Understanding Challenges in the Data Pipeline for Development Data

Fahad Pervaiz

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2019

Reading Committee:

Richard Anderson, Chair

Kurtis Heimerl

Abraham Flaxman

Program Authorized to Offer Degree:
Paul G. Allen School of Computer Science & Engineering

University of Washington

Abstract

Understanding Challenges in the Data Pipeline
for Development Data

Fahad Pervaiz

Chair of the Supervisory Committee:
Professor Richard Anderson
Paul G. Allen School of Computer Science & Engineering

The developing world is relying more and more on data driven policies. Numerous development agencies have pushed for on-ground data collection to support the development work they pursue. Many governments have launched efforts for more frequent information gathering. Overall, the amount of data collected is tremendous, yet we face significant issues in doing useful analysis. Most of these barriers are around data cleaning and merging, and they require a data engineer to support some parts of the analysis. This thesis aims to understand the pain points of cleaning development data. It also proposes solutions that harness the thought process of a data engineer to reduce the manual workload of the tedious process of cleaning such data. To achieve these goals, two research areas are critical: (1) to discern current data usage patterns and to build a taxonomy of data cleaning in the developing world; and (2) to create algorithms to support automated data cleaning, which target selected problems including matching transliterated names. With these goals, this thesis will empower regular data users to easily do the necessary data cleaning and scrubbing for analysis.

TABLE OF CONTENTS

Page

List of Figures iv

List of Tables vi

Chapter 1: Introduction 1

Chapter 2: Background 4
2.1 Types of Datasets 4
2.2 Challenges in Development Data 6
2.3 The Development Data Pipeline 7

Chapter 3: Identified Problems in the Data Pipeline 12
3.1 Methodology 13
3.2 Findings 15
3.3 Discussion 26
3.4 Conclusion 30

Chapter 4: Case Study: Improved Cold Chain Information System 32
4.1 Summary 32
4.2 Introduction 32
4.3 Developing a CCIS for Laos 38
4.4 Field Experience 48
4.5 Discussion 50

Chapter 5: Case Study: Developing Cold Chain Data Standard 52
5.1 Summary 52
5.2 Introduction 52
5.3 Immunization 54

5.4 Cold Chain Information System 55
5.5 Software Context and Challenges 57
5.6 Data Standards 63
5.7 Discussion 65
5.8 Conclusion 67

Chapter 6: Case Study: An Assessment of SMS Fraud in Pakistan 68
6.1 Summary 68
6.2 Introduction 68
6.3 Related Work 70
6.4 SMS Fraud Background 72
6.5 Data Collection 73
6.6 Data Analysis 80
6.7 Results 83
6.8 Qualitative Analysis 88
6.9 Discussion 91
6.10 Conclusion 95

Chapter 7: The Scalability of SMS Reporting Systems: Integrating with National Health Information Systems 96
7.1 Summary 96
7.2 Introduction 96
7.3 Background 98
7.4 Case Studies 102
7.5 Barriers to Responding 113
7.6 Discussion: Lessons at Scale 117
7.7 Conclusion 123

Chapter 8: Name Resolution for Data Cleaning 125
8.1 Summary 125
8.2 Introduction 125
8.3 Related Work 126
8.4 Methodology 127
8.5 Evaluation 129

8.6 Discussion 133
8.7 Conclusion 134

Chapter 9: Conclusion 135

Bibliography 139

LIST OF FIGURES

Figure Number Page

2.1 Various stages of the data pipeline along with a list of challenges at each stage. 8

3.1 Summary of specific challenges grouped into categories. Challenges in bold text were mentioned by three or more participants. 15

4.1 (Top) fridgetag 30DTR with no alarms in last thirty days, (bottom) fridgetag 30DTR with two high temperature alarms, yesterday and two days ago. . . . 36

4.2 Typical rural health center staffed by approximately four health workers . . 39

4.3 Lao PDR, the NIP office is located in the capital Vientiane. 39

4.4 System architecture: Health workers send data via SMS to an android phone that syncs with cloud system. Cold chain manager and SMS moderator manage these systems using their respective web interfaces. 42

4.5 Vaccine refrigerator labeled with A for reporting. 48

4.6 Five valid messages that all have the same semantic meaning. each message tells the system that refrigerators A and B had zero alarms and that the current stock levels for pentavalent and pneumococcal are 20 and 30. 49

6.1 Screenshots from Safe SMS app, showing how to label a conversation 75

6.2 Number of conversations that were available on each user's phone, the ones they uploaded and the conversations that were labeled by them 80

6.3 Presence of different features with important ones highlighted 87

7.1 Wheel (Close up of the SMS report job aid for DSS) 106

7.2 A BHU dispenser explaining how he uses the SMS reporting wheel to create a SMS message 108

7.3 A 30 day temperature recorder (30DTR) in a Lao refrigerator showing two high alarms in the last two days 108

7.4 The Lao SMS Immunization Manager (SIM) showing a list of incoming SMS reports of October 2014 112

7.5 An example message showing a sample of special characters accidentally typed during a training in Laos 122

8.1 Sensitivity of string matching algorithms against the Niger transliterated locality names 130
8.2 Sensitivity of different heuristics defined, using a combination of string matching algorithms, against the Niger transliterated locality names 132

LIST OF TABLES

Table Number Page

3.1 Distribution of participants by role and organization type. 14

6.1 Summary of User Labeled Data. 81

6.2 Summary of Fraud Types. 83

6.3 Examples of fraudulent messages that were collected. English translations are given for messages sent in Roman Urdu. 84

6.4 Summary of Heuristic Results 86

8.1 List of Heuristics. 127

8.2 Examples of failed matches 133

Chapter 1

INTRODUCTION

Global development organizations and governments in developing regions increasingly rely on data to inform policies that are intended to improve health, education, employment, human rights, and economic development. Significant amounts of data are collected and analyzed by a wide array of stakeholders (*e.g.*, government agencies, non-governmental organizations, global development donors, and social enterprises) to conceptualize, implement, evaluate, and support policy decisions. Often these stakeholders have different goals and strategies for data collection and analysis. For example, while government agencies often collect a wide range of data to get an overview of different development indicators, non-profit and non-governmental organizations gather data to identify insights on a specific topic or to measure the impact of a specific intervention. For these reasons, attempts to collect data are often disorganized and in silos, which results in the availability of copious amounts of poor-quality data that is inconsistent, isolated, and lacks structure and standards, making it hard to clean and analyze. In many instances, data is collected without much consideration and planning, and often it remains little used or forgotten, resulting in time and cost intensive replicated efforts to collect the same data for different purposes.

Although it is desirable to combine those datasets that cover similar domains, merging, transforming and cleaning datasets containing different schema and types is a non-trivial process that requires a substantial amount of effort. Data processing also involves multiple stakeholders, both within and outside the organization, and often data goes through multiple processing stages, including importing, merging, rebuilding missing datasets, standardizing and normalizing, duplicating, and exporting. Processed datasets then undergo cycles of analyses and visualizations. Many of these stages, from data collection to data visualization,

and processes within these stages are isolated from each other, making it easier for people to work on data independently. However, this isolation also means that people who work on one aspect (*e.g.*, data cleaning) might have no control over processes in other stages (*e.g.*, collection) and may not fully understand the context in which the data is collected, cleaned, transformed, or analyzed.

Several Information and Communication Technology for Development (ICTD) researchers have investigated challenges in collecting development data [45] and designed new tools that are more suitable to gather development data [115, 61, 23, 21]. However, the research that examines challenges in different stages of the data pipeline is largely absent. While some researchers have provided taxonomies for dirty data (inaccurate, incomplete, and inconsistent data) in the context of systems in the developed world [81] and identified challenges such as schema reverse engineering or lack of constraints on data types [131], no work presents a collective list of challenges in various stages of data pipeline for data collected in low-resource environments of developing countries. These datasets and the underlying contexts introduce new challenges in the quality of collection, processing, and analysis due to barriers surrounding literacy, infrastructure, culture, partnerships, and funding. Thus, there is a need to build a checklist of challenges that classifies different barriers that are encountered while processing development data.

This thesis takes a deeper dive into the challenges in development data by examining various projects that collect and analyze data with different objectives. I learned about challenges of development data first hand from a series of projects on which I worked. To gain a more comprehensive view of challenges for development data, I conducted an interview study to understand list of pain points in the development data pipeline. This pipeline is complex with various stakeholders involved at different stages, making it harder to streamline data tools. To handle this complexity, development data systems require a diverse and overlapping set of tools to handle specific challenges and needs.

In this thesis, I first present the background around data challenges, which I substantiate with related work. This is followed by a qualitative study, conducted to understand and

emphasize the most frustrating pain points in the development data pipeline. Subsequent chapters recount case studies implemented in the domains of health and finance for developing countries, and each exposes a wide range of data challenges. Chapter 7 outlines two case studies that illustrate how errors originate within human-typed data that is transmitted over SMS. All this work is already published [122, 6, 119, 120, 121]. Ultimately, in the conclusion I argue that significant work remains in order to streamline the data systems and pipeline for developing world data. My analysis of the case studies as related to the findings from the qualitative study form the foundation for new recommendations for addressing the challenges within development data systems and pipelines.

Chapter 2
BACKGROUND

In this section, I provide the background for why development data is different from other types of data and how this distinction complicates data processing. Beyond this I establish the stages of the data pipeline and offer examples of the issues that each stage aims to resolve.

2.1 Types of Datasets

Within international development there are two predominant types of datasets: administrative data and survey data. Administrative data contains indicators that are tracked and reported regularly such as once a month. Often, employees at low level facilities (*e.g.*, primary health center) fill out paper forms at the end of every month to submit to higher-level authorities, who in turn aggregate data from other low level locations, prepare reports, and send the data and reports upstream. The data propagated in this fashion eventually finds its way to a country-level database. In contrast, survey data captures a snapshot of the current state of affairs for specific questions. The data is often collected and reported by a trained surveyor of a central authority. Government agencies generally rely on administrative data supplemented by surveys, while non-governmental agencies typically conduct their own surveys and partner with government agencies to access relevant administrative data. At times, there is a need to merge the two types of datasets for more detailed analysis, such as combining crop surveys with malaria patient data to evaluate if stagnant water in fields results in the rise of mosquito-driven diseases. To elaborate more on complexities, we provide some real-world examples for merging administrative datasets.

- The District Health Information System [109] in Pakistan tracks patient visits from all health centers for about 20 diseases. The data is collected using paper-based forms that

are sent to supervisors within the administrative hierarchy. After manually aggregating data from multiple sites in the administrative division, the data is eventually converted into an electronic format—an excel file without any macros—by manual transcription at the Tehsil (sub-district) level[1] [108]. The electronic records are sent through the administrative hierarchy, until they are integrated into the national level records nearly a month after the initial data collection.

- Cold Chain Equipment Management (CCEM) [116] is a Microsoft Access based tool that records the status of refrigerators and vaccines to track any vaccine stock outs and refrigerator malfunctions for better delivery of immunization. To configure CCEM for a country, all facilities and refrigerators are added to a Microsoft Access file which is then distributed to all the district offices. To combine the country-level data, merging of these files created difficulties in past deployments of CCEM because some districts made local changes to the fields in the Access files.

- Zambia's health information system tracks cases for several diseases. This dataset has different administrative hierarchies for different years. In the last few years, Zambia has created 11 additional districts. Since the data is aggregated over districts, it is hard to do temporal analysis as some of the districts have split into smaller units.

- Cold Chain Equipment Inventory (CCEI) is a DHIS2-based open source information system for refrigerator and cold chain tracking. It was deployed in Laos by UNICEF in collaboration with the Lao National Immunization Program [6], independently of another DHIS2-based disease tracking system deployed by the health ministry. After the success of CCEI, there was an attempt to merge the two systems. However, the administrative hierarchy in the two systems was different and also had different transliterations of the same words. It took a month of troubleshooting to create a

[1]Tehsil is an administrative unit in some South Asian countries that usually has one city along with a number of towns and villages

mapping between the two hierarchies and spellings.

Government collected data quality varies from country to country in the developing world. Some countries have strong National statistic offices that collect and generate reliable data, others have weak collection agencies with only old census data, and some countries have no data collection agency at all.

2.2 Challenges in Development Data

As demonstrated through the examples above, data collected in low-resource environments (i.e., development data) often has a unique set of characteristics, such as sparse data, inconsistent values, and unstructured and unknown coding schemes which arise from practices in collection, as well as limitations in infrastructure and organizational factors. This leads to collection and curation of messy datasets that have high variability in formatting and integrity, contrasting with consistently structured datasets from contemporary, conventional information systems [8]. The data pipeline challenges in the international development domain can therefore be much more time consuming to resolve.

Data-driven decision making for international development is becoming more common. Unfortunately, underdeveloped communities have not attained a modern data infrastructure for a myriad of reasons, including limited education, lack of access to computing devices, irregular power supply, and constrained bandwidth. Moreover, further issues arise due to limited or insufficient training of data collectors. While constraints such as access to electricity and computing devices are manageable, other critical challenges such as ensuring quality training of people commissioning, collecting, curating, and analyzing datasets is not straightforward [107]. This is especially important if data is flowing up a hierarchical chain instead of being reported directly at the national level. This increases the likelihood of errors manifesting in the data as more people at all levels are involved with typing and aggregation.

Development data poses processing challenges because typically the data is entered on paper forms, especially in rural areas, which clerical staff who digitize it a later point [92].

The systems used for data entry often do not have any constraint checks, like a plain excel file, which results in spelling errors, inconsistent values, and formatting issues. In many cases even the organization of such datasets is inconsistent, *e.g.*, having values such as name or date as part of the field names [160]. This is in sharp contrast to the data entry in an information system where checks from the system help mitigate standard errors and inconsistencies.

One of the main problems is the ever-changing requirements for collection. Over time, more and more indicators are added to the collection process with novel requirements from new partners. This causes a huge burden for lower level staff to fill out more forms. For example, Pervaiz et al. [121] revealed that the staff of a basic health facility in Punjab, Pakistan must fill out seven different forms with redundant information at weekly, fortnightly, and monthly intervals. This presents a major obstacle for staff to complete all the forms accurately and keep up with the latest guidelines to record additional indicators.

All of these requirements lead to inconsistent, poorly recorded, incorrectly formatted, missing, and fabricated data. Additionally, we have witnessed reorganization of administrative hierarchies in developing countries. Since all the data collection is tracked and identified through this hierarchy, temporal analysis becomes very difficult. The fluctuating nature of the developing world's data collection environment makes data from such environments very different from the rest of the world.

2.3 The Development Data Pipeline

Figure 2.1 shows the major stages in development data pipeline: data collection, data cleaning, and data analytics. We situate our research in a body of related work examining and addressing barriers encountered by researchers, practitioners, and data analysts in different stages of the pipeline.

Data collection provides an early opportunity to identify errors and collect clean data, however, it is a prime entry point for mistakes and errors into the pipeline. Due to lack of resources and poor infrastructure in developing regions, large scale data collection typically happens on paper forms. For example, at rural health facilities in Pakistan, all attendance

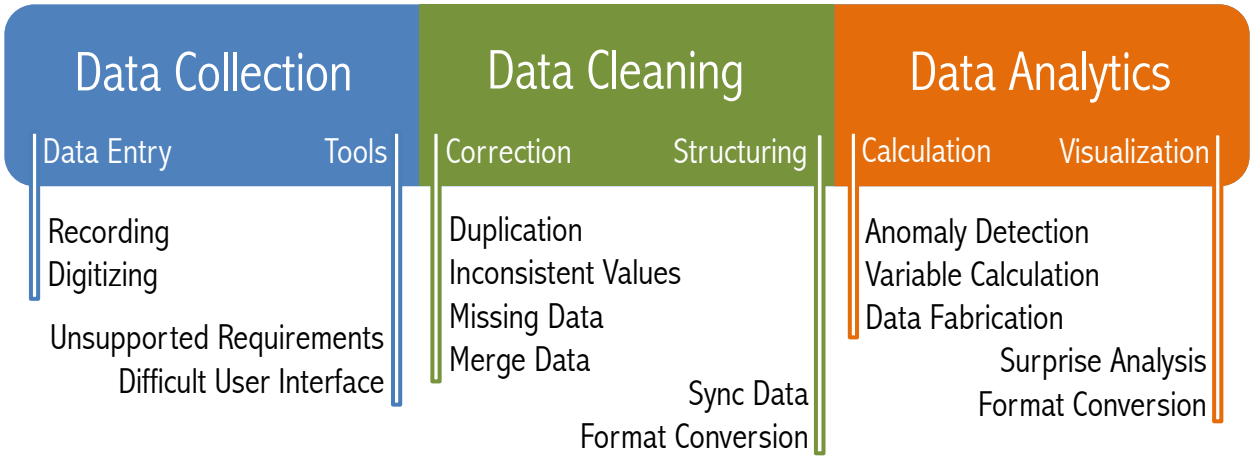

Figure 2.1: Various stages of the data pipeline along with a list of challenges at each stage.

records are kept in large paper based log books. Field workers fill out these forms manually and send them to a local data collection center. The data on the forms is then manually transcribed by data entry workers into a computer, often with poor accuracy [45]. Several researchers have designed solutions to address this gap by integrating the paper collection process with the digital collection process. These solutions range from using a smartphone camera to digitally capture a paper form [115, 44] to using cloud and crowd for processing forms [33, 91].

Several mobile-based tools are designed to cater to the needs of remote digital data collection in low-resource settings. Open Data Kit (ODK) [61] and CommCare [153] are mobile data entry tools that can be configured with any form to collect data. ODK is a popular tool in the development community, with over 210,000 downloads on Google Play store from over 130 countries. DHIS2 [79] and OpenMRS [145] are customizable information systems with dashboards and reporting tools. These also have mobile data collection components, although they are mostly used with web-based collection forms. Even with improved mobile tools, digital data collection faces challenges [32, 114]. These tools can also introduce new issues in the data, such as duplicate entries or the generation of multiple versions of databases due to some unsupported requirement. For example, ODK-1 currently

does not allow editing of an existing form. If a user wants to add a new field to an existing form, then a new form must be created. This action creates a new database that will require merging with the original form's database to get a single consistent dataset. Several tools and methods have been proposed to improve the error rates at the collection stage [118]. For example, Shreddr [33] is a tool that takes an image of a paper form and uses the crowd to digitize it. In this way, the errors being introduced at the transcription step are reduced or completely eliminated because multiple people transcribe the same entry, reducing the probability of transcription errors. Usher [31], a mobile data collection system, gives users the option to select values from previous entries, thus reducing potentially reducing spelling inconsistencies.

Data cleaning is the critical, yet it is the most time-consuming component of the data pipeline. It is well known that 80% of data processing time is spent on cleaning and aligning data [39]. This ranges from removing dirty data (*e.g.*, missing values, wrong or inconsistent values, and unexpected formats [81]), data errors, duplicate values, and inconsistent data structures to merging the data from multiple sources. Several researchers have studied the data cleaning process, laying out the scope of the problems and current approaches with respect to data warehousing [131, 81, 66]. However, a fundamental limitation of these works is that they target well-structured databases, which are the backbone of information systems. Our work, on the other hand, focuses on datasets from resource-constrained settings in the developing world where even the structure of datasets is very messy and disorganized [160]. Fixing incorrect or inconsistent values requires domain knowledge, making it hard to automate the process unless the tool is domain specific. Also, finding a domain expert for manual processing presents challenges to organizations and governments that have limited resources.

Tools have been built to simplify the cleaning process. For example, duplication is an issue in which the same entity is present in different parts of the data and could be represented in various formats [34]. Several works have designed solutions to remove duplicates mostly by defining a similarity function to find a cluster of matching entities [99, 85, 34, 55],

however, the problem remains intractable since some entities might have different representations in different datasets (*e.g.*, one data might have 'John Doe' while the other may have 'Doe, John'). Some tools use pattern detection, from a user's interaction with data transformation, and heuristics to infer the intended structure and data domain [75, 133]. Wrangler and Potter's wheel are interactive tools that show users the transformations as they progress. Their goal is to engage regular users in the cleaning process without making it overly complicated [67]. Similarly, Google Refine [89] includes the user in the cleaning process but relies on learning by example. It displays only a snapshot of the transformations to the user for verification. Although these tools work well for datasets that have well-designed schema and structure, often they yield unsatisfactory results when used for international development datasets.

Different development datasets use different structures to collect and store data for specific purposes [95, 119]. The data pipeline consists of various tools that expect data to be in a certain format and structure. The presence of several tools, which are intended for different stages of the pipeline, poses structural and formatting issues for data processing since each tool has a unique format that hinders the smooth flow of data across tools [72]. Information systems like DHIS2 and OpenMRS solve some of these problems by providing more complete solutions without the need for multiple tools. However, there are several disjointed data collection efforts in developing regions that do not rely on setting up one large system. Prior work has proposed a tidy format that represents data in a manner that is easy to import for further analysis [160, 95]. Having all tools follow the same tidy formatting standard is not practical due to the well established formats the tools follow. Also, the tools are in competition with each other. Organizations handle this by writing custom scripts for data conversion.

The lack of standard identifiers that could be used as a primary key makes it harder to merge datasets from different data sources. The foremost strategy is using a common attribute that is present in both datasets. Usually this is the name of a location or person, since national ID or location code are not commonly part of datasets. Local names that

are transliterated into Roman characters often results in multiple spellings [51], and because of this, the probability that multiple datasets may have different spellings for names is high, which makes this approach less rewarding. For example, 'zango algabit ii' and 'zongon algabid' are two spellings in different datasets representing the same town in Niger. A second strategy, known as probabilistic record linkage, takes a wider range of possible identifiers and calculates a probability of the two records being the same [14, 96].

There have been a range of systems that aid data exploration through a visual interface [75, 133, 60]. These tools support what is called the sensemaking model [25]. This helps the layman user to better understand data for cleaning, integration, and other calculations [30]. Moreover, feedback to a field worker through visualization has been proven very useful and effective in motivating the worker [46].

Development data goes through several stages of the pipeline before it is analyzed and presented to policy makers. Several researchers have studied individual challenges in different stages of the data pipeline, however, until now, no single researcher has evaluated the stages of the pipeline as a whole in one consolidated work. It is important to highlight a summary of the burdensome barriers and study how challenges in one stage will impact the processing in another stage. This thesis aims to establish a checklist of challenges in the data pipeline and how challenges at different stages impact other stages of the pipeline, as well as identify the underlying causes for these challenges in order to understand how these challenges propagate.

Chapter 3

IDENTIFIED PROBLEMS IN THE DATA PIPELINE

Development data cleaning is highly distributed because of multiple stakeholders involved in the process. Stakeholders will work on different stages of the data cleaning process, some in collaboration and others in isolation. Data goes through multiple stages of the processing pipeline, allowing the process to be easily segregated. This compartmentalization also means that those doing data cleaning might not have any control over data collection and may not understand the context in which data was collected.

ICTD literature demonstrates copious amounts of research on the data collection stage of the pipeline, yet this research neglects to analyze the processing that occurs after collection. Building a taxonomy for data cleaning will help identify the biggest gaps in this process. Although some researchers have provided dirty data taxonomies with respect to developed world data [81] where the challenges are around schema reverse engineering or lack of constraint on data types [131], to my knowledge, there is no work on data cleaning taxonomy for developing regions.

This chapter explores the relationship between existing development data collection and cleaning processes through an analysis of interviews with stakeholders from various organizations. I interviewed eight employees of international development organizations and five government workers in Pakistan. This approach offered insight into both the global perspective, as the people from development organizations had worked with projects from well over 100 different countries, as well as a more in depth administrative perspective from a specific country. The goal of this chapter is to compile the pain points expressed by the practitioners and identify the gaps that have the greatest impact. From my analysis, I propose a basic taxonomy of development data cleaning and identify areas where support tools can help achieve

better cleaning of such development data. In the last decade, incredible improvements have been made on the data collection side of the pipeline, however, I argue that insights on how data goes through various processing stages for analysis will make efforts to create better tools for data cleaning and achieve better analytics more feasible considering the complexity of the challenges involved.

