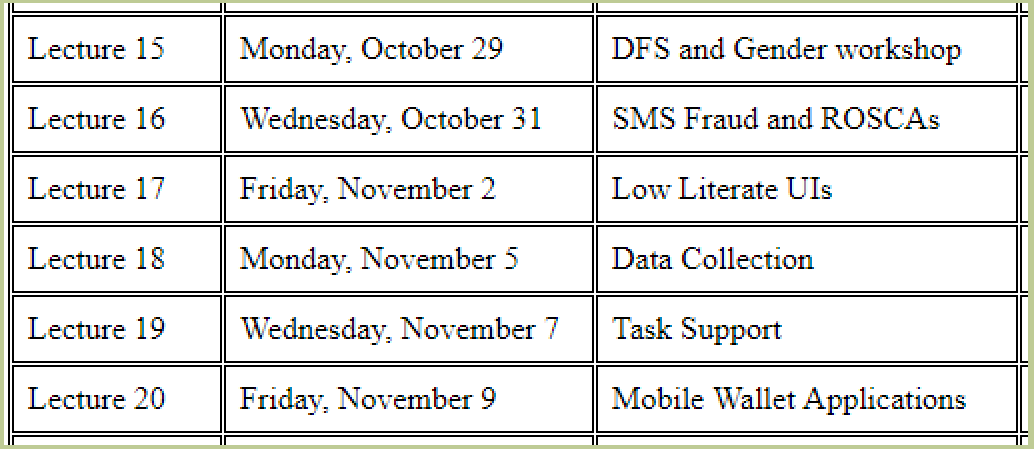
Data Collection

Lecture 18: CSE 490c

OPEN DATA KIT

Announcements

- New Homework Available
  - Paper on Low Literate user interfaces
  - Programming Assignment 2 due Monday
- Lecture Schedule



Lecture 15	Monday, October 29	DFS and Gender workshop
Lecture 16	Wednesday, October 31	SMS Fraud and ROSCAs
Lecture 17	Friday, November 2	Low Literate UIs
Lecture 18	Monday, November 5	Data Collection
Lecture 19	Wednesday, November 7	Task Support
Lecture 20	Friday, November 9	Mobile Wallet Applications

Topics

- Data Collection
- Open Data Kit
- Data Integrity
- Security

Who collects data?

- NGOs
- Civil Society Organizations
- Governments
- Researchers

Data Collection Problem

- Data collectors performing surveys
- A survey is a form with a fixed set of fields
- Advanced version of surveys
  - Skip logic
  - Variable entries (e.g., for each child)

- Paper based approach
  - Create blank forms
  - Fill them in
  - Send them to a central location

Mobile Data Collection Requirements

- Data entry on mobile device
- Submission of data to a server
- Mechanism for installing forms on device
- Offline data entry
- Run on low cost devices
- Low cost software
- Support for large forms

Technology Choices (c. 2008)

- Basic Phones (SMS)
- Feature Phones (Java Phones)
- IVR
- Personal Digital Assistants (PDAs)
- Laptops
- Smart Phones
  - iPhone
  - Android
  - Other OS (Blackberry, Symbian, Windows Mobile)
- PAPER !!!!!

Smart Phone History

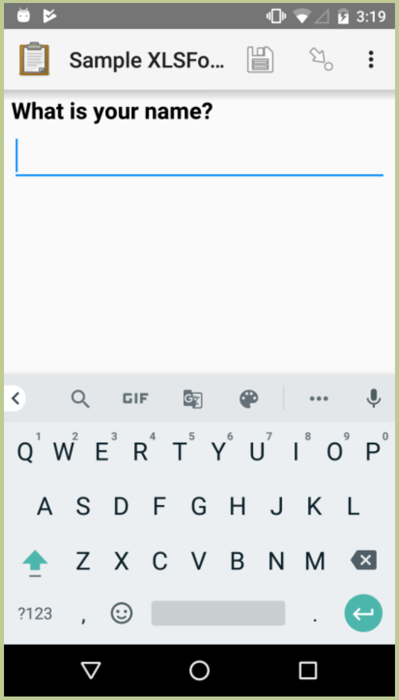
- Oct 2003, Andy Rubin launches mobile OS project for digital cameras
- Jul 2005, Google acquires Android Inc.
- Nov 2007, Google announces Android and Open Handset Alliances
- Sept 2008, first commercial Android Device
- Sept 2005, Apple and Motorola release ROKR E1, the first mobile phone using iTunes
- Sept 2006, ROKR killed, iTunes references unnamed phone
- Jan 2007, iPhone Announced
- June 2007, iPhone released

