PPOL 6818 - Spring 2025 - Week 7

Béatrice Leydier

Advanced Coding and Survey Deployment in SurveyCTO

Georgetown University Initiative

guide

on Innovation, Development and Evaluation

Major steps of good data collection

0. Outcomes definition

week 4: research design

week 2 : project structure, documentation

1. Stocktaking project codebook & previous surveys
2. Survey writing
3. Survey coding

last week

4. Survey deployment
5. Data management
6. Data exporting

this week

7. Data monitoring and quality checks

Major steps of good data collection

0. Outcomes definition

week 4: research design

week 2 : project structure, documentation

1. Stocktaking project codebook & previous surveys
2. Survey writing

week 12 : survey optimization

3. Survey coding

last week

4. Survey deployment
5. Data management
6. Data exporting

this week

7. Data monitoring and quality checks

week 11 : data quality

Data Management

17 89

GEORGETOWN UNIVERSITY

Principles

- **Safe**
    - user access and management
    - device access and management
    - encryption
    - de-identification
        - HIPAA 17 identifiers (safe harbor)
        - microdata anonymization (statistical disclosure control)
- **Streamlined**
    - no duplication of information
    - server datasets, inputs and outputs of data collection
    - workflow for data corrections

Goals

- **Streamline your data pipeline**
    - sampling frame
    - survey responses
    - audits and corrections
- **Keep your data safe**
    - user, device management
    - encryption

Code in SurveyCTO

Reminder

GEORGETOWN UNIVERSITY

SurveyCTO : Coding Process

1. **Set up your survey Google Sheet**

    a. Add a Cover Sheet tab, add a Changes Log tab

    b. Check the formid in the Settings tab

1. **Upload and test your survey regularly**

    b. Debug from the SurveyCTO error prompts

    c. Test coding from the SurveyCTO testing interface

1. **Use help**

    c. Help tabs of your survey Google Sheet

    d. SurveyCTO help: https://gui2de.surveycto.com/help.html

Set up your survey Google Sheet

1. **TAB: SETTINGS**
   - Change form title and ID
   - Version control (keep as default)
   - Default language (for forms in several languages)

1. **TAB: SURVEY**
   - Understand default variables
   - Define new questions
     - Variable type
     - Label
     - Relevance
     - Constraints

1. **TAB: CHOICES**
   For the select_one and select_multiple var types

1. **Add two TABS: Cover Sheet, Changes Log**

Default variables

Those variables automatically appear at the beginning of the Excel form definition and shouldn’t be deleted. They won’t show up to enumerators (hidden field) but will be useful for data collection monitoring.

start	starttime	record the date and time the survey was started
end	endtime	record the date and time the survey was ended
deviceid	deviceid	record the unique ID of the device used to fill out the survey
subscriberid	subscriberid	record the subscriber ID associated with the device's SIM card, if any
simserial	simid	record the serial number associated with the device's SIM card, if any
phonenumber	devicephonenum	record the phone number associated with the device's SIM card, if any
username	username	record the username of the user filling out the form
caseid	caseid	record the unique ID of the case for which the form was filled out

All forms also come with three HELP TABS (one for each tab) describing in details each variable type, syntax, operations, etc.

Main variable types

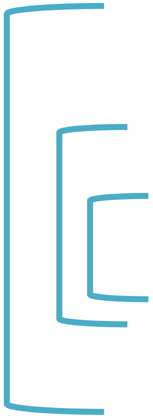
- note
- text
- integer, decimal
- select_one, select_multiple
- datetime, date, time
- image (also: audio, video, file) – incl. pictures ('new') and signatures
- barcode
- calculate

Constraint, relevance, and other options

- **Constraints** to limit the possible range of answers
  - Tailor constraint error message
- **Relevance conditions** to establish skip patterns
  - Can be combined with groups
  - One trick: “required” notes with certain relevance conditions
- Other **options**
  - Hint to give more explanations about the question (italic, subtitle)
  - “Required”
  - Appearance
  - Media to add files