3.1 Methodology

In the summer and autumn of 2017, I conducted semi-structured interviews with representatives at three international development organizations and the provincial government of Punjab, Pakistan who were involved in commissioning, collecting, cleaning, analyzing, and visualizing datasets in resource-constrained settings of developing regions. Among the three international organizations, the first was an international private foundation with projects commissioned in over 45 countries in different areas (*e.g.*, health, education, financial services); the second organization was a global health organization that has worked in over hundred countries and have access to several datasets; the third organization was a global health research institute that access to data from almost all developing countries. Since there is an acute shortage of data scientists and experts [12], particularly in the space of international development, I used convenience sampling to recruit experts in these organizations and the government by leveraging my network of past collaborators. I then used snowball sampling on the networks of participants to recruit more participants that could provide insights about challenges in different stages of data pipeline.

Participants worked at multiple departments in the organizations and the government. Some participants, as illustrated in Table 3.1, were managers for various projects and supervised a team of data experts while others were individuals working on a specific data collection project as a domain expert or general data expert. Participants had varying degrees of experience with data processing. The sample of managers and individual contributors with varying experiences allowed me to get a well rounded prospective as certain details like issues around partnerships were dealt by managers while trivial cleaning tasks

Organization Type	Manager	Analyst	Total
Non-Government	3	5	8
Government	2	3	5
Total	5	8	13

Table 3.1: Distribution of participants by role and organization type.

were handled by individuals on the project. Similarly, decision making tasks like designing a schema, defining the tools for collection, and setting up guidelines for processing were handled by individuals with more experience. All interviews were conducted at the workplace of participants.

I conducted face-to-face interviews with representatives at the Non-governmental organizations in English and with Punjab government officials in Urdu. I am fluent in English and Urdu. The interviews lasted about 40 minutes, on average. I took detailed notes on paper for all interviews and requested permission to audio record participants. While representatives at the international development organizations gave the permission, government officials in Pakistan expressed hesitation in being recorded. I thus audio recorded only participants working at the organizations. The interview questions focused on the tools, workflows, and processes participants use for data collection and processing, barriers and frustrations they experience in the data pipeline, challenges they face in managing and using legacy datasets, opportunities and challenges in building and maintaining partnerships, and features they wish to see in the next generation of data processing tools. The study was approved by my institution's Institutional Review Board (IRB).

I transcribed audio recordings and translated field notes to English. I used an iterative approach to generating interview questions. After each interview, I discussed results with other researchers and updated my questions. I subjected my data to thematic analysis as outlined by Braun and Clarke [20]. All researchers, involved in the project, participated in the coding process and iterated on the codes to identify themes until consensus was reached.

Data Collection	Data Entry	• **Human error** • **PDF Extraction** • Requirement Gathering • Data Collectors not trained well	• Biases in reporting • Bad Handwriting • Vague questions • Device error
	Tools	• **Dashboard view is lacking** • **Tool is slow for big data** • Create filter for reports	• Adapting paper form in digital tool • Can't run data query easily
Data Cleaning	Correction	• **Replace values** • **Unit conversion** • **Remove duplicates** • Fix spelling errors	• Map multiple fields • Clean open text • Fix identifiers • Normalize aggregates
	Structuring	• **Merge data from different sources** • **Restructuring data format** • **Code conversion/terminology mapping**	• **Splitting the aggregation** • Connect data from same source
Data Analytics	Calculations	• **Test data for accuracy** • **Fill missing data** • Identify outliers	• Adjust values to remove biases • Calculate derived variables
	Visualizations	• **Eyeball data for data accuracy** • **Calculate daily/weekly report manually**	

Figure 3.1: Summary of specific challenges grouped into categories. Challenges in bold text were mentioned by three or more participants.

3.2 Findings

My analysis revealed several specific challenges in the data pipeline for international development datasets. Although the exhaustive list of challenges is long, Figure 3.1 highlights the major challenges reported by my participants that were problematic or time consuming for them. These challenges are sorted in three stages of the data pipeline, described earlier in the chapter. The bold text indicates challenges that my analysis reported as major themes. I discuss the three stages below along with partnership as a fourth theme that spans all three stages.

3.2.1 Data Collection

Data collection, either directly by surveying participants or importing data collected in the past, is the first step in the data pipeline that generally requires careful commissioning, planning, schema designing, as well as training of enumerators. The process of data collection in the past decade has shifted from digitizing paper forms to using digital tools. Hence, I split the issues into two categories, one related to the data gathering process itself and the other being the issues with tools.

Data Entry: The most common frustrations that the participants brought up were human errors in collecting data from the field. Participants highlighted multiple reasons, for example, spelling mistakes and entering incorrect data, that contributed to human errors. Participants noted that these errors stems from lack of good training, lack of motivation to enter values correctly, and limited textual and digital literacy skills impeding data collectors to understand and enter data correctly. A participant working at an international development organization explained several human errors in selecting identifiers in a paper-based household survey being conducted in Latin America:

> *We have signs in the form to pick identifiers for each facility or segment or household. It always happen that they don't pay attention and input the wrong thing. Then it's a challenge to fix it. We look at the start time of the survey and then try to match up the linked surveys. Sometimes we send emails to supervisors in the field for asking what's going on.*

Participants expressed extracting data from legacy textual formats as another common source of frustration. They needed access to legacy data for building longer time series datasets for a specific region, a process that they reported as extremely challenging due to the lack of access to raw structured datasets. Participants stated how external organizations often give access only to prepared reports instead of raw structured data, either because the data collection is done by a local organization who is willing to share only PDF reports or

because the raw data does not exist at all since it was either purged after the reports were created. A participant in the international development organization stated:

> *"The most extraordinarily challenging dataset was our 'India Vital Registrations' constructed from particular series of PDF reports. Extracting this data was terribly difficult because these reports were for each state in India for each year. PDF reports were very long, and had slight variations in the format and page breaks. We had to get an army of people for PDF extraction and those [extracted numbers] might not all be right."*

Several participants highlighted how the data entry operators for their projects struggle with understanding unclear handwriting while transcribing paper-based forms. Often, transcriptions of forms with poor handwriting leads to more errors when data entry operators have no or limited context in which the data is gathered by the data collectors. Participants also highlighted how some questions in survey forms are poorly worded and lacks local context, making it difficult for data collectors to input correct value. A participant described how his form had a vague question about the availability of electricity at health centers in rural regions and a 'yes' or 'no' response, that led to inconsistent responses from data collectors since the intermittent availability of electricity was coded differently by different collectors; some answered it as a simple 'yes' or 'no', while others vaguely answered as 'majority of the day' or were specific like 'more than 8 hours'. Participants noted how collectors are often not well trained due to limited time and resources for training. Sometimes the training material is not well thought out since its developed by technical people who take things for granted and miss important details. A participant, who worked on analyzing the refrigerator's upkeep and capacity for proper vaccine storage at health centers, stated:

> *"Translating paper based data into an electronic tool is very challenging. The number one problem is training of the data collectors by technical people. The data collectors who note the serial numbers [of vaccine refrigerators] for the age of*

the equipment has a huge implication on the data cleaning. People's handwriting is a huge factor. This all could be easy if data collectors are well trained."

Some participants mentioned how training is often not well designed and implemented or yield unexpected issues when data collectors are not well-informed about how to use technology. For example, a participant noted how GPS devices sent erroneous data and showed a health facility in a different district and at times in a different country. This happened because the collector did not wait long enough to get the accurate measurements even after being trained on how to tune the device. Some participants highlighted how the respondents of their surveys fabricated data for personal or political reasons, leading to collection of data with huge biases. For example, a participant experienced men over-reporting their height and women under-reporting their weight in a global obesity survey. Another participant noticed supervisors in a health center submitting incorrect data about the population they serve or the number of medicine they dispensed to get more funding for their facility. Several participants echoed sentiments like *"numbers can be inflated or deflated for political reasons."*

Tools: Participants also highlighted challenges in transitioning data collection system from paper to digital. A participant noted how data collectors with years of experience with paper-based tools find it challenging to map the questions when paper-based forms are converted into phone- or tablet-based forms. Participants also noted a range of issues when the data collected digitally or gathered by transcribing paper forms is exported to electronic formats in specialized systems for analysis. Three participants felt that the tools they were using slowed down while processing large datasets. A participant working at the government stated how he *"cannot use excel because it takes long time to load and use R to first remove duplicate entries to process datasets in excel."* Many participants noted the absence of easy to use filters and detailed statistics in the dashboard of tools that were built specifically for their projects. For example, although a participant had access to high-level preset statistics, the system did not let him perform calculations to identify poorly performing districts or

health centers, forcing him to export the data to excel and creating new reports. Like this participant, other participants noted how some tools provide no access to raw data, as a result participants resort to workarounds such as creating reports periodically, and downloading and merging them to recreate raw datasets. There are several factors that contributes to unusability of a tool that goes beyond the affordance of the tool. An NGO participant, who is fluent in most MS Office tools found it harder to work with MS Access due to its lack of user friendly design and relied on a tool that she is more comfortable to work with. She stated:

> *"I can't run these queries easily in MS Access tool so I have to export data into excel but everything else related to the facility vanishes since the export operation emits the supporting data on facility. I have to work both with MS Access and Excel in parallel. My approach to data cleaning is limited by my abilities in the tool. I sort refrigerators by age and pull it out in excel, and simultaneously look at records in MS Access to make sense."*

3.2.2 Data Cleaning

Data cleaning, that includes data correction and structuring, is part of the data pipeline that focuses on fixing basic issues in the data.

Correction: Participants mentioned replacing values, converting units, and removing duplicates during the data correction stage. For example, government participants expressed how the data received from health facilities in rural regions was imprecise. They had to manually correct human errors by following up with the supervisors in rural regions when daily or weekly numbers in the report do not add to the monthly reported numbers. They reported carefully inspecting the data to find inconsistencies. This exacting and time-consuming process often did not work when dealing with large datasets. Participants working at the international development organizations recounted using a set of well-defined internally created instructions (*e.g.*, replacing missing data with non values) for data cleaning so that

datasets are better supported by data analysis tools. A participant stated:

> *"When they have missing data and they mark it with like NA or NON or something and we want to replace that with the right value for us. Like whatever Python stores it as missing value to make sure that it becomes an integer column."*

Most participants working with the government expressed frustrations in removing duplicates in the incoming data. Some of them reported doing it manually, for example, by looking up a person's national identifier in attendance data that should appear once within a day but ends up appearing multiple time for the same day. When datasets did not have such identifiers, which often was the case, removing duplicate entries was an extremely challenging process. Some participants used 'Remove Duplicates' feature in Microsoft Excel to clean their data, but reported being unsatisfied by how little customization they could do while using this feature. Several participants also performed unit conversions on the datasets and considered it an important step that required more context as well as access to documentation. A participant stated:

> *"Sometimes we get data as a ratio of patients to catchment population, so the numbers are actually in rate and you have to know that from the documentation that they are not numbers. You have to multiply that rate by a population to get numbers. Some form of metric conversion is really common while cleaning datasets."*

Some participants discussed other challenges in data correction steps. For example, a few participants reported how health facilities and houses are assigned a standard numeric identifier. They reported how errors made by workers at these health facilities while inputting the standard identifier propagates throughout the system when concatenating datasets with some overlaps using these identifiers. Several participants reported challenges in cleaning values from open text fields that they support so that data collectors have flexibility in inputting information. A participant noted:

> *"Data collectors have to record measurements sometimes in milliliters, sometimes in other units, and at times in shorthand because doctors don't always use the same terminology. So, we have to leave a giant open text box that we go clean later but its a big pain."*

Even with predetermined codes open text fields have other challenges, for example, misspelling of codes. A participant highlighted how correcting spelling errors is non-trivial:

> *"The biggest problem is when we have excel based data source that ingest data from a system that does not enforce data integrity. For example, one dataset was related to breastfeeding. Participants had to write 'breastfeeding' in one of the text-boxes, and they misspelled 'breastfeeding' in 47 different ways."*

Structuring: Data structuring is an integral part of data cleaning process. All participants acknowledged merging data from different sources as the biggest challenge they face in the entire data pipeline. Data merging process includes syncing data from various current systems as well as merging from datasets collected in past. One of the big challenge in data merging is *name resolution problem* which arises because of transliteration of local language names into English or French [128]. People employ various transliteration systems which results in multiple spellings for the same location. Moreover, geographic boundaries as well as names for administrative units change over time due to geopolitical factors. These issues can be avoided by using standard identifiers for each facility or person or location, but errors in entering identifiers were also reported as very common. A participants working at an international development organization stated:

> *"Definitely the thing we discussed is that when we have different data sources and we need to merge them, it comes up so often that even the administrative unit has been spelled differently."*

Participants highlighted how even after successful merging of datasets based on the name of region, facilitiy or person, there are challenges in mapping various codes and identifiers

to match international standards. A participant explained how different countries either use different versions of international classification disease (ICD) codes as cause of death or devise their own coding scheme, making it difficult to compare datasets for cross-country analyses. Similarly, another participant highlighted how numeric scales could confuse data collectors, resulting in inconsistent representations. The participant shared an anecdote where her organization decided to use '0 to 5' as codes to represent varying degree of availability of electricity at a health center. While data collectors in some districts used '0' to record 'unknown', ennumerators in other districts used '5' to signal 'unknown'. These issues were also reported as prevalent for other fields like GPS coordinates and dates, for example, some countries follow 'mm/dd/yyyy' format while others use 'dd/mm/yyyy'. These cases often require manual follow-ups and metadata lookups outside the dataset to determine a correct mapping, increasing the complexity of merging such datasets and requiring countless man-hours to untangle these differences.

A few participants also highlighted other structuring issues. For example, participants found it challenging to split aggregated data to get numbers for units downstream the administrative hierarchy (*e.g.*, splitting national-level dataset to get data at district or provincial level). A few participants used coping strategies such as splitting the aggregated numbers proportionally based on population or other indicators that properly represent the ratio among subregions, a technique that comes with its gaping limitations. Participants also reported facing difficulties in connecting data from different forms that belong to the same survey. This issue is largely due to the implementation of digital survey applications like ODK or SurveyCTO that generates a separate dataset for each form, making it complicated when multiple forms are filled for same entity like household survey and immunization survey. It is cumbersome to connect the two datasets without a common identifier.

3.2.3 Data Analytics

Calculations: Several participants reported using strategies to test data coming from cleaning stage for its accuracy. Mostly, they reported relying on other available datasets to calcu-

late the approximate range of values to verify the data. The verification is based on either a simple distribution model or a complex model. A participant elaborated this strategy:

> *"For verification, we rely on other data for same country, region, age, gender, or time period. By this, we can see if the numbers fall under the same realm of plausibility. We check for outliers, things that looks really crazy."*

This strategy could serve development practitioners only when similar data is available, which may not often be the case. Another strategy used by participants is to fill missing values by deriving data from other sources. A participant shared how his team used this strategy by creating a complex model:

> *"We had all sorts of other problems because entire state years were missing in a random fashion. It looked like a sparse matrix. Fortunately, we had access to national level data by year. We tried filling missing values by matching state level data with national level data using a very complicated model."*

Less frequently stated frustrations were with calculating derived variables (*e.g.*, calculating age from date of birth) that consumes time. Some participants mentioned doing calculations to find outliers in the data to clean obvious errors. Other participants reported evaluating biases in the data that could arise due to over or under reporting and made adjustments to minimize the effect of those biases. These biases arise due to complex socioeconomic, cultural, and political factors in various contexts, for example, misreporting by health facilities to receive more funding for vaccines, self-reporting by subjects to avoid societal shaming, or using different reporting practices for noting cause of death. A participant highlighted biases in health data measuring obesity levels:

> *"It is challenging to adjust the self reported data on weight, quantify the biases, and remove them. We leverage measured data that we have for other similar countries in the same time period. Once we have measured data source from the*

> *past and a self-report dataset in the same country, we match those up controlling for change across time and see what is the difference. Often times, the self-reported data is under measured."*

Visualizations: Participants often used visualizations for cleaning data. Several of them reported using graphs, either on a dashboard or self-generated, to eyeball the data for any anomalies. A participant explained:

> *"Once we have clean data that is approved by the analysts, I scan the data spatially. I look at visualizations to make sanity checks. If there are data points [for health facilities] out in the middle of the ocean then there is something wrong with the data."*

Like this participant, other participants also noted *"plotting the data on a time series"* to look for outliers and *"weird data points."* While several participants appreciated using visualizations to identify data anomalies, some participants working at the government found it time consuming and annoying to create visualizations *"every week requiring countless man hours,"* and wished for easy to create visualization components integrated in the data cleaning and analysis tools.

3.2.4 Partnerships

Partnership is the keystone for development where international development organizations work together with local governments, non-profits, non-governmental organizations, as well as for-profit social enterprises. These partnerships are pivotal to understand local contexts, as one participant working on a large-scale project to eradicate polio highlighted:

> *"A country changed the strategy from giving oral polio vaccine to giving intravenous polio vaccine and they updated their immunization cards. It became a problem because we were not aware. It is important to have a working relationship with the Ministry of Health. If you don't have that relationship, you would*

not know that you have to adapt your survey, and will consequentially lose a lot of data. I think, context is the most valuable thing you could ever have when you are working on a project like this."

While many participants underscored the importance of partnerships, they often attributed these partnerships as a source of frustration around data collection and processing. Availability of several datasets with varying credibility from different partners complicated data processing for some participants. Several participants complained about the lack of easy access and limited permissions on datasets shared across organizations. They noted how some datasets either have limited data (*e.g.*, high level statistics in a report format) or only a subset of raw numbers (*e.g.*, immunizations done at a health center without supporting data such as the expected number of vaccinations). Additionally, partnerships were also reported as a source of stress when the objectives of partners were misaligned. A manager working at an organization stated:

"Part of the problem is that this [data cleaning] is all very fragmented [among partners]. Its all very chaotic and you are just trying to do it and I think it is a sad situation. There is a lack of network of all resources and skills that is needed for data processing."

Participants reported how this struggle often translate into non-cooperation between partners. For example, a participant reported that members of the partner organization were unwilling to use MS Access tool that she configured because of disagreements about the selection of technology. Another participant noted how the partner organization declined stewardship of the data once the project was concluded, since they did not want to give advantage to their competition.

Participants expressed several key differences in the challenges faced by international non-governmental organizations as compared to a local government. They noted that a government has more control over the incoming data as they dictate the reporting requirements and has easy access to data entry sources that are mostly owned or funded by the government.

This results in the government's ability to manually follow up with local facilities to fix the errors and get the needed metadata for further data processing. Participants observed how the same convenience is not generally available to non-governmental organizations since they rely on the existing government data reporting network and often lack funding to support their own data collection process. Even when they have funds, the data is often provided by the government health centers using processes beyond their control. This hampers their ability of data processing despite having better resources and ability to build complex verification models, compared to simpler verification processes and tools used by the government. Participants working at the international organizations expressed their inability to control low-level cleaning tasks (*e.g.*, removing duplicates and replacing values) that were often dealt by local workers employed or funded by the government, resulting in errors being propagated in other stages of the data pipeline.

3.3 Discussion

Compared to data collected in developed regions, data collected in low-resource environments often presents data collection, cleaning, and analysis challenges that are substantially difficult to address due to poor availability of tools and resources, lack of planning, and uncoordinated efforts of multiple stakeholders working in the space. This is especially challenging in the context of last mile problems when limited training has been provided to the people responsible for data collection and curation. Poorly designed forms, schemas that lacks structure, enumerators with limited training and skills contributes to collection of data that has incorrect values and inconsistencies. The use of diverse and misconfigured tools, that often expect data in different formats and standards, results in the creation of messy and poorly formatted datasets.

Government institutions play a critical role in data collection and processing. Semi-functional governments lead to uncoordinated attempts to digitize data where data is collected only by some government offices and is often left to gather dust instead of being properly analyzed. Moreover, there are usually multiple organizations pushing for data com-

pilation in a given region; some compile the data themselves while others partner with or fund local organizations for this task. Often the same set of data enumerators collect data for multiple partners on various projects. The lack of communication and coordination among different development agencies result in the availability of excess data for some indicators and no data for other indicators. Due to this, multiple overlapping datasets exists with varying degrees of accuracy, making it hard to produce reliable analysis from such datasets.

Data collection is the first stage where imprecise data is introduced, which can be prevented with better training. The reporting person or surveyor often does not understand the significance of following given instructions while entering data. This could be avoided by coming up with a framework or guidelines for training material to better target the common scenarios that cause bad or malformed data entries. Moreover, the data entry person could be motivated with examples of how their data will be used to help the problem and how their mistakes could lead to mishaps like vaccine outages or disease outbreak in their region. This will introduce sense of ownership and credit in making the impact on public health or relevant domain to data entry. Lastly, investing more time in generating a form that limits entries by listing valid options and showing informative error with on the spot validation will significantly reduce trivial data correction tasks in later stages of pipeline [44, 61].

Three key issues emerged from my analysis as the source of major frustration for development data: (1) extracting data from legacy textual formats, (2) merging data between existing large data sources, and (3) validating data accuracy. Since a large amount of historical data and data provided by external organizations is often available only as PDF reports, extracting data from textual files is a critical and indispensable step in the data pipeline for development datasets. Extracting numbers automatically from such PDF reports is a non-trivial undertaking due to lack of standardization and details being tangled in the text of the reports. Even if a data analyst writes a tailored script for a specific set of reports, examining numbers in the right perspective requires countless man hours in extraction, cleaning, verification, and analysis. A promising step in this direction is leveraging crowd as a resource (*e.g.*, Captricity [33]) to extract and digitize data from PDF reports while preserving

the privacy of the data being digitized. Moreover, the extracted data could be maintained and shared with the community to avoid repeated attempts of the same extraction by other organizations.

Participants noted data merging as the most frustrating process mainly due to the name resolution problem. Although several inter-governmental organizations (*e.g.*, World Health Organization) has built a guidance resource for creating country wide master location list to avoid the name resolution issue [111], the implementation and adoption by local governments has been limited. There is a need to examine the hesitations and barriers that governmental agencies and other organizations face in creating and using these lists. With the recent advancements in machine learning, natural language processing (NLP) algorithms also holds a promise to accurately match transliterated names with basic variations in spellings of location and facilities names. However, it is worth noting that these algorithms need language and domain based training data to be effective, and collection of such training datasets could itself be a challenge.

Merging issues also arise in the mapping of codes and terminologies within datasets that are about the same domain. This is mainly due to lack of standardization and numerous stakeholders working with their own definitions. Although data standards have been designed and proposed for various types of development data, achieving consensus and the resultant wide adoption has been a distant reality. Perhaps a machine learning model could be build to predict the unit or variable based on the distribution in the datasets, however, this might be a non-trivial undertaking due to changing reporting norms that may alter such distributions over time. Current practice is to look up the coding information in any documentation present on the data, which is a strategy occasionally successful. The lack of good documentation on data, which often is the case when data is obtained from an external organization, hampers efforts to even train a model due to missing context and unavailable raw ground truth.

Data validation is at the forefront of the pipeline once the data is collected and imported. Validating datasets collected in resource-constrained regions has its own challenges. Most of

the participants recounted how they eyeball data in tabular or visual format and attempt to manually correct errors. This tedious process could be made simpler by designing validation algorithms and sophisticated models that rely on historical data to highlight data anomalies. Such automation, along with some manual tuning, could reduce the cognitive load of findings errors manually, leading to more coverage in less time during the verification processes.

Even though there are several other challenges in different stages of the data pipeline, the three issues discussed above accounted for the most frustration, resulting in a significant amount of manual labor to extract, merge and validate datasets. To significantly reduce the burden of data analysis, careful investigation of the errors in different stages of the pipeline and building appropriate solutions to tackle those specific problems is the need of the hour. Although Machine Learning algorithms seems an attractive option to tackle some complex cleaning processes (*e.g.*, identifying outliers, addressing the name resolution problem, removing duplicates), special caution is needed while using them to ensure that biases in data collection and reporting do not propagate into training of models.