Open Data Kit

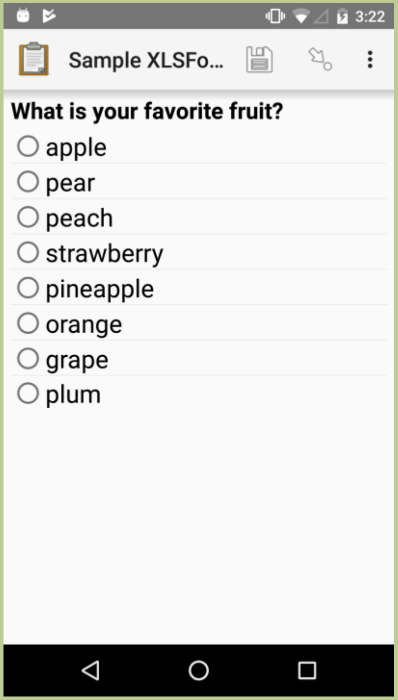
- ODK 1.0: Forms based data collection
- Project launched in April, 2008 while Gaetano Borriello was on sabbatical at Google
- CSE Grad Students Waylon Brunette, Carl Hartung, and Yaw Anokwa joined project as Google interns and brought project back to UW
- Maintained at UW with grad students and professional staff
- Transitioned out of UW as open source projects, 2018

Android app ODK Collect

- Android app for surveys
- Multiple question types
- One question per screen
- Forms loaded from server
- Completed forms uploaded to server
- Forms use XML format from the ODK XForms specification



What is your name?

