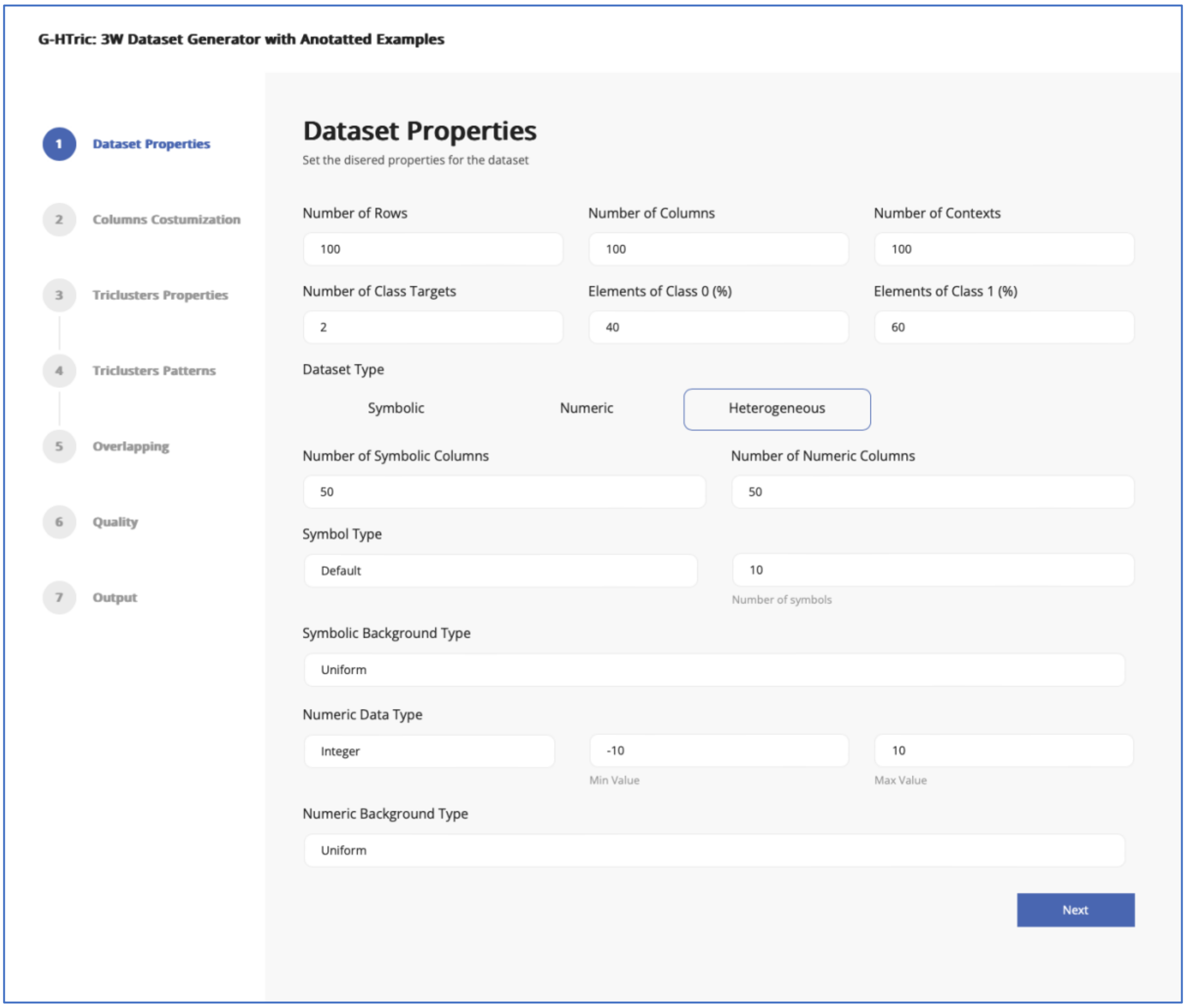
G-HTric: 3W Dataset Generator with Annotated Examples and Triclustering Ground Truth Solutions

DOCUMENTATION

This file introduces the use of *G-HTric's* user interface by presenting the parameters and the application behavior. Each of the following sections shows the different stages of the generator.