One of the other difficulties with development data is the fragmented nature of its collection. This is, in part, caused by the donor based funding model that supports specific projects for bounded periods of time. As a result, data is collected only during the funded period. Even when the funding is resumed shortly, data standards may change, leading to collection of inconsistent data. To gain funding opportunities, several non-governmental agencies control the data they collect. Although some funding opportunities explicitly request collected data to be publicly available, there is a scope to improve data sharing practices and stewardship in the area of international development. Often times, the collected data is shared with various partners without providing supporting data like codes or the populations served by the health facility. These details are critical for both cleaning and analysis, and absence of these details lead to time consuming and erroneous data processing. This could be avoided with streamlined expectations in partnerships and proper documentation by the collecting government departments as well as third party surveyors.

Eliminating the root causes of dirty data is a significant undertaking because of the

challenges outlined by my analysis (*e.g.*, different goals and strategies of agencies involved in data collection and analysis, non-cooperation between partners, lack of training and limited digital literacy of enumerators). Combined efforts from multiple stakeholders are needed to improve the collection, cleaning, and analysis of development data. For example, technology architects can play a role by designing tools that address the current data pipeline problems, data architects can design well-structured schemas and forms after taking into consideration local context and culutral norms, and policy makers and government can enforce creation of lists with standard names or identifiers to be used by all organizations working in a country. In parallel, the community could explore tools and processes to mitigate root causes of dirty datasets. The challenges of working with dirty data will persist in the long run in some form due to historic data coming from legacy systems and archived reports, and until current bad data collection practices, poor form and schema designs, and lack of standardization and documentation remains unaddressed.

This chapter is a preliminary effort to compile the complex issues in collecting and processing development data. My work has some limitations. I could only recruit participants working at large-scale international development organizations. I speculate that less tech-savvy organizations might have issues that are distinct than what I reported. Similarly, I interviewed officials working at the government of Punjab, a province in Pakistan that has used digital technologies in a wide array of domains to improve governance in the last six years. Other provincial and national level governments, who are behind in the technology adoption pipeline, might face different challenges. Future work should consider including perspectives of a wide array of stakeholders, with varying degrees of experience and exposure to international development and technology, to expand this checklist of challenges into a rigorous taxonomy of data processing pipeline.

3.4 Conclusion

Collection, cleaning, and analysis of development data is a complex undertaking because of changing data structures, involvement of multiple organizations, and disconnected data

collection efforts. Some of these development data problems are similar to data collected elsewhere but diverse set of stakeholders with limited technical skills and resources makes it more complicated. There is a push by various non-profit organizations, non-government agencies, and international development donors to standardize the tools to streamline data processing for everyone involved. However, my analysis demonstrates how different projects have different needs, and one generic tool cannot solve everything. There is a need to build specific tools for transliterated name resolution, legacy data extraction from textual formats, and quick data validation. Moreover, best practices should be established by publishing documentation on collected data, open access to supporting data like national health facility list, and better stewardship of data to restrict repeated efforts to collect the same data.

This chapter creates a checklist of major pain points experienced by data analysts working with poor-quality datasets collected in low-resource environments of developing regions. Overall, there is a need to design usable and modular technical tools and establish standard data collection procedures to make it simpler to process development data that is inherently complex in nature. Data collection, for example, can be improved with more resources spent on training the collectors and curating data standards that are scalable. This will lead to well formatted high quality data. Using natural language processing algorithms, new tools needs to be created for matching transliterated names to make it easier to merge datasets that don't follow proper standards. To make it easier to use legacy datasets, more usable toolkits are required that can automatically convert PDF reports into other digital forms. Furthermore, we need data models and APIs that can enable even less technical user to validate data for errors and anomalies. All these technical solutions will help improve major pain points in the development data pipeline. The reliance on these tools could be dramatically reduced if we collectively adopt better practices for data collection, curation, analysis, documentation, and sharing.

Chapter 4

CASE STUDY: IMPROVED COLD CHAIN INFORMATION SYSTEM

4.1 Summary

In this chapter, I demonstrate how to implement an information system that is scalable to other regions while having a low burden in keeping data up to date via a distributed data collection process over SMS. I recognize that legacy data issue is unavoidable as information system formats will evolve to cater to new requirements through scale. Through this case study, I witnessed the format of data collection change as local officials found errors that had originated due to language barriers. Second, I found that having a system that produces value for stakeholders at various levels aids in quicker adoption. This work establishes better guidelines for following processes to have a successful data processing information system.

4.2 Introduction

4.2.1 National Immunization Programs

Immunization is recognized as one of the most successful public health interventions in history. Vaccines save millions of lives every year from preventable diseases. An example of the success of immunization is the near eradication of polio. The number of cases per year has declined from an estimated 400,000 in 1980 to under 300 in 2012 [52]. Robust global organizations support immunization, both in terms of donor funding, as well as in global governance and coordination. In almost all developing countries, routine immunization is part of the public health system and is administered centrally by a department inside the Ministry of Health, which I refer to as the National Immunization Program (NIP). Vaccines are distributed nationally and are available for free or at low cost in public health facilities.

Vaccines are imported into the country to the national vaccine store, and then distributed through a hierarchy of vaccine stores until they reach health facilities. At health facilities, vaccines are stored until they are used for immunization or are sent on to secondary facilities. Different schemes are used for immunization, such as outreach delivery, where vaccines are carried to a remote site for use, or static delivery, where people come to the facility for immunization. To ensure that the vaccines remain viable, it is critical that they are kept at appropriate temperatures during transit and storage. This is done with refrigerators and freezers at storage locations, and refrigerated trucks and cold boxes for transit. These are collectively referred to as the vaccine cold chain.

4.2.2 Logistics Challenges

There are two basic logistical problems for vaccines. The first is the distribution of vaccines; ensuring that every health facility has an adequate supply of all vaccines in the routine and supplementary immunization schedules. A stock out is said to occur when a facility has insufficient vaccines on hand to perform scheduled services. Stock outs mean that immunization sessions must be cancelled or children who arrive do not receive vaccinations. Since missed vaccinations are often not made up, overall coverage is reduced. There are multiple causes of stock outs including overall shortages of vaccines in the system (for example, if insufficient stock comes in to the national level), delays or mistakes in ordering at different levels, over allocation of stock to some facilities (which means there is not enough to get to other facilities), travel delays or lack of transport, and incorrect forecasts of demand.

The other major challenge is maintaining the cold chain to keep vaccines in a safe temperature range. The safe range is generally considered to be 2C to 8C with the freezing of vaccines the biggest concern. Some exposure to temperature above 8C is acceptable, although this varies across vaccines. The WHO guidelines are that vaccines should not be exposed to temperatures of less than-0.5C for more than one hour, or temperatures of more than 8C for more than 10 hours. These conditions are referred to as alarm conditions. A countries' NIP generally requires that facilities store vaccines in dedicated vaccine refrigerators. WHO

maintains a list of refrigerators that have been certified as suitable for vaccine storage, the PQS list [110]. This list currently consists of about 45 models. The reason for this number of models is that there are a range of sizes, as well as different energy sources (electricity, solar, gas, kerosene) for the refrigerators as well as several manufacturers. Many of these refrigerators are specially designed for areas with poor power infrastructure and extreme temperatures. For example, an Ice Lined Refrigerator (ILR) is an electric refrigerator that maintains a layer of ice, so that vaccines remain cold during power interruptions. There are many practical difficulties that countries face in keeping their cold chain equipment functioning. Common problems are extended power cuts or lack of fuel. Equipment maintenance is often poor, so that refrigerators can get either too hot or too cold, for example, a refrigerator might have improper thermostat settings, causing the temperature to plunge below freezing during cooling cycles. When refrigerators fail, it can often take a very long time to have the refrigerator repaired or replaced. This can be caused by the non-availability of spare parts or delays in finding service personnel. In this case a facility may be left without vaccine storage for extended periods of time.

4.2.3 Information Systems

One of the big challenges in strengthening immunization logistics systems is lack of information, especially detailed information from the health facility level that can be used for mid-level management decisions. There is generally very little information available from individual health facilities on the level of vaccine stocks or the quality of the cold chain equipment. In many countries there is a lack of basic information such as whether health facilities have adequate equipment to store their vaccines, or if they have stock on hand. When this information is available, it often comes from aggregate reports which are collected at the top level of the health system, or from cold chain assessments which are conducted at infrequent intervals. What is generally lacking is detailed information that can directly support management to improve the stock and cold chain equipment at the facility level. This brings me to the focus of my work; I am interested in developing Cold Chain Information

Systems (CCIS) that make information available to health system managers, so that they can take action based on data to improve the functioning of their health systems. I define a CCIS to consist of the following components:

- A national level database of health facilities and associated cold chain equipment that allows updates to be performed in a distributed fashion.

- A mechanism for bi-directional information flows that allows facility information to be communicated up the health system hierarchy from facilities, and also allows communication of summary information back down to facilities.

The database for a CCIS is a Cold Chain Equipment Inventory (CCEI). This includes a registry of all of the health facilities and vaccine storage facilities in the country, along with information on the facilities, such as the size of the population they serve, and the power sources available at the facility. The database also tracks all of the vaccine refrigerators and other cold chain equipment, with information on the models, the power sources used, and storage capacity of the equipment. It is necessary to track this information to determine if the equipment is sufficient for storing the required vaccines. The information that is most important for immunization logistics is information on stock as well as status of the storage equipment. The system would allow reporting of low stocks or equipment problems, which could trigger either immediate remedial actions, or institute changes that would improve performance later on.

Currently, information for most CCIS is manually collated from paper forms. Paper based solutions cannot provide real time information and recently there have been efforts to collect data via SMS [149] in other supply chain domains. SMS is attractive because of its low cost and wide accessibility in the field [150]. One such study, SMS for Life, used text messages to track stock levels for anti-malarial drugs in Kenya [9, 151]. Multiple projects, including Tanzania's Integrated Logistic System (ILS), Kenya's Mission for Essential Drugs and supplies (MEDS) and Benin's Cola Life uses SMS to track orders, vaccines, medicine and other clinical supplies [105].

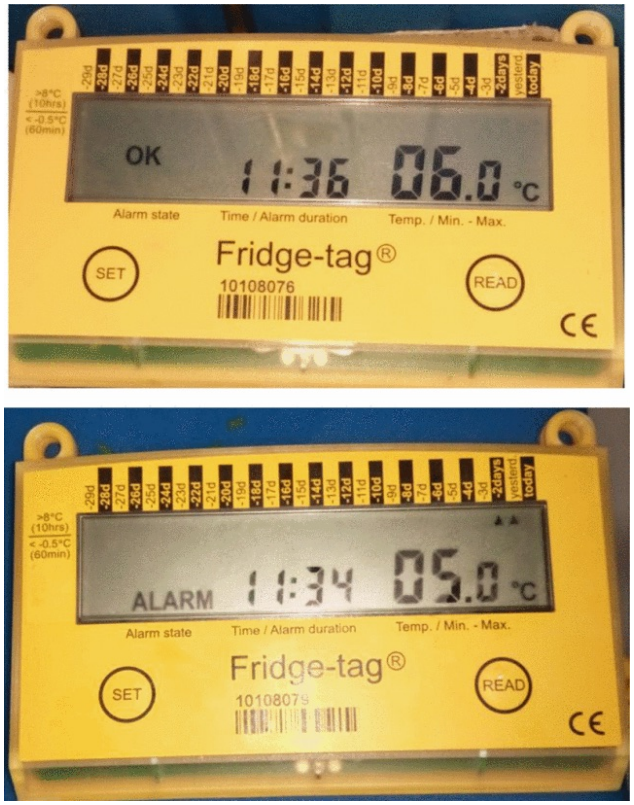

Figure 4.1: (Top) fridgetag 30DTR with no alarms in last thirty days, (bottom) fridgetag 30DTR with two high temperature alarms, yesterday and two days ago.

There have been studies and efforts in making successful vaccine distributions [68, 104, 167]. Having an immunization information system enables better decision making for quality and impact [167]. Cold Chain Equipment Manager (CCEM) [5] is being used in several countries to maintain country wide inventories of vaccine refrigerators and cold rooms. CCEM is used to track and resolve shortages in vaccine storage capacity which is a common bottleneck in the vaccine supply [167]. Nexleaf developed Cold Chain Monitor [106] to automate SMS temperature reporting. OpenLMIS [130], VaxTrac [50] and Jeev [80] are among other vaccine coverage tracking system being used in different parts of the world.

4.2.4 Data Reporting

The goal is to enable information to be collected from health facilities to allow improvement in vaccine stock distribution and the vaccine cold chain. This means that we need to have information from health facilities communicated to managers in a timely manner. For vaccines, this simply means reporting the quantity of vaccines on hand. These numbers can be compared with the vaccine requirements to determine if the facility is under or over stocked. Action taken on this information can include adjusting the supply of vaccines that are delivered, or sending emergency stock to make up for a short fall. For reporting information on refrigerator status, we base our reporting on the FridgeTag 30 Day Temperature Recorder (30DTR). This electronic temperature logging device is placed in a refrigerator and record temperatures over the month. The device tracks daily high and low temperatures, and also shows an alarm if temperatures are out of range for an extended period of time. A high alarm is registered if the temperature is above 8C for more than 8 hours, and a low alarm is registered if temperature is less than-0.5 C for more than an hour (see Figure 1.) The number of high temperature and low temperature alarms over the last 30 days can be read from the device. The reporting that I am interested in is the number of high alarms, and the number of low alarms over the month. A functioning refrigerator should have zero alarms, so the presence of any alarms indicates a problem, and if there are multiple alarms, especially freeze alarms, then action needs to be taken to repair the refrigerator. The value of the 30DTR is that it gives a very good measure of whether or not the equipment is working well.

The FridgeTag devices have been widely distributed by UNICEF and individual countries. However, the experience in the field is that when these devices are deployed, information is not reported back to management, the devices just sit in the refrigerators, and the information is not acted on. One of the main contributions of this work is to develop a system that allows the 30DTR information to be reported back to the district level to support actions. I noted that there are multiple systems for automatic reporting of temperatures, where regular reports of the temperature are sent to a server via the GSM network. This includes devices

developed by BeyondWireless [163], Nexleaf [106] and University of Washington [28]. Automatic reporting can be viewed as complementary to the work I am doing, and a backend system could be developed to receive both automatic and manual SMS reporting. I focused on the manual reporting of FridgeTag results because costs are much lower for a FridgeTag than real time devices, and there are hundreds of thousands of FridgeTags deployed worldwide with no adequate reporting system.

4.3 Developing a CCIS for Laos

4.3.1 Laos

This project was initiated in Laos following a series of discussions between UNICEF and the Lao Ministry of Health. UNICEF chose to implement this project in Laos as part of a much larger effort in strengthening the vaccine cold chain. Laos was selected as a representative country that would easy to work in for an initial project, being a small country with a stable political environment.

Laos is a landlocked Southeast Asian country located between Thailand and Vietnam. The population of Laos (2014 est.) is slightly under seven million. Laos is categorized as a lower middle income country, with a per capita income roughly the same as Pakistan's. The economy is expanding with annual growth of about 8% in recent years.

The public health system for Laos is a standard, hierarchical system, with three levels: national, province, and district. The country has 17 provinces and 145 districts. I noted that since Laos is a small country, province and district populations are low, and the provinces behave much like districts in larger countries. Immunization services follow this hierarchy, with vaccines stored at the national store, provincial stores, and district stores. A substantial amount of the system management takes place at the provincial level. For example equipment repair is an activity managed by a provincial staff. In rural areas, the primary health facility is a health center which provides basic services (including immunization) and has a staff of about four.

Figure 4.2: Typical rural health center staffed by approximately four health workers

Figure 4.3: Lao PDR, the NIP office is located in the capital Vientiane.

Laos has a fairly good electrical power system due to substantial hydroelectric resources, and almost all health facilities are on the grid. On my facility visits, health workers did report that there will be occasional day long outages. Since electrical power is generally available, almost all health facilities use electric refrigerators, although there are a few facilities in remote regions using solar powered refrigerators.

4.3.2 Requirements and Specifications

The requirements and specifications for this project were defined in collaboration between UNICEF, Lao NIP, PATH (a Seattle based global health NGO), and University of Washington. The project was initiated with stakeholder meetings in Laos in March and October 2013. The initial concept for the project was reporting and monitoring of fridge tag alarm data from all health facilities each month. In the initial meetings, Lao NIP emphasized the value of including stock level data in addition to the temperature monitoring data. The idea of monthly electronic reports was very exciting to local stakeholders and many different schemes were discussed for expressing complicated reports via SMS. However, stakeholders also realized that any implementation of the project would need to be deployable and therefore accessible to people unfamiliar with SMS.

From the collaborative design process three overarching requirements emerged for the project:

- Scale: The system must be able to scale to the national level. This meant that the all health facilities must have the capability to make reports, the system backend must be able to handle a large influx of messages during reporting time, and managers must be able to view aggregated vaccine and refrigerator reports.

- Low burden: There must be a balance between collecting as much information as possible and decreasing the monthly reporting burden on individual clinics. Requiring an excessive amount of information creates an arduous task and increases the likelihood

that the reports will not be submitted. Additionally, complicated monthly reporting requires longer training and delay system deployment.

- Use of data: the motivation for the project was to strengthen the immunization program by making better data available to managers. There needs to be consideration of how data can be acted on, as well as who needs to receive original data or summary reports.

The first, major design decision for the system was to select SMS reporting as the mechanism for monthly data submission from all health facilities. The desire was to have a system that could be deployed using technology accessible at health centers, and it was assumed that almost all health workers would have access to mobile phones for sending SMS messages. UNICEF and NIP conducted field studies evaluating the feasibility of SMS based reporting. An important finding from these visits was that the SMS cost of 100-150 Kip (USD 0.02) per SMS was not a problem if a handful of SMS reports were sent per month. However, these findings also revealed that few health workers had much experience with SMS. One of the reasons related to the Lao script. Mobile phones in Laos generally do not support the Lao script. For text messaging, the common solution is to transliterate Lao into the Latin script using a style of messaging referred to as Karaoke. The feasibility study determined that it would be practical to use SMS with the following conditions and caveats:

- Since all health workers have access to basic phones, the project could rely on personal phones and is not required to distribute phones; however, access to data connections or smartphones could not be assumed.

- The price of SMS messages would not be an issue as long as messages could be sent to a Lao number.

- The reading of Latin scripts would be difficult for some health workers. However, there is no practical alternative. The system would need to be robust to accommodate

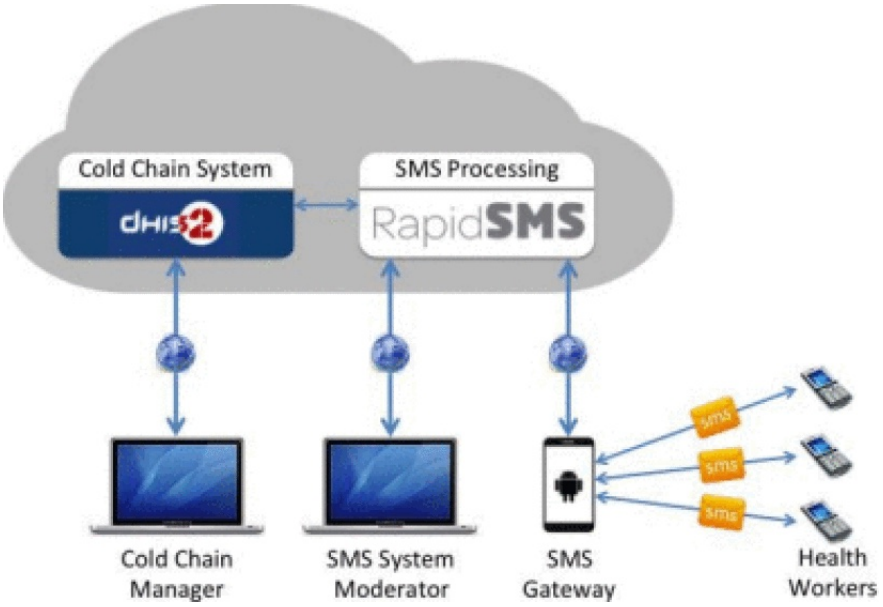

Figure 4.4: System architecture: Health workers send data via SMS to an android phone that syncs with cloud system. Cold chain manager and SMS moderator manage these systems using their respective web interfaces.

common errors in text entry.

- Acknowledgments must be sent back to health facilities upon the receipt of message reports notifying the sender of any errors or actions taken.

The final reporting specification extended beyond the fridge tag alarms to support monthly vaccine stock levels, the reporting of refrigerator failures, vaccine stock outs and other events that can occur in the middle of the month. These events, that need immediate attention, can quickly be forwarded to the appropriate individuals to facilitate repair and attention. During the design discussion multiple mechanisms for registering users with the system and associating them with a health facility were proposed. SMS could not be the sole method to register or unregister a user because that would over complicate the message syntax. The SMS management system would need to have a web accessible management console for UNICEF and NIP staff to moderate users and messages.

4.3.3 System Components

The final system design consisted of three components: a backend database responsible for maintaining the cold chain inventory; a SMS processing system responsible for translating incoming messages into actions on the backend database, and SMS to HTTP gateway to upload incoming SMS messages to the system and send out responses generated by the system. See Figure 4.

Cold Chain Inventory Data Model

It is critical that health information systems are built on top of common standards. This is necessary so that common solutions can be applied across multiple countries, and so that different systems can exchange data. With this in mind, I built the system on top of the Cold Chain Equipment Inventory (CCEI) data model. The data model was developed by my research team in collaboration with the UNICEF Cold Chain Logistics Working Group. The model gives standard definitions for a national cold chain equipment inventory, identifying specific information that is needed both from the health facilities and from the cold chain equipment. The model was designed in a collaborative fashion, beginning with an initial set of inventory definitions that was put together based on several existing software tools for cold chain logistics. The document was then collaboratively edited by the UNICEF Cold Chain Logistics group. Based on that effort, a more formal set of definitions was developed that was circulated individually to about 15 immunization cold chain experts, from UNICEF, WHO, and NGOs including PATH and CHAI, who provided very detailed feedback. The data standard was built upon existing standards where possible, such as using the WHO Performance, Quality and Safety (PQS)/Product Information Sheets (PIS) catalogs [110] and defining several of the fields with respect to ISO standards.

The core of CCEI is the model of facilities and their associated equipment. The full standard contains several other components such as a representation of the administrative hierarchy, the association between facilities and assets, and some localization information.

The full model can be found online [144] and also discussed in chapter 5.

Backend System

The back end of the CCIS is built on top of the District Health Information System (DHIS2), a widely used health indicator and reporting system. DHIS2 was initially developed by HISP (Health Information Systems Program), as an open source tool for ministries of health to use to collect and analyze health indicators. In 2014, it was deployed in 46 countries as an integral part of health system reporting systems. The software is a Java web server which may be deployed at the country, province, or district level and easily integrate into existing workflows. It includes a data layer, service layer, and web presentation layer which facilities to support user report generation, integration with mobile devices, and a GIS module. The core data model for DHIS2 is designed around an organizational hierarchy, which associates data elements with organizational units at every level of the hierarchy. Once data has been entered, the system allows aggregation of the data elements over the hierarchy in to regional reports of health indicators. The DHIS2 system has been designed to allow full country level customization of data elements making it a very flexible tool.

The CCIS module is an extension of DHIS2 to support cold chain equipment and stock inventory. The extension of DHIS2 to support a cold chain equipment inventory required significant software development. Development of the cold chain module took roughly one year to complete. The technical extension of DHIS2 was to allow organizational units to maintain a collection of assets, with properties associated with each asset. The introduction of a new data model required low level additions to the source code. A new user interface was also developed to support the management of equipment, including the viewing of assets across different administrative levels. The cold chain module was built on top of version 2.14 of DHIS2, and is available as a download on the DHIS2 download site. There were two very important reasons for building the CCIS on top of DHIS2. The first was that the DHIS2 platform provided a vast amount of functionality, so it was a practical choice that reduced system development effort. The second and more important reason is that by building on

top of a system that is already extensively deployed, it is easier to sustain the system, as well as integrate it into a countries' existing health information system.

