

How does FormShare stores my data?

In a nutshell FormShare stores your submissions as relational data by creating an independent database for each ODK form the you upload.

The following guide provides details on how FormShare does this

What is relational data?

One of the main differences in terms of data management between FormShare and the rest of applications in the market is that FormShare stores your submissions as relational data in a relational database. A relational database is a collection of data items with pre-defined relationships between them. These items are organized as a set of tables with columns and rows. Tables are used to hold information about the submissions to be represented in the database.

Let's use a household survey as an example. Our database might have a table containing household information, with columns representing variables like head of household name or gender, while each row contains data for each individual household. These tables can be linked using keys. If we collect information about each crop grown by a household, the database might have a table containing crops information, with columns representing variables like crops name, total sold, and market value, while each row contains data for one individual crop but each of them connected to one household in the household table using the household ID as key to link them.




One of the main advantages in storing data this way is that complex queries can be performed on the data using SQL language. For example, you can calculate the total income per crop per household:

```
SELECT household_id, crop_name, SUM(total_sold x market_value) AS income  
FROM table_crops GROUP BY household_id, crop_name
```

This would return a table like this:

household_id	crop_name	income
001	Maize	16720.58
001	Beans	7570.34
002	Maize	21836.45
002	Sorghum	32567.20

Other advantages are:

-  Duplicate keys are blocked from entering the database. For example, it cannot be two households with ID 001 in the database.
-  Data cleaning get automatic checks. For example, a data cleaning technician cannot change a crop to a name that does not exist in the table of possible crops.
-  Data cleaning is audited no matter the tool used. For example, a data cleaning technician could use R, STATA, Excel, or the web interface to change the market value of Maize and the database will record the name of the technician making the change along with the date and time, previous value and new value.

FormShare **is the only** data management system for ODK in the market that stores submissions relationally in a relational database.

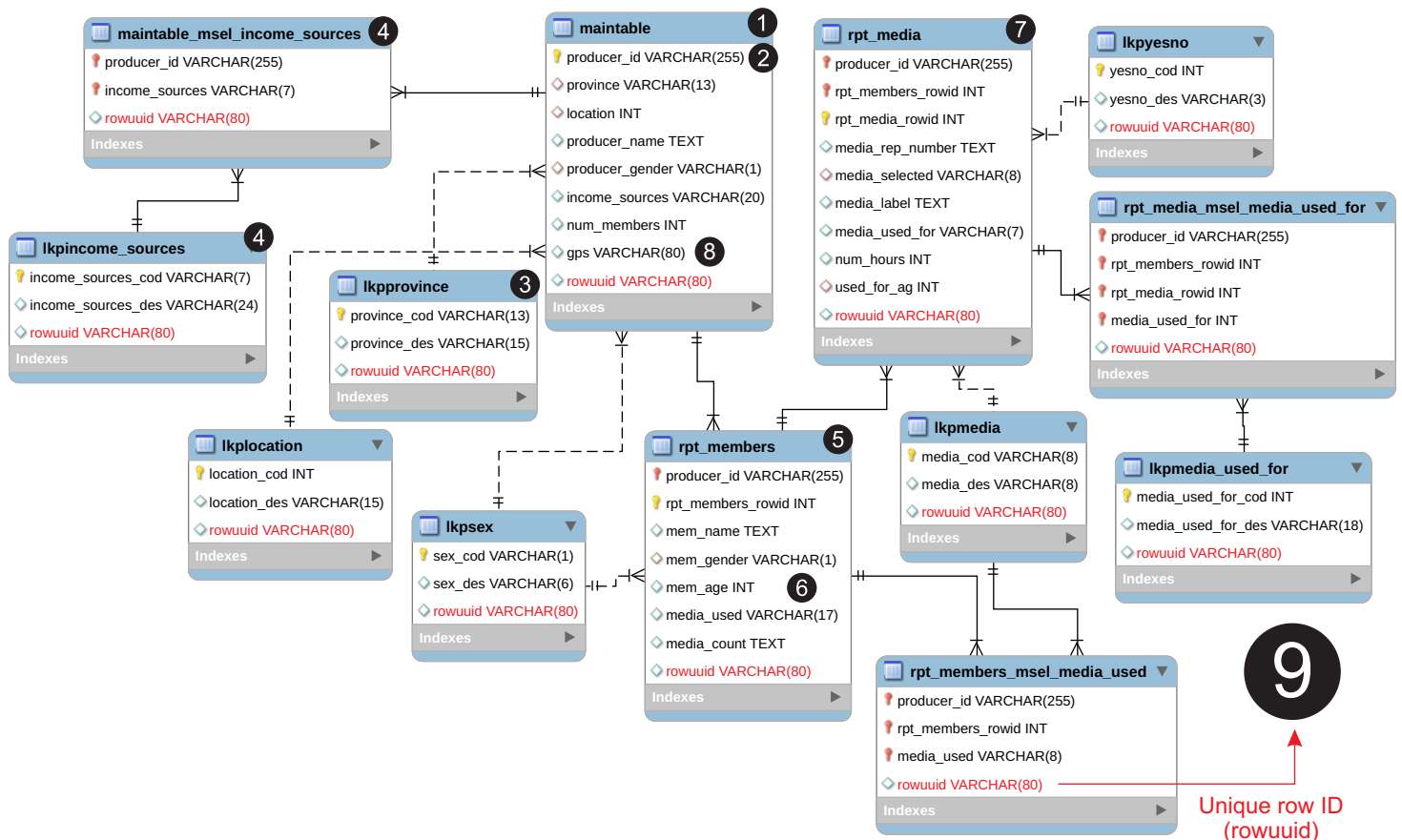
How data is stored? One ODK form = One database