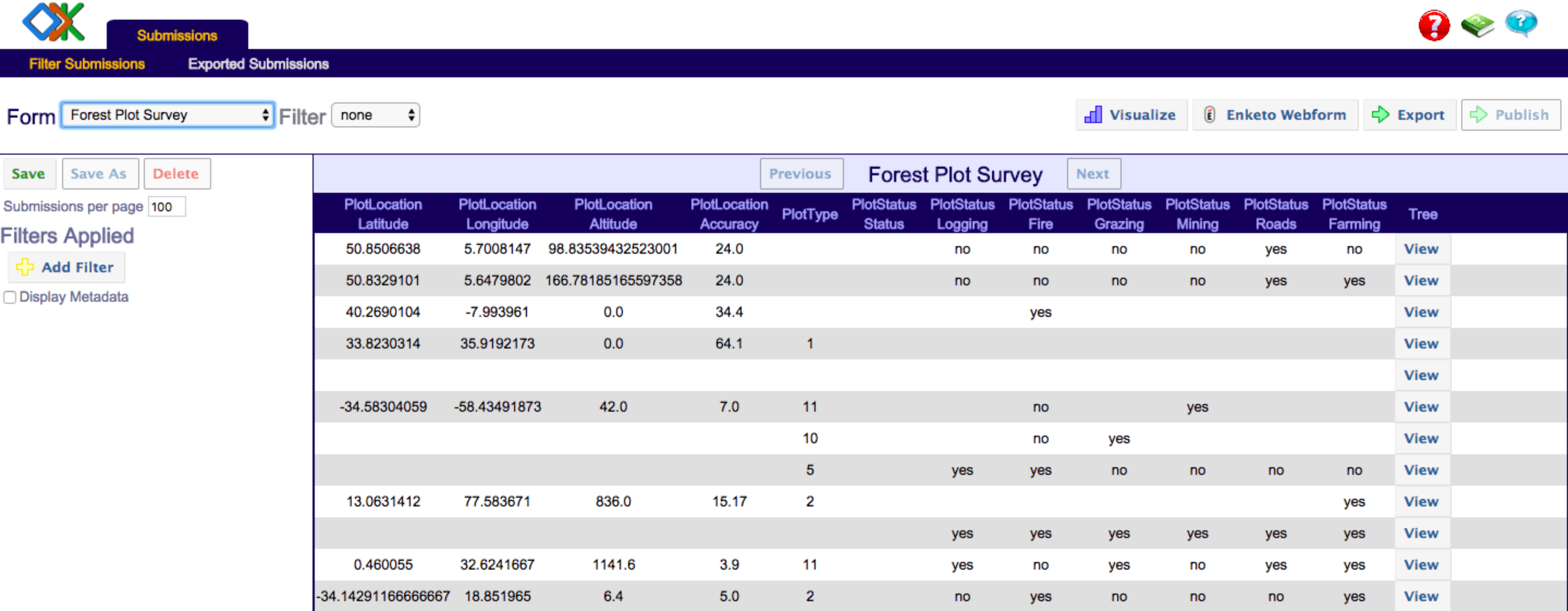


What is your favorite fruit?

○ apple
○ pear
○ peach
○ strawberry
○ pineapple
○ orange
○ grape
○ plum

Backend Server
ODK Aggregate

- Open Source Java application that stores and presents XForm Survey data
- Can be hosted on AWS, Azure, or other local or cloud server



Form: Forest Plot Survey Filter: none

Submissions per page: 100

Filters Applied

□ Display Metadata

Forest Plot Survey

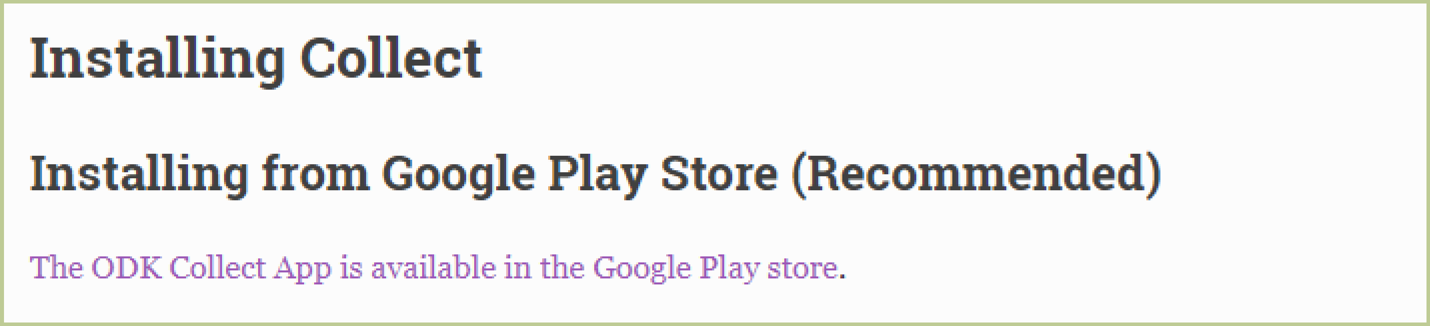
PlotLocation Latitude	PlotLocation Longitude	PlotLocation Altitude	PlotLocation Accuracy	PlotType	PlotStatus Status	PlotStatus Logging	PlotStatus Fire	PlotStatus Grazing	PlotStatus Mining	PlotStatus Roads	PlotStatus Farming	Tree
50.8506638	5.7008147	98.83539432523001	24.0			no	no	no	no	yes	no	View
50.8329101	5.6479802	166.78185165597358	24.0			no	no	no	no	yes	yes	View
40.2690104	-7.993961	0.0	34.4				yes					View
33.8230314	35.9192173	0.0	64.1	1								View
												View
-34.58304059	-58.43491873	42.0	7.0	11			no		yes			View
				10			no	yes				View
				5		yes	yes	no	no	no	no	View
13.0631412	77.583671	836.0	15.17	2							yes	View
						yes	yes	yes	yes	yes	yes	View
0.460055	32.6241667	1141.6	3.9	11		yes	no	yes	no	yes	yes	View
-34.14291166666667	18.851965	6.4	5.0	2		no	yes	no	no	no	yes	View