1 – Dataset Properties



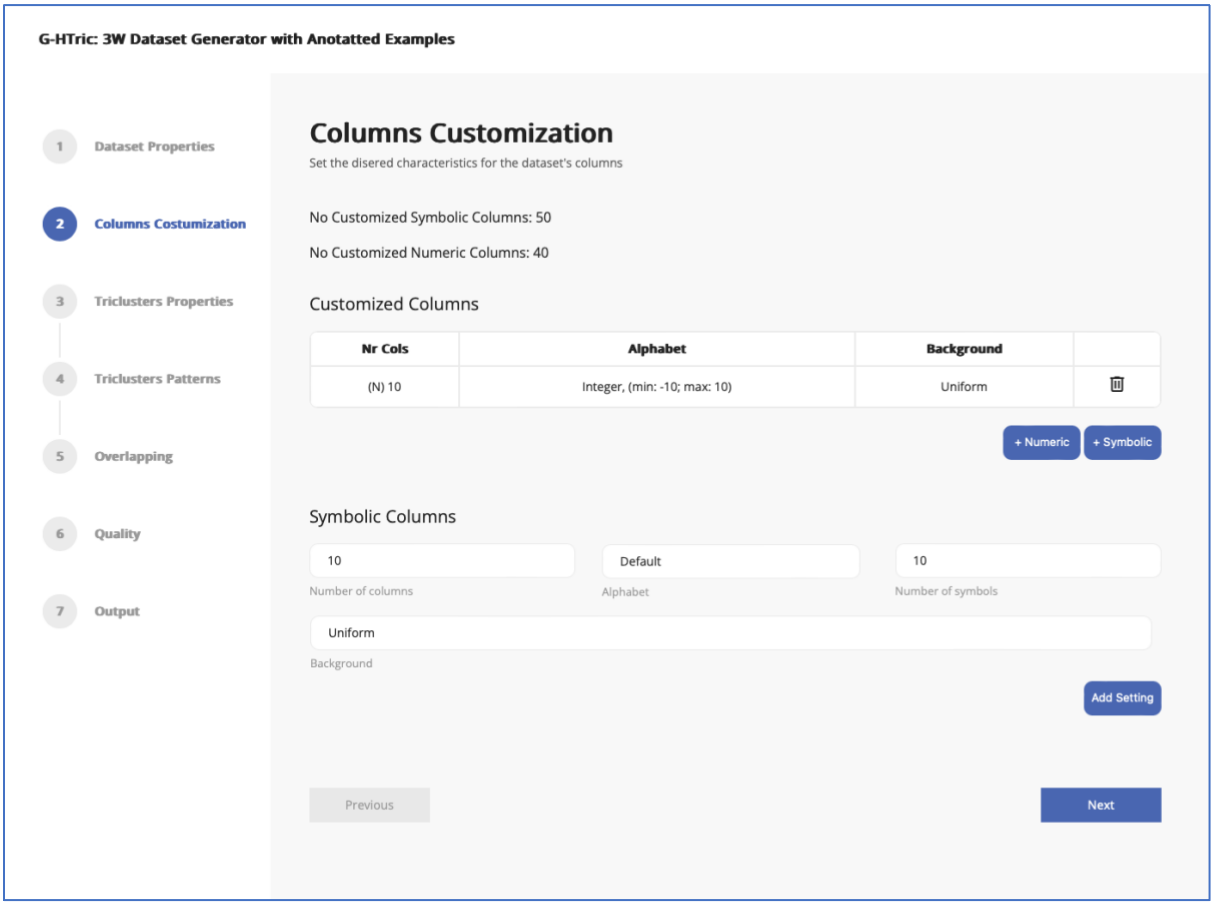
The first step is to define the set of properties that will characterize the dataset. This can be done at the first step of the application form, as depicted in the above figure. For example, the dataset can be composed of 100 rows (observations), 100 columns (variables), and 100 contexts. This is defined in the parameters - *Number of Rows*; *Number of Columns*; and *Number of Contexts*.

The parameter *Number of Class Targets* defines how many labels will be generated as class targets. The parameter *Elements of each class* receives the percentage of observations that will belong to each of the class targets. **N.B:** If no annotated dataset is desired, this parameter should be defined as 0 (zero).

The parameter *Dataset Type* sets how will be composed the dataset to be generated. It can be symbolic, numerical, or heterogeneous (having symbols and numbers). If the user chooses a Symbolic dataset, the user will have to indicate (in parameter *Symbol Type*) if the user desires to use the system default symbols or if wants a custom alphabet. If a *custom alphabet is chosen, a set of symbols has* to be indicated. The order of the symbols in the input determines the ordering of the alphabet. If selecting a numeric dataset, the user can define if the numeric alphabet is represented by either real-valued or integer values and define the allowed range of values. If the user wants a heterogeneous dataset, the user must define both parameters mentioned before for symbolic and numeric components of the dataset as the default. Customizations for some columns can be imposed on these properties in the next step (available for heterogeneous datasets).

The parameters *Symbolic Background Type* and *Numeric Background Type* allow the user to choose how the background values of the dataset are distributed. It can be Uniform, Normal, Discrete or Missing. If choosing Normal or Discrete, the user must define additional parameters. If Normal, Mean and Standard Deviation must be specified. If Discrete, the user should complete the probability for each symbol. Customizations on specific column distributions can also be imposed in the next step (available for heterogeneous datasets).