For each ODK form that you upload into FormShare, the system creates a database to hold its submissions. This database is called “repository”. At first, forms are uploaded in a “testing” stage (without a repository) and then you can create a repository for them to store “real” data. The reason for a testing stage is that it is easier to replace a form without a repository because FormShare does not need to alter the underlying database, however it is possible to merge new versions of a form into a common repository.

Lets explore the below ODK Form:

	type	name	label	
1	select_one province	province	Province	Variables outside repeats
	select_one_from_file locations.csv	location	Location	
2	text	producer_id	Producer ID	Variables outside repeats
	text	producer_name	Name	
3	select_one sex	producer_gender	Sex	Variables outside repeats
	select_multiple income_sources	income_sources	Sources of income	
4	integer	num_members	Number of household members	Variables outside repeats
	begin repeat	rpt_members	Information for each member	
5	text	mem_name	Member name	Variables inside repeats
	select_one sex	mem_gender	Sex	
6	integer	mem_age	Age	Variables inside repeats
	select_multiple media	media_used	Media used	
7	calculate	media_count		Variables inside repeats
	begin repeat	rpt_media	Information for each media	
8	calculate	media_rep_number		Variables inside repeats
	select_one media	media_selected	I'm going to ask you about each media. Define media number \${media_rep_number}:	
9	calculate	media_label		Variables inside repeats
	select_multiple media_used_for	media_used_for	Purpose of using this media	
10	integer	num_hours	Number of hours in a day using the media	Variables inside repeats
	select_one yesno	used_for_ag	Used to obtain agriculture information	
11	end repeat	rpt_media		Variables inside repeats
	end repeat	rpt_members		
12	geopoint	gps	GPS	Variables outside repeats

This ODK Form will generate the following repository:



- Variables outside repeats:** They are stored in a data table called "maintable". This is the primary table of the repository.
- Primary key:** To control duplicate submissions the repository needs a **primary key**. When uploading a form you need to select a variable who's data will not duplicate across the whole sample that you expect to have. For example, **producer_id**
- Single selects:** Single select whether their options come from external files or not create a lookup table. The lookup table will be called "lkp_**[listname]**". For example, if you have "select_one province" FormShare will create the lookup table called "lkp_**province**". Each lookup table has two columns: "**[listname]_cod**" storing option names/codes and "**[listname]_des**" storing option labels/descriptions. For example, the lookup table "lkp_province" will have the following columns: "**province_cod**" and "**province_des**". The lookup table will store all the options and be linked to the data table using that list name. For example, the column "province" in "maintable" will be linked to the lookup table "lkp_province" using "province_cod". The data tables storing the submission data will save the option name/code while the label/description of the option will always resides in the look-up table. The **primary key** of all lookup tables is the "**[listname]_cod**" column. For example the **primary key** of "lkp_province" is "**province_cod**".

