SurveyCTO

Help - Exporting and publishing data - Overview

Below, please find a few topics to help you get started accessing and using your data.

Exporting, publishing, and using your data

SurveyCTO makes it easy to not only collect data, but also to export, publish, and start learning from that data. Everything you need to get started is located on your server console's Export tab.

The quickest and easiest way to access and start using your data is to go to the *Your data* section of the Export tab. There, you will find two simple options available for each of your survey forms:

1. **Download**: directly download .csv data in the most common and convenient formats. This option exports your data into either a single wide-format .csv file that includes extra columns for both repeated data and multiple-choice questions, or into a series of long-format .csv files. Media files included in your data (e.g., photos or audio recordings) are included in the .csv files as URLs that you can use to download those files.
2. **Explore**: get started exploring your data using SurveyCTO's built-in *Data Explorer*. Summarize individual fields, summarize relationships between fields, and drill down to browse individual submissions. Start learning from your data as soon as it starts coming in.

Exported form data always includes a fully-unique identifier for each and every filled-out form (also known as the submission's "key") in the *KEY* column; use this identifier to uniquely identify and track all form data. All exported data also includes the version number of the form definition used on the device itself, in the *formdef_version* column; use this column to make sense of the data if your form has changed over time (for more on updating forms, see *Updating an existing form*).

If you have encrypted your form data with your own encryption keys, you will need a form's private key file in order to download or explore its data. If you don't have the private key, you will only be able to see data for fields that were explicitly marked as *publishable*. (And, to access a file attachment – like a photo or audio recording – you will need the private key, even if the field is flagged as *publishable*.)

While the above options for direct download and exploration are meant to help you get started working with your data, a richer, more full-featured range of additional options are also available to you. You can learn about them in the following help topics:

- *Exporting data with SurveyCTO Sync*
- *Publishing data to the cloud*
- *Advanced publishing with datasets*

- *API access*

For more about the format of exported data, see *Understanding the format of exported data*. For more about using the *Data Explorer* to explore your data, see *Using the Data Explorer to monitor incoming data*.

Understanding the format of exported data

SurveyCTO defaults to exporting data in .csv files. These are formatted in a comma-separated-value format supported by just about any spreadsheet, database, or statistical analysis software. For the main .csv file exported for each of your forms, the first row (also called the "header row") contains field or column names, and each additional row contains data for a specific submission (also called a "filled-out form" or "instance"). Every file includes columns that correspond to the fields in your form, plus the following extra columns:

- **KEY**: a unique identifier for each submission, automatically assigned by SurveyCTO. You can use this key to uniquely identify and track each submission through your entire data pipeline.
- **instanceID**: a duplicate copy of the submission's key (can be ignored).
- **formdef_version**: the version number of the form definition used to fill out the submission. Use this column to make sense of data when your form has changed over time (for more on updating forms, see *Updating an existing form*).
- **SubmissionDate**: the date and time that the submission was submitted to the SurveyCTO server. This may well differ from the date and time at which the form was initially filled out.
- **review_status** (maybe): if the review and correction workflow is enabled for the form and you have chosen to export not-yet-approved submissions, then this column indicates the review status for each submission.
- **review_comments** (maybe): if the review and correction workflow is enabled for the form, then this column includes any comments added to each submission.
- **review_corrections** (maybe): if the review and correction workflow is enabled for the form, then this column includes a history of corrections made to each submission. (When you download data directly from your server console's Export tab and corrections have been made to that data, the download page will also include a *yourformid_correction_log.csv* file that contains a complete record of corrections.)
- **review_quality** (maybe): if the review and correction workflow is enabled for the form, then this column includes the quality classification for all reviewed submissions.

Except for some special cases described below, each field in your survey will be exported as its own separate column. The name of each column (the value in the header row of exported .csv files) will simply be the name that you gave the field in your form definition – unless you are exporting data with *SurveyCTO Sync* and have selected the option to include group names in .csv column headers. In that case, column headers will include both field names and enclosing groups (e.g., *consented-agriculture-crops* for a field named *crops* that's inside groups named *consented* and *agriculture*).

Dates and times (like the *SubmissionDate* column discussed above) are always exported relative to the time zone of the exporting computer – *not* the time zone of the data-collection device. This is done as a convenience so that data collected in multiple time zones can be easily analyzed relative to a single fixed time zone. (In the case of data published directly from the server to the cloud, the time zone used is always UTC.)

GPS locations captured with *geopoint* fields are exported as four columns rather than just one; for example, for a field named *gpsloc*, the four columns would be: *gpsloc-Latitude*, *gpsloc-Longitude*, *gpsloc-Altitude*, and *gpsloc-Accuracy* (altitude and accuracy will be in meters – and set to exactly 0.0 when unavailable). GPS data collected with geoshape or geotrace fields export into single columns, each of which includes a list of points that represent the polygon or polyline saved; the list of points is separated by semi-colons, and each point has a latitude, a space, a longitude, a space, an altitude, a space, and an accuracy.