Group and repeat groups

- Group questions for
  - Skipping patterns (relevance condition on the group)
  - Appearance options (several questions on the same screen)
    - field-list
    - list-nolabel
  - Nested groups are OK: they follow the following structure □



- Repeat groups
  - When a set of question is repeated a n number of times, but n is not known in advance (depends on a survey answer), e.g. household roster
  - Calculate fields specific to repeat groups
    - index()
    - indexed-repeat(repeatedfield, repeatgroup, index)
  - AVOID nested repeat groups

Operations and expressions

Details are in the **help-survey** tab of the form. NOT the Excel syntax.

- Arithmetic: + | - | * | div | mod | = | != | > | >= | < | <=
- Logic: and | or | not()
- Functions (not exhaustive): pulldata() | string-length() | coalesce() | min() | substr() | round() | regex() | if() | int() | date() | once(random())

${var} to refer to the question *var* in this form (can be replaced by . if we're in that same line)

NB resource to debug regex: https://regex101.com/

Calculate fields

- Hidden fields to the enumerator
- Proper variables in the survey (like survey questions)
- Used most often when
  - Creating intermediate variables for complex constraints, relevance conditions
  - Pulling data from other datasets
  - Generating random numbers
  - Using automatic random audits

Pull data dynamically

- **From a previous answer to this survey**
  - A direct answer or a calculate field
  - To display in another question or to use in a constraint or relevance option
  - Using **${var}**
- **From an external dataset**
  - To pull a list of choices in a select_one or select_multiple dataset
  - Function : search()
  - Filter down the search() based on previous answers

Free tips

- Groups and repeat groups
  - automatically fill out the label for the end line from the label for the begin line
- Use Excel functions for efficiency
  - filter questions by type
- Variable names
  - Start and end with a letter
  - Only letters, numbers, underscore (_)
  - Reverse english order
  - Less than 30 characters (Stata)

Datasets and Dynamic Surveys

GEORGETOWN UNIVERSITY

Why use server datasets

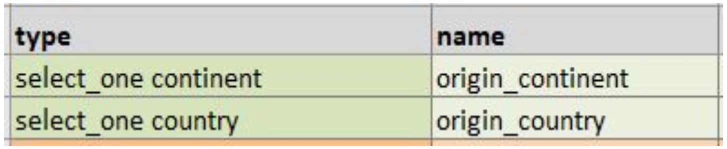
- **DRY = Do Not Repeat Yourself**
  - same information in one place
  - information queried for different tasks (forms)
  - example : list of respondents (sample), list of options

- **Automated updates**
  - same information updated by different tasks (forms)

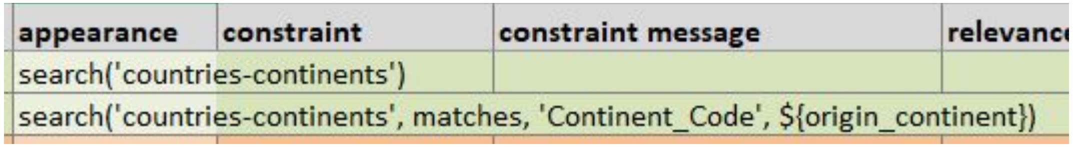
- **Dynamic workflows**
  - queries and updates : audit forms, case management

Datasets to list options

- Survey tab, appearance column:
  - `search(‘dataset’)`
  - `search(‘dataset’, matches, ‘var’, ${value})`