2 – Columns Customization



In this step, the user can impose specific characteristics for a number of columns that do not follow the default alphabet/distribution specified in the previous step.

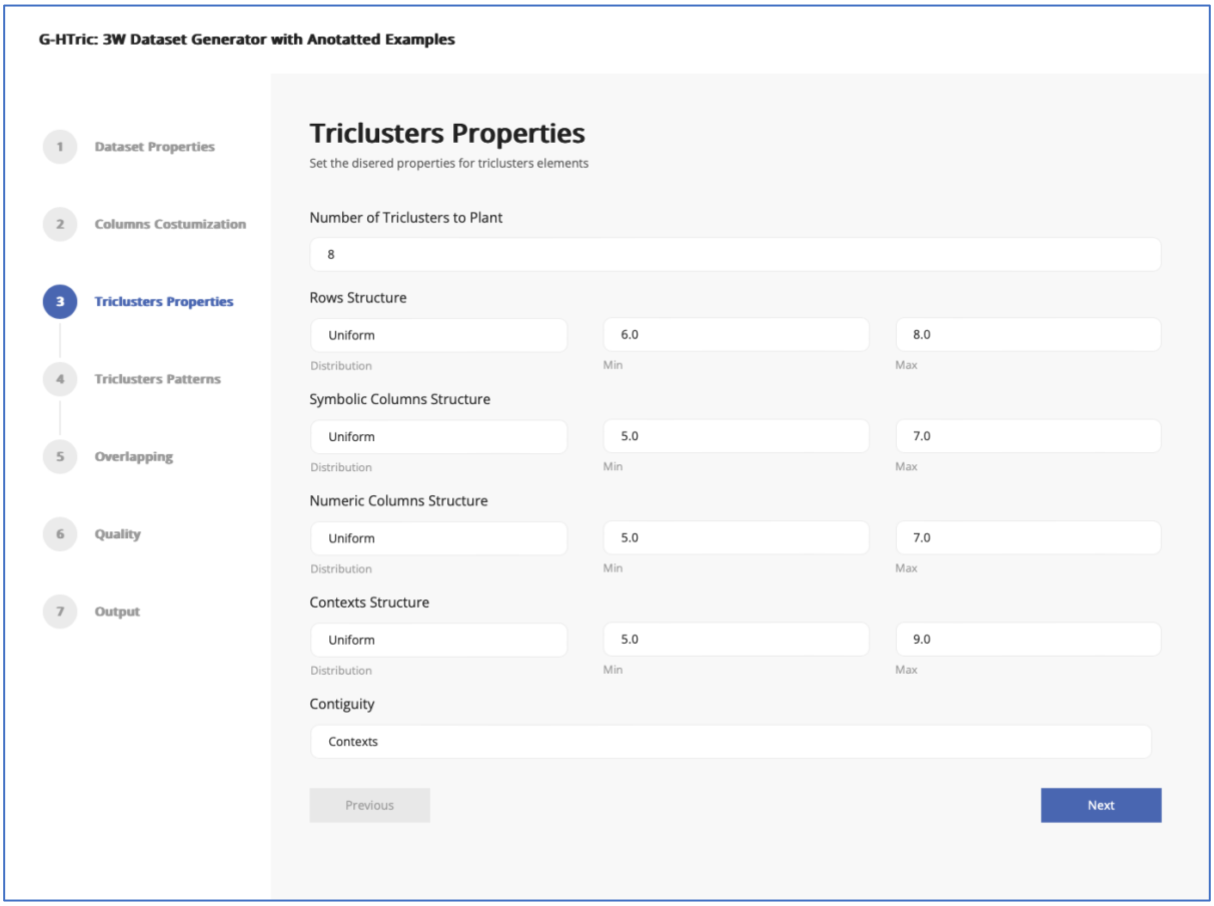
- By clicking on "+Numeric," the user can add new characteristics for a set of numeric columns to be generated. The user should define how many columns, data type (integer or real), maximum and minimum values, and the background distributions for this specific set of numeric columns.

- By clicking on "+Symbolic," the same can be done for a set of symbolic columns. In this case, the user should indicate if he wants to use the system's default symbols or a custom alphabet. If the user chooses a custom alphabet, a set of symbols has to be indicated. Also, the background distribution for this set of columns can be defined as previously.

In both options, the user should finally click on "Add setting" to save the introduced settings.

The user can introduce as many settings as he/she wants since "no customized columns" are available. The number of symbolic/numeric columns available depends on each type's initial number of total columns. It is indicated at the beginning of the page updating as the user saves the desired settings.

