 5 characters**" of 'I' which, for this example, is used to denote "Internal". False positives may increase exponentially as by "I" is a vowel, a high probability of the letter "I" being used in metadata within five characters of our field that is not associated with our field. By this virtue, uniqueness in the Fieldvalue becomes equally important. So again, let us consider how to donate a value uniquely. Example:

| **FROM** | | **TO** | |
|---|---|---|---|
| **RANOGA Fieldname** | **RANOGA Fieldvalue** | **RANOGA Fieldname** | **RANOGA Fieldvalue** |
| RANOGAClass | P | RANOGAClass | RCP |
| | I | | RCI |
| | C | | RCC |
| | S | | RCS |

Above we are merely taking key letters from both Fieldname and Fieldvalue, concatenating them together to produce a unique nomenclature.

The same metadata search, 'RANOGAClass' "within 5 characters" of "RCP", now results in only one match. DLP, CASB and other downstream technologies may extract metadata differently, and as such, may use a different raw text to read values. Extracted data rarely looks like Classification=I, which in itself would be unique, but is, unfortunately, not reality.

It is essential to be mindful of existing cybersecurity solutions within the organization when constructing Fieldname(s) and Fieldvalue(s). Taking a holistic approach when configuring your security ecosystem not only increases accuracy in metadata lookups, but also provides additional return on investment for the existing solutions, and a swifter time-to-value.

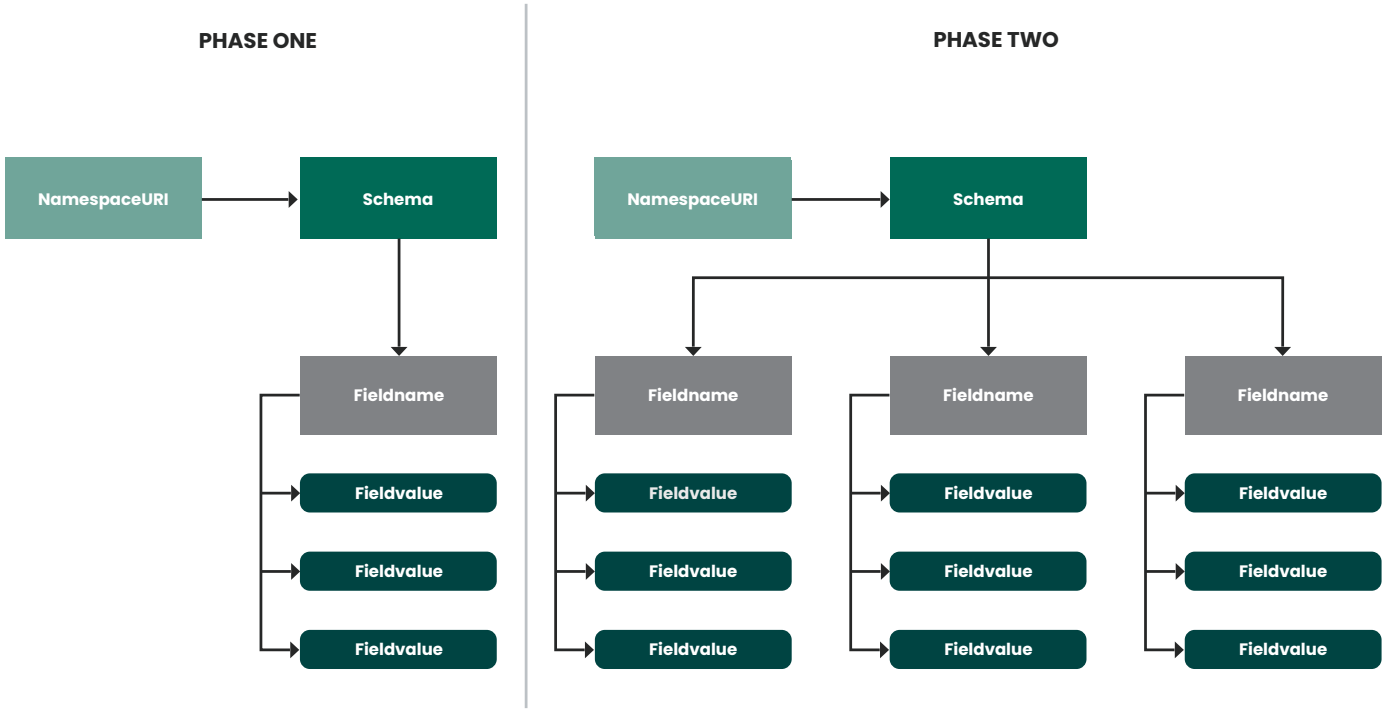
Note: It is recommended for best practice to avoid the use of special characters, including but not limited to, (!@#$%^<*>&,-.[/

DCS supports Alphanumeric and underscore (_) for FieldValue

## MULTIVERSE

Like any building, the foundation is the essential part, and it is within the definition of schema and metadata those foundations are set. If we continue the analogy of a building, it is easier to add more to the building, adding foundations to extensions and add-ons, than to remove a foundation or replace it with something entirely different. It is, of course, possible, only more time-consuming and costly to do so later.