Form Creation
ODK Build and ODK XLSForm

- ODK Build
  - Interactive forms designer
  - Model of one question per screen
  - Set parameters for individual questions
  - Upload forms to Aggregate, then transfer to device
- ODK XLSForm
  - Surveys generally require lots of iteration in development
  - Better to have a source code model
  - Storage format for forms in Excel
  - One row per question
  - People deploying ODK generally comfortable with Excel

Hello World

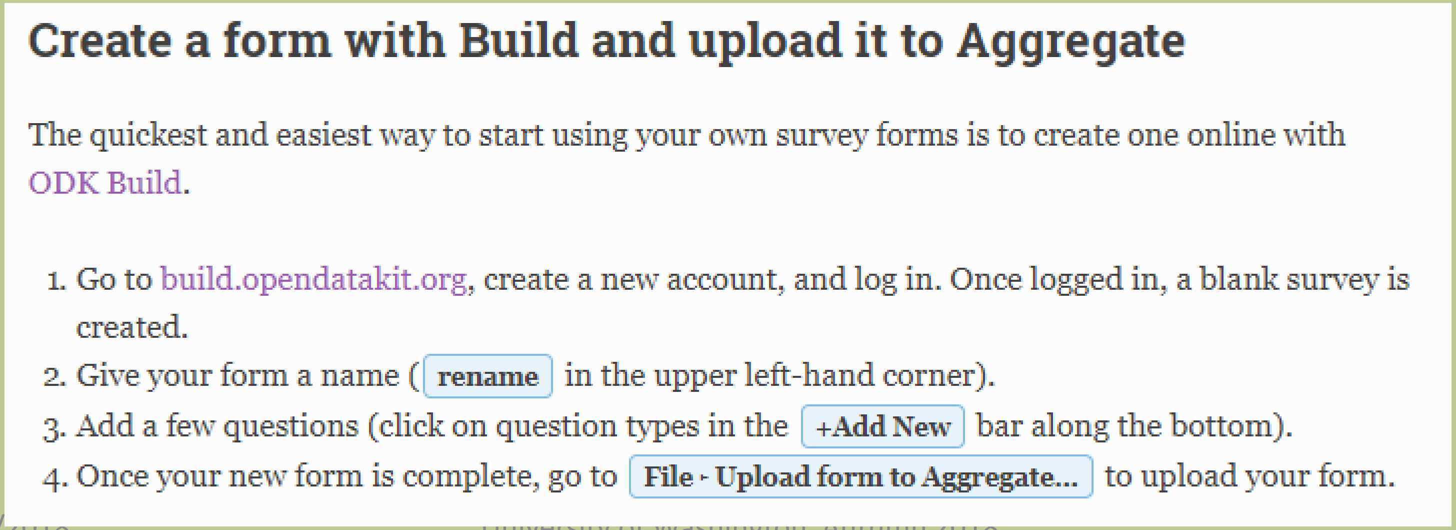
- Easy enough that I can use it



Installing Collect

Installing from Google Play Store (Recommended)

The ODK Collect App is available in the Google Play store.