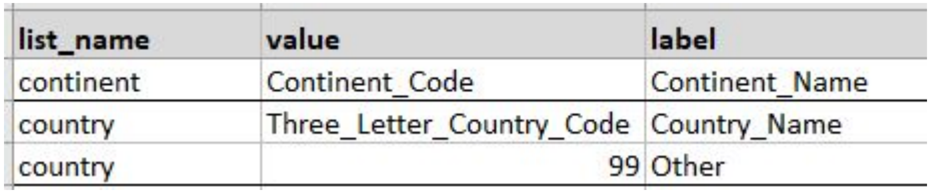


type	name
select_one continent	origin_continent
select_one country	origin_country



appearance	constraint	constraint message	relevanc
search('countries-continents')			
search('countries-continents', matches, 'Continent_Code', ${origin_continent})			

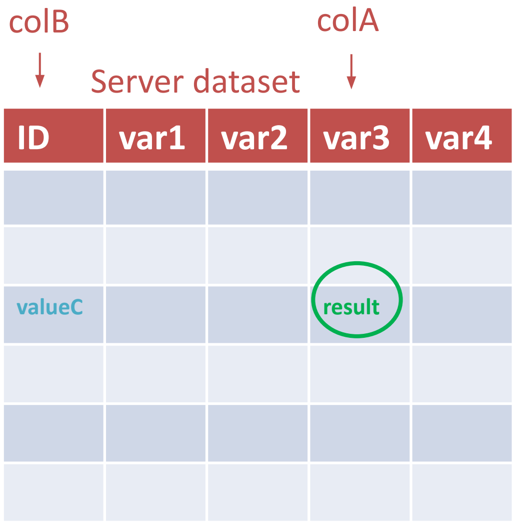
- Choices tab:
  - Value col in dataset
  - Label col in dataset



list_name	value	label
continent	Continent_Code	Continent_Name
country	Three_Letter_Country_Code	Country_Name