Multiple-choice fields that allow the user to choose more than one option (captured with *select_multiple* fields) are exported as a space-separated list of selected option values (like "1 3" if only options with values 1 and 3 were selected). When you download data directly from your server console's Export tab, a separate 1/0 column is also added for each choice; for example, a single field named *crops* that has choice values 1, 2, and 3 would be exported with extra *crops_1*, *crops_2*, and *crops_3* columns, with submission values of 1 if selected or 0 if not selected. *SurveyCTO Sync* will export these 1/0 columns as well, if set to do so in the preferences – but its default behavior is only to export the single space-separated list of selections.

Fields inside repeat groups can have multiple values for a single submission, so they require the exported data to be structured a bit differently. Repeated data can be exported in either "long" or "wide" format:

- **Wide**: in wide format, additional columns are simply added to the primary .csv file in order to accommodate repeated data. For example, if a household roster included a repeated field named *age*, then the exported .csv file would include columns *age_1*, *age_2*, and so on; if the maximum number of household members in the data was 21, then there would be 21 *age* columns and 21 columns for every other field in the roster's repeat group. The server console's Export tab defaults to exporting in this format, and there is a preferences option in *SurveyCTO Sync* to include wide-format exports as well.
- **Long**: in long format, a separate row is added for each repeat instance. Since the main .csv file is structured to have only one row per submission, *SurveyCTO Sync* exports each repeat group into a separate .csv file – so a form with repeat groups will export as a primary .csv file plus additional .csv files for each repeat group. Each row in a repeat group's .csv file will have its own unique *KEY* value, plus a *PARENT_KEY* value that can be used to link to the primary submission or parent-group .csv file.

Exported .csv files are encoded in Unicode format so that they can support the widest possible range of characters and scripts. Unfortunately, Microsoft Excel defaults to importing .csv files as Latin text, so some accents or other scripts may be distorted; see *Looking at data in Microsoft Excel* for details on how to safely import your data into Excel.

SurveyCTO will also export data in other formats. See the help topics on *SurveyCTO Sync* and publishing data to the cloud for more details.

Using the Data Explorer to start visualizing and exploring data

You can use statistical software like Stata, SPSS, or R to visualize and analyze your data, or other software like Microsoft Excel or Google Sheets (both of which have free statistical-analysis add-ons). But to get a quick start exploring and visualizing your data, you can use SurveyCTO's built-in *Data Explorer*.

Using the *Data Explorer*, you can easily summarize data submitted for individual fields, summarize the empirical relationships between fields, and drill down to browse individual submissions. With it, you can start learning from your data right away.

There are two ways into the *Data Explorer*:

1. Click the "Explore" action for any form in the *Your data* section of the Export tab.
2. Click the "Monitor form data" action for any form in the *Form submissions and dataset data* section of the Monitor tab.

The *Data Explorer* is mostly the same whichever path you take, but you can save and maintain separate workbooks for monitoring vs. exploration/analysis. Also, when you enter via the Monitor tab, quality-check results are summarized with your data.

To learn more, see the full help topic in the *Monitoring* section...

Using Stata

SurveyCTO can automatically generate Stata .do files that import, merge, and partially process your exported data. You can use these auto-generated .do files as they are, or you can use them as a starting-point for your own back-end processing code.

To download a Stata .do file template for one of your forms, go to the Design tab of your server console, scroll down to that form in the *Your forms and datasets* section, and then select *Download* and *Stata .do template*. Because you can export your data in different ways, you'll need to say a few things about the format of your exported data: whether it's in long or wide format, and whether it includes group names in the column headers (see the help topic on data export formats for more). You'll also need to choose which language to use for labeling in Stata, if your form has multiple languages. Finally, you'll need to download and unzip the .do file(s).

For most forms, SurveyCTO will output two files: import_FORMID.do and FORMTITLE_corrections.csv (where FORMID and FORMTITLE are replaced by the form's ID and title, respectively). If your form has repeat groups and you selected the "long" export format, then additional .do files will also be output and automatically called from the main import_FORMID.do file.

To try the .do file out, first export your data into .csv format and save it to the same directory as the .do file template. Be sure that the export filenames are the SurveyCTO defaults, so that the .do file will be able to find them okay; if you export from the server console and use your browser to save, it might add a "(2)" or "(3)" on the end, so be sure to catch and remove those suffixes. Once you have the .do file(s) and .csv file(s) in the same place, try running the main import_FORMID.do file.

If you get an error about not being able to find the .csv file, the problem could be the "working directory" that Stata is using. If you get such an error, your best bet is probably to add a "cd" command to the top of the .do file, to change the working directory to the one in which you have saved your .do and .csv files (e.g., "cd ~\Files" on OSX or "cd C:\Files" on Windows).

Another possible source of errors concerns variable names: since Stata only allows variable names to be up to 32 characters long, you can run into trouble. If you have very long group and/or field names, the first 32 characters could fail to uniquely identify a variable. If you run into this problem and you're using *SurveyCTO Sync* to export your data, you can choose not to export group names in your .csv column headers (see *Data export options*), then re-generate your Stata templates and re-export your data; that will shorten many field names because they will no longer include enclosing group names.