Definition in Excel

type	name	label
text	producer_id	Producer ID
select_one province	province	Province
text	producer_name	Name

list_name	name	label
province	CE	Central Kenya
province	CO	Coastal Kenya
province	EA	East Kenya
province	NA	Nairobi
province	NO	Northeast Kenya

Construction of a lookup table based on a list name

Data from ODK Collect

producer_id: "001"
 province: "NO"
 producer_name: "Peter Oyango"

Data stored in FormShare

maintable		
producer_id	province	producer_name
001	NO	Peter Oyango

lkpprovince	
province_cod	province_des
CE	Central Kenya
CO	Coastal Kenya
EA	East Kenya
NA	Nairobi
NO	Northeast Kenya

FormShare analyze the selects in a form and tries to re-utilize them as much as possible. One lookup table can be linked to many data tables.

Definition in Excel

type	name	label
text	producer_id	Producer ID
text	producer_name	Name
select_one sex	producer_gender	Sex
begin repeat	rpt_members	Information for each member
text	mem_name	Member name
select_one sex	mem_gender	Sex
end repeat	rpt_members	

list_name	name	label
sex	M	Male
sex	F	Female

Construction of a lookup table based on a list name

Data from ODK Collect

producer_id: "001"
 producer_name: "Peter Oyango"
 producer_gender: "M"
begin rpt_members
 M1 mem_name: "Peter Oyango"
 mem_gender: "M"
 M2 mem_name: "Mary Otieno"
 mem_gender: "F"
 M3 mem_name: "Joshua Oyango"
 mem_gender: "M"
end rpt_members

Data stored in FormShare

maintable		
producer_id	producer_name	producer_gender
001	Peter Oyango	M

lkpsex	
sex_cod	sex_des
M	Male
F	Female

rpt_members			
producer_id	rpt_members_rowid	mem_name	mem_gender
001	1	Peter Oyango	M
001	2	Mary Otieno	F
001	3	Joshua Oyango	M

- 4 **Select multiple:** ODK Collect stores “select multiple” variables in one field with selected options separated by space. FormShare stores it as independent rows in a separate table linked to the data table using such multi-select and to the associated lookup table. For example:

Definition in Excel

type	name	label
text	producer_id	Producer ID
text	producer_name	Name
select_multiple	income_sources	Sources of income

list_name	name	label
income_sources	onfarm	Agriculture on-farm
income_sources	offfarm	Agriculture off-farm
income_sources	nonag	Non agriculture activity

Construction of a lookup table based on a list name

Data from ODK Collect

producer_id: "001"
producer_name: "Peter Oyango"
income_sources: "onfarm offfarm"

1 producer
2 Income Sources
1 row
2 rows
Link

Data stored in FormShare

maintable (Parent data table)		
producer_id	producer_name	income_sources
001	Peter Oyango	onfarm offfarm

maintable_msel_income_sources (Child multi-select table)	
producer_id	income_sources
001	onfarm
001	offfarm

lkpincome_sources	income_sources_des
onfarm	Agriculture on-farm
offfarm	Agriculture off-farm
nonag	Non agriculture activity

Each multi-select table is named in the following way: “[data-table-with-the-multiselect]_msel_[multiselect-variable]”. For example the multi-select variable called “income_sources” is outside any repeat therefore it will be stored in “maintable” thus the multi-select table storing the options as independent rows will be called “maintable_msel_income_sources”. The **primary key** of a multi-select table will be the combination it's parent's **primary key** plus and the multi-select variable. For example the **primary key** of “rpt_members” is “producer_id” + “income_sources”.

- 5 **Repeats:** Repeats create data tables. The name of the data table is the same of the repeat. For example, the repeat “rpt_members” will create the table “rpt_members”. Repeats at the same level of variables outside a repeat will become “child data tables” of “maintable”. For example, “rpt_members” is a child table of “maintable”. The **primary key** of a repeat data table will be the combination it's parent's **primary key** plus an sequence column called “[repeat_name]_rowid”. For example the **primary key** of “rpt_members” is “producer_id” + “rpt_members_rowid”. The sequence will start in 1 and increment for every row.