country	99	Other

- Dataset
  - Attached with survey or loaded into the server

Datasets to pull data



colB

Server dataset

colA

ID	var1	var2	var3	var4
valueC			result	

In the survey: calculate field

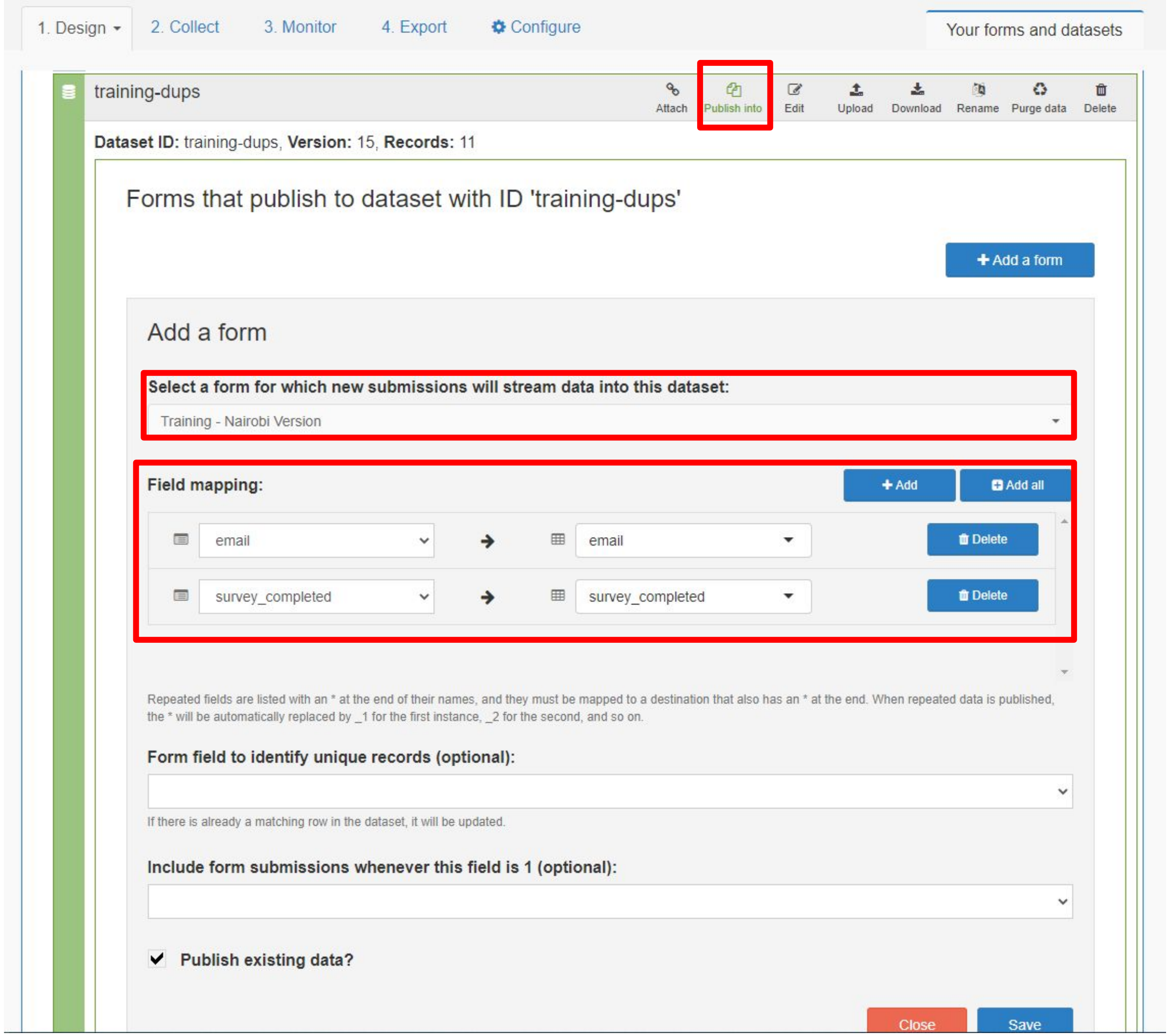
pulldata(dataset, colA, colB, valueC)

```
pulldata(serverdataset,
        'column in server dataset to pull data from',
        'column in server dataset to identify the record with',
        ${value in this survey that identifies the record})
```

