Mobile SMS survey data management and preservation

Inna Kouper
Charitha Madurangi, Kunalan Ratharanjan, Tom Evans, Beth Plale

Indiana University

This work is supported in part by NSF grants BCS 1026776 and SES 1360463
Mobile phone penetration

Mobile phones in Africa

Cell Phone Ownership Surges in Africa

Adults who own a cell phone

<table>
<thead>
<tr>
<th>Year</th>
<th>U.S.</th>
<th>S. Africa</th>
<th>Ghana</th>
<th>Kenya</th>
<th>Tanzania</th>
<th>Uganda</th>
</tr>
</thead>
<tbody>
<tr>
<td>2002</td>
<td>10</td>
<td>8</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>2007</td>
<td>64</td>
<td>89</td>
<td>83</td>
<td>82</td>
<td>73</td>
<td>65</td>
</tr>
<tr>
<td>2014</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

Note: U.S. data from Pew Research Center surveys.
Source: Spring 2014 Global Attitudes survey. Q68.

PEW RESEARCH CENTER

http://www.pewglobal.org/2015/04/15/cell-phones-in-africa-communication-lifeline/africa-phones-7/
Project: Agricultural Decision Making and Food Security in Africa

The project examines how small-scale farmers adapt to food and climate variability.

It integrates physical models with real-time environmental data and weekly farmer decision making in individual fields.

Farmers are asked weekly about their decisions to plant, grow, and harvest and about weather conditions.
TextIt SMS Platform

- Cloud-based commercial SMS service
- Builds surveys through GUI
- Provides summaries and minimal analytics
- Data can be downloaded manually or via API
Text messaging (SMS) to collect data

**Pros**
- High-frequency, automated data collection
- Large sample sizes
- Relatively low cost

**Challenges**
- Set up learning curve
- Data may be incomplete due to non-response or lack of SMS credits
- Typing increases errors
- Some programming expertise required
- Possibilities depend on platform
Data completeness and errors

Did it rain this week? Please answer “yes” or “no”:

- Yes, yes, YES, YESS, YAS
- No, no, No!, N0, NO, no rain
- 50
Data retrieval via API

```
"count": 23,
"next": null,
"previous": null,
"results": [{
  "flow_uuid": "0f6ded36-0ed4-4c10-86f5-eaa6b0fd0f9c",
  "flow": 29541,
  "run": 2110573,
  "contact": "7ed0ad89-969d-4cce-b012-c7d91d9ae731",
  "completed": true,
  "values": [{
    "category": {
      "base": "No"
    },
    "node": "b7f97b23-3c3c-4bba-b34c-11d997dba5e2",
    "time": "2015-04-29T10:09:49.894Z",
    "text": "NO",
    "rule_value": "NO",
    "value": "NO",
    "label": "mmf harvest"
  }
}]
```
Our Approach: Data Management and Preservation Pipeline
Database Design

- **Raw** preserves original data
- **Split** organizes data into flows (surveys), runs (responses within one survey) and contacts (respondents)
- **Pre-processed** minimizes workload for queries
Storage and Access Architecture

Web Page  

REST tool  

Backend API  

Raw Databases
- zambia
  - Raw files
- kenya
  - Raw files

Split Databases
- zambia
  - flows
  - runs
  - contacts
- kenya
  - flows
  - runs
  - contacts

Processed Databases
- zambia
  - contact_responses
  - ......
- kenya
  - contact_responses
  - ....
Data Retrieval Pipeline

- Pipeline is fully automated
- Data from TextIt is ingested weekly
- Metadata from TextIt is stored in addition to data (survey responses)
- Data is restricted to users within university
- Data can be downloaded via dashboard or http request
Metadata Improvements

**Retrieved from TextIt**
- IDs
- Flow names
- Total number of runs (responses)
- Number of completed runs
- Variable labels

**Added by the team**
- Country
- Season
- Creator
- Date created
- Run start and end date / time
- Flow type
- List of questions
- Farmer contact and location
Data / Metadata GUI

- View and modify metadata details for flows and contacts
- Download inactive contacts for a given time period
- Download farmer response details for given question and time period
Monitoring Dashboard

Zambia – Farmer SMS Data

Weekly Overview [May 15th 2017 - May 22nd 2017]
- Flows Overview:
  - Total Flows: 436
  - Open Flows: 43
  - Started Last Week (May 15th 2017 - May 22nd 2017):
    - Flows: 4
    - Resend Flows: 2

Respondents (Contact) Overview:
- Total Contacts: 3047
  - Active Last Week (May 15th 2017 - May 22nd 2017): 327
  - Inactive Last Week (May 15th 2017 - May 22nd 2017): 891
  - No response for a month (April 23rd 2017 - May 22nd 2017): 453

Daily run completion rate [May 14th 2017 - May 22nd 2017]

Flow completion rate [May 14th 2017 - May 22nd 2017]

- Flows details
- Farmer contacts details
- Flow completion rates
- Inactive / non-responsive contacts
What have we achieved?

• Automated pipeline enables consistent long-term preservation of raw data and full control over it.

• Pre-processing and transformations help to quickly retrieve subsets of data for analysis.

• Improved metadata facilitates search, access, and future re-use.

• Organized storage enables future visualizations and integrations.

• Interface and dashboard makes interactions with data easier, no technical skill required.

Curation, curation, curation!
Challenges

• **Security** – need to add proper authentication and access options for research team and the public

• **Data cleaning** – some can be automated, but most is still manual

• **Maintenance**
  - Changes in commercial platforms (e.g., APIs) require modifications of backend and data
  - Staff is needed for ongoing curation and technical maintenance
  - Data preservation is not central to research projects

• **Analysis and dissemination** – still done outside of the pipeline, not reproducible
Future Work (Questions)

• What standards in data documentation and preservation can help to improve this work?
• How can SMS data be integrated with other types of data (e.g., sensor data or household interviews)?
• What analytical products are most useful and for what types of stakeholders?
• How can we measure the impact from curation?