4.3.4 SMS Processing System

The SMS processing engine receives messages, processes them, communicates with the CCIS backend, sends responses and notifications and allows for administrative moderation. I built the the SMS processing engine on top of RapidSMS which is an open source Django server application designed for customized scalable mobile web services. RapidSMS provides the basic features of a messaging system: integration of an SMS to HTTP gateway, management of contacts, and message logging. The most important feature of RapidSMS is a message routing system with hooks for custom applications to process incoming and outgoing messages. The SMS processing system consists of three modules built on top of RapidSMS.

The first module is the message parser which is programmed to interpret message reports and system commands. A more detailed overview of the message syntax supported by the system is given below. The parser and message syntax were designed to be as robust as possible in order to accept slight variations of correct messages. For example all whitespace, capitalization and punctuation is ignored. The second module is the administration and user management console. This is a web front end for local NIP officials to manage the SMS system from. Besides the basic user and facility operations the administrator module also allows for moderation. Even though the system is designed to be automated, I expect a small number of users might have difficulty with the message syntax. From the administrator console messages that cannot automatically be parsed can be handled individually. The last module in the SMS processing system is a connection to the CCIS backend. This connection imports an organizational hierarchy and basic inventory so that users phone numbers can be associated with health facilities and basic errors such as incorrectly labeled refrigerators can be detected as soon as possible. The CCIS connection is also responsible for updating the CCIS module in DHIS with the monthly vaccine reports and fridge alarms. This is a fundamental piece of the entire system and the link between two of the major components.

4.3.5 SMS Gateway

The final component of the system is an SMS to HTTP gateway that connects the RapidSMS server to the Lao telecommunication system. Local partners are currently in discussions with several Lao telecommunication operators to get a short code that is accessible from Lao mobile phones and has support for an HTTP gateway. While waiting for the short code system is using an Android SMS to HTTP gateway so that health facilities can SMS monthly reports immediately via a Laos based number. The gateway is an application running on an Android phone that forwards all incoming SMS messages to a URL over an available mobile data or wifi connection. The server processes the HTTP request and responds with a JSON document containing an SMS response. The Android application parses the HTTP response and forwards the outgoing message to the original phone number. This method allows people with basic phones in Laos to seamlessly interact with the RapidSMS and CCIS servers through a local Lao phone number via SMS.

There are several existing Android HTTP to SMS gateway applications already in existence. For this project I examined SMSSync by Ushahidi and EnvayaSMS. Both applications have a well-documented REST API to receive and send SMS messages on a local number in any country. They have the ability to forward messages based on prefix keyword, blacklist spam numbers and basic HTTP authentication. Both applications also allow a polling interval for the phone to check the server URL for any pending messages that need to be sent. This allows for messages to originate from the server. A major limitation of Android based HTTP to SMS gateway is that each Android application is limited to 100 SMS messages per hour by the Android OS. EnvayaSMS provides a simple method around this limit by providing expansion packs that increase the limit to 500 messages per hour. This plus additional logging features and more robust recovery from network errors lead us to choose EnvayaSMS as the HTTP to SMS gateway for this project.

Besides the message limit there are other inherent limitations for an Android based HTTP to SMS gateway. The Android phone represents a single source of failure for the

whole system. Currently the phone stays at the NIP offices in Vientiane and it is NIP's responsibility to make sure the phone is charged, has a data connection, and is recharged with sufficient airtime to send outgoing messages. This is a large amount of work to add to already busy workloads of NIP officials. The Android phone is a source of constant attention and possible errors and I looked at methods to eliminate it. The ideal solution is to get a Laos based mobile short code that works across networks and have begun the discussion necessary with telecommunication operators to establish one.

4.3.6 Message Syntax

A unique feature of this system's design is the robust message syntax. I carefully designed a syntax that would not be dependent on whitespace, punctuation or special characters. Even so, I knew that with limited SMS experience, health-workers might still have difficulties so I also strove to make the syntax as flexible as possible. I made several assumptions about message content that allowed me to simplify the syntax. A message is a concatenation of keyword plus argument pairs. The crucial aspect of the message syntax is that all keywords are two letters while all other labels are single letters or digits. This means that I can ignore whitespace for separation, as long as every keyword has at least one argument when multiple keywords are chained. Additionally, for this deployment all of the numeric data being collected is integers, so I also assume that there will be no floating point numbers and therefore it is safe to ignore commas and periods. I were also able fit the number of high and low alarms into a single digit each because UNICEF determined that reporting a maximum of nine alarms per month was sufficient, as more alarms in a month would indicate a serious problem and no additional information is gained from the exact number of alarms.

There are three classes of keywords the system uses, these are for monthly reports, spontaneous reports, and administration tasks. The monthly reports use two keywords: ft (Fridge Tag) which takes as an argument the 30DTR information for each refrigerator at the facility and sl (Stock Level) which records the stock of up to five vaccines. When entering refrigerators into the CCIS database a single letter is associated with each refrigerator at a health

Figure 4.5: Vaccine refrigerator labeled with A for reporting.

facility. The refrigerator is labeled with this letter (Figure 5) so that when composing messages the correct refrigerator can be referenced. Similarly each vaccine is assigned a unique letter to be used in the message syntax. Currently, health workers are reporting Pneumococcal vaccine (with a P), and Pentavalent vaccine (with a D). Other codes have been set aside to allow reporting of all routine vaccines. Figure 6 shows several variations of similar monthly report messages that are all valid. Spontaneous reports are for events that need immediate action such as a refrigerator needing repair or a vaccine stock out. The administration tasks are more advanced features but are provided for completeness and include keywords for help pages, changing response language, and registering a new phone number.

4.4 Field Experience

4.4.1 Deployment Status

The project was active in three districts in three separate provinces in Laos. The plan was to start in these districts, and as the system is established, expand to the remaining districts

1. ft a 0 b 0 sl d 30 p 20
2. a00b00sld30p20
3. A 0 0 B0 SL D30 P20
4. ft a0,b0;sl d30; p20
5. Ft0 SL d 30, p 20

Figure 4.6: Five valid messages that all have the same semantic meaning. each message tells the system that refrigerators A and B had zero alarms and that the current stock levels for pentavalent and pneumococcal are 20 and 30.

in each of these provinces, and eventually expand to the remaining provinces. The project began in October 2013 by having health centers report by SMS, with the results being manual interpreted. An initial version of the SMS gateway and SMS processing system was deployed in January 2014. The SMS reporting was being done from all of the health facilities in the districts, along with the district and provincial vaccine stores. Reports were being received from 27 facilities all together.

The initial results were favorable, there was a high rate of reporting, and acceptance of the system from the health workers. The process of reporting on stock levels and fridge tag alarms was making information visible that was not previously available. In particular, there was some information on faults with refrigerators being reported which is actionable. This work was at the feasibility stage, so results were being fed back into operational improvements, and it was too early to assess impact. In April 2014, all participating sites were visited with interviews of health workers, so there was substantial ongoing assessment. When the full systems is established, there was a plan to do an impact study based on key performance indicators such as number of stock outs and rates of equipment repair. My research team's involvement in the project ended in 2015 so I am not sure about the current operational status of this system.

4.4.2 Challenges

Field visits have identified various challenges for the project. I had anticipated that there would be some adjustments necessary, and are investing significant effort in field monitoring and ensuring I get feedback from participants. Here are a set of issues relating directly to the SMS deployment:

- Almost all phones in Laos support the Thai script, but few phones support the Lao script. When a survey of 23 health workers was conducted, I found 100% could read Lao written in Thai script, 65% could read Lao written in Latin script, and 22% could read English script. Unfortunately, for political reasons, it is not possible to use the Thai script, so sending the text messages where Lao is transliterated into the Latin script.

- The SMS commands are a mix of Latin characters and digits. Since many health workers were not familiar with English characters, there were difficulties with visually similar characters. I saw confusion in sending 0 instead of 0 and $ instead of S.

- There are four cellular carriers in Laos. There appear to be some issues in sending SMS messages between carriers.

- The SMS format evolved over the period of deployment as new sites were added. This introduced a challenge for the system to provide backward compatibility of SMS format.

- There are difficulties in tracking the phones used to submit messages. The system do not require a health center code in the message, but associate the message with the facility that the phone is registered with.

4.5 Discussion

As the project is in the initial feasibility phase, I focused on various operational issues. However, there two other very important components of this project that I need to mention.

The first is training, health workers need to receive training in the process of reporting the monthly data through structured SMS messages. It turns out that about 33% of the health workers also need training in how to send SMS messages. The project is currently investing in various training aids and technologies to support training. The training of health workers is considered to be biggest component of scaling the system to a national scale project. The other big issue is use of data. District and provincial managers have expressed great interest in the system, and standard operating procedures are being developed on how the data that is reported will be acted up. It is recognized that the system will only be of value if leads to people using the data to address problems with stock, or to repair or allocate cold chain equipment.

This project was launched in Laos to develop a system that will help strengthen the Lao immunization system. However, the goal of the project is to develop a system that can be deployed across multiple countries, since there are common problems to immunization systems. In terms of system design, my research team is considering issues such as localization early in the project. There are country specific issues such as developing an SMS gateway solution that will have to be repeated in each deployment. Deployments of SMS solutions in other countries have had to focus more on payment issues that we have. The specific challenges around choosing a script for SMS may not come up in many countries, since Laos is a small country with its own script, which is not a large enough market for handset manufactures to implement the script.

Chapter 5

CASE STUDY: DEVELOPING COLD CHAIN DATA STANDARD

5.1 Summary

Data standards are an integral part of an information system that relies on data coming from various stakeholders' efforts. This chapter describes the complex of tools being used to track cold chain data for different purposes by different stakeholders. Most importantly, it presents a data standard for cold chain that will work for a given scenario across stakeholder requests. The data standard has been built with the consideration that future needs will evolve and require the extension of this model to track additional information. This effort contributes to my data challenges taxonomy in that I determine that the existence of multiple stakeholders with different data requirements creates a barrier to establishing acceptable standards. Nonetheless, if the standards are built with a specific application in mind, then the development stakeholders can be encouraged to reach consensus given that the narrow scope will limit the variation and diversity in scenarios that the standards must support.

5.2 Introduction

Health information systems are critical for strengthening public health in developing countries [3, 113]. Accurate and up-to-date information supports management and decision-making, and allows resources to be directed to where they may have the most impact. The health information system collects data through the health hierarchy and the data is maintained centrally. Reports and analysis from the data can be generated either at the national level or in a distributed manner. The health information system often supports multiple health domains, where each health domain is a programmatic area around a disease, inter-

vention or service.

However, countries face many challenges in deploying effective health information systems. Many reports are still submitted on paper forms, with the resulting information stored in standalone databases or spreadsheets. Network infrastructure, although improving rapidly, still can be unreliable, especially at the periphery. The software environment for a health system is often complex. In addition, organizational issues may make change difficult since some groups may lose control over data and new working relationships need to be established. IT resources are also often limited, making it difficult to maintain and support systems. Finally, decisions on the adoption of health systems software can be highly political [141].

Recognizing the complexity of existing information systems and organizations, how does a country successfully implement a modern health information system? To explore how information system needs for a health vertical tie into the overall health information system, we look at one particular domain, immunization and the vaccine cold chain. We want to understand how addressing programmatic needs for use of data can align with centralizing information services. In this paper we describe work that we have done to support over a dozen countries in building information systems for their vaccine cold chains, and then draw some general lessons to inform work in other program areas. One of our main findings is that data standards are needed to provide a bridge between the multiple implementation options, as well as providing a link between the software applications and practitioners.

Literature in standardization argues that standards emerge through use rather than consensus [40]. Given the rapidly growing complexity around the standards makes achieving a consensus very difficult. This is mainly because of the bureaucratic processes of standardization bodies are too slow [41]. Informal bodies can contribute tremendously in shaping the standards in today's complex space [148]. Shared standards can be successful and largely adopted if enough actors are convinced [139]. If the standards are simple and flexible to changes then it can be adopted in complex health information systems [15].

We argue that a standard can emerge through consensus if the community is specific. In

cold chain domain, the consensus is achievable since the programs are fairly uniform. The standard need to define backbone of the cold chain space and have flexibility to extend for specific implementations. This is not only to incorporate future expansions but also to reach an agreement within the community of cold chain experts.

5.3 Immunization

Importance of immunization, logistical challenges, and needs of cold chain information system has been explained in chapter 4. I am going to discuss some details and concerns here again for context. Immunization is well recognized public health intervention and requires distribution of vaccines to all urban/rural health facilities while keeping it refrigerated during transit and storage at different distribution centers. This variety of cold storage is collectively referred as the "vaccine cold chain" [93].

Immunization logistics is concerned with the distribution of vaccines. Essential problems include maintaining adequate stock levels and ensuring that vaccines maintain safe temperatures. A logistics information system manages information about vaccine shipments, vaccine use, and the fixed assets in the system. In this work, we focus just on the information systems associated with the physical cold chain, which consists of an inventory of the cold chain storage equipment along with associated information about the health and storage facilities. Even though this basic equipment and facility information is fundamental, it is often unavailable or out of date.

Perhaps the most basic question about a country's vaccine cold chain is whether or not it has sufficient capacity to store the required vaccines. However, in many countries, the answer to this question is not known, as the Ministry of Health does not know how much cold storage equipment is available for vaccines. This question becomes especially important with the introduction of new vaccines such as the rotavirus and pneumococcal vaccines. These new vaccines take up more space, and are more expensive and more sensitive to heat than older vaccines, which increase the importance of having sufficient capacity in the cold chain. It is also important to understand the quality of the cold chain, including the

working condition and age of equipment. Since many health facilities do not have access to regular grid electricity, there are vaccine refrigerators with other power sources including gas, kerosene, and solar power. Knowing the distribution of power sources of equipment is critical for estimating overall costs (for example, gas and kerosene can be ten times as expensive as grid electricity) and planning for upgrades. Information about the cold chain is also important for management of existing equipment, and acquisition and allocation of new equipment, which often takes place at an intermediate level, such as at the district level.

5.4 Cold Chain Information System

A cold chain equipment inventory (CCEI) is a data set consisting of information about vaccine storage devices (refrigerators, freezers, cold rooms, freezer rooms, cold boxes and vaccine carriers) along with information about the health facilities. Basic information about the cold chain equipment includes age, model, working condition, and the health facility it is located in. We discuss the data fields in more detail below. When we talk about a cold chain equipment inventory we generally are referring to a national inventory, although in some large countries, such as India, inventories might be done by state. The inventory generally focuses on facilities in the public health system as in most low and middle-income countries vaccines are distributed through the public system.

We now present a number of use cases for the cold chain equipment inventory. In this discussion, it is important to consider the stakeholders, which we divide into three broad categories: Global: donors and intergovernmental organizations such as UNICEF and WHO; Implementers: NGOs, consultants, and academics; and Country Staff: Ministry of Health (MOH), National Immunization Program (NIP), and logisticians. Although all of these stakeholders are aligned on the goal of strengthening immunization systems, they have differences in priorities and emphasis.

Cold chain equipment inventories are used to support all of the following activities:

- Assessment: Evaluation of the vaccine cold chain to determine if there is sufficient

working storage capacity to ensure that vaccines are kept safe until they are used. One of the important areas for assessment is when new vaccines, such as pneumococcal and rotavirus vaccines. New vaccines often have large packaging requirements, which can overwhelm existing storage capacity.

- Quantification of need: Determine how many new refrigerators are needed to meet storage demands for vaccines. This quantification can be based on scenarios such are addition of new vaccines and retirement of different types of refrigerators. This may be expressed as a multi year plan that identifies needs over several years.

- Re-engineering supply chains: There is interest reorganizing how vaccines are distributed within countries by doing things such as removing intermediate levels of storage and changing delivery timings. Estimating the cost requires having the underlying cold chain inventory data, and possibly applying sophisticated computer modeling with systems such as Hermes [84].

- Market shaping: Cold chain inventory data provides important information to estimate the global demand for different classes of vaccine refrigerators, such as solar powered refrigerators. One application of this is to demonstrate sufficient demand so that manufacturers will be able produce and market certain models.

- Cold chain management: The cold chain inventory support many tasks on managing equipment from allocation of refrigerators to health facilities, to planning equipment upgrades, and tracking maintenance.

- Immunization information systems: The cold chain inventory can be a component of other information systems for immunization to support other activities such as managing vaccines in a logistics management information system.

These use cases impose different requirements on a cold chain equipment inventory, and are motivated by the different stakeholders. For example, assessment is a use case that is

driven at the global level, and may be a requirement before funding is released to support vaccine introduction. Management on the other hand is activity at the country level, where information is used to make decisions about individual pieces of equipment, and is not a direct concern to the external stakeholders.

5.5 Software Context and Challenges

A wide range of systems are used to manage cold chain inventories. The context differs between countries, so it is natural to see different approaches taken. Inventories are sometimes managed by a standalone, local application without web support, and other times are part of a larger database or a component of an application for another purpose. In this section we give an overview of various approaches for maintaining cold chain inventories. One thing to note is that the cold chain inventories fall into a common pattern of health domain software systems where there are simultaneously spreadsheet tools, single machine database tools, and web-based database tools

5.5.1 Spreadsheet Solutions

The most common and basic approach to representing a cold chain inventory is to track the information using spreadsheets. There are advantages to this approach: spreadsheets are simple to use and software is widely available. However, spreadsheets are generally single-user documents and there are challenges in maintaining multiple versions. Further, the functionality of spreadsheets is limited with respect to analysis of the data. A prime concern about spreadsheets, raised to us by a World Health Organization (WHO) official, is the difficulty in linking information across spreadsheets (e.g. associating refrigerators with health facilities).

There is a range of spreadsheet approaches used for cold chain inventories that can be modeled using a hierarchy:

1. Simple spreadsheets. The most basic approach is to maintain the information as lists.

We have seen many different ways this information is stored. For example, in Laos, separate spreadsheets existed (in different formats) for each manufacturer of equipment, in additional to extra sheets for facility information and populations

2. Inventory spreadsheets. The next level up the hierarchy is spreadsheets that are designed specifically for a cold chain equipment inventory. An example we have worked with is an inventory for three states in India that consists of seven separate spreadsheets for equipment from different types of facilities, along with another two spreadsheets for vaccine logistics for these facilities. The ID numbers from the original inventory forms are used to link equipment to facilities. Due to the complexity of the survey forms, the spreadsheets were quite large, with some having over 300 columns.

3. Excel-based cold chain tools. At the top of the list are a group of cold chain analysis tools built in Excel. WHO maintains a group of tools for national and regional immunization managers that support activities such as tracking immunization coverage and managing stock levels. Some of these tools, such as District Vaccine Data Management Tool (DVD-MT) provide sheets for an inventory of cold chain equipment. The DVD-MT tool provides a well-structured inventory that includes many of the fields we recommend in our own data model (discussed further below).

5.5.2 Single Machine Applications

Moving up from spreadsheets are applications using local storage, frequently implemented using Microsoft Access. The most widely used cold chain equipment inventory application is CCEM [5], which was developed by PATH in collaboration with UNICEF, WHO, and USAID. CCEM is a Microsoft Access application with the following functionality:

1. Equipment Inventory. As an Access application, the database is represented with a set of interlinked tables. The main tables are for health facilities, refrigerators, the administrative hierarchy and refrigerator types.

2. Report Generation. This is the key functionality for users, with domain specific reports and charts.

3. Modeling. CCEM was designed to support the development of multi year equipment acquisition plans. CCEM has a simple modeling engine that determines an equipment allocation to satisfy requirements, and allow schedules for adding or removing equipment over several years.

4. Inventory process support. By presenting a clearly defined data model, CCEM has provided an entry point for countries to begin monitoring their cold chain inventories. This schema in turn makes it easy to generate forms from the data model to facilitate the collection of useful inventory information. Together these features have made creating and maintaining a cold chain inventory a more tractable problem, which has been one of the contributions of CCEM.

5.5.3 Web Based Applications

The basic requirements for an inventory tool are to allow updates and generation of reports from multiple sites. This suggests a web-accessible database. There are many possible ways to implement this. Generally speaking, the sizes of the databases involved are modest, so this is not an inherently difficult problem. The only web-based cold chain inventory tool of which we are aware was developed by UNICEF for use in India. It is currently undergoing pilot use in several states of India. The system tracks cold chain equipment at the facility level, and also maintains information about human resources and training.

Multiple other applications support logistics and the immunization system. These systems frequently maintain information about the cold chain, even though this is not the main purpose of the application. Pakistan's Vaccines Logistic Management Information System (vLMIS) is a web-based national system designed to give latest data on key vaccine logistics and cold chain indicators essential for better decision-making. The system is designed

with access to users from federal, provincial, and district levels with responsibilities varying from data entry to analysis and decision-making. This project is designed and developed by USAID Deliver project and John Snow, Inc. (JSI). Currently, it is running in 54 polio endemic and adjacent districts, with multi-year expansion plan to reach reporting from all districts. Other examples are the logistics management system OpenLMIS [157] and the vaccine stock management system VSSM [112]. Both of these applications track limited information about the vaccine cold chain. Currently, the cold chain inventory applications are completely separate from these logistics management systems, but there are obvious synergies when interoperability issues are addressed.

An alternative to building a custom inventory system is to build on top of a more general platform. DHIS2 is an open source health indicator reporting system developed by the Health Information Systems Program (HISP) and used in roughly 30 countries. HISP has been active since 1994 in developing health information systems with the goal of making health data useful at all levels of the health system [16, 17].The motivation behind developing DHIS2 was to improve the quality of health data and use information for action based on different tools [98]. DHIS2 is a three-tier application that uses Hibernate to manage the data layer, allowing multiple database implementations to be used, including PostgreSQL and MySQL. The service layer uses the Spring framework, and the web presentation layer uses Struts 2, which includes Jasper Reports, a GIS module, and JQuery.

DHIS2 allows system administrators to design the reporting units, indicators, validation rules, and data entry forms. This is significant since it makes DHIS2 a generic tool that can be easily adopted for countries implementing their health information system. The core data model for DHIS2 is designed around abstract data sets, data values and date elements associated with organizational units. The organizational units are then organized into an organizational unit hierarchy.

The facility model for CCEM matched the organizational units for DHIS2, so the facility data could be handled by the existing mechanisms. The extension that was necessary was for assets, where a collection of assets was associated with each organizational unit and had

their individual properties. Instead of directly implementing the cold chain assets types such as refrigerators or cold rooms, generic equipment types and equipment attributes were used. The asset model also relies on catalogs so that fixed properties of a type of equipment could be represented separately from the instance. To handle this generically, catalog types and catalog type attributes were included. This level of indirection allows new types of assets to be added by the system administrator without updating the DHIS2 code. For example, diagnostic equipment could be added to an instance of the inventory module just by adding appropriate equipment and catalog items.

With the completion of the asset module for DHIS2, we have a version of the cold chain tool running on a web-based system. We used one of our existing country data sets for the test version, and implemented reporting for 30-day temperature alarms and recorded equipment maintenance. The base module can be used for cold chain inventories for other countries and it is possible to extend the system to handle other data associated with the cold chain such as automatically collected temperature data [29].

5.5.4 Challenges

There are several challenges with cold chain equipment inventories. The logistical challenge to build an initial cold chain inventory can be enormous, since this is often done by having trained teams of workers visit all the health facilities in the country to collect information. Since countries typically have about one health facility for every 10,000 people, the number of health facilities is large, for example, Kenya has over 5,000 health facilities for a population of 44 million. Travel to all facilities is expensive and time consuming due to difficult roads.

After an inventory is constructed, the main challenge is keeping it up to date. This is, in fact, the major criticism of cold chain inventories: they are generally not kept up to date, and so the investment is squandered. We know of several countries (for example, Malawi, Nicaragua, and Uganda) that have kept inventories updated by maintaining a standalone database and having updates done centrally. However, in many other countries that we are familiar with, the inventory remains static after it has been collected. One of the arguments

for a web-based implementation of a cold chain inventory system is that it should be easier to maintain the latest state by allowing distributed updates. Some PC based implementations of inventories approach the update problem with a feed forward files to allow the merging of updates from multiples versions of the database for different regions.