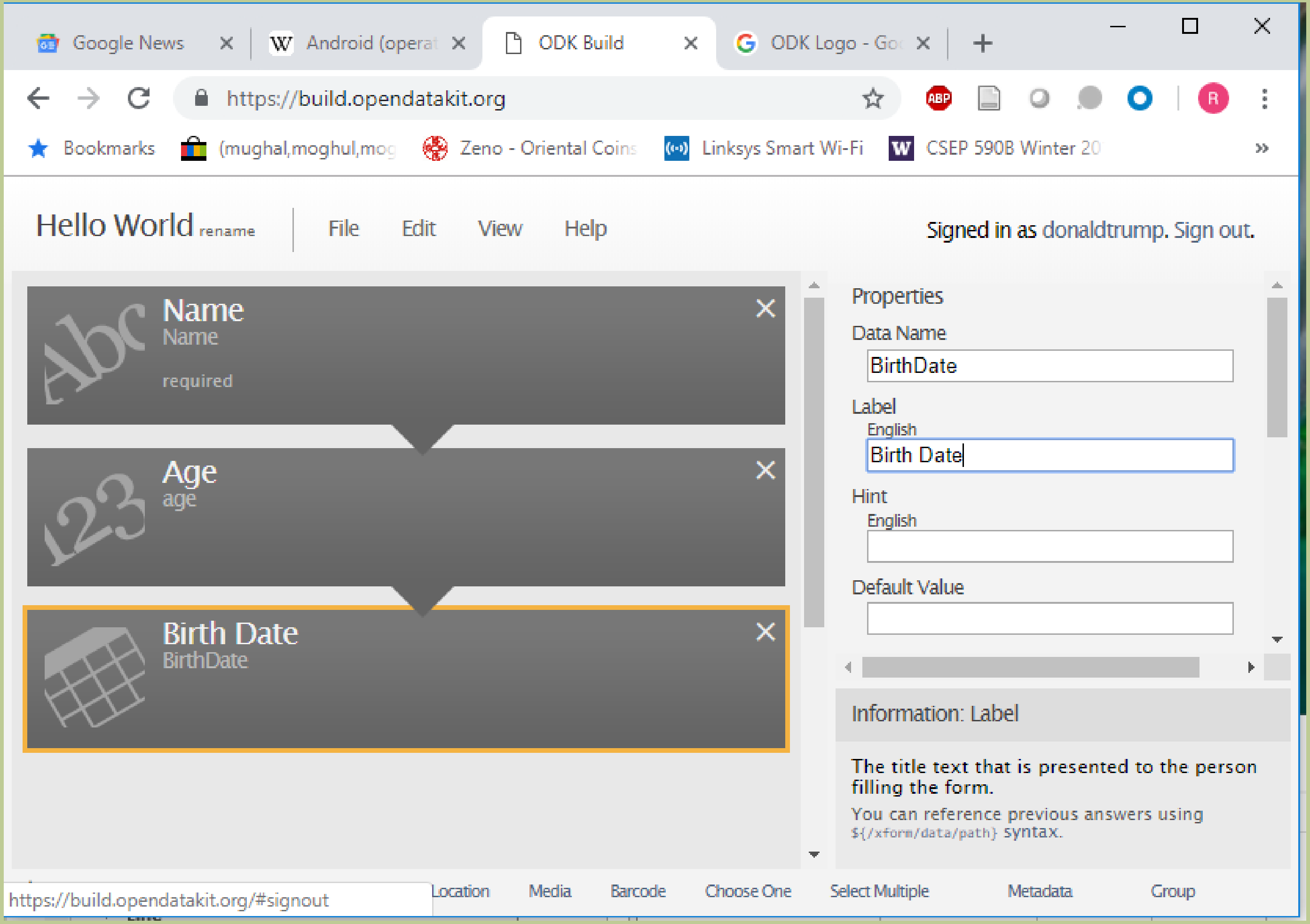


Create a form with Build and upload it to Aggregate

The quickest and easiest way to start using your own survey forms is to create one online with ODK Build.

1. Go to build.opendatakit.org, create a new account, and log in. Once logged in, a blank survey is created.
2. Give your form a name (rename in the upper left-hand corner).
3. Add a few questions (click on question types in the +Add New bar along the bottom).
4. Once your new form is complete, go to File › Upload form to Aggregate... to upload your form.