Often times the colB and valueC correspond to the same variable (respectively in the server dataset and in the survey) - it is a key that identifies the records. This key is used to know which observation to pull in the column A from the server dataset.

In the SurveyCTO console: attach dataset to the form

Datasets to push data



Dataset ID: training-dups, **Version:** 15, **Records:** 11

Forms that publish to dataset with ID 'training-dups'

Add a form

Select a form for which new submissions will stream data into this dataset:
Training - Nairobi Version

Field mapping:

email	email
survey_completed	survey_completed

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

Form field to identify unique records (optional):

If there is already a matching row in the dataset, it will be updated.

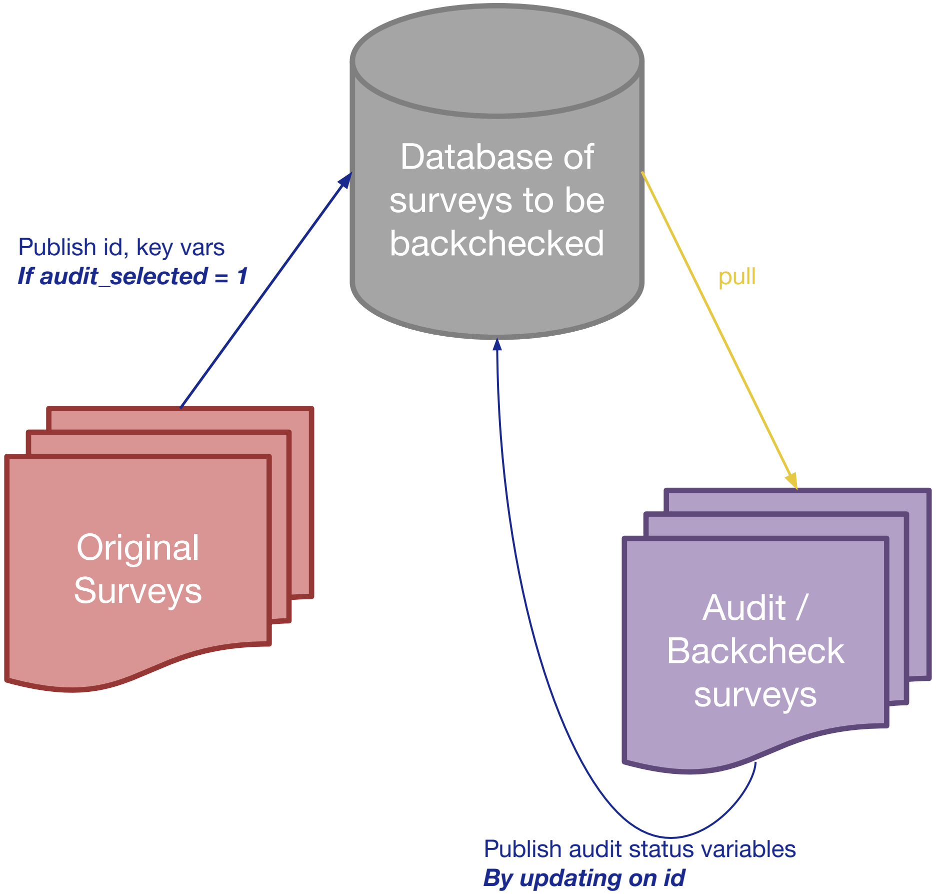
Include form submissions whenever this field is 1 (optional):

☑ **Publish existing data?**

Backcheck or audit coding

1. Randomly select surveys
   - Calculate fields: once(random), audit_selected
   - Calculate field: audit_status
2. Push selected surveys
   - Publish if audit_selected=1
   - List of variables to publish
3. Pull list of surveys to audit
   - Select_one with the search() command
4. Push (update) audit_status

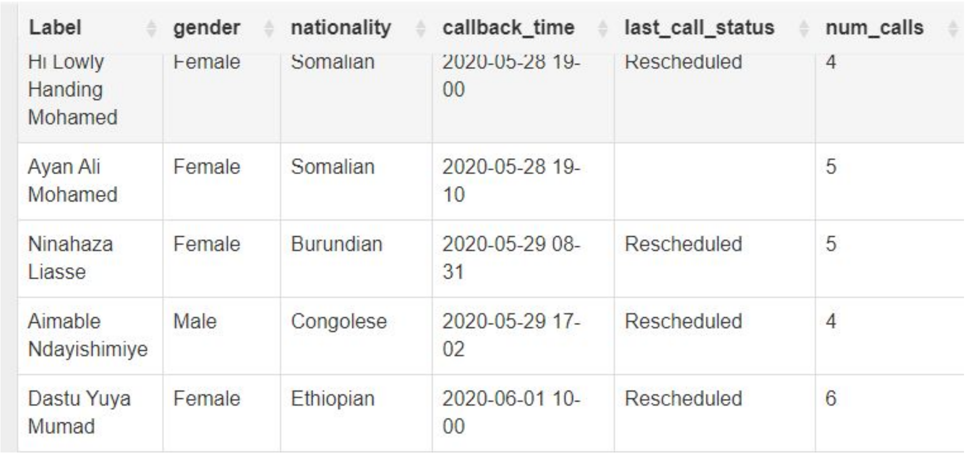
Backcheck or audit coding


Database of surveys to be backchecked
Publish id, key vars
If audit_selected = 1
pull
Original Surveys
Audit / Backcheck surveys
Publish audit status variables
By updating on id