The challenges surrounding implementation of an inventory model include maintaining data quality and allowing the system to be easily updated. Ensuring that the data is accurate can be extremely difficult, especially if the data is recorded passively and is not being used to provide feedback to people involved in the immunization system. Finding some means to keep the information up-to-date is the biggest challenge around inventory implementation. This relates both to the technology and the procedures that are in place for updates to be received and processed. The updating process is further complicated if the inventory is not managed centrally and is instead represented by disconnected data sources.

Inconsistent data models introduce a huge barrier in merging the inventories for better analysis. These issues start from a basic problem of merging separately maintained datasets on health facility logistic details and cold chain details. This makes a simple query of looking up available refrigerators with vaccines required a daunting and challenging task. Inconsistent data models also makes it challenging to feed the data to generic modeling and planning tools for better decision making.

Several organizations, including PATH, CHAI, Village Reach, JSI and local governments, are working on improving the data reporting for immunization. These efforts occur at regional or country level in various parts of the world. Most of these projects are running independently, which makes it harder to expand them at later stages and merge with other efforts. Major difficulties arise from mismatch in data being collected or the data format being used in various technologies supporting these projects.

We propose that with the appropriate technology, the opportunity exists to have cold chain inventory data promptly updated to reflect changes, as well as to incorporate additional information gained through routine reporting. Analysis and visualization tools at all levels could support planning and management tasks to ensure that appropriate equipment

is acquired and that the cold chain is of sufficient quality and capacity for immunization programs. Further, the cold chain inventory system could be tied to other information systems, such as those used for stock management, and could also serve as a backend for new applications that support features such as automatic temperature monitoring.

5.6 Data Standards

Through our work on multiple tools and country deployments, we became convinced that a major gap was a lack of common data standards for Cold Chain Equipment Inventories (or CCEI). This was reflected by difficulties in building country inventories based on available data, the fragmented tools used to work with inventories, and the confusion that people had between the inventory and the tool used to store it. Our hope was that the existence of a data standard would:

1. Regularize the process of data collection for cold chain equipment inventories, including making it easier to generate data collection tools,

2. Provide a mechanism for cold chain data to be shared between applications and allow applications to interoperate,

3. Give a basis for common analysis tools to be used across cold chain data sets,

4. Support the structured representation of inventory data, thus increasing quality of the data.

The CCEI data model was developed in a collaborative manner. The initial model and data definitions were based upon those inside CCEM. Since the CCEM tool had been developed through a series of stakeholder workshops in Uganda and Panama, and in consultation with UNICEF and WHO, its definitions were already in close agreement with the data reporting supported by WHO. The model was refined over a period of 18 months through multiple rounds of review. An initial review was conducted by members of UNICEF Cold

Chain Logistics group. Based on that effort, a more formal set of definitions was developed that was then circulated individually to about 15 immunization cold chain experts from UNICEF, WHO, and NGOs including PATH [117] and CHAI [27], who provided very detailed feedback that was incorporated into a final version. The project built upon existing standards where possible, such as using the WHO Performance, Quality and Safety (PQS)/Product Information Sheets (PIS) catalogs and defining several of the fields with respect to ISO standards.

The CCEI standard was established with a core to represent information about health facilities and refrigeration equipment. It is expected that the standard will be extended to include additional modules. Examples of modules that are under consideration, and are of interest to different stakeholders are transportation, equipment maintenance, and temperature monitoring.

One of the requirements for CCEI was to include sufficient information to assess the quality and capacity of a national cold chain, and to be a basis for estimates of the equipment that would be necessary to upgrade the cold chain for introduction of new vaccines. This requirement influenced the selection of indicators associated with the health facilities, such as recording the population associated with a health facility, and identifying the energy sources available for refrigerators. Since many of the assets in the cold chain are standard equipment, the inventory also includes an official catalog of models of equipment available. This means that a refrigerator can be represented by just a model name, and information for the refrigerator (such as capacity) can be pulled from the model database.

The data model is a set of facilities and a set of assets. Assets come from a group of predefined types (such as refrigerators and cold rooms) and there is a reference catalog that gives the properties of specific models of equipment. The most detailed information is associated with facilities, where location information, including position in the countrys administrative hierarchy, is stored along with information on the population served by the facility, the power infrastructure, and process of vaccine distribution. In addition to specifying the basic inventory models, the CCEI data model includes standards for the administrative hierarchy

and country localization.

Although we consider the outcome to have been a success, there were multiple challenges faced in constructing the standard. As with any standards efforts, there were disagreements over details of data elements and a balance between completeness and not complicating the standard. One of the challenges throughout the process was conveying to the stakeholders what a data standard was and distinguishing between the process of specifying the data definitions, and their use in various applications. Scoping the data definition was also problematic, as there were many suggestions of other components of the immunization system that could potentially be included. One solution to this was to state that in the future, extensions to the standard could be consider with the inclusion of additional modules. Finally, there is potential overlap of aspects of this standard, with other efforts, such as work on facility registries.

5.7 Discussion

We now present a number of observations from our experience with cold chain equipment inventories.

We believe that there is a need for multiple architectural approaches in the implementation of cold chain equipment inventories based on different contexts in countries. We consider a web based system utilizing a database for the cold chain inventory as the obvious architectural choice for many reasons, including reliability and the ability to support distributed updates. However, we has seen a continual demand for our stand alone Microsoft access tool, where it has been used in countries including Philippines, Indonesia, Pakistan, and Georgia for inventories. We came to understand that the autonomy that a PC application allows, so that it can be used without needing to engage a larger organization, is an important feature in some situations. Tasks such as creating a cold chain inventory can be managed by a small group of health officials, so being able to use laptops, without engaging IT staff removes an institutional obstacle.

It is important to be able to support the migration between inventory systems. This was

one of the motivations for the development of the data standard so that applications could import from / export to a specific format to migrate data. We used an excel representation of the CCEI model for this. We have used this to move data from multiple CCEM inventories, including Malawi, Kenya, and Ghana to DHIS2 systems, with other countries, such as Uganda, also interested in moving from CCEM to their national DHIS2 system. One of the successes in developing the data standard is that it has allowed other organizations to build tools with cold chain inventories and include the data from existing CCEM deployments.

Pakistan Cold Chain Equipment Inventory maintenance effort was started by UNICEF at small scale by targeting 55 polio high-risk districts in the country. They initiated the project by using CCEM as their primary tool that provided them the convenience to run it on their local PC or laptop. They collected data on paper forms from 2083 health facilities and then entered the data on local machines without the need of Internet. This database was then incorporated in a web-based national Vaccine Logistic Management Information System (vLMIS), developed by John Snow Inc at a later stage. vLMIS has scripts to extract data from CCEM files and add it to this online system.

The starting point for building a cold chain equipment inventory is almost always an existing dataset represented in excel - so there is a migration from an excel based inventory to one of the other tools - which again is accomplished by representing a standard form in a sheet. There are multiple challenges to doing this with difficulties around the structure of data (such as the choice of fields) as well as the quality of data, including issues such as having to deal with a wide range of spellings of the same terms. The representation of names of facilities can be a particular challenge.

Cold chain equipment inventory is a small yet critical dataset in immunization and health domains that is easily ignored. We expect any data collection system in this space will eventually be consumed by wide scoped national health information systems. Hence it is important that an agreed upon data standard exist so that data can be prepackaged and moved into popular information systems. Given that all the small scale and intermediate tools in this domain also support this standard, the interoperability of these individual

systems and scaling-up will be a much easier and smooth task.

5.8 Conclusion

The underlying goal of this effort is to make cold chain equipment inventories more useful to the management of country immunization programs. A major gap was the lack of a common model for inventories, leading to ad hoc representations, difficulties in moving between inventory tools, and the confusion between the actual inventory and the software that was used to represent it. We developed a data model, which has been shared with the relevant technical communities, and has been used to align a collection of existing software tools. We have the aspiration to see this evolve into an open data standard, but recognize that it is not at that level yet. The validation of the standard will be through its use in multiple tools and if it becomes a common representation used for by countries when storing and sharing their inventory data - in other words, if it becomes a defacto standard. Formalization of the model into a true open standard will require an organization that hosts the standard and a structure for community agreement in updates to the standard.

We have discussed multiple use cases that we encountered while working in this space where a data standard could play a critical role. These use-cases include spreadsheet to database data migration and coexistence of single machine and web-server applications. We also discussed some future work in standardizing analysis tools that can be achieved given the data standards are deployed in information systems.

Chapter 6

CASE STUDY: AN ASSESSMENT OF SMS FRAUD IN PAKISTAN

6.1 Summary

This chapter aims to enhance the understanding on the issues of collecting crowd source data with regard to financial issues. This chapter illustrates the challenge in acquiring consistent data from the public due to lack of trust or motivation. Moreover, it takes a significant amount of time to review data and establish ground truth. Machine learning could potentially solve these issues, but such approaches require a lot more data than what is often collected. Hence, I claim that review needs to be done by people involved in the collection process. Spelling errors due to transliteration present barriers to query the data based on names or nouns. I could not automate the review process for this data due to its unstructured and open text. These challenges illustrate the issues regarding open text fields in survey data, which could result in a variety of input values for the same item.

6.2 Introduction

Financial inclusion, consisting of access to the formal economy, including banking, loans, and credit, has been recognized as an important development objective [2, 137, 49]. As the world economy moves to be more digital through services like electronic banking, credit cards, mobile money, digital payments, and other mechanisms, development organizations have begun promoting digital financial services (DFS) as a primary means to achieve financial inclusion. With this ongoing promotion, attention, and adoption, research has shifted to assessing the myriad barriers to the uptake of DFS. These include network and service outages, insufficient agent liquidity, complex user interfaces, poor customer recourse, inadequate data privacy,

non-transparent fees, and fraud that targets the customer [94, 70]. Fraud, in particular, is especially problematic because low-income and marginalized populations are more sensitive to financial loss.

While fraud is a problem in many countries, anecdotes regarding SMS-based fraud are particularly common in Pakistan and have a wide circulation among DFS researcher networks [42, 103, 155]. Similarly, for those who have spent time in Pakistan, personal observations of incoming SMS messages attempting to initiate fraud are seemingly common. While it may be surprising that SMS is the primary vector of these attacks, as mobile money is not directly implemented with SMS, many components of mobile money systems, such as transaction receipts, use SMS [125]. Similarly, both SMS and mobile money are associated with basic mobile phones. Therefore, the perception exists in the digital financial service community that SMS fraud is widespread and directly targets mobile money operators and users [100].

To investigate the scope and scale of the problem of SMS fraud, specifically in Pakistan, I developed and deployed a SMS data collection application (app) at a University in Pakistan, with help of my research team. This app gathers user SMS and sends it to be anonymized and analyzed at our in-country research server. Given the sampling bias inherent in such a method, we also extended this data set with an advertised SMS-forwarding service, widening our participant base outside of the university. Lastly, to better understand the experience of fraud, we conducted interviews with low-income rural and urban Pakistanis who would be most harmed by fraud. These combined datasets provided a wide base to better understand the practice and impact of SMS fraud in Pakistan. In particular, I wanted to address the following questions:

- What does SMS Fraud *look like*? What sorts of attacks are *common*?
- Are *DFS* a common component of SMS-based Fraud?
- Can we *easily* and *accurately* detect and classify SMS Fraud in Pakistan?
- What are the unique properties of SMS fraud in Pakistan?

- How is fraud experienced among *rural and disadvantaged* Pakistanis?

Our results show that there is a large ecosystem of fraudulent SMS in Pakistan, which I categorize into ten different types and three different classes. Surprisingly, financial services are largely *not* a component of this ecosystem, despite the push for DFS in Pakistan. I also demonstrate that a simple classifier can detect a large majority of fraud messages with a small false positive rate. Lastly, I show that the SMS fraud ecosystem in Pakistan has a number of unique features, including language-based targeting and that rural users, just coming online, may be primed to be victims of future SMS-based fraud.

6.3 Related Work

6.3.1 Spam

The detection, avoidance, and blocking of spam is one of the most studied problems in computer science. Spam can take many forms, including email [36, 57], web sites [59], and even voice [90]. SMS, as a communication medium, has similarly been targeted by attackers. Several works have looked into adopting email spam filtering techniques for SMS spam detection [56, 43]. Sophisticated algorithms are needed as the spam and SMS traffic evolves with time [102, 76, 134]. Unfortunately, these systems are highly language-dependent, although there have been efforts for SMS Spam filtering using non-context based features [165]. These filtering systems are often deployed within cellular networks, blocking at the SMSC [74].

My research departs from this robust body of prior work in that I do not focus on solving the problem of SMS spam in Pakistan. Indeed, it is likely that an off-the-shelf filtering algorithm is already present (but disabled) in the telecom messaging center. Our discussions with telecom representatives indicated that the government considered enabling such filtering as a form of "censorship". As such, the problem is largely structural rather than technological. This provides a unique opportunity to explore an active fraud and spam SMS ecosystem.

6.3.2 Fraud

Digital fraud, much like spam, is a heavily studied area. Scammers use mediums like email [71], voice [97], and SMS to reach a wide audience with socially engineered attack messages [4]. To successfully scam people, scammers deploy various strategies to make their messages appear to be official and trustworthy [97, 152]. Several works have looked into how various versions of the Nigerian email scam or similar fake lottery scams have spread over email and what happens once a victim reaches out to a scammer [71, 24]. This chapter extends prior work by exploring attacks present in a new geographic region (Pakistan) and the unique properties of attacks found there.

6.3.3 Digital Financial Services

Digital financial services (DFS) are a set of services that provide access to formal banking solutions through mobile technology [77]. These services, such as mobile wallets or credit solutions, rely on existing cellphone communication channels like USSD, SMS, or data. As a nascent technology in many markets, the technical aspects of DFS systems do not appear as the subject of many studies. Reaves et al. [135] examined and summarized vulnerabilities present in developing world DFS apps due to insecure connections or data leakage. Similarly, Castle et al. [26] expanded the threat model and pointed out SMS as a vulnerable communication channel in DFS due to the lack of number verification that can lead to SMS spoofing. Phipps et al. [126] explored the potential for ThinSIM-based attacks on mobile money systems. My work provides a supporting view into ongoing attacks on DFS systems, finding that they are largely absent in the current ecosystem.

6.3.4 Security in the Developing World

Numerous researchers have explored security in developing contexts [10, 156]. Common themes include social mismatches between technologies and cultures [78], providing security in light of infrastructure failure [37], and the application of developed-world best practices

to other areas [78]. This research is firmly in the same area, detailing the experiences users in a developing country (Pakistan) have with attempted SMS-based attacks.

6.4 SMS Fraud Background

To begin, I first define fraud in SMS and then provide examples of SMS fraud that motivated this study.

6.4.1 Definition

For this chapter, I define fraud as an act where one person is attempting to deceive another person to get money or other items of value.

Items of value may include credentials such as account numbers, PINs, or personally identifying information that can be used to acquire other things of value. SMS *fraud* thus refers to fraud that are initiated or executed through SMS. I distinguish SMS *fraud* from other forms of unwanted SMS messages such as unsolicited advertisements or *spam*. I discuss this distinction in more detail in the Data Analysis section.

6.4.2 Examples of SMS Fraud

To familiarize the reader with the context for this work, I present and discuss several examples of SMS fraud gathered from online sources prior to embarking on our work in Pakistan.

One of the most common types of SMS fraud is what I refer to as the *lottery fraud*: A fraudster sends a message informing the recipient that he or she has won some money, and that the person must contact a certain number to receive it. Once the victim calls back, the fraudster convinces the victim that they must pay a fee to obtain the prize money. A payment is made by some mechanism, but the prize money is never delivered [161]. This is also a staple among email fraudsters who announce winnings of implausible lotteries such as the BillGates lottery. Typical examples include:

Ghana (Twitter): *Valued customer, ur number is one of our lucky winner of Gh12,000 on Airtel Wo Mner3 Nie promo! Call Mr Owusu to cash it out on 026263xxxx*[1]

Kenya (Twitter): *CONGRATULATIONS from SAFARICOM MAISHA NI MPESA TU! Promotion, You have Won. Ksh 100,000.00 your secret code 555555 Call (078349xxxx) for more information. NB: DO NOT PAY Anything.*

A second common SMS fraud is *receipt fraud.* In this case, a fraudster sends a fake receipt for funds added to a targeted subscriber's account and then calls the subscriber to ask for that money back, stating that it was accidentally sent to the subscriber's mobile wallet. The receipt format mimics the form of a legitimate mobile money receipt. Depending on the sophistication of the fraudster, there may be an attempt to spoof the sending address, although the majority of the examples I have seen appear to originate from a mobile number. The receipt fraud is a direct attack on individuals, as opposed to the lottery fraud which relies on sending a large number of messages to collect responses. The receipt fraud is directly linked with mobile money, so it can be classified as digital financial service fraud, while lottery fraud is not. Examples of receipt fraud are as follows:

Kenya (Twitter): *MPESA LHR9VQ7DKE Confirmed.You have received Ksh9,730.00 from BEN ONYANGO 070671xxxx on 15/8/17 NEW M-PESA balance is Ksh*(Pending)*Pay Bills via M-PESA.*

Uganda (Twitter): *MTNMobMoney. Y'ello. You have received UGX973,000.00 FRrom: KTM LTD. Token ID: 79864532991. Remember to get secret code from sender to access your funds.*

6.5 Data Collection

Gathering a broad and representative sample of SMS fraud in Pakistan required a range of different data collection methodologies. My research team began with the development

[1] I obscure the final digits of phone numbers for privacy reasons, though the numbers themselves are likely no longer valid.

and deployment of an app-based solution that would provide "ground truth" of SMS spam and fraud rates by recording and transmitting all SMS messages received by 246 study participants. Following this, to ensure that we discovered all types of fraudulent messages, we set up and advertised a fraud SMS forwarding phone number for users outside of the study, with a total count of 518 messages received. In order to broaden our study to the voices of non-smartphone users, we conducted eight interviews outside of Lahore with low-income, basic phones users.

IRB approval was obtained for all aspect of this study, with rigorous safeguards in place to ensure anonymity, security, and privacy for the participants. Additionally, when the smartphone app users opened the app for the first time, they had to go through several screens explaining the research objectives and the data that would be uploaded. On the last screen, they were asked to consent to the data upload and the privacy agreement by checking a box on that screen.

6.5.1 Smartphone App-based Collection

We began our data collection by developing an app for Android phones, the most commmon smartphone platform in Pakistan. This platform records some received SMS, as decided by the user, and forwards them to research servers for collection and analysis. Users set the app as their default SMS handler, allowing it to investigate all messages passing through the phone.

Our application, known as "Safe SMS", was built as an extension of the existing "QKSMS" open-source SMS handling app [13]. Our version was extended to provide the following new features:

1. Upload messages and labels to a secure research server;
2. Allow users to label a message as either *Fraud*, *Spam*, or *Ok*;
3. Allow users to mark a conversation as private so it is never uploaded; and

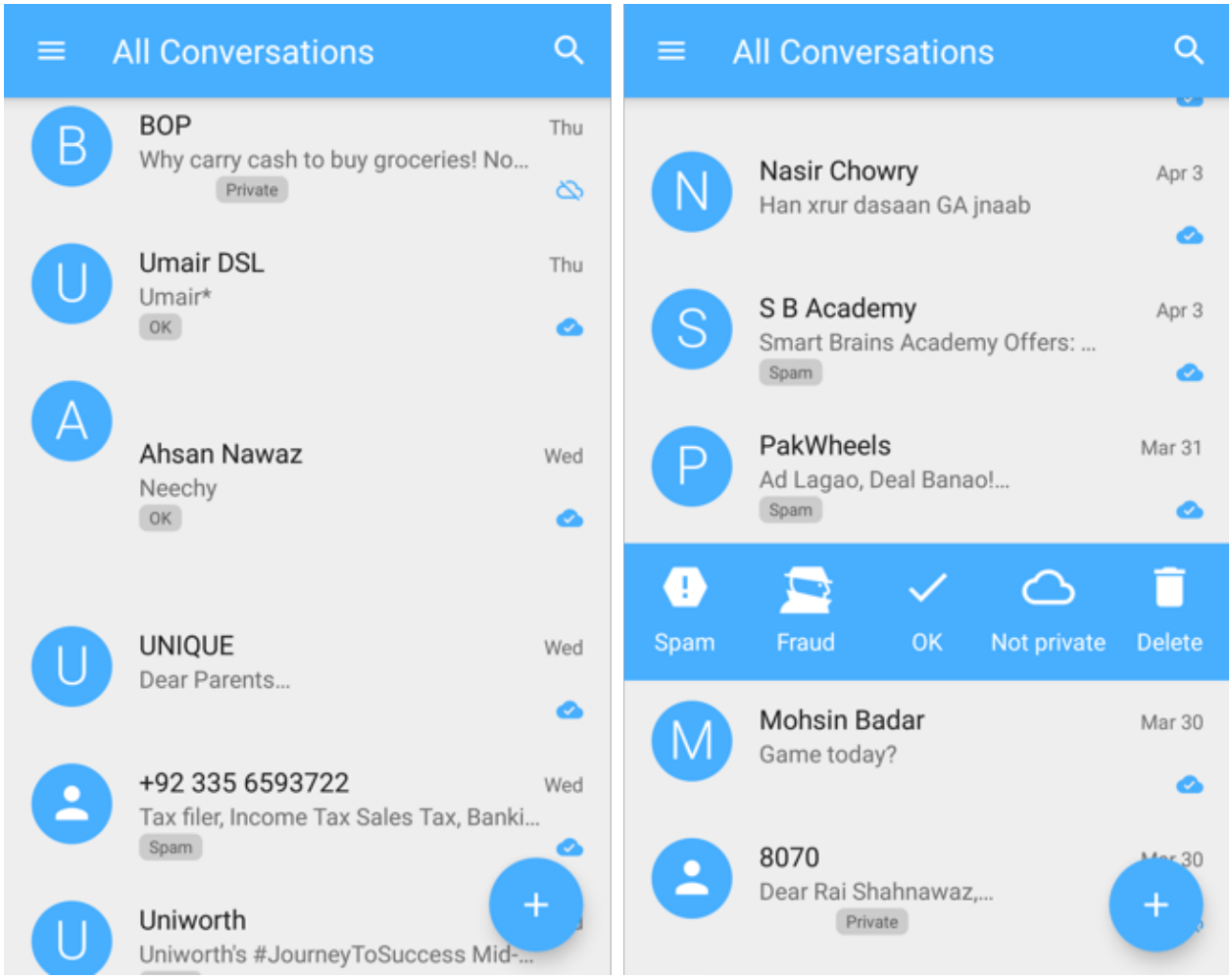

Figure 6.1: Screenshots from Safe SMS app, showing how to label a conversation

4. Offer an in-app tutorial explaining the research project, organizations, and the additional features.

The app kept a database (DB) of supplemental variables in parallel to the Android system-wide SMS DB. The local app DB tracked a user's marked label for a given thread, the upload status of messages, the SMS header of messages received after a user agreed to the app's consent, the location of the device when an SMS was received, and the list of threads that a user wanted to maintain as private.

Furthermore, all the data was sent a over secured connection to make sure the data was sent as encrypted to our verified server. The data upload happened in the background after a user visited to the upload screen and manually pressed the upload button. The user received a notification once the upload was complete.

Labels: We introduced labels as a way to ascertain the user's perspective about the nature of the messages. Users could label a conversation with one of three labels: 1) *Fraud*, indicating that the message is an intentional effort to defraud the receiver, 2) *Spam*, indicating that the message was a broadly sent advertisement, and 3) *OK*, indicating a normal conversational SMS. The goal of this labelling was not to provide a "ground truth" of the content of the messages (instead I tagged them manually) but to explore if our definition matched the perception of our participants. The app also had a screen to sort and view all of the unlabelled messages or messages with specific labels, to encourage correct labelling and reduce satisficing [82].

Private Messages: Given the amount of personal information present in SMS communications, supporting privacy was extremely important. We did this through an upfront option to mark a conversation as "private". Marking a conversation as "private" excluded any message in the conversation with that person (with the phone number as the identifier) from being uploaded to the research server. Similarly, if any message was uploaded in previous sync and later marked as "private", the app would send this change to the server resulting in the data being deleted. Data analysis was done at the conclusion of the study, so no intermediate data (uploaded but then marked "private") was used in the analysis.