Creating a form



Properties

Data Name: BirthDate

Label
English: Birth Date

Hint
English:

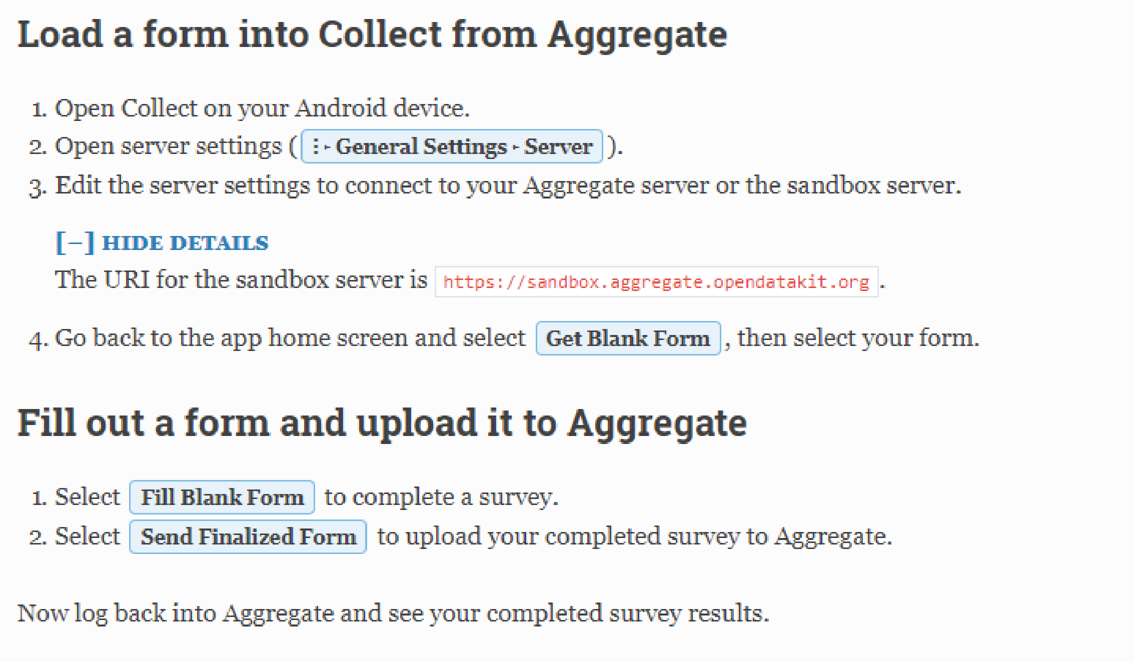
Default Value:

Information: Label

The title text that is presented to the person filling the form.

You can reference previous answers using ${/xform/data/path} syntax.

Collecting Data



Load a form into Collect from Aggregate

1. Open Collect on your Android device.
2. Open server settings (⋮ ▸ General Settings ▸ Server).
3. Edit the server settings to connect to your Aggregate server or the sandbox server.

   [–] HIDE DETAILS

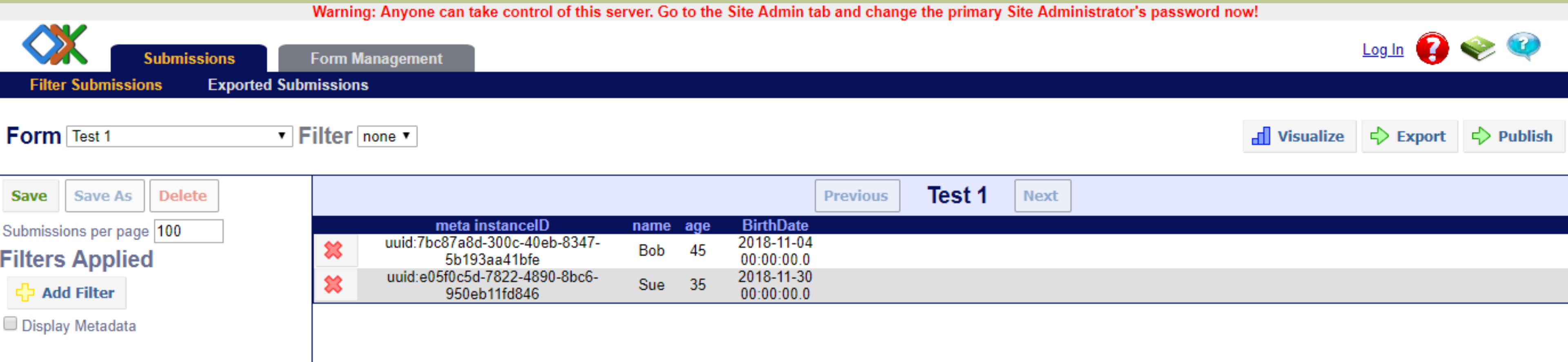
   The URI for the sandbox server is `https://sandbox.aggregate.opendatakit.org`.