3 – Triclusters Properties

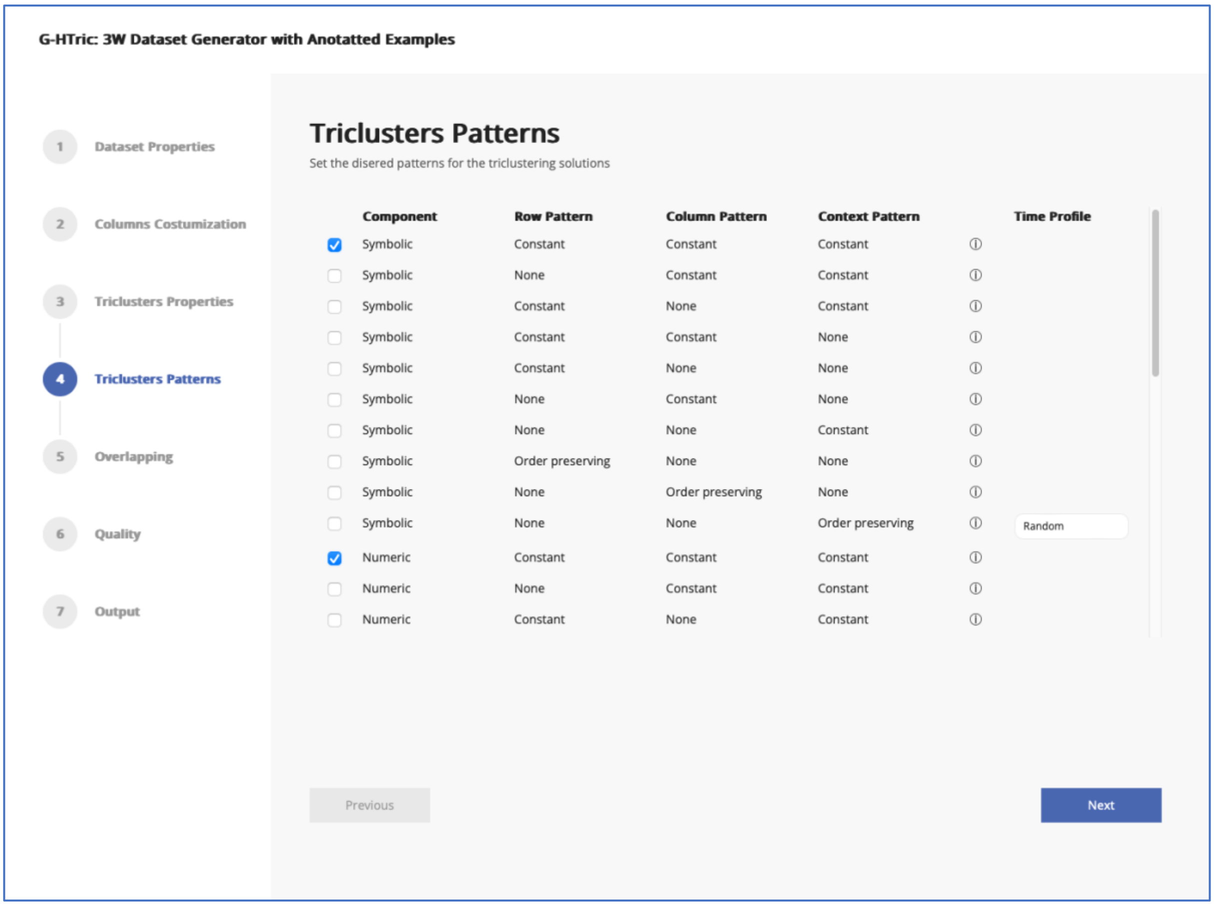


This next step defines the amount and the structure of the planted triclusters on the dataset to be generated. The number of triclusters in the dataset can be defined through the parameter *Number of Triclusters to Plant*.

The following three sets of parameters define their structure: *Row*; *Column*; and *Context*. They are defining their distribution and its respective parameters. The user has two types of distribution available: Normal and Uniform. The interface dynamically adapts the individual parameters to ask for Mean and Standard Deviation for the first type and Min and Max for the second. Note that if the user chose to generate a heterogeneous dataset, he/she should also define the structure for both symbolic and numeric columns.

The last parameter, *Contiguity*, enables the selection of whether the planted triclusters should be contiguous across the column or context dimension. **N.B.:** In heterogenous datasets, contiguity can only be imposed across contexts.

4 – Triclusters Patterns



We now focus on the set of patterns that will be expressed by the set of triclusters planted. If the user chooses a Symbolic or Numeric dataset, the number of patterns chosen will be uniformly distributed across the set of triclusters available. Otherwise, the number of mixed triclusters composed of symbolic and numeric patterns chosen is randomly defined, and the remaining ones are distributed between symbolic and numeric triclusters.

