**Technical Documentation: Classification REGEX Generation**

**Overview**

The Classification REGEX Generation process is responsible for generating REGEX (Regular Expression) terms based on the information provided in an Excel file. This document outlines the necessary guidelines for formatting the Excel file and explains the three sheets required in the file.

**Excel File Guidelines**

**File Name**

The Excel file must be named **Classifications\_from\_Glossary\_Terms**.

**Sheets**

The Excel file should contain three sheets, each serving a specific purpose:

1. **TERM\_INFO**: This sheet holds the essential information required for creating classifications. It includes the following columns:
   * **Classification\_Name**: This column contains the name of the classification to be used.
   * **Associated\_Glossary\_Terms**: This column contains the name of the glossary term upon which the associated classification will be based.
   * **Keywords**: In this column, you can provide various phrases that may be used to represent the glossary term.
   * **Functional\_Domain**: This column specifies the functional domain to which the associated classification and glossary term belong. It helps in organizing classifications based on their relevance to specific domains.
   * **Classification\_Level**: Indicate in this column the level to which the classification will be applied, such as table or schema. This allows for granular control over where the REGEX terms will be used.
   * **Applied\_To**: Use this column to specify the elements to which the classification will be applied. This may include table names, column names, or any other relevant entities.
2. **ABBREVIATION\_MAPPINGS**: This sheet is required and is used to provide a mapping of abbreviations and their corresponding full forms. This mapping helps in expanding abbreviations within the Keywords column of the TERM\_INFO sheet. The columns in this sheet are as follows:
   * **Abbreviation**: Enter the abbreviation in this column.
   * **Word**: Specify the full form corresponding to the abbreviation in this column. This is the word that is being abbreviated in the abbreviation.
     1. **Each entry must ONLY have one word**. There can be a one to many relationship from **Word** to **Abbreviation**, but there cannot be a many to many/one. For example, “Storage Location” cannot be an entry in **Word**. It must only be one word, not two.
3. **IGNORE\_WORDS**: This sheet is required and serves to exclude specific words from being considered as part of the REGEX terms. This is useful when certain words are frequently used but are not relevant to the glossary terms. Furthermore, this helps when an occurrence of a REGEX match is incorrect. The sheet contains three columns:
   * **Classification\_Name**: The classification in reference that should ignore the words listed in the appropriate **Words\_to\_Ignore** row.
   * **Keyword\_in\_Reference**: The keyword of the classification that is causing the incorrect match.
   * **Words\_to\_Ignore**: Enter the words that should be ignored in the REGEX term generation process.

**Generating REGEX Terms**

1. Ensure that the Excel file adheres to the guidelines mentioned above and contains the necessary information in the specified sheets.
2. Provide the validated Excel file to the Purview architect responsible for generating the REGEX terms.
3. The Purview architect will run the necessary code using the Excel file as input.
4. The code will process the information in the **TERM\_INFO** sheet and generate REGEX terms based on the glossary terms, keywords, functional domains, classification levels, and scopes provided.
5. Using the **ABBREVIATION\_MAPPINGS** sheet, the code will utilize it to expand any abbreviations found in the Keywords column of the **TERM\_INFO** sheet.
6. Using the **IGNORE\_WORDS** sheet, the code will exclude the specified words from being considered in the REGEX term generation process.
7. The output of the code will be a set of REGEX terms that can be used for classifying and identifying elements within the specified functional domains and scopes.

**Conclusion**

By following the guidelines and using the appropriate Excel file format, the Purview architect can efficiently generate REGEX terms for classifying various elements in the data environment. These REGEX terms aid in better organization, searching, and analysis of data based on the associated glossary terms and their equivalents.