As organizations grow, they can meet more complex requirements by employing multi-field schema. To start the journey of classification, immediately going to a multi-field schema may be tempting, but it also complicates the design and increases the scope of the metadata and its use cases and can be exceptionally challenging for many. It is imperative to train end-users to make relevant choices with the various values that are available.




Fortra recommends organizations start with a simple structure that will bring clarity to the “big picture”, and enable them to make more informed business decisions. Introducing a smaller schema than may be initially envisaged to shorten time-to-value and help them sort their unstructured data into manageable containers. Using schema as a method of defining risk can help to mitigate exposure potential.

At any time, if a multi-field schema is envisioned, it may be necessary to consider the naming convention of fields and values differently. Here we will build a second field to represent the departments: Purchasing, Invoicing, Care, Sales. Example:

| Fieldname | Fieldvalue(s) |
|---|---|
| RANOGAClass | RCP |
| | RCI |
| | RCC |
| | RCS |
| RANOGADept | RDP |
| | RDI |
| | RDC |
| | RDS |

In a document, this could be represented as

| Name | Value | Type |
|---|---|---|
| RANOGAClass | RCI | TEXT |
| RANOGADept | RCI | TEXT |

Within downstream technologies this could be interpreted as RANOGAClass=RCI, RANOGADept=RDI

Note: Downstream technologies may represent the value pairing differently, e.g. no equals sign

In Fieldnames, as previously described, it is best practice to avoid the use of SPACE and special characters within the value to avoid possible misinterpretation from downstream solutions.

## MULTI-VALUE

Fortra's Data Classification Suite allows multiple values within a Fieldname. A common usecase is for a field to contain values for "Releasable to", and related the list of "Country Names" for the values. By this method a sensitive document could had a controlled release to multiple countries or multiple departments.

Using the fields we have previously used within this whitepaper, again assuming we restricted the selection to two values per field, this could be represented as:

| Name | Value | Type |
|---|---|---|
| RELEASABLETO | CAN,USA | TEXT |

A comma separates the second value (,). This system informs RANOGA that this document contains metadata that corresponds to each risk category.

Note: Allowing any combination of multi-value is possible, however as with the example above, PUBLIC and INTERNAL in the same multiselect may be determined by others as inherently conflicting, thus it is the responsibility the organization to clearly lay out the governing principles in a supporting policy document

## FIELD TYPES

There are two metadata types that DCS can write; TEXT and DATE. These can be used in multiple different usecases including but not limited to holding a risk weighting for the data, type in free text and date fields that may contain vital information like retention and legal hold dates for documents.

## ORDINAL RANKING

An Ordinal scale is a quantitative data that is listed in a particular order. Data can be named, grouped and ranked without establishing the degree of variation. Ordinal scales in the DCS schema taxonomy often rank the sensitivity of information based on the impact of the realised risk should the data be disclosed.

| Fieldname | Fieldvalue(s) |
|---|---|
| RANOGAClass | RCP |
| | RCI |
| | RCC |
| | RCS |

In a strict sense, RCP (RANOGA PUBLIC) is a lower value than RCS (RANOGA SECRET), and in logic, we often code to detect the "highest" value. It's possible to reorder the Fieldvalues, so at some future point, the list might read:

| Fieldname | Fieldvalue(s) |
|---|---|
| RANOGAClass | RCP |
| | RCC |
| | RCI |
| | RCS |

The effect of this would not be that all documents that had been classified would require new metadata, but that it may be evaluated differently within Data Classification Suite depending on any business logic applied.

Some organizations may feel that obfuscation of both Fieldname and value are optimal for security. Example:

| Fieldname | Fieldvalue(s) |
|---|---|
| CorpIncClass | CC_1 |
| | CC_2 |
| | CC_3 |
| | CC_4 |

Whereas this is a perfectly valid solution, the issue may occur when values are latterly changed in ordinal position to support a business process change. Let us again assume that CorpInc wants to state that items represented at "Internal" as CC_2 should be treated as more restricted than Items of CC_3.