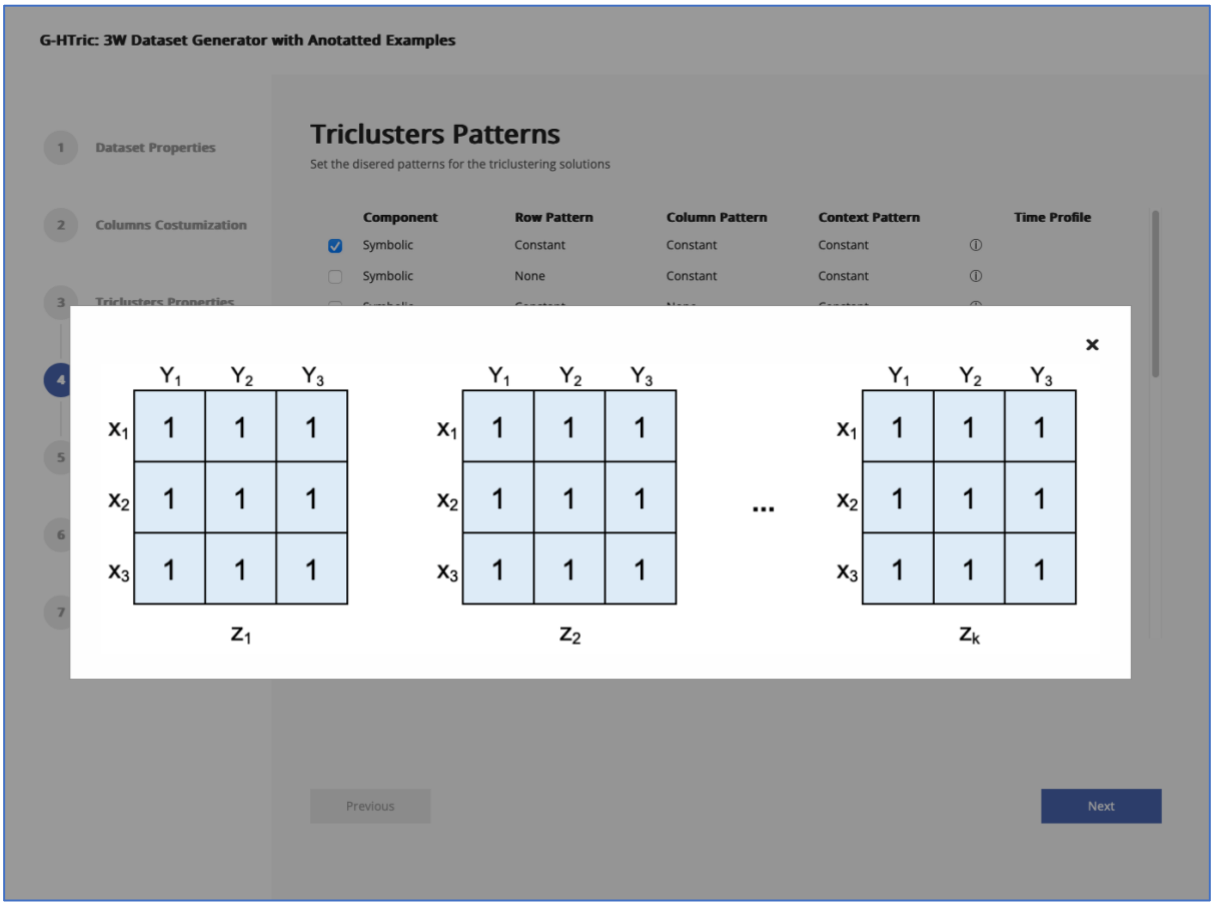
For example, if the user sets four patterns and the dataset has eight triclusters, two triclusters will be assigned to each type.

The user should select combinations of patterns for the three triclusters' dimensions available on the user interface. The available patterns are also documented in the Table below. The Symbolic components are available for Symbolic and Heterogeneous Datasets while the Numeric components are available for Numeric and Heterogeneous Datasets.

Component	Rows Pattern	Columns Pattern	Contexts Pattern
Symbolic	Constant	Constant	Constant
Symbolic	None	Constant	Constant
Symbolic	Constant	None	Constant
Symbolic	Constant	Constant	None
Symbolic	Constant	None	None
Symbolic	None	Constant	None
Symbolic	None	None	Constant
Symbolic	Order Preserving	None	None
Symbolic	None	Order Preserving	None
Symbolic	None	None	Order Preserving
Numeric	Additive	Additive	Additive
Numeric	Constant	Additive	Additive
Numeric	Additive	Constant	Additive
Numeric	Additive	Additive	Constant
Numeric	Additive	Constant	Constant
Numeric	Constant	Additive	Constant
Numeric	Constant	Constant	Additive
Numeric	Multiplicative	Multiplicative	Multiplicative
Numeric	Constant	Multiplicative	Multiplicative
Numeric	Multiplicative	Constant	Multiplicative
Numeric	Multiplicative	Multiplicative	Constant
Numeric	Multiplicative	Constant	Constant
Numeric	Constant	Multiplicative	Constant
Numeric	Constant	Constant	Multiplicative
Numeric	Order Preserving	None	None
Numeric	None	Order Preserving	None
Numeric	None	None	Order Preserving