Tutorial: Given the extensive additions to the default "QKSMS" app, we built a tutorial in our app, "Safe SMS", that introduces the user to our team and research agenda, as well as how to mark a conversation as private, how to label conversations, and how to upload data. The tutorial initiates when the app is launched for the first time on a device. A user can initiate the tutorial again by pressing the question button which is available from the menu on all screens.

The app was uploaded to the Play store, which is the official app store for Android

smartphones. This made it easy for us to distribute the app with a single URL in which the app install page on the Play store opens. We setup a research server in Pakistan with firewall and security certificates in place to ensure all communication between the app and the server was encrypted and secure.

App Testing

We conducted a series of iterative design sprints to shake out any bugs or confusion about our design. This was done in two rounds: first internally with seven members of the research team and then externally with eight recruited outside users from our home university. The external testers were asked to install the app and upload some SMS conversations. After this we interviewed these users for feedback. The users were tested with improvements implemented from the previous phase's feedback.

Our findings pointed to the the privacy implications of the app. Users were uncomfortable with the app's repeated requests for permissions during the install. The app recorded the location of the users to be able to categorize the user as rural or urban. We removed this feature to require fewer permissions. Second, users were uneasy about the messages being uploaded in the background. To resolve this, we introduced a screen that lists all the unsynced messages, showing users that only selected messages were being uploaded. The user could search and filter the conversation based on the marked labels to make it easy to select all intended conversations. This was also the step where a user could further exclude sensitive or personal messages from being uploaded, even if they were not initially marked as private. This could introduce bias in our data as a user might select only the fraud related messages for upload, which relies on their perception of fraud. Several users reported confusion about label icons and how to track which conversations had been labeled. We removed all the label icons and only made them visible in the menu that had full text. Moreover, we changed the view that lists all the conversations. We added label names, sync status, and a privacy tag on each conversation so the user could scroll through the list quickly to look up the message status rather than having to select and view each individual conversation.

App Deployment

Once the system was generally accepted by our testers, we expanded to a wide-scale deployment outside of our research group. We advertised the app at a local university in Pakistan with the initial target being students, whom we considered among the most text-savvy. Also, we encouraged faculty and researchers at the university to advertise the research project in their research labs and class rooms as a means to drive adoption. Some faculty requested that we provide a five minute talk in their class about the purpose of the app as well as how to use it.

We became concerned that only asking university students to test the app could skew our results in unpredictable ways. To resolve this, we used personal connections in different rural and urban localities to generate a more diverse set of participants. In urban areas, we approached 125 individuals from different offices and incubation centers. We also forwarded the app to 40 people through our personal contacts, akin to a convenience sampling. In rural areas, we reached out to 20-25 individuals from various villages through similar personal connections. In the end we had 246 users who downloaded and installed our app.

6.5.2 SMS Forwarding Service

Although the app provides a robust base to explore fraudulent and spammy SMS, our participant selection methodology could incidentally introduce a sampling bias and cause us to miss other classes of messages. To resolve this, we set up a Pakistan-based phone number where anyone, including those not participating though the installation of the app, could forward fraudulent messages through SMS or WhatsApp. This allowed us to gather fraudulent SMS from a broader audience, with any kind of phone (basic, feature, or smart), people with limited technical skills (forwarding an SMS is more broadly understood as compared to downloading and installing an app), and people from outside of the initial geographic range of our study. While it does not eliminate sampling biases, as our goal was to gather as wide a range of data as possible, this intervention gave us insight from a much broader sample.

The forwarding service was implemented on a smartphone with an active SIM and an enabled WhatsApp service. Our advertised number could be used to send SMS or WhatsApp messages to this device from anywhere in Pakistan. The phone received all the forwarded SMS messages as well as all the message screenshots or texts sent over WhatsApp. The SMS messages were uploaded to our "Safe SMS" app server under a specific user that represented this forwarding service. All the WhatsApp messages were human transcribed from images into a database on the same server.

In Punjab province, home to over half of Pakistan's population, the service was widely advertised through social and print media both in Urdu and English languages. The advertisement asked citizens to forward the fraudulent messages for a public good, and they were advised to append "Forwarded from: .." at the beginning of the message or take a screenshot of the SMS and send that over WhatsApp or SMS. Over the course of seven weeks, we received 746 fraud messages from 351 users.

One shortcoming of the forwarding technique was that it lost additional data around the forwarded SMS, such as the sender's number, the receipt time, and any messages previously received from that number. Although we requested that people send the number from which they received the fraudulent SMS number, some only sent the fraudulent SMS content.

6.5.3 Interviews

In order to investigate further how fraudsters successfully defraud people, we conducted informal interviews with eight participants. We focused on conducting interviews with low-income and illiterate people with the assumption that fraud schemes may target them. Three of these participants were part of a low income population within a major city while five were from a rural village, both in Punjab. The interviewees were randomly approached in markets and public areas of the village, and may not have been aware of the SMS fraud research prior to us approaching them. All of the village participants were from the home village of one of the researchers. Due to privacy concerns, we took field notes but did not record conversations.

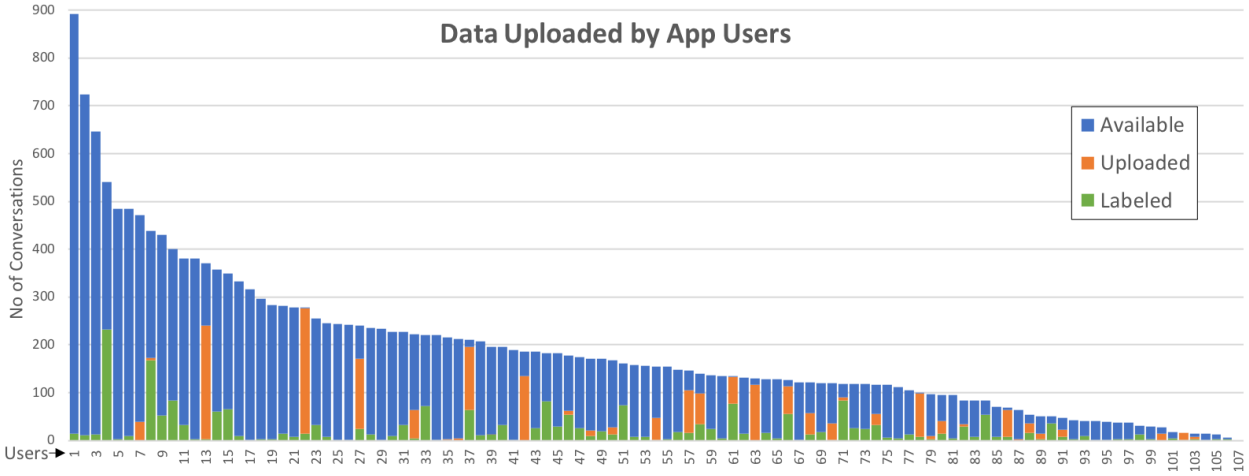

Figure 6.2: Number of conversations that were available on each user's phone, the ones they uploaded and the conversations that were labeled by them

6.6 Data Analysis

Our "Safe SMS" smartphone app had 246 installs with 106 of those users uploading some data. Out of those, 100 users self labeled part of their data with the tags *Fraud*, *Spam* and *Ok*. Figure 6.2 shows the number of conversations available on each user's phone and the number of messages they uploaded. A conversation is a thread of messages sent and received between the unique pair of a user and an external phone number. Collectively, users uploaded 4057 conversations that consisted of 52,169 total messages.

6.6.1 Authoritative Researcher Labels

While we did ask participants to label messages, I never intended to use these to evaluate the ecosystem as users could interpret the labels in different ways. Instead, I sought to create an authoritative, "ground-truth" set of *researcher labels*. To do this, the research team labeled all the conversations. The data was labeled by two reviewers including myself, irrespective of whether a user labeled it or not. Those who labeled the messages are native speakers of the local languages (Urdu and Punjabi) in which the messages were composed. After each

Table 6.1: Summary of User Labeled Data.

User Label	Conversations (Messages)	Actual Labels for Conversations (Messages)					
		Fraud	Ok	Spam	Spam with Status	Status	Unknown
Unlabeled	1869 (33233)	20 (22)	816 (28198)	788 (2403)	21 (819)	188 (1481)	36 (310)
Fraud	148 (215)	**111 (113)**	3 (8)	33 (93)	0 (0)	1 (1)	0 (0)
Ok	264 (9661)	1 (1)	**152 (8525)**	62 (615)	9 (114)	40 (406)	0 (0)
Spam	1776 (9060)	12 (12)	29 (146)	**1641 (6228)**	22 (1471)	62 (1182)	10 (21)
Total	4057 (52169)	144 (148)	1000 (36877)	2524 (9339)	52 (2404)	291 (3070)	46 (331)

reviewer went through the labeling exercise individually, they merged their labels, discussing any mismatches and inconsistencies to reach consensus.

Through the labeling process I discovered that the original categories of *fraud*, *spam*, and *OK* were insufficient. As such, I expanded the categories to include the following:

- **Ok:** Conversation between two real people.
- **Fraud:** SMS that deceive a user to defraud them of their money.
- **Spam:** A generic advertisement for products and unsolicited public messages.
- **Status:** A system notification for a specific user about their account status, action taken on a service, or delivery update.
- **Spam with Status:** The same number sends spam with advertisements as well as sends account status updates. This is common for services where a user has an account, like a telco, a bank, or a ride sharing service.

A user might consider a *Status* message as *Spam* or as *Ok*, so I created the more specific label. Moreover, some spam could be considered fraud. A misleading advertisement about credit or insurance that over promises, and, hence, can be regarded as fraud, would still be labeled as spam. For example, this message is misleading because permanent residency processes take longer than one month but I labeled it as spam:

> Australia Permanent Residency Approval in 1 Month BA/MA Are Eligible 2018 Relax Policy 100% Success Embasy Fee Also Return If Rejection 03211818190 Natasha.

In general, users have a reasonable understanding of fraud and spam messages. Table 6.1 lists the conversations that users labeled with the corresponding messages' count in brackets. Users labeled *Fraud*, *Ok*, and *Spam* with 75%, 58%, and 92% accuracy respectively. I adjusted the numbers to remove the confusing messages with labels *Status* and *Spam with Status* that could be considered as either spam or OK by the user. I found that accuracy of user labeling increased to 75%, 71%, and 97% for *Fraud*, *Ok*, and *Spam*, respectively. This illustrates that we can rely on user labeling to a certain degree, but more importantly indicates that our definition of labels is reasonable given that overall 93% of user-labeled conversations agreed with our labels. These users were sophisticated considering that they were motivated to contribute, comfortable with determining a label, and spent time marking data.

6.6.2 Forwarding Service

Our SMS forwarding service significantly expanded our fraud message corpus. We received 152 messages from the app, 228 from people forwarding the fraud message to a central number, and 518 from text or screenshot of the text via WhatsApp. While these numbers are impressive, 39% of forwarded messages did not include the address of the sender. Hence, we did not have the sender's information for 289 out of 898 fraud messages.

6.6.3 Data Collection Challenges

We had challenges in data collection due to lack of trust and our strategy to incentivize using social and public good rather than offering monetary benefits. This hindered mass adoption of the app resulting in 246 users out of which only 106 uploaded any data. Moreover, the forwarding service received only limited data since people tend to delete fraud and spam messages especially those received on feature or basic phones.

Table 6.2: Summary of Fraud Types.

Fraud Scheme	Fraud Type	Count	Roman Urdu	Urdu	English
ARY Jeeto	Lottery	557	553	4	0
BISP	Lottery	215	127	88	0
Waseela-e-haq	Lottery	46	46	0	0
UK Award	Lottery	40	0	0	40
Easyload	Damsel	31	27	4	0
Scholarship	Lottery	4	4	0	0
Bank Service	Steal Creds	2	2	0	0
ATM Card	Steal Creds	1	0	0	1
Pak Army	Lottery	1	1	0	0
Zong	Lottery	1	1	0	0

6.7 Results

6.7.1 A Rough Taxonomy of SMS Fraud in Pakistan

In consolidating the fraud messages gathered from various methods, I found ten repeated fraud schemes as shown in Table 6.2. I classified these schemes into three categories. The most common type is the lottery type where the fraudster announces that the user has won or received something from a lucky draw or a scholarship or public program. The second type, I call "Damsel in Distress", in which a fraudster poses as a vulnerable young woman who is in need of help. The fraudster appeals to the user to send a few mobile credits that she will return later. Finally, I saw a few instances of a fraudster trying to steal credentials by stating that the victim's bank services, like SMS notification or ATM card, had been disabled. The fraudster invites the victim to call to reactivate the services, but provides a number through which the fraudster can intercept the call and obtain the victim's personal information.

Each scheme has a generic story that can be broadcast to anyone. Nonetheless, I observed fraudsters who change the formatting with new lines or adjust the wording or spacing slightly.

Table 6.3: Examples of fraudulent messages that were collected. English translations are given for messages sent in Roman Urdu.

Fraud Scheme	Example	English Translation
ARY Jeeto	ARY JEETO PAKISTAN K show me apne is 03047227028 se 8038per SMS kiya tha Qrandzi me ap ki 1adad bike or 5lakh cash nikla he ap is no 0303769xxxx pr call karen	In ARY JEETO PAKISTAN show, you sent SMS from 03047227028 to 8038. From the lucky draw, you got 1 bike or 5 hundred thousand cash. You should call this no 0303769xxxx
BISP	BENAZIR INCOME support ki taraf se apko Rs.25200 mubarek ho.apka ye number0323759xxxx BISP mein Register tha.ap is number per 0306709xxxx rabita karen..	From BENAZIR INCOME support, congratulations on Rs.25200.Your this number0323759xxxx was register in BISP.You should contact this number 0306709xxxx..
UK Award	CONGRATS! YOUR MOBILE NUMER HAS WON 500,00 POUNDS IN THE 2018 PEPSI PROMO. TO CLAIM YOUR PRIZE. SEND UR NAME, AGE AND MOBILE NUMBER TO: ppeawd@hotmail.com	
Easyload	mery is number 0304946xxxx py 50 ka Mobilink load karwa do main bad me wapis kar don gi call nne.pleasa.saba	Send Mobilink load of 50 to my this number 0304946xxxx. I will return it later call nne.pleasa.saba
Bank service	Dear Customer! Ap Ki HBL Ki SMS ALEART Service Khtm Ho Rahi Hai Dubara Free SMS ALEART Active Krne K liye Is Pr Visit Krain . Visit www.smshbl.com	Dear Customer! Your HBL SMS ALEART Service is ending. To make Free SMS ALEART Active again, visit this . Visit www.smshbl.com
ATM Card	Dear Coustmores,your ATM card has been blocked Because you did not have an update yet. If you want your ATM card to work properly, then contact this	

There are several techniques that fraudsters employ to appear legitimate. Sometimes they will add a user's number to the message to make it specific to that user. I also found a message that claimed to be from Zong telecom service and provided the actual website address in the message, however, the call back number which was provided was not Zong's customer service number. A few UK award messages were sent from a number with a +44 country code (the UK's actual code), while the majority of messages with that scheme were sent from numbers that originated from Pakistan. Some fraudsters are willing and able to make their messages appear more authentic.

Most fraud requests their victim to make a call. The call, in turn, initiates the fraud and makes it more challenging to trace. Moreover, up to five fraud messages in our dataset, complete with unique call back numbers within the message, would originate from a single number. This suggests that a huge cache of unique SIMs are being used in order to conduct fraud. This was an unexpected finding, in that the Pakistani government has regulated SIM registration and required biometric verification linked to a person's national ID card to receive a new SIM. In addition to this, an individual can only have up to eight SIMs (five for voice and three for data). Therefore, fraudsters with large caches of SIMS might indicate a black market for registered SIMs linked to the identities of people who are unaware that the SIMs had been obtained for such purposes.

6.7.2 DFS-Specific Fraud

Surprisingly, fraud over SMS in Pakistan is not related to DFS. It contributes to more general call based fraud since the majority of fraud schemes ask victims to call instead of directly send money. The easyload fraud, where the fraudster asks for money directly, requests that victims send airtime credits instead of money over a mobile wallet. There was only one instance in that scheme where the message asked for "Jazz Cash" instead of easyload. I also know from our interviews, discussed in next section, that the call to lottery based schemes will result in the victim being asked for small fee in the form of airtime credits through prepaid scratch card numbers. There are a few sophisticated frauds where a fraudster tried

Table 6.4: Summary of Heuristic Results

True Label	Total Messages	ID'd as Fraud	Percentage
Ok	36877	35	00.10%
Spam	9367	10	00.11%
Fraud	898	891	99.22%
Status	3070	16	00.52%
Spam/Status	2404	0	00.00%

to obtain bank account details or credentials, but those are very rare in our dataset.

6.7.3 Detecting Fraud

Specific features of messages allow us to distinguish fraud from regular and spam messages. Figure 6.3 shows the percentage of different labels that were positive for each feature. A fraud message is never sent from a short code and typically has a call back phone number in the message. Relatively, fraud messages are likely to have congratulatory words, phrasing related to receiving something, and terms about a lucky draw. Congratulatory words included various spellings of "congratulations" or "mubarak" (Urdu for congratulations) both in English and Urdu. Words related to receiving included "won", "awarded", "nikla" (Urdu for got), "mila hai" (Urdu for received) and "aye hain" (Urdu for came). Furthermore, I consider certain keywords related to the type of fraud, like "lucky draw", "qrandazi" (Urdu for lucky draw) and "load" as these types of fraud aim to get money through easyload or announce that the victim has won a lottery. Even with some features better correlated with fraud, there is no one feature that can distinctively indicate fraud.

I explore several heuristics based on the insight I gained from the data. The one that was most effective for detecting fraud is as follows:

$$f(x) = (N(x)\ OR\ M(x))\ AND\ (C(x)\ OR\ R(x)\ OR\ L(x))$$

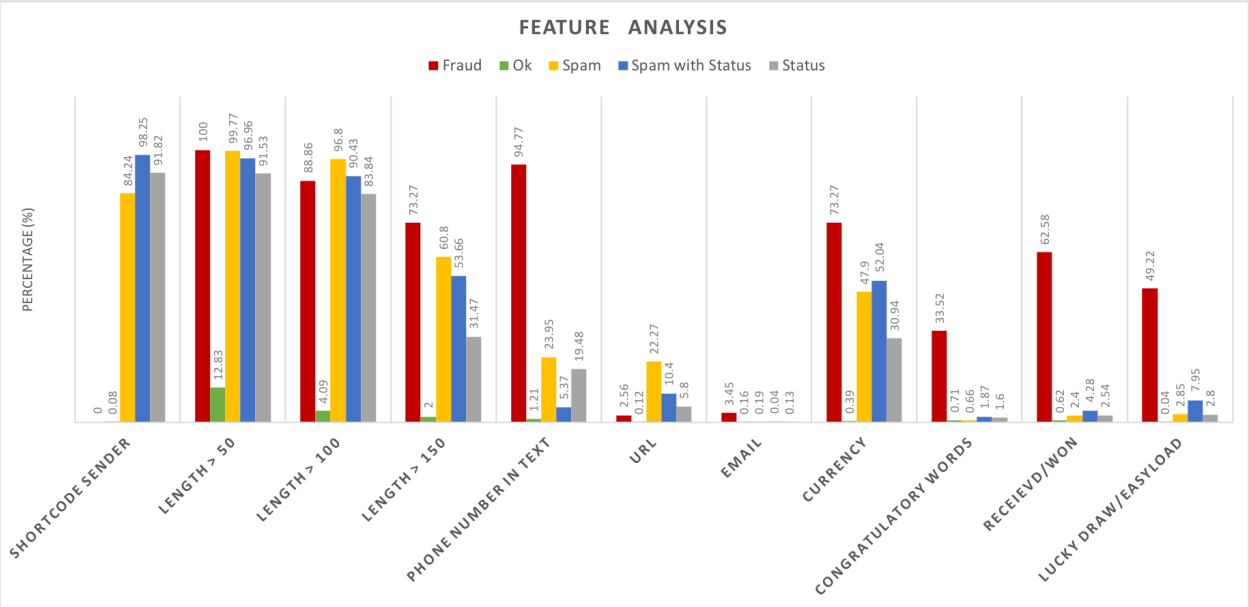

Figure 6.3: Presence of different features with important ones highlighted

where N, M, C, R and L are functions that return as true if the phone number, currency, congratulatory words, receiving words, or lucky draw-related words are present. The results of the heuristic are shown in Table 6.4. It detects 99.22% of fraud messages as fraud and less then 1% of other labeled messages as fraud. This is promising in that a simple heuristic-based, unsupervised algorithm can detect fraud without knowing specific details about different fraud schemes. This makes it generic enough that if a new lottery scheme appears, the algorithm would likely be able to detect it. The heuristic has its limitations as fraud vocabulary evolves, and, eventually, it could be difficult to differentiate fraud from spam. This can be avoided by building heuristics specifically for spam detection, an area that is well researched. Moreover, I believe that the heuristic for fraud has to be able to evolve.

6.7.4 Fraud Specific to Pakistan

In Pakistan the majority of fraud messages are composed in Roman Urdu, while only one scheme is consistently sent in both Arabic Urdu and Roman Urda. This exception is the Benazir Income Support Program (BISP) scheme. BISP is an unconditional cash transfer program for low-income women, which is arguably why it is sent using both scripts. Two other schemes send a few messages in Arabic Urdu, yet they more commonly use Roman Urdu. The "UK Award" scheme and the "ATM Card" fraud are sent in English only, as they are intended for a different audience with ATM cards or have email savvy. This illustrates that fraudsters adjust the language and scripts of messages depending on their intended audience.

Furthermore, my intuition is that some of these fraud messages are hand-typed on mobile phones because of the tremendous amount of variation in the transliterations of words. This generally arises due to multiple authors transliterating into roman letters. For example, "Mariam" and "Mariyam" represent the same name in Urdu. "Chak", "check", "chek", and "cheak" are various spellings that refer to a bank check or cheque. I had difficulty running my queries for standard Urdu transliterations like "Mubarak" (congratulations) because of the huge variety of transliterations found in the data, which included, "mubarik", "mobarak", "mubarek", "mubarak", "mubark", "mobarik", and "mubraka".

6.8 Qualitative Analysis

Here I describe the findings based on our interviews in low-income areas of Punjab. These results, by in large, reinforce my quantitative analysis, however, the five interviews with people from rural areas were insightful for understanding who may be most at risk of being defrauded.

6.8.1 Urban Participants

The interviews with three participants from urban areas revealed that all SMS users received promotional spam messages as well as fraud messages. Each participant was aware of fraud SMS messages and none of them had ever responded or fallen for fraudulent schemes circulated over SMS. They were reluctant to share their fraud messages, explaining that they were not sure about our intentions for asking to view those messages. Overall, they were aware of fraud over SMS and took a suspicious approach to SMS activity received from unknown numbers.

6.8.2 Rural Participants

The five participants from the rural village were semi-literate with education ranging from 8th to 10th grade, with the exception of one who attended college. Nearly all of them reported that they are low-income or belong to low-income families. They could read and write messages in Urdu and transliterated Urdu, they had smartphones with internet capabilities.

Four out of the five rural participants reported that they had not received spam or promotional messages. The exception was the person who went to a college in a nearby city. Nonetheless, all of them reported that they had received multiple fraud messages. I hypothesize that the fraud messages are being sent to a wider list of (or potentially even randomly generated) numbers as opposed to spam messages that target phone numbers collected from various services that urban dwellers intentionally or unintentionally subscribe to like pizza delivery, weekly deals, or online shopping order notifications.

All of the rural participants had experienced fraud or had witnessed someone fall victim to fraud (or at least engage in making an initial response to fraud). Three participants reported successful fraud attempts, while two discussed how they responded at first but them realized the fraudulent nature of the communication. The following are reports from the five rural participants:

Rural Participant One reported that he received a call from a regular number (non-

shortcode number) and was told that he had money waiting for him from Benazir Income Support Program (BISP). He was asked to pay Rs. 500 as a registration fee. The money was requested in the form of teleco prepaid credits by sending an SMS with the number from the balance reload scratch card. The participant sent the SMS but subsequently did not hear back from the fraudster. Later he tried to call the number but it was disconnected.

