Help - Exporting and publishing data - Advanced publishing with datasets

Below, please find information on using SurveyCTO datasets for more advanced data workflows.

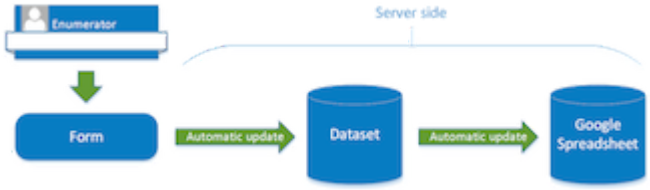
Introduction to advanced dataset usage

SurveyCTO datasets help you to organize and manage your data. Broadly speaking, you can use datasets to:

1. Provide pre-loaded data as an input into one or more survey forms.
2. Organize, combine, monitor, export, and publish subsets of the data submitted by one or more survey forms.

You can combine these two functions for cases where you want to treat the data submitted as part of one or more survey forms as pre-loaded data for one or more other survey forms. As you can imagine, there are a great many possibilities.

Datasets are similar to spreadsheets. They are organized around rows (also known as "records"), on the one hand, and columns (also known as "fields") on the other. SurveyCTO datasets (also known as "server datasets") are constructed and maintained on the server. They can be used to pre-load data into survey forms, and form submission data can publish into these datasets (however: in the case of encrypted forms, only fields explicitly flagged as *publishable* – and thus left un-encrypted – can be published into a server dataset). Server datasets can also be monitored for data quality or published to the cloud so that incoming data streams out to, e.g., some kind of outside visualization or dashboard.



For a small working example, see the "dataset basics" sample form.

Your SurveyCTO subscription may not allow you to use the full range of dataset features because SurveyCTO employs a tiered subscription model in which different subscription levels include access to different features. Following are the key dataset features and whether or not they are supported under your current subscription:

1. Manually upload server datasets: **Supported**
2. Stream data from form submissions into server datasets: **Supported**
3. Publish server datasets to the cloud: **Supported**
4. Configure automated quality checks for server datasets: **Supported**

To learn more about your current subscription and options for upgrades, go to the *Manage Subscription* section of the SurveyCTO website.

Manual use of server datasets

The *Your forms and datasets* section of the Design tab has all of the options you need to manually manage server datasets for the purposes of attaching pre-loaded data to your survey forms. See *Pre-loading data into a form* for a full discussion. You can upload new or revised data; download, rename, or purge existing data; and manage to which forms each dataset will be attached as pre-loaded data.

When you purge a server dataset's existing data, all rows of data will be deleted – but the existing columns in that dataset will remain (albeit currently empty). If you want to completely eliminate old columns that are no longer desired, you will need to delete the dataset entirely, then re-create it.

And when you upload data for a dataset, you always upload a .csv file. Please note the following:

1. The first row of your .csv file should include short, unique names for each column. These column names should not themselves include commas or quotes. Any uploaded column names that do not correspond to fields already in the dataset will be added to the dataset as new fields.
2. If your data contains non-English fonts or special characters, you will need to save your .csv file in Unicode/UTF-8 format. If you cannot directly save or export your .csv file in that format, you can use *SurveyCTO Sync* to re-encode it: just choose *Re-encode .csv file* from the *Tools* menu, select your file and the current encoding for which your file's text appears correctly in the preview window, and then save the re-encoded .csv file.

When you upload new .csv data for an existing dataset, you can choose whether to append the new data to the dataset's existing data, merge with existing data, or replace all existing data. If you choose to merge with existing data, you specify the name of one of your .csv columns to use in uniquely identifying dataset records; incoming .csv rows will either update an existing row of the dataset – if a row with a matching value in that column already exists – or insert a new row into the dataset. When specifying a column with which to merge, there must be both (a) a column in the incoming .csv file with the specified name, and (b) a field in the existing dataset by that same name.

Aside from attaching server datasets to survey forms, you can also download or export these datasets for your own back-office or analytical purposes. To download a dataset's current data, click on *Download* and then *Download data* on your server console's Design tab; that will give you a .csv file with all of the dataset's current data. To export a dataset's current data, simply select that dataset from the list of forms and datasets when exporting data with *SurveyCTO Sync*. Exporting dataset data follows the same process as exporting form data. Please note that if you are automatically publishing form submissions into your dataset,

then your download or export may not reflect data published within the last few minutes. Dataset .csv files are only updated once 5-10 minutes have passed since the last update to the dataset; this prevents datasets – and the forms to which they are attached – from updating with every single form submission.

And finally, you can click *Attach* to manage the list of forms to which any of your server datasets are attached. An attached dataset's data will be available as pre-loaded data that can be pulled into calculated form fields, used to dynamically populate multiple-choice option lists, or even pre-loaded as default values for user-editable survey fields.

That covers the basics of using manually-managed datasets for attaching pre-loaded data to survey forms. See below for discussions of more advanced dataset techniques.