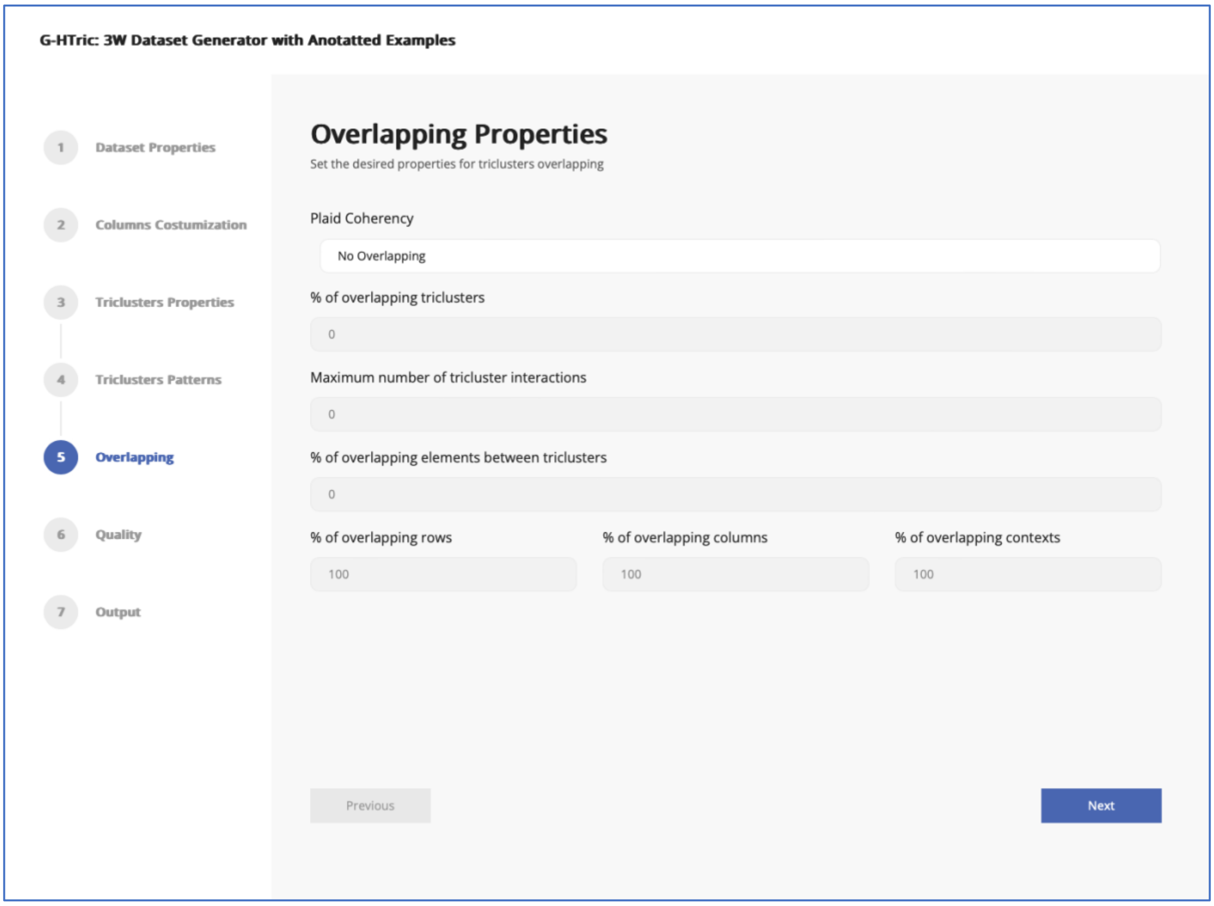
The Order Preserving pattern on contexts, allows the user to select whether the generated temporal pattern can have an arbitrarily number of increases and decreases along time, or follow a monotonically increasing or decreasing pattern.

The interface makes available an example image for each pattern to described it and help the user choosing the patterns desired (by clicking ⓘ symbol).