Like the first participant, Rural Participant Two received an SMS about him collecting money from BISP. After he called the number as requested in the message, he was told that he needed to pay a fee to initiate the disbursement process and that the fee should be paid in the form of easy credit reload. He discussed the scenario with an imam at his local mosque who warned him that this was fraud. The fraudster kept calling him even though he told the fraudster that he suspected fraud. Eventually, the participant gave the number to a trusted person to report to authorities.

Rural Participant Three received a BISP fraud message in Arabic Urdu script. The fraudster informed him that he had received Rs 30,000, but to receive the money he was required to pay Rs 2,000 using Telenor's Easypaisa, which is a telco mobile wallet. The fraudster sent him a number to dial for whomever has an Easypaisa account. The participant went to an Easypaisa agent for help to make the transfer from his account. The agent explained to him that dialing the given code would transfer all the money from his account. This made the participant realize that this was a fraud.

Rural Participant Four informed us that he had learned about different frauds from his friends who had been victims. He recounted how three of his friends had received an urgent message from a women who claimed that she had run out of credits on her phone and desperately needed to increase her balance. She promised that she would return the credits. All his friends who responded to these pleas and sent money never received any money in return. When they called the number, it played a recorded message in a woman's voice that repeated the same plea as had been previously texted.

Similarly, Rural Participant Five did not experience fraud himself but instead narrated an account of another person he knew who had been a victim of fraud. This acquaintance

was told that he had won 400,000 rupees. The fraudster asked for Rs 2,500 in the form of scratch card numbers for balance reload in order to initiate the payment. The victim went to Participant Five because the he worked at a general store which sells scratch cards. They only had 12 cards worth Rs 100 but the participant went with him to nearby village to buy the additional 13 cards. The victim sent the card numbers to the fraudster and they responded by giving him an address in a major city and the name of the person he should meet. He went to that address and found nobody with the given name there.

6.8.3 Summary of Interview Findings

Our interviews revealed that urban dwellers have a more suspicious attitude toward fraud messages and are familiar with fraud schemes. On the other hand, people from rural areas are more vulnerable due to lack of awareness about fraud schemes. They generally do not receive spam messages but regularly experience fraud SMS. The fraudster typically collects from their victims in the form of scratch card numbers for prepaid airtime. The most common scheme appears to be the BISP fraud.

6.9 Discussion

6.9.1 The Fraud Ecosystem

Unexpectedly, SMS fraud in Pakistan was overwhelmingly represented by lottery fraud schemes relevant to local programs and products. In these schemes, fraudsters aim to collect money in the form of airtime credits, which are more universal than mobile money wallets. Another fraud that was prominent in our dataset was the "damsel in distress" scheme, in which a woman pleads for help in the form of easyload credits. This scheme uses a simple narrative and requires no sophistication to convince the victim. The third most frequent fraud scheme, of which we obtained a few examples, is credential stealing where the victim is told their services are disabled and that they must call to reactivate them. This scheme is targeted at more tech-savvy bank users who would enable SMS notification, use ATM cards

for their account, and potentially have an online login. While I did not see this third type of fraud frequently, it alludes to the potential for the expansion of SMS fraud in Pakistan.

My analysis suggests that mobile users in urban areas are well aware of fraudulent schemes and are skeptical of schemes. The SMS app users were able to label SMS messages with over 70% accuracy in each category, while the verbal requests to low-income urban dwellers to share their fraudulent messages were unsuccessful. People's suspicions aid them in staying vigilant against fraud attempts, but a fraudster can overcome this by using more sophisticated and developed narratives. I did see those who are semi-literate and live in rural areas are most vulnerable to the current attacks. All of our rural interview participants had been the victim of fraud or knew someone who had been the victim of fraud.

Fraudsters use simple, socially engineered messages to defraud and convince their victims to call them. The majority of SMS fraud becomes a call-based fraud which contributes to the regular fraud ecosystem. They do not appear to use automation, and they appear to hand type the messages because the messages are visually similar with slight variations and errors. The fraudsters have access to a surprisingly large cache of phone numbers, some of which are employed to send the fraud SMS and others to list as call back numbers for their schemes. I found that fraudsters take advantage of a loophole in Pakistan's Biometric Verification System (BVS) to obtain verified SIMs. These SIMs are linked to real people, mostly in rural villages, who have no idea that their information is being misused. Most of the misuse of the BVS system happens at franchises and small outlets, and because of this we would advise telcos to enforce stricter practices at these centers.

We had several discussions with stakeholders, including officials from telecommunication organizations, Pakistan Telecommunication Authority (PTA), and other security experts. Our conversations revealed that telcos have no interest in processing and filtering messages of additional equipment costs extra, and they did not see how this would lead to a return on investment. Many countries have a regulatory system that targets spam traffic, and have deployed spam filters in their GSM networks. PTA does not enforce any such regulation, stating that this would be a form of censorship. Moreover, the PTA officials mentioned that

more fraud happens when a fraudster directly calls their victims rather than by sending an SMS and waiting for the victim to call back. They explained that fraud over phone calls can reach upwards to millions of rupee. They also claim that educated people are vulnerable to responding to these frauds. Fraudsters operate in gangs, and once a victim responds to an attack, multiple attacks will follow.

6.9.2 Mitigation Strategies

Education is the most effective strategy to mitigate fraud attacks. From our data, I observed that the PTA educated the public about BISP fraud by sending universal notices. Several banks also sent messages warning their users to never give out their passwords. All these messages arrived in threads where banks or the PTA were spamming users with other non-educational messages, so the educational part may have been overlooked. Moreover, these are very specific examples of fraud awareness, and I would recommend a more broad approach that would address existing and potential schemes. A regular schedule for informing the public about a current list of fraud schemes could also alleviate people being defrauded by new schemes.

Disabling fraudulent phone numbers quickly is key to stopping all call initiated fraud. This strategy would especially help the most vulnerable. To collect active numbers, a government authority could establish a service where people could forward a potential fraud message. It is very easy to identify a message of a known fraud scheme with the heuristic I developed. Phone numbers identified in this way could be automatically reported while the other unknown fraud-type messages could be evaluated case by case. Our SMS forwarding service that we advertised widely was the major source for our fraud corpus. During interviews, our participants requested a way to report fraud messages but without having to forward the messages. This indicates that a mitigation strategy could rely on reaching out to public for supportive data.

Preventing fraud via a fraud detection app at the user end is a more robust and distributed approach. The smartphone app with my defined heuristic could be effective in warning users

as fraud schemes evolve. The heuristic is a simple algorithm that does not rely on heavy machine learning and would work locally on a user's phone instead of requiring heavy phone processing or cloud computation. The challenge from our experience is that either users do not trust lesser known apps or they lack the expertise to download and install a new app. At the same time, users who lack trust for a new app or use apps like Truecaller, a caller identification app that warns them about malicious phone numbers, typically recognize fraud SMS. I recommend explaining to users how to install a fraud detection app as well as the app's purpose.

6.9.3 Fraud in Marginalized and Rural Communities

People who are relatively new users of mobile services are most vulnerable to SMS based fraud. They lack awareness about how messages could be spammed to random numbers and how fraudsters use this to their advantage to spread their fraud schemes to a large audience. Hence, these users trust these messages. As demonstrated in our interviews, they will engage in the fraud until someone in their community alerts them. The number of vulnerable newcomers is expanding as more people are coming online in this digital and mobile age. The majority of these newcomers are part of marginalized and rural communities. If members of their community are experienced users, then they are more likely to learn quickly about how to avoid SMS scams.

The marginalized community that is at risk includes urban dwellers who might be educated but less tech-savvy. They can recognize a simple lottery scheme but are not aware or suspicious enough to pick up sophisticated attacks that are trying to steal banking details or credentials. For instance, they might not recognize that instead of "https://www.hbl.com", the given URL is "www.smshbl.com". This could lead to them signing into a fake website and hence giving away their bank credentials. Moreover, they might not know how to install a caller identification app like Truecaller, which is effective at identifying malicious numbers.

6.10 Conclusion

This chapter examines the SMS fraud ecosystem in Pakistan using data collected through a smartphone SMS app, a SMS forwarding service, and interviews. It identified ten fraud schemes that represent the following three categories of fraud: lottery, "damsel in distress", and stealing credentials. It concludes that people in rural regions are the most vulnerable to these frauds. The majority of these frauds require the victim to call the fraudster, and most ask for money via airtime credits instead of through DFS. The chapter also presents a heuristic that can detect fraud with a very high accuracy and is generic enough to be adapted to new fraud schemes as they evolve. Finally, it presents a mitigation strategy that is drawn from the data analysis, experiences of defrauded individuals, and discussions with various stakeholders.

This chapter presents struggles with data in financial domain that is different from health related datasets. The project was set up to collect well organized data with help of the crowd but it encountered issues due to data in open text format, various spellings because of transliteration, and lack of motivation to send labelled data. It establishes that crowd could be used to get some labeled data but it will still require a fair amount of work to remove noise and determine accurate labels. Secondly, the data indicated that these communities regularly write transliterated local language to communicate with each other so variety of spelling due to transliteration is a common issue that we will encounter for open text form fields or general open format data collection.

Chapter 7

THE SCALABILITY OF SMS REPORTING SYSTEMS: INTEGRATING WITH NATIONAL HEALTH INFORMATION SYSTEMS

7.1 Summary

This chapter digs deeper into the challenges to gathering cleaner data during the collection stage when data is being collected over SMS at a large scale. These barriers were identified while studying two national SMS data collection systems. Based on this case study, I argue that the health worker can be motivated to engage with the system and send accurate data if the system responds with feedback. Moreover, I propose that the usability of SMS format should consider both ease of learning the format for new users as well as quickly format the reporting SMS for expert users. Nonetheless, better training could help overcome the usability issues and result in faster adoption with cleaner data reported. I found that errors in SMS reporting are linked to variable quality of training across regions that is due to multiple teams involved who are responsible for training their respective regions to scale the training process.

7.2 Introduction

This paper presents two case studies of implementing national scale health data reporting systems in collaboration with ministries of health. The systems have a common architecture and technology - they both support submission of data from mobile phones through encoded SMS messages. Our interest in pursuing this work is to gain a greater understanding of sustainability, scalability and challenges for ICTD systems. Working with governments on ICT systems creates opportunities for broad deployment, and also has potential for sustained

use if the systems are fully adopted. However, there are a multitude of organizational, political, and technical challenges in this information system domain.

The projects we describe are a disease surveillance system being implemented across Pakistans Punjab province, and an immunization information system being implemented in Laos. Both of these systems are intended to be deployed at large scale, with the Punjab system to be deployed to all 1519 primary health facilities in Punjab, and the Laos system to be deployed to all 1132 health facilities and vaccine stores in Laos. The projects were deployed in 2013 and 2014 - the Punjab system was initiated in January 2014 and has achieved reporting from 377 (25%) of the health facilities in the province. The Laos system was started as a pilot in 35 facilities in October 2013, with a gradual expansion planned to reach the entire country by the end of 2016. We make no assessment at this stage as to whether these projects are successes or failures as both projects are works in progress. Both projects have faced challenges, and adjustments are being made reflecting learnings from deployment. Lessons coming out of this work point to the importance of aligning the reporting tasks with the needs of the health worker, ensuring the system is robust to accommodate local adaptations, and the challenge of managing stakeholder requirements.

A second theme of the chapter is SMS reporting systems. Many ICTD researchers and practitioners have used SMS technology for data collection and information dissemination. The attraction of SMS reporting is that it can rely on the lowest common denominator mobile technology, and fits with a model of reporting from personally owned mobile phones. Both of these projects have health workers use their own, personal phones, as opposed to provisioning workers with either computer access or more sophisticated devices. Core challenges include health worker perceptions of the cost of reporting and maintaining up to date knowledge of structured formats for reporting. Since the reporting environment is dynamic with personnel turnover and changing circumstances, it is deemed necessary to provide mechanisms for human intervention into the reporting system such as a management interface.

The background section presents some information on the context of public health systems

in developing countries, followed by related works on Health Information Systems and on SMS reporting systems. Section 4 of this chapter describes our two case studies, first the surveillance system in Punjab, Pakistan, and then the immunization information system in Laos. Results from the case studies are then synthesized in the discussion section of the paper.

7.3 Background

This work is in the context of national scale Health Information Systems (HIS). Across developing countries, efforts are underway to improve information systems with the belief that this will strengthen the health system. The reasonable expectation is that better health system data will lead to improved decision making and management. There are many commonalities for health systems in low and medium income countries (including the two we are focusing on). We provide a summary as these have direct implications to the systems we are deploying.

- The health system is run by the Ministry of Health (MOH), which is politically well connected. There are multiple departments inside the MOH which have significant autonomy and are isolated from each other.

- The public health system operates a hierarchy, with managers at each level.

- The MOH works with a constrained budget, and is often dependent on donor funds.

- Global organizations such as World Health Organization (WHO), United Nations Children's Fund (UNICEF) and GAVI Vaccine Alliance have significant influence.

- There is decent infrastructure at the top levels of the health system, although technical capacity in IT may be limited.

- The peripheral health facilities can have problems with infrastructure such as lack of electrical power. Computer and internet are generally not available, but health workers have their own personal mobile phones.

- There is a well established system of paper reporting of health data, although it may take months for data to traverse from the health centers to the national level. Information flow is strictly upwards.

The focus of this paper is the deployment of SMS reporting systems as parts of a national scale HIS. This entails both implementation of facilities that support the upper level of the health system, as well as the last mile component. These two aspects have different characteristics: at the national level, the main challenges are organizational while at the periphery, it is necessary to understand the users and meet their needs. We discuss related work for large scale information systems first, and the later address the work that informs development of SMS systems and reporting.

Much of the ICTD literature focusing on national scale IT systems and scalability is framed with respect to IT failure [48, 63, 64]. This legitimately reflects the challenges of large scale IT systems in the developing world. The fundamental critique by Heeks is the tendency to take system designs from the developed world, and deploy them without appropriate understanding of developing world context, as he refers to as the design-reality gap [64] or the design-actuality gap [63]. This gap is often manifested between hard rational design and soft political reality. The failure literature also makes a positive contribution to the design and evaluation of large scale systems by establishing frameworks to analyze risks of failure. Challenges of deploying health information systems have been described in the body of work emanating from the HISP network [16, 19, 141]. The richness of this body of work comes from the process of deploying the District Health Information System (DHIS) in many countries in conjunction with an action research program overseen by University of Oslo. Key arguments from this include the political aspects of adoption of health information systems, as well as the complexity of managing competing health domains inside a single system, which

is often expressed in disagreements over health indicators [142]. HISP have also published influential work on health systems architectures based on development experience from DHIS and DHIS2 [141, 140], which complements more general enterprise architecture work for health information systems [101]. Several authors, including Ammenwerth [3] and Panir[113] have undertaken critical surveys of health information systems. One of the underappreciated aspects of the systems implementation in this domain is the complexity of stakeholders, Hayes points out that this is growing more complicated with the increasing role of consultants in developing country information systems projects [62]. Another challenge that grows out of development funded projects is the role that incentive payments have in influencing behavior. As Sanner points out, this is significant issue for sustainability of projects, as this impacts the transition from an initial donor deployment to a system that can be managed by the government [143].

An important theme in the health information systems literature particularly relevant to this work is the last mile problem - or how HIS integrate with peripheral health units. There are multiple approaches to the last mile problem spanning the technology spectrum. These range from interactive voice systems (IVR) [86] and SMS based reports [132] to direct integration with mobile phones [17] or data reporting with smart phones [61]. No single channel is an ideal solution for the last mile problem and the advantages and disadvantages of each must be analyzed based on the particular context of an individual deployment. Patnaik et al. explored how mobile electronic forms, SMS, and voice compared when used by health workers in India [118]. They identified three primary categories which affect the choice of data channel: how easily operations integrate into existing practices and technology, the effectiveness of data reporting, and the ongoing cost of operation. For both projects in this paper an important fourth category was added: ability to reach national scale.

Based on the constraints listed above for HIS in developing countries we find that SMS is the most suitable data channel for reaching national scale for several reasons:

- With no special applications needed, universal access on even the most basic phones

reduces start up costs and allows for immediate wide-scale deployment [159].

- SMS systems can process higher volumes of incoming messages per phone line than voice.

- Lower ongoing costs to send encoded information over SMS rather than voice.

SMS has been shown to be a very effective channel when one-way communication is required [47], however, Patnaik et al. showed that a major complication with SMS as a reporting channel is an error rate of 4.5% for SMS vs. an error rate of 0.45% for voice. Similar error rates have been observed by large scale SMS reporting projects. The SMS for Life project, which recorded stock levels of anti-malarial medications at 129 health facilities in Tanzania, reported an average error rate of 7.5% [9]. A major factor in the accuracy levels of SMS formats is training and without training, SMS formatting error rates as high as 35% have been reported [38]. However, other studies have shown that when health workers use SMS for data reporting error rates have a tendency to spike when new services are introduced but quickly stabilize[11] and improve with feedback[7].

The two countries we are working with are Pakistan and Laos. The countries are both lower-middle income countries by the World Bank classification. In 2013 Laos ranked 133rd in per capita GNI, while Pakistan ranked 134th. For the human development indicator, which captures a broader range of indicators than just income, in 2014 Laos was ranked 139 and Pakistan was ranked 146. Thus, the countries are fairly well matched in terms of development. One dimension where they differ radically is population, Pakistan has a population of 188 million, while Laos has a population of 6.7 million. The province of Punjab in Pakistan alone has a population of 101 million (and if it were a separate country, it would be the 10th largest in the world). To simplify our language in this paper, we are going to refer to both the deployment in Laos, and in Punjab as national scale deployments. From a geographic perspective Punjab and Laos are both similar in size at a little over two hundred thousand km2. The large population of Punjab is split almost evenly between rural

and urban populations and so rolling out a province wide reporting system requires bridging the gap between medium and low resourced regions. Both of the projects considered in this paper are to build reporting systems to collect information from health facilities, so the key measure of size is the number of health facilities. For Punjab, the reporting system will eventually cover 2,754 health units, while in Laos, the project will eventually cover 1,132 health facilities, so the disparity in health facilities is not as great as in population.

7.4 Case Studies

In this paper we introduce two SMS based health data reporting projects that aim to scale to the national or provincial level and examine ICTD issues particular to scaling projects to this level. Both projects are highly integrated with the health system and local governments and involve many stakeholders from international organizations like WHO and UNICEF to local statistical officers. In both cases SMS was chosen as the data delivery channel because every mobile phone can send and receive SMS, even over weak rural networks.

The first case study we introduce is the Punjab Disease Surveillance System (PDSS) which is a daily disease reporting system launched by the Punjab Information Technology Board (PITB) in Punjab, Pakistan in January 2014. PDSS is a replacement for WHO sponsored disease reporting system that collected weekly statistics[1]. The second case study we examine is a partnership between UNICEF and the Laos Ministry of Health (MOH) aimed at strengthening the vaccine cold chain system. The Laos SMS Immunization Manager (Laos SIM) has been implemented by the National Immunization Program (NIP), a division of the Laos MOH, which began pilot tests in November of 2013.

7.4.1 Methodology

In both the PDSS and Laos SIM projects we worked closely with the implementing partners during the design of the system and have conducted additional follow up assessments

[1]http://www.emro.who.int/pak/programmes/communicable-disease-a-surveillance-response.html

throughout the project. In Pakistan, we assisted PITB with formative work during the design phase. We continued to have follow up conversations about the progress of the project periodically. One of the authors spent two weeks in September 2014 to conduct qualitative fieldwork at 9 rural health facilities and have detailed technical conversations with implementing personnel at PITB and WHO. Health facilities were divided into three categories based on their reporting rates in August 2014: high compliance (18 - 31 days reported), low compliance (5-17 days reported) and no compliance (0 - 4 days reported). 3 health facilities were picked from each these categories and interviews were conducted for 45 minutes to an hour per facility. In addition we have SMS logs for the project from May 2014 through September 2014. In Laos we have worked closely with UNICEF and NIP in establishing the requirements and constraints and iterating on the system design. Multiple field visits have taken place for training at the first set of pilot sites and coordination with NIP officials. Observational data was collected, along with qualitative interviews on the use and issues around SMS reporting.

7.4.2 Punjab Disease Surveillance System

Motivation

Tracking the spread of diseases at early stages is very critical for providing timely and effective responses[83]. WHO has established 19 priority diseases which government health systems should monitor on a daily basis since they are highly communicable and can spread rapidly if not contained. While systems exist for disease tracking in the developed world these cannot easily be transferred over to developing regions where reduced connectivity and lack of access to technology become hurdles. In case of Punjab, which has both large densely populated urban centers and small rural areas, creating a system that works across the whole province involves unique design constraints within the field of ICT4D.

Background and Design Rationale

The Punjab Disease Surveillance System is an information system for tracking patient visits at government hospitals and health centers. The objective of this system is to help decision makers analyze and visualize data on reported cases in real time and at fine grain levels. The data is reported daily, directly from each health facility. This was deployed by Punjab Information Technology board (PITB), which is a government department working on technology interventions for other public departments. The project began in 2012 as a web based information portal set up at every hospital in collaboration with the Department of Health. PITB hired dedicated data entry personnel at these hospitals who entered data on laptops with a USB modem. Since these operators are separate from hospital staff there is no added reporting burden for hospital staff. However, it is impractical to hire separate data operators for basic health units (BHU) which reside far from urban centers and have many few cases to report. BHUs are often on the front lines of new outbreaks and getting timely statistics is important to prevent the spread to larger cities.

However, a major goal of PDSS was to reach the basic health unit (BHU) and rural health center (RHC) in addition to well connected hospitals. At this point we began collaborating with PITB in the design of a reporting system that would (1) be immediately deployable in rural Punjab, (2) scale to the provincial level and (3) require minimal technology costs or upgrades. A survey of digital reporting tools indicated that options requiring mobile data access would require devices that health staff did not currently have and a voice based systems could not economically scale to thousands of reports per day. PITB determined that the only feasible data channel for BHUs and RHCs would be SMS. In this section we will discuss the design, implementation, and deployment of the SMS based PDSS.

Implementation

Although SMS is a ubiquitous data channel, sending structured data via SMS requires a well designed message syntax. As Patnaik et al. discovered SMS reporting has a significantly

higher error rate over voice reporting. Selecting manually entered SMSs as a data channel is therefore an optimization for scale and cost. In order to overcome this, the design of the reporting syntax must take into account the end users workflow and technical literacy.

In designing the PDSS SMS syntax we looked for existing and proven SMS reporting examples. One such example is the SMS reporting wheel from InSTEDD[2]. This wheel encodes the date, disease and number of cases in a paper based job aid (Fig. 7.1). The data types are represented by three concentric cardboard circles. The circles rotate individually and have a two digit code plus third checksum digit. Each data entry on the wheel has a corresponding three digit code and the wheels can be aligned to form a nine digit report. There were three major motivations for using the SMS wheel: (1) numbers are easier to enter on T9 keypad mobile phones, (2) the wheel simplifies translation from data entities to encoded numbers and (3) the checksum provides error detection for mistyped numbers.

The use of the wheel as a job aid has had both positive and negative consequences for the project which will be discussed in section 5. The original design envisioned health works sending a single SMS per day with a comma separated list of each nine digit code for diseases seen that day. Typically each BHU has just three staff members, a dispenser, nurse and doctor, while an RHC has one or two more staff. It is the responsibility of the health units dispenser to send the daily reports via SMS (Fig. 7.2).

Training was conducted by inviting district Health Officers along with 3 trainers to the central level. PITB along with MOH lead this training with examples and exercises. These trainers then trained all dispensers at BHUs and RHCs, in their respective districts. Reporting wheels were given out to all dispensers at these trainings.