Definition in Excel

type	name	label
text	producer_id	Producer ID
text	producer_name	Name
select_one sex	producer_gender	Sex
begin repeat	rpt_members	Information for each member
text	mem_name	Member name
select_one sex	mem_gender	Sex
end repeat	rpt_members	

• Same level

Data from ODK Collect

producer_id: "001"
producer_name: "Peter Oyango"
producer_gender: "M"
begin rpt_members
M1 mem_name: "Peter Oyango"
mem_gender: "M"
M2 mem_name: "Mary Otieno"
mem_gender: "F"
M3 mem_name: "Joshua Oyango"
mem_gender: "M"
end rpt_members

1 producer
3 members
1 row
3 rows
Link

Data stored in FormShare

maintable (Parent data table)		
producer_id	producer_name	producer_gender
001	Peter Oyango	M

rpt_members (Child data table)			
producer_id	rpt_members_rowid	mem_name	mem_gender
001	1	Peter Oyango	M
001	2	Mary Otieno	F
001	3	Joshua Oyango	M

- 6 **Variables inside repeats:** Because repeats create data tables, all the variables contained by the repeat will be part of that data table.

- 7 Nested repeats:** Repeats inside another repeat will become “child data tables” of it's parent repeat. For example, “rpt_media” is a child table of “rpt_members”. The **primary key** of a nested repeat data table will be the combination it's parent's **primary key** plus an sequence column called “[repeat_name]_rowid”. For example the **primary key** of “rpt_members” is “**producer_id**” + “**rpt_members_rowid**” therefore the **primary key** of “rpt_media” is “**producer_id**” + “**rpt_members_rowid**” + “**rpt_media_rowid**” The sequence will start in 1 and increment for every row.

Definition in Excel

type	name	label
text	producer_id	Producer ID
text	producer_name	Name
select_one sex	producer_gender	Sex
begin repeat	rpt_members	Information for each member
text	mem_name	Member name
select_one sex	mem_gender	Sex
begin repeat	rpt_media	Information for each media
select_one media	media_selected	I'm going to ask you about ...
integer	num_hours	Number of hours in a day ...
end repeat	rpt_media	
end repeat	rpt_members	

Data from ODK Collect

```

producer_id: "001"
producer_name: "Peter Oyango"
producer_gender: "M"
begin_rpt_members
  M1 mem_name: "Peter Oyango"
  mem_gender: "M"
  begin_rpt_media
    MD1 media_selected: "internet"
    num_hours: 5
    MD2 media_selected: "radio"
    num_hours: 3
  end_rpt_media
  M2 mem_name: "Mary Otieno"
  mem_gender: "F"
  begin_rpt_media
    MD1 media_selected: "radio"
    num_hours: 6
  end_rpt_media
  M3 mem_name: "Joshua Oyango"
  mem_gender: "M"
end_rpt_members
  
```

Data stored in FormShare

maintable (Parent data table)		
producer_id	producer_name	producer_gender
001	Peter Oyango	M

rpt_members (Child and parent data table)			
producer_id	rpt_members_rowid	mem_name	mem_gender
001	1	Peter Oyango	M
001	2	Mary Otieno	F
001	3	Joshua Oyango	M

rpt_media (Child data table)				
producer_id	rpt_members_rowid	rpt_media_rowid	media_selected	num_hours
001	1	1	internet	5
001	1	2	radio	3
001	2	1	radio	6

- 8 Geopoint outside a repeat:** If you record the GPS position as part of your data, place the “geopoint” variable outside any repeat. FormShare will detect it and use it to display your submissions in a map and to generate products like KLM.
- 9 Unique row ID (rowuid):** This is perhaps the most important feature in a FormShare repository. Each row of data in any table (data, lookup, or multi-select) in any repository has a unique row ID. This unique ID allows FormShare to identify a row in the system and determine to which table, repository and form it belongs to. The unique Row ID is commonly used in API data cleaning to update data in the repository.