G-HTric: 3W Dataset Generator with Anotatted Examples
Dataset Properties
Columns Costumization
Triclusters Properties
Triclusters Patterns
Set the disered patterns for the triclustering solutions
Component
Row Pattern
Column Pattern
Context Pattern
Time Profile
Symbolic
Constant
Constant
Constant
Symbolic
None
Constant
Constant
Previous
Next

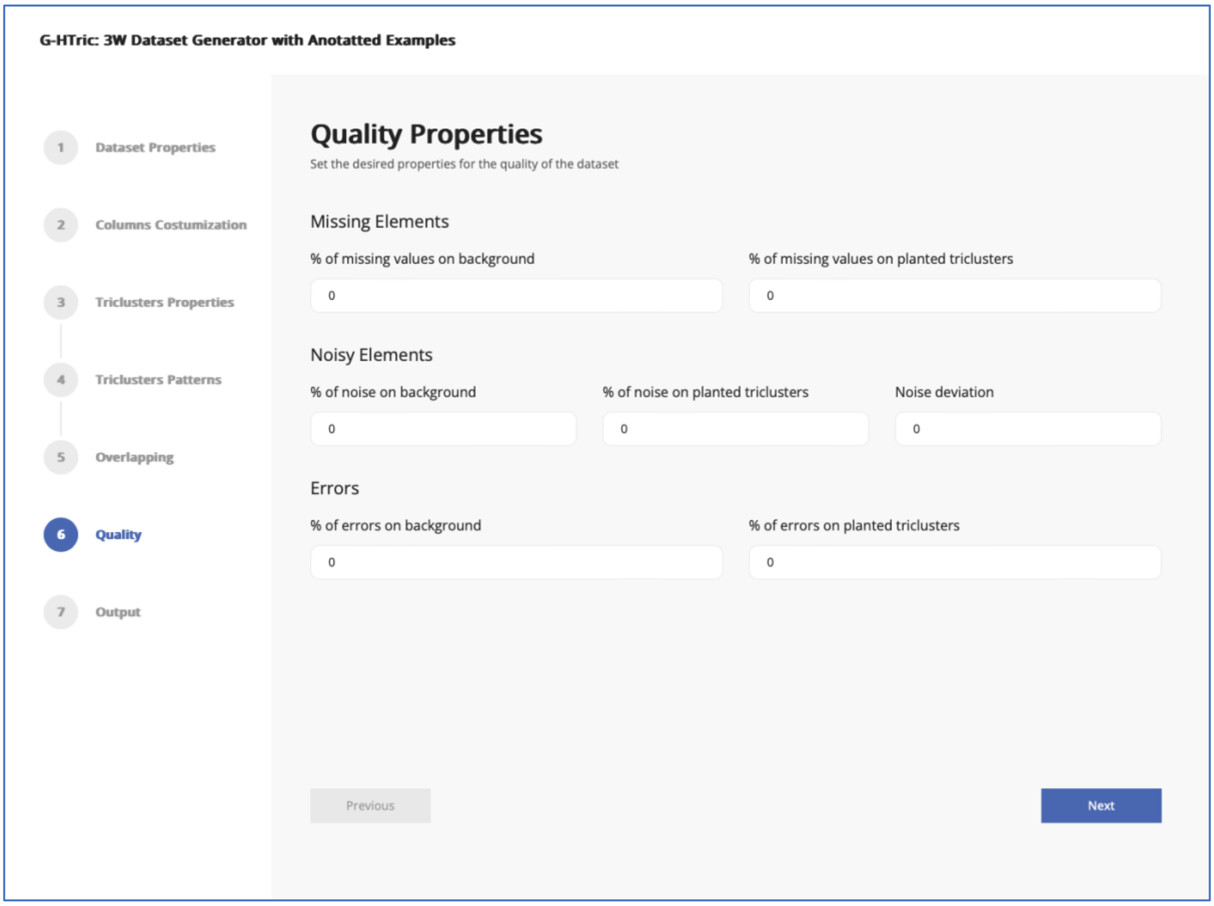
5 – Overlapping



The Overlapping step, shown in the Figure above, allows the user to define the number of triclusters that are allowed to overlap and how their interactions are expressed. The first parameter controls this interaction – *Plaid Coherency*, which makes five types of possible overlapping interactions available: **Additive, Multiplicative, Interpolated, None** and **No Overlapping.**