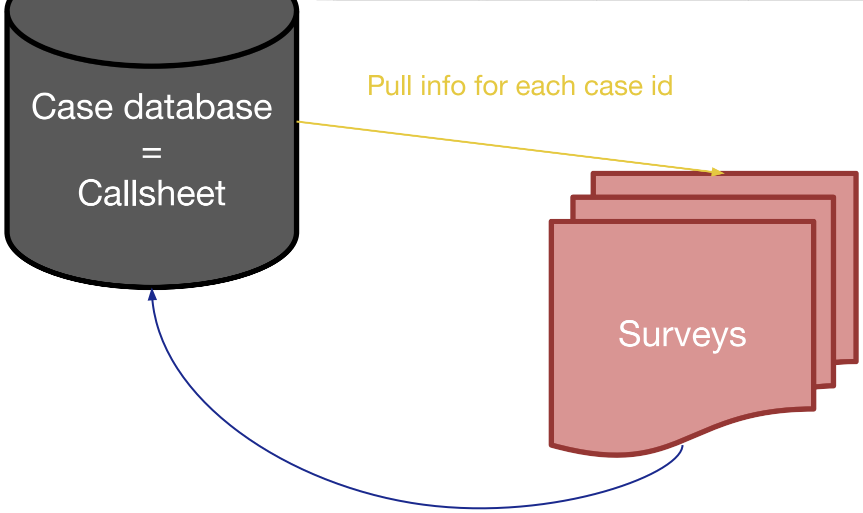
- Publish operations happen on the scto console, export tab (in the server dataset: Publishing > From forms…)
  - Server dataset has a unique id (id)
  - Both forms publish to the same server dataset
    - Original survey publishes: unique id (key, but renamed), audit_selected, audit_status, other relevant identifying variables
    - Audit survey publishes: audit_status
  - Cannot use KEY as the unique id (because both surveys will have each have their own KEY variable)
    - Rename the key published from the original survey into, for example **id** or **key_selected**

- Pull operations happen in the survey code
  - Pull list of surveys to be audited/backchecked with a **select_one** question and a **search()** appearance function
  - Use the search() function to sort out the list of surveys to be audited
    - audit_status = 0
    - Other filters if/as appropriate

Case management



Label	gender	nationality	callback_time	last_call_status	num_calls
Hi Lowly Handing Mohamed	Female	Somalian	2020-05-28 19-00	Rescheduled	4
Ayan Ali Mohamed	Female	Somalian	2020-05-28 19-10		5
Ninahaza Liasse	Female	Burundian	2020-05-29 08-31	Rescheduled	5
Aimable Ndayishimiye	Male	Congolese	2020-05-29 17-02	Rescheduled	4
Dastu Yuya Mumad	Female	Ethiopian	2020-06-01 10-00	Rescheduled	6



Case database = Callsheet

Pull info for each case id

Surveys

Publish survey status variables
By updating on id

- Publish operations happen on the scto console, export tab (in the server dataset: Publishing > From forms…)
  - Case database has a unique id (id)
  - The unique id is used to make sure that the data collected via the survey is linked to that same observation in the database
- Pull operations happen in the survey code
  - Pull data for variables linked to the case needed for the survey

Encryption and Data Safety

GEORGETOWN UNIVERSITY

Why keep your data safe

- **Research credibility**

- **Public trust**

- **Direct harms and risks**

User and device management

- Roles, teams
  - user management
- Devices
  - password protection
- Cloud storage
  - no download, duplication
  - connectivity
- Communications protocol
  - ‘encryption at rest’
  - emails

Data anonymization

- HIPAA criteria
  - safe harbor
- Incidental identification or statistical disclosure (microdata)
- IRB compliance
- Aggregation for reporting, publication

End-to-end encryption

- Create a public/private key pair
  - public key is the lock
  - private key is the key to open the lock

- Implications
  - do not lose the private key (back up)
  - dynamic server usage

Practice : Dynamic Surveys

1789

GEORGETOWN UNIVERSITY

Georgetown University Initiative

gui2de

on Innovation, Development and Evaluation

www.gui2de.org