Even if you do not plan to customize them much, you should familiarize yourself with the Stata code contained in these templates. Most broadly, each template does the following:

1. Imports, labels, and formats all incoming data.

2. If using long format, organizes data for repeat groups (if any) into separate .dta files (linkable via the *key* and *parent_key* variables).
3. Merges with any previously-imported data, dropping any duplicates (by default, previously-imported data is respected and not overwritten, but see this help topic about overriding that behavior if you allow un-approving data in your review and correction workflow).
4. Applies data corrections, if any.
5. Saves the revised Stata dataset.

In more detail, each template:

1. Initializes Stata ("clear all", "set mem", etc.). Depending on your memory requirements and version of Stata, you may need to revise this code.
2. Initializes filenames and locations in local macros. If you later want to change your .csv or Stata directories, you can update these macro definitions.
3. Lists any names of repeat groups in local macros, both as they would appear in .csv filenames and as they would appear in imported .csv file headers.
4. Lists any names of text, note, date, and date-time fields in local macros. The fields are listed with the names as they will come into Stata, from the exported .csv headers. SurveyCTO tries to make these lists as accurate as possible, based on the form definitions. However, it is possible that you might need to tweak them.
5. Imports the primary incoming .csv file.
6. Drops any note fields, since they do not contain data.
7. Converts any date and date-time variables from text format into Stata's internal date/time format. That way, they sort and filter properly. Please note that the default code uses the clock() and date() functions to parse incoming dates, and we automatically assume MDY or DMY date ordering based on the regional settings of the computer outputting the template. If your computer's regional settings are different from the computer that exported the .do template, you may need to search for "date" and "clock" in the .do file and adjust the MDY or DMY to match your computer's regional settings. (When all else fails, use a text editor to open the raw .csv file exported by SurveyCTO Sync and see what format the dates are in. Then make sure the date() and clock() calls specify the correct format.)
8. Converts all text fields to text format. By default, this includes "calculate" fields – but you can destring them later if you want, or remove them from the *text_fields* macro to not convert them at all.
9. Labels variables and *select_one* values. Note that Stata can only label numeric values, so the template will only label *select_one* values when all possible values are numeric. Note also that variable labels are truncated at 79 characters, but the template also adds a "note" to each field with the full text of the label (as found in the form definition).
10. Merges with any previously-imported data, dropping any duplicates. Because the Stata process is designed to run repeatedly – each time importing .csv files that likely contain both old and new data – the import process defaults to never overwriting existing data with incoming .csv data; that way, you can always update or extend existing data in the Stata file without fear of it being overwritten, and the import process will only add new data to the existing data file. However, if you use a review and correction workflow and allow un-approving data, then the default behavior will mean that you potentially miss changes in data that happen after submissions have been approved. To re-import data that was previously imported – and catch potential corrections made after the initial approval –

change the *overwrite_old_data* local macro at the top of the template from 0 to 1. See this help topic for more on advanced correction workflows.

11. Saves the updated data file.

12. If using long format, runs additional .do files to process any secondary .csv files for repeat-group data, essentially applying all of the above logic to each repeat group (saving each as its own .dta file).

13. Outputs the codebook and all variable notes to the Stata console.

14. Applies corrections, if any. If you wish, you can enter data corrections into the FORMTITLE_corrections.csv file (where FORMTITLE is the title of the form).

    SurveyCTO gives you an empty corrections template along with the Stata template, and you can then edit this template and add a row for each correction you would like to make to the data. If you want to use Microsoft Excel to maintain the list of corrections, you should maintain a .xls or .xlsx file in which all cells are set as *Text* format; otherwise, Excel will do funny things like assume that door number "4/16" is April 16 and encode the value as a very long number instead of "4/16". If you maintain your corrections in .xls or .xlsx format, you can either "Save as" .csv format for the Stata template, or you can update the Stata template to import your .xls or .xlsx file directly.

    If there are any corrections in the corrections .csv file, they will be applied in sequence, row by row. Each row should indicate the *key* of the row to correct, the name of the field to correct (as it appears in the header of .csv exports), the corrected value, and any notes you might wish to maintain in the corrections file (optional).

    If there is any error applying a correction, an error message will be output to the Stata console, including the row number of the offending correction. For example, if you entered a correction with a value of "John" for a numeric field, or if you had a typo in a field name, there would be an error applying the correction.

    Note that the Stata code for applying corrections is tricky. In essence, each correction is output as Stata code, into a temporary .do file, and the template executes the temporary .do file in order to apply the corrections. Thus, you have auto-generated Stata code outputting and running its own auto-generated Stata code.

15. Saves the corrected data file.

The auto-generated Stata templates are mostly meant to get you started on your back-end data-processing. Feel free to revise, extend, and delete whichever parts you don't need or want.

If you use *SurveyCTO Sync* to download and export your data, see also the help topic on using Sync with Stata.