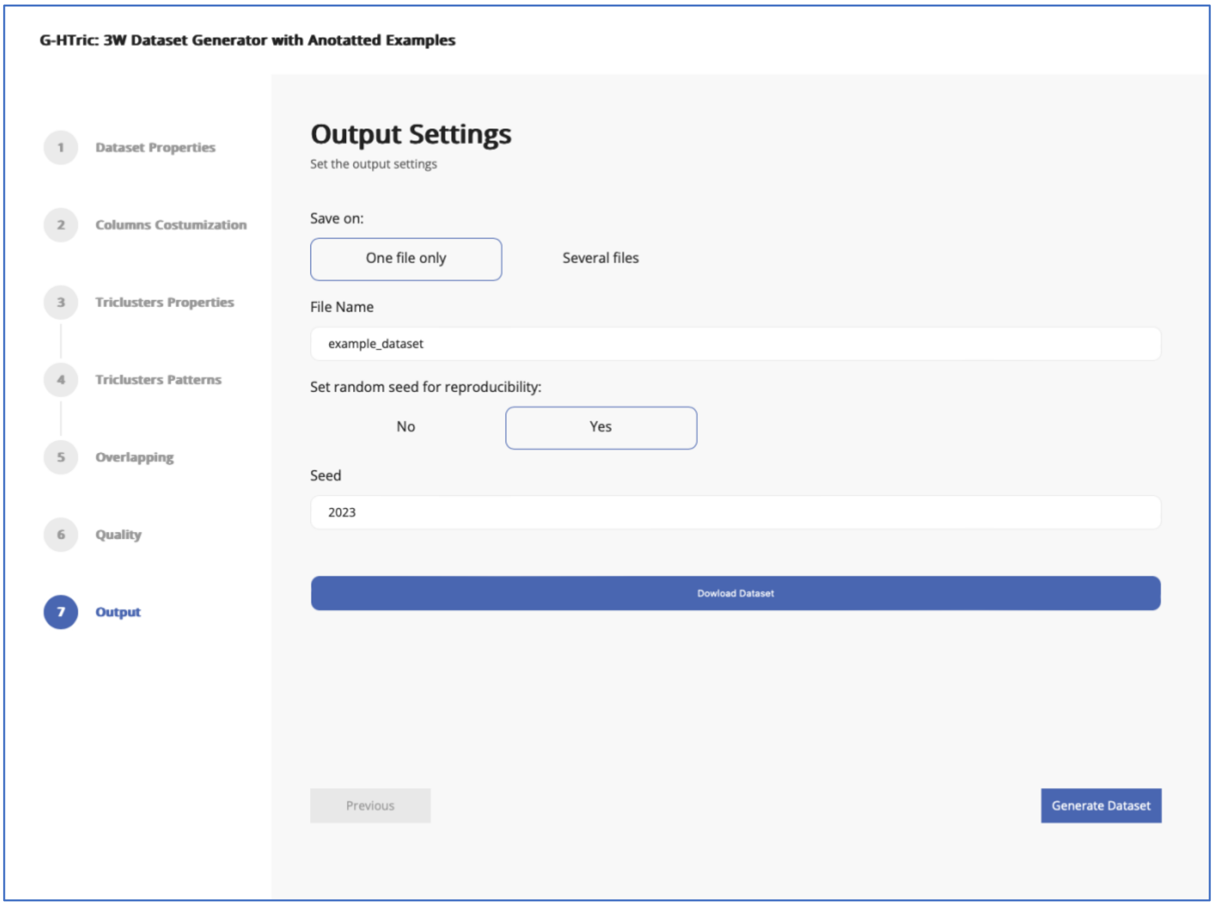
The second step is to set the amount planted triclusters that can overlap. This is done through the parameter – *% of Overlapping Triclusters*. Then, the user must define how the overlapped triclusters will interact with each other. This is done, first, by defining the maximum number of subspaces that can overlap simultaneously, using the parameter – *Maximum Number of Triclustering Interactions*. Then, the user specifies how many elements two overlapped triclusters can share, using the parameter – *% of Overlapping Elements Between Triclusters*. The last three parameters allow the introduction of restrictions on the number of *rows*, *columns*, and *contexts* that a set of overlapping triclusters can share.

6 – Quality



The Quality step, illustrated in the above figure, controls properties from the dataset and the triclusters. Here, the user can define the amount of missing values, noise, and gross errors on both the dataset's background and planted triclusters. For noise, the parameters are *% of Noise on Background* and *% of Noise on Planted Triclusters*. These parameters control the maximum amount of noisy elements, just as above. The *Noise Deviation* defines the maximum deviation of expectation. This means that the noisy value will be, at maximum, at a distance of this value from the original value. The last setting defines the proportion of errors on the dataset.

7 – Output



The last stage before generating the new dataset is defining how the output will be stored, as shown in the figure above. The first parameter – Save On – allows the user to decide whether the dataset should be stored on a single or multiple files. Multiple files are worth it when the dataset has large dimensions since it can be divided into small chunks across several files. The second parameter – *File Name* – sets the prefix of the name of all three output files. The last parameter – *Random seed* – allows the user to select a random seed for the specific dataset generated to enable the reproducibility of the same dataset with the same input parameters.

After completing all the input parameters, the user can generate the dataset by clicking the button "*Generate Dataset*". While generating the dataset, the interface will display a loading image. When generated, it displays a button "*Download dataset*" to download a zip file containing three files. The first file will contain the dataset in a *TSV* format, with the values separated by a tab delimiter. The remaining two files will contain information about the triclusters planted on either *TXT* format, with some statistics and *JSON*.