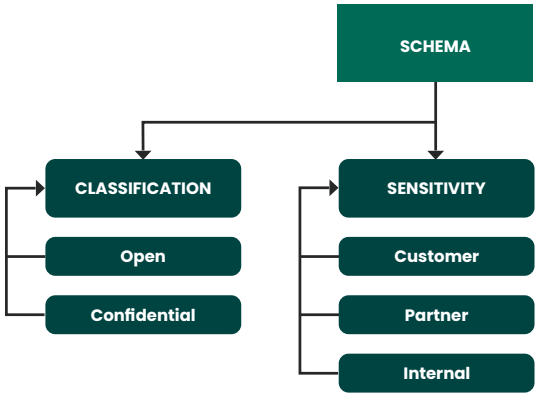
| Fieldname | Fieldvalue(s) |
|---|---|
| CorpIncClass | CC_1 |
| | CC_2 |
| | CC_3 |
| | CC_4 |

When determining the values, it is also essential to recognise the way humans may interpret these values. The risk is that, without this being correctly documented, administrators, at a later date, may be tempted to change the order of the values back to a more "eye natural" position of CC_1, CC_2, CC_3, CC_4

For this reason, Fortra recommends choosing the value in a manner that will reduce the temptation to change the ordinal position based on anything other than the business requirement.

Note: Once metadata is written, it is absolute. The ordinal rank of the metadata and associated value within Data Classification Suite does not affect the way metadata is written, nor how it will be represented to downstream solutions

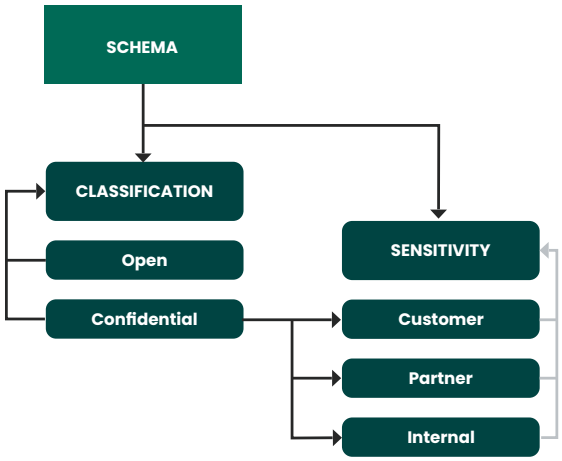
## NamespaceURI




Here we have two fields. The first we will use to hold the risk value of the data, and the second to hold risk exposure.

In a traditional view, the end user would be presented with both fields, they could select **Open** or **Confidential** and select one (or more) from the Sensitivity field. This may not bring the clarity the business is looking to achieve. To help end users make the best choice it may be an advantage to link them together.

Let us now change the model to add in a conditional view.




Here the end user will now only be presented options for **Sensitivity** if they select **Confidential**.

This can allow the end user to realise how the risk artifacts work together in a frictionless manner.

## Visual Presentation of Metadata

We know now that metadata is targeted at both business intelligence and downstream solutions, however, if **RANOGA** chose the short form for both Fieldname and Fieldvalue, how could this choice be more easily expressed to the human reader?

Data Classification Suite offers additional capability to have a "display text", i.e. text that is displayed to the user when selecting a field and can also be used to add visual markings to the document. Because the design of this information is tailored explicitly for human consumption, the Display Text can contain spaces. Example:

| RANOGA Fieldname | RANOGA Fieldvalue | RANOGA Display Text |
|---|---|---|
| RANOGASens | RSCU | Common Use |
| | RSPC | Partner Confidential |
| | RSRH | Restricted Handling |

This distinguishes machine-readable metadata values from the selectable options presented to employees. Display text can be unique in different schema even if the Fieldnames and Fieldvalue are the same, which allows the Fields and values to be presented in different languages for different users in multinational organizations.

## Region A Configuration:

| RANOGA Fieldname | RANOGA Fieldvalue | RANOGA Display Text |
|---|---|---|
| RANOGASens | RCP | Public |
| | RCI | Internal |
| | RCC | Confidential |
| | RCS | Secret |

## Region A Configuration:

| RANOGA Fieldname | RANOGA Fieldvalue | RANOGA Display Text |
|---|---|---|
| RANOGASens | RCP | 公衆 |
| | RCI | 内部 |
| | RCC | 機密 |
| | RCS | 秘密 |

Because the documents both have the same metadata schema, the documents will retain the correct classification when shared across different regions of the business.

This method may be characterized as supporting intentional clashes but is done so with forethought and understanding as a way to empower **RANOGA** to drive value and cohesion in a frictionless workflow.

Note: Within Data Classification Suite it ispossible to extend the Visual display with the Field Description. This can help  provide guidance to the end user on how and when to use the field, which can provide significant benefit for new users