4. Go back to the app home screen and select Get Blank Form, then select your form.

Fill out a form and upload it to Aggregate

1. Select Fill Blank Form to complete a survey.
2. Select Send Finalized Form to upload your completed survey to Aggregate.

Now log back into Aggregate and see your completed survey results.



Warning: Anyone can take control of this server. Go to the Site Admin tab and change the primary Site Administrator's password now!

Form: Test 1 | Filter: none

Submissions per page: 100

Filters Applied

□ Display Metadata

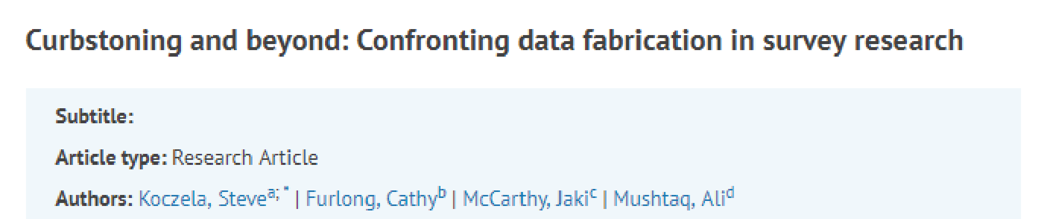
Test 1

	meta instanceID	name	age	BirthDate
✖	uuid:7bc87a8d-300c-40eb-8347-5b193aa41bfe	Bob	45	2018-11-04 00:00:00.0
✖	uuid:e05f0c5d-7822-4890-8bc6-950eb11fd846	Sue	35	2018-11-30 00:00:00.0

Examples

- Forest mapping in the Amazon
- Berkeley Human Rights Center
    - Post conflict assessments
- D-Tree International
    - IMCI Protocol
- Verbal Autopsy

Data Integrity



Curbstoning and beyond: Confronting data fabrication in survey research

Subtitle:

Article type: Research Article

Authors: Koczela, Steve[a; *] | Furlong, Cathy[b] | McCarthy, Jaki[c] | Mushtaq, Ali[d]

- Traditional problem with surveyors
  - What if enumerators cheat
  - This even has a name: curbstoning
- Data collectors make up data instead of doing surveys
- Methods for detection
  - Made up data often is not random enough
  - Consistent omission of data
    - E.g., Missing informant phone numbers
  - Made up data may not have appropriate means
    - E.g., Across sample, 40% of households might be away, while faker only identifies 20% of households being away

Automated Quality Control for Mobile Data Collection, Birnbaum (2012)

Tools for detecting bad data

- Compare distribution of each collectors value with composite of other collectors
  - Multinomial means and variances
  - Very accurate if number of fakers is low

- Big brother
  - Record question times
  - Record GPS locations

Computer Security for Data Collection Technologies, Cobb, Sudar (2016)

Security

- Threat Model
  - CIA Goals: Confidentiality, Integrity, Availability
  - Adversaries
    - Governments, Thieves, Hackers, Partners, Enumerators
  - Potential Threats
    - Unauthorized access
    - Entering fake data
    - Coercing enumerators
    - Theft
    - Legal access to data
    - Instability of application
    - Information leakage on device
    - Fake ODK applications

Interview Study

- What do users care about
  - Data Loss
    - Encryption not used, because it risked data loss
  - Integrity
    - Enumerators answering “no” to shorten interview
  - Exploited data
    - Generally less of a concern, but there are some very sensitive ODK deployments
- Importance of device management
- Different levels of technical expertise
- Ethics board considerations
- Context: Comparison with Paper

Digression: About Names

- What does the Open in ODK mean
  - Open (Data Kit) vs. (Open Data) Kit

- ODK 1.0 and ODK 2.0
  - ODK 1.0 and ODK 2.0 are different projects that address different use cases
  - Naming suggests that the latter is replacement for the former