Publishing form data into server datasets

As discussed in the help topic just above this one, you can manually create and manage server datasets from the *Your forms and datasets* section of your server console's Design tab. And if you attach a server dataset to a form, that dataset's data will be available as pre-loaded data that can be pulled into calculated form fields, used to dynamically populate multiple-choice option lists, or even pre-loaded as default values for user-editable survey fields. But you might also want to go in the other direction: publish data from a form's submissions into a dataset (which could then, in turn, be attached to one or more forms).

If you publish a form's submissions to a server dataset, then any submissions received for that form will automatically add to or update the existing dataset. If you want, you can configure multiple forms to publish to a single dataset, which can then serve as an attachment to one or more other forms. This allows for extremely powerful workflows, depending on your needs.

To publish form data to a server dataset, first find the dataset in the *Your forms and datasets* section of the Design tab (or add it there, giving it a title and unique ID). Then click *Publish into* and select the form from which you would like to publish data.

After that, you will need to decide exactly which form fields to publish, to exactly which fields in the dataset (we call this process "mapping"). There is an *Add all* button to simply add all form fields at once (or, in the case of encrypted forms, all form fields that have been marked as *publishable*), plus an *Add* button to add fields one-by-one. For each added field, SurveyCTO will default to publishing to a field name in the dataset that matches the field name in the form – but you can override this default in cases where you want the column name in the dataset to be different.

Note that fields that are inside repeat groups will be listed with an * at the end of their names, and they must be mapped to destination fields that also end in *. When repeated data is published, the * will be automatically replaced by *_1* for the first instance, *_2* for the second, and so on.

Other options

Finally, you have a few other options available:

1. You can indicate one of your form fields to use in uniquely identifying records in the dataset. If you do, then new form submissions will either update an existing row in the dataset – if a row with a matching value in the corresponding field already exists – or insert a new row. If you don't specify a unique identifier, then new submissions will simply publish as new rows. You would likely want to use a unique identifier if you were using a single dataset to merge data from multiple forms. (This unique identifier has to be one of the form fields that you listed in the field mapping. I.e., you also have to publish whatever field is being used to uniquely identify rows.)
2. You can indicate one of your form fields to use as a kind of filter. If you do, only submissions for which the specified field contains the value *1* will be published. For example, if you wanted to only publish data for children, you might select an "ischild" field as the filter; in your form, that could be a *calculate* field with a calculation expression like "if(${age} < 18, 1, 0)".
3. You can check *Publish existing data* if you want to publish existing form submissions. If you don't check this option, then only new submissions that come in (after you configure publishing) will publish to the dataset.

Publishing

As form submissions come in to the server, they will be published to your datasets according to the mappings that you have configured – but there will be a brief delay. The datasets, the forms to which they are attached, and the record counts on the Export tab will only update 5-10 minutes after the most recent submission has been received. This means that, as data-collectors are actively uploading their submissions, the datasets will not update again and again and again, but will instead update only after 5-10 quiet minutes have passed. So, if you download a form or a dataset and find that it doesn't have up-to-the-minute data, please wait a few minutes for the dataset to update.

If you would like to further publish your datasets on to Google Sheets or Fusion Tables for real-time monitoring or visualization, see the help topic below.

Publishing server datasets to the cloud

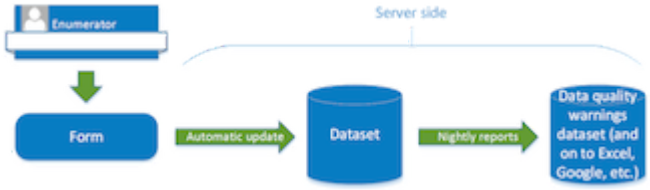
You can publish any server dataset to Google Sheets or Fusion Tables, much in the same way that you can publish form data directly to Google. You even configure the publishing in the same *Advanced: publishing form and dataset data to the cloud* section of the Export tab.

See the help topics on publishing form data for details:

- *Introduction to cloud publishing*
- *Publishing to Google Sheets*
- *Publishing to Google Fusion Tables*

Monitoring server datasets for data quality

You can configure automated quality checks to monitor the quality of incoming dataset data, much in the same way that you can configure quality checks directly for form data.



There are just two things to keep in mind:

1. You manage your dataset quality checks in the *Automated quality checks* section of the Monitor tab (the same place where you configure quality checks for form data).
2. But before you configure any dataset quality checks, you first have to go into *Options* to specify a dataset field that uniquely identifies dataset rows. This is because the quality checks can always use the auto-assigned *KEY* value to uniquely identify form submissions, but different datasets may name their unique identifiers differently. When a quality check results in a warning for a particular row, the value in the ID field you specify will be included in the report so that you know which row triggered the warning.

Aside from those two differences, quality checks for datasets are configured in the very same manner as quality checks for forms are configured. See *Using quality checks to monitor the quality of incoming data* for details.