Once the reports have been decoded, data is stored in a central server, which is accessible through a web dashboard and integrates with the provincial hospital reporting system. This dashboard is accessed by PITB, Health department and WHO officials. Currently it displays specific graphs to show the reporting performances of each district. PITB and WHO work

[2]http://instedd.org/technologies/reporting-wheel/

Figure 7.1: Wheel (Close up of the SMS report job aid for DSS)

together to export the data in excel format for analysis and produce weekly epidemiology situation report bulletin distributed to DOH officials.

Statistics

The system began receiving disease case reports in January 2014. By May 2014 17,730 correctly parsed messages had been received. We have done a detailed analysis of all messages received by the PITB shortcode from May 2014 through September 2014. This four month period represents the state of the system after five months of operation which we argue represents a long term adoption of a new reporting task. In this period 22085 messages received were intended for the PDSS system. Of those messages received 77% (16913 messages) were correctly identified and parsed by the system. There are 2943 BHUs and RHCs in 36 district of Punjab. PDSS is operational in 13 districts, which has 1519 primary health facilities. Currently, the project has achieved reporting from 377 of these facilities. Second roll out will happen in 3 to 4 months with better quality job aids, trainings focused on known problems,

cost free SMS line and automated feedback responses. The aim of second phase is to achieve 80% to 90% data reporting. The final phase will then target all 36 districts in Punjab.

7.4.3 Laos SMS Immunization Manager

This project is part of Laos CCIS that is the information system discussed in detail in chapter 4. The focus of this chapter is to dig deeper into the SMS data collection component of the system. I restate parts of the motivation, design rational and implementation below for context.

Motivation

This project is a collaboration with UNICEF and the Lao National Immunization Program (NIP). The motivation for the project was to build a system to collect data on the status of the vaccine cold chain to ensure that vaccine refrigerators remained in good working condition. It is important that vaccines are stored in a fixed temperature range (usually between 2°C and 8°C) so that the vaccine keep their potency. In order to track cold chain issues UNICEF has distributed temperature logging devices. These devices have 30 day temperature recorders (30DTR) that track the refrigerator temperature for thirty days, and record if the temperature has been too high, or too low for an extended period of time (Fig. 7.3). A low alarm occurs if the refrigerator temperature is lower than 2°C for more than one hour and a high alarm occurs if the refrigerator temperature is higher than 8°C for more than eight hours. Both of these types of alarm indicate that the refrigerator should be serviced in some manner.

UNICEF has distributed 30DTR refrigerator tags in many countries. However, the status of these devices is rarely reported to higher levels. As a result, the issues with the cold chain remain unfixed. For example a district technician can easily repair a refrigerator with multiple low alarms each month, but only if they are notified about the malfunction equipment. The primary aim of this project is to "close the loop" on the 30DTR, by building a mechanism for the refrigerator status to be reported to the higher levels of the health system.

Figure 7.2: A BHU dispenser explaining how he uses the SMS reporting wheel to create a SMS message

Figure 7.3: A 30 day temperature recorder (30DTR) in a Lao refrigerator showing two high alarms in the last two days

The hope is that reporting this information will lead to action in repairing and servicing equipment, and the data will also support better management decisions in acquisition and allocation of equipment.

The basic system that was envisioned for Laos would support data reporting from all health centers (966 sites) and all vaccine storage locations (166 sites including both district and provincial vaccine stores) on a monthly basis. This information would go into a backend database with a full inventory of the country's cold chain equipment. An important component of the system would be the ability to forward notifications to people who could act on them and to send summaries to managers at different levels.

Background and Design Rationale

UNICEF, the Lao MOH, and the Lao NIP began discussions to collaborate on a national fridge-tag reporting system in early 2013. Laos, with a population of 7 million and stable political environment was seen as a feasible location for an initial country-wide 30DTR reporting deployment, which could then be adapted for deployment in other countries. Prior to design of the system, multiple field visits took place, with interviews of health workers at all levels of the system. The initial field studies showed that almost all health workers had access to basic, SMS capable phones, but very few health workers had access to data enabled smart phones. Internet access was rarely available at the health facility level. Based on these results it was determined that SMS would be the best channel for data collection at national scale. The decision to rely on personally owned mobile phones was motivated by sustainability concerns, including keeping the cost and complexity of the project as small as possible. However, these field visits also revealed that health workers had limited experience with SMS and so the reporting format would have to be both simple and robust.

UNICEF and NIP decided on an incremental deployment of the system with more reporting districts added every three months. Meetings with all stakeholders were held to develop a simple reporting format for 30DTR month reporting. During these meetings NIP expressed a desire to also report stock levels of five major vaccines. This would allow them

to have a more complete picture of the national cold chain at minimal added complexity and reporting burden.

Trainings were conducted at all health facilities in three districts. NIP and UNICEF officials traveled to each facility demonstrating the message syntax. During this visits, the refrigerators were recorded for the national database, and given letter codes to be used in reporting. Health centers would generally have a single refrigerator, while the district and provincial vaccine stores would have between two and ten refrigerators. In November 2013 the pilot group of health facilities began sending SMS reports to a phone at NIP. This first stage was designed to test the feasibility of using SMS and with just 20 health facilities reporting messages were manually transcribed from the phone and entered into the database.

An initial syntax for reporting was designed which also included keywords for functionality that could be added later. From the beginning of the project it was recognized that the SMS messages could both send in monthly reports, giving refrigerator status, as well as event reports such as stock outs, or refrigerator failure. There were competing views on the syntax of the SMS system, one desire was to make it as simple as possible, especially for reporting from a health center, while the other desire was to make the syntax regular so that it could be extended easily and would easy to process. During the pilot stage the final message syntax was agreed upon for two different use cases. Both syntaxes ignore whitespace and other non-alphanumeric characters allowing for as much variability in input messages as possible. The first syntax is based on two letter operation codes and designed to be extendable. Each operation code is followed by a message to parse and multiple operation codes can be included in the same message. The second syntax is a shortcut syntax and designed for simple messages from rural health units reporting monthly alarm and stock data.

Implementation

In February of 2014 we released a prototype for an automated SMS reporting tool that would be rolled out in the next stage of the project. This system was developed on top

of RapidSMS[3] but used a custom message parser since the default parser in RapidSMS cannot handle multiple operation codes in a single message. To get messages from the Lao telecommunication network into RapidSMS we used EnvayaSMS[4] which is an Android SMS to HTTP gateway. Using EnvayaSMS allowed for quick deployment of the prototype and could easily be switched to a third party SMS to HTTP gateway as the project scaled. Android SMS to HTTP gateways like EnvayaSMS can scale to 500 outgoing messages per hour, which would probably be sufficient for a nationwide deployment of the Laos SIM project. However, NIP is currently in the process of setting up access through a Telco hosted SMS gateway to allow for higher throughput, as well as a more robust gateway that supports additional payment options.

With this RapidSMS based prototype automatic replies are sent after every submission confirming a successful message or reporting any errors found. This two-way interaction with the system not only provides confirmation that messages are received but also gives rapid feedback. This feedback helped to standardize the messages as some users self corrected messages based on the automated response. For example one user sent "SL0 P550D4250" which resulted in a parse error because the 0 should be in front of the sl (stock level) tag. Two minutes later the same user sent in the correct message "0SL P550D4250" based on the automatic error message.

Based on feedback from this initial prototype we developed an SMS moderator interface which was rolled out in the next phase of the deployment in August 2014. This is a custom module built on top of RapidSMS that allows NIP to directly manage incoming messages (Fig 7.4). The Laos SIM moderator dashboard looks like an email inbox of incoming SMS messages. It categorizes messages into report submissions and non-reports and flags reports that failed to parse correctly. The moderator can manually fix messages that have obvious and unambiguous errors. The dashboard also enables moderators to link phone numbers to facilities, change a user's preferred language, send messages to a contact and view a history

[3]www.rapidsms.org

[4]www.sms.envaya.org

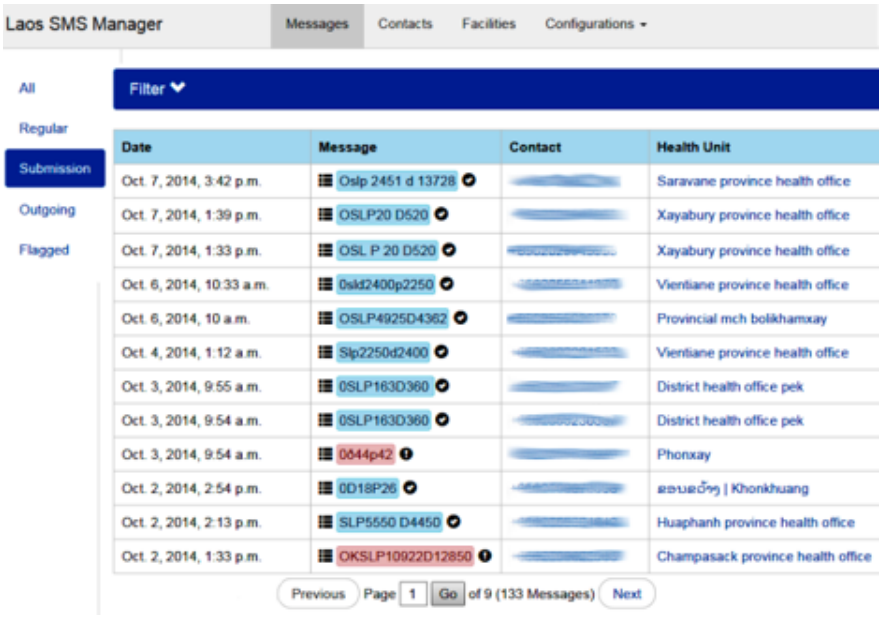

Figure 7.4: The Lao SMS Immunization Manager (SIM) showing a list of incoming SMS reports of October 2014

of messages sent by facility and phone number. The moderation interface was requested by the project manager base in Laos, as she faced significant challenges in managing incorrect message submissions and tracking the different phone numbers used by health workers.

In total 443 messages have been received for the Laos SIM deployment between November 2013 and October of 2014. From November 2013 through March 2014 twenty facilities were reporting and from March through October 2014 thirty-five facilities were reporting. All messages were received in the first five days of the month. We conducted a detailed analysis of all messages received from August through October 2014. During this time 149 messages were received 144 of which were report messages. Only 16 of the report messages had parsing errors and all but one of these were trivially fixed by the NIP moderator. As the Laos SMS cold chain SMS reporting system begins to scale to all 18 provinces in Laos the moderation feature will move from national NIP offices to provincial health facilities.

7.5 Barriers to Responding

In this section we identify three barrier that exists in system adoption from the health workers side. SMS cost, usability issues - either with the job aids or SMS syntax, and a lack of real time feedback. Since all users are employees of the health system systems there is motivation for overcoming these barriers, however, limitations in what the health works are willing or capable of doing still exist. We also identify two major methods organizations can use to overcome these barriers to reporting: practice based training and encouraging local adaptations.

7.5.1 SMS Cost

The cost of SMS plays an important role in adoption. Both projects have been working to establish toll-free shortcodes, however, this requires negotiations with each telecommunication industry - an ongoing process. Until toll-free numbers are established health workers must bear the cost. During field visits we found that rural health works will front the cost if its their job, but only to a limit.

In Laos, most health workers did not have a problem with paying 150 kip ($0.02) for one SMS per month. One did complain saying *"If its the duty then I have to do it but I find it very hard because I have to use my personal money and sometimes I dont have the balance"*. On the other hand in Pakistan, most health workers complained about the cost of SMS, 1.20 Rupees ($0.01), as they are sending much more frequently. One health worker said *"I used to send the data on daily basis. Then my friends told me that you spend so much, daily Rs 2.52. They send data for 10 days in one message. Since then I am sending 10 days data in one message. That cost me Rs 4.20. We were told that we will get refund but have not received any so I am trying to save some money as it cost quite a bit"*. However some did not find the cost prohibitive saying *"I do not care about the cost of SMS since its part of my job and I want to do my job properly"*.

Obtaining toll-free shortcodes is highly country specific since it relies on existing infras-

tructure and third party operators. Even though the telecommunication sector is established in both Pakistan and Laos there is a no support for free incoming SMS messages. Most mobile carriers provide shortcodes for premium paid services with mobile terminated billing. This means it is trivial to get a shortcode charging more than standard but difficult to get ones charging less - even if the shortcode operator has ability to cover the cost of incoming SMS messages.

In Laos, the main telcom suggested the project distribute their own SIM cards to all health workers since they cannot guarantee message delivery between service providers but can provide free in group messaging to a subset of their own subscribers. Such a distribution would add an additional burden to scaling the system, and would not be able to reach all health facilities, so for now the Laos SIM project is continuing to use the existing Android SMS gateway.

7.5.2 Usability Issues

Data collection through a new interface always comes with usability issues that must be overcome and SMS is no exception. The intended users had varying degrees of familiarity with SMS, with the Punjab health works being much more familiar with the format than Lao health workers. However, even for those health workers familiar with SMS, only a few had used the medium for submitting reports and understanding the specific SMS format presented a challenge.

In Punjab, the reporting wheel was used as a job aid to assist in message generation. But using the reporting wheel requires familiarity and precision and can be time consuming as one says *"It takes a lot of time to send message. I am not an expert but it takes approximately 10 to 15 minutes for one day report"*.

In contrast, the SMS syntax used in Laos encodes more data with fewer characters. This requires a more detailed understanding of the format and has the potential to generate more errors. Interviews with health workers indicated that they all felt they were able to master the format, and observations showed that they could enter the message in under a minute.

Generally, training is considered as a good solution to handle this barrier but delivering that training properly to thousands of health workers involved is difficult.

7.5.3 Two-Way SMS and Automated Feedback

Previous SMS mHealth initiatives have identified the importance of two way communication in the system [11, 38, 47] and a lack of these feedback in Punjab has been a major barrier to continued reporting. Many health works expressed concerns about the message correctness and delivery and were frustrated when not getting replies. One health worker in Punjab said *"It has been there for 9 months but we did not receive any feedback. It seems like we are wasting our time and higher officials are not taking any interest in this. Obviously if someone is send messages for 30 days, there has to be a feedback. It has been 9 months and we should know if there is any reaction or action against it."*. Automated feedback is fundamental to the continued operation of an SMS service and we have recommend that the Punjab DSS initiate feedback.

Two-way interaction is also useful for reminding health workers to send in reports [9, 47]. This is a common case in Laos reporting happens once a month. Currently, a mediator from UNICEF reminds them if reports are late but we have recommended automated reminders be adopted. Reminders for a monthly reporting cycle are fairly straightforward to implement, however, in the case of daily reporting the risk of spamming phones is high. Exploring how reminders can added to the Punjab DSS is an active area of research. One possibility is to move towards weekly feedback, as one health workers suggested: *"You should convey the report weekly through a message that tells how many days we reported and which days we did not report"*. Another health worker was very interested in getting a performance report relative to others saying *"We have monthly meeting at District Office where everybody gets their performance report for that month. We should be told in front of everybody what is our reporting performance for this system"*.

7.5.4 Training

In both projects implementers felt that the key to overcoming usability issues was training. In Pakistan the training model closely followed the hierarchy within the department of health. District level staff were trained at a central location on how to generate messages with the reporting wheel. Each district was then responsible for organizing a training for the dispensers at RHCs and BHUs in the district. District follow ups revealed that there was no standard method for district trainings. Some districts conducted training with hands-on practicing while others simply showed the reporting wheel and a brief explanation. The diversity in district trainings can be observed in the received messages some of which to not follow the intended conventions. A common theme is to add extra context to a message that is already encoded in the nine digit numbers like the following message:

Disease surveillance report. *BHU 66/12L 27-8-14 426251512 28-8-14 550086567 29-8-14 684086189*

In Laos, each health center in pilot stage received a training visit where NIP staff demonstrated the SMS syntax through paper based practice. These site visits often revealed inadequate understand of the 30DTR devices, which staff at UNICEFF extended into training on general vaccine management. This highlights the fact that a well managed training program can serve multiple purposes. During the next phase, the geographic spread of provincial stores made site visits impractical and so trainings occurred over the phone. Error rates at provincial stores were initially high, but have reduced since the addition of automatic feedback. Health workers in Laos told us that the most useful aspect of trainings were hands on practice. As one health worker said *"I understood [the format] after the training obviously needed practice but after practice I am very comfortable with it"*.

7.5.5 Local Adaptations

Training is one strategy to overcome barriers to reporting. A second mechanism facilitating reporting is to promote local adaptations that naturally emerge. The health workers

submitting the SMS reports are the experts on how usability issues affect them and what resources exist to overcome them. The implementing partners should encourage and spread such innovations.

In Punjab, issues with the job aid made it difficult to align all three wheels properly. Health workers at several facilities created three column chart that they found easier to use. PITB is going to distribute those chart with better quality reporting wheels in the next phase. Health workers, in both Laos and Punjab, are familiar with paper so will write the SMS reports reports down before submitting. This makes it easier to double check the syntax with colleagues and helps with fixing any errors they system reports. The flyer explaining the SMS reporting system in Laos now has a table in the bottom half for logging these messages.

7.6 Discussion: Lessons at Scale

The deployment of both the Punjab DSS and Laos SIM offers a unique perspective SMS reporting systems scalability. Understanding the issues created by ICTD projects at this scale is an important aspect of sustainability in development. We believe that general themes emerging from both projects offer important lessons and contributes to the ICTD community by demonstrating the impact of universal mobile communication systems for reporting.

In the subsections that follow we identify six aspects of scale that have influenced the design, deployment and success of these systems. Where appropriate we identify potential obstacles and effective methods of overcoming them. In addition we identify places where each project could benefit by adapting procedures and practices from the other. Although the design of any reporting system is highly dependent on the context and objectives of individual projects by synthesizing our results from two different projects we believe that this analysis will be more generally applicable.

7.6.1 Automation: Hybrid SMS Systems

When scaling any reporting system a major goal is to automate as much of the system as possible, however, in analyzing both projects we find that full automation is not feasible.

In fact from the data collection perspective SMS reporting systems almost always require manual entry; since the technology to send automated SMSs a more effective channel, such as mobile data, likely exists.

Manually entered SMS data requires a very specific format to be parsed and errors will inevitably occur. These errors can often be easily and unambiguously fixed by a human. In our analysis of the PDSS messages 23% of messages were not formatted correctly and the majority of these could unambiguously be interpreted by the district statistician. Similarly in Laos 11% of messages used incorrect syntax and all but one of these was corrected by the moderator through the SIM interface.

A second reason incoming messages may not be processed is if valid messages are received from unregister numbers. Both of the systems rely on having phones registered to associate messages with health facilities. There are multiple reasons why messages may be sent from different phones including health workers switching simcards or using someone elses phone. Personnel turnover is another reason messages may come from unregistered phones. In this case someone must determine which health facility to associate the new number with. In Punjab the current method of registering new users is a very manual process and involves IT personnel at PITB; however, in Laos, if the system receives a valid message from an unregistered number it will automatically send a message back asking for facility codes and identifying information. Any response to this message must be approved by the moderator for the number to be registered. In this way the messaging system is designed to reduce the amount of moderation work required.

7.6.2 Quality Assurance and Data Management

It is apparent that a basic level of moderation is required for a successful SMS reporting system. As projects scale across a province or nation the moderation burden dramatically increases. The most effective method to handle moderation at scale is through distributed data ownership. The introduction of a central reporting system is a paradigm shift not only in data collection but also information flow. In both Punjab and Laos data usually flows

through hierarchical levels of the health system. A SMS reporting system bypass the middle of this existing hierarchy as data flows directly from the lowest levels to the central health ministry. Designing a distributed moderation system not only restores traditional data flows but also creates a sense of ownership in the project at all levels. The SMS moderation interface should include the task of follow ups on non-reporting, incorrectly formatted SMS and invalid reported numbers.

PDSS aims to reach reporting from approximately 2943 facilities in 36 districts in Punjab province while Laos SIM aims to achieve reporting from 1132 health facilities from 148 district in 18 provinces. In order to achieve this scale PDSS has begun to involving the Statistical Officers (SO), who is responsible for health statistics at the district level. The role of the SO will be to advocate for proper adoption and assess data accuracy. Program director of the system at PITB points out that involving SOs will play an vital role in system improvement.

"These SO will be enforcing the data and will be reviewing the accuracy of the reports."

In Laos, districts are small, with approximately eight health facilities, so the management responsibility is being distributed to provincial offices where the relevant skilled staff and resources exist for proper quality assurance. On the average a province will be responsible for 64 facilities so SMS management should not be a major burden.

7.6.3 Stakeholders: A Fine Balance

Large scale deployments generally engage several decision making stakeholders. These stakeholders, with different backgrounds, strengthen the deployment by helping in refining the design and supporting the implementation. Once the project has rolled out, the definition of success or the suggestions for design updates, to stir the project in right direction, is very diverse and sometimes conflicting. This is bound to happen as every stakeholder has a different view for the project. In this mess, finding the right balanced definition to categorize project as success or lead to success is hard. Project could be successful for one stakeholder while a complete failure for others.

PDSS has PITB, Health Department and WHO as primary stakeholders. While PITB

defines the project as leading to success, WHO considers this project as unfruitful due to inconsistent data reporting. Health Department takes this project as yet another data collection project that might fail. Since PITB views the issues being faced as resolvable, they consider this project leading to success. One PITB manager says *"It will take time for this system to run smoothly but eventually it will be adopted. The project has energy and will carry on."*

Stakeholders of Laos SIM consist of NIP, MOH and UNICEF. The different organizations have had different priorities for the data system. One early issue that came up was with respect to reporting of vaccine stocks and cold chain status. The project was initiated to focus on reporting refrigerator status and it was only because NIP also wanted vaccine stock reporting that the system was extended. The use of the vaccine stock reporting data has turned out to be problematic, as the reported values depend on the relationship between the reporting data and the monthly delivery date for vaccines. Another case where pressure has come on the project is the possible extension to support the reporting of other commodities.

One of the long term challenges that the Laos SIM system will face is the complexity of the different departments of the MOH. As a large organization some functions need to cross different sub-organizations, and these may not align with the desired system architecture. A key component of the refrigerator reporting system is to be able to be able to forward action messages to initiate service. The complexity is that NIP is in charge of vaccines and immunizations, but another organization Medical Products Supply Center (MPSC) is in charge of the all of the vaccine refrigerators. Thus, the repair technicians are not direct employees of NIP, so there is a negotiation across the sub-organizations.

7.6.4 Analysis: A Chicken and Egg Problem

End goal of these projects are analyzing the data and get some decisive reports about the region. Evaluating the project, immediately after the deployment, based on these end goals is not practical. These deployments have to go through multiple iterations to cater the barriers to reporting before high quality data is available. This could take some time and

should be considered while the project is progressing. There are certainly insightful reports that are extracted based on the interim system data, which might already be better and detailed information than previously known to decision makers.

UNICEF wants to look at the immunization rates and vaccines stock deficiency in Laos country. WHO wants to calculate outbreaks in areas of Punjab and mark the areas that do not have health concerns. These are high level question that will be answered once these project are fully functional. Stating that an area has no health concerns or stock outs without high quality data is not possible. Failure to answer these main questions does not mean that project is a failure. It could mean that more time is required to evolve. So projects should be evaluated subjectively to look at the progress it has made in achieving better interim data based reports.

7.6.5 Design Inertia: Retraining Is Slow

Once the technology has rolled out, it is hard to iterate the design. SMS format for Laos SIM evolved over time as it was simplified for better understanding. These changes complicates the system as some health workers still send data using old format. Therefore, adopting these changes in the technology is hard while keeping backward compatible. The difficulty in communicating with the health workers and updating the system was unscored by the fact that changing the reporting phone number led to almost all of the data being missed in one reporting cycle. Updating already trained personnel about new format is a slow task when hundreds or thousands of health workers are involved. PDSS was launched with poor quality reporting wheels and replacing all of them is a hard task as thousands have already been distributed and tracking who still has old wheel is a long task.

Design iterations are expected to happen as not everything can be anticipated beforehand and usability issues will surface after deployment. The faster scaling takes place, the harder it gets to iterate the system design, which is a critical exercise to make the project successful. It will work best to scale the system at a slow pace so redesigning is easy to incorporate into the system as lessons are learned about realities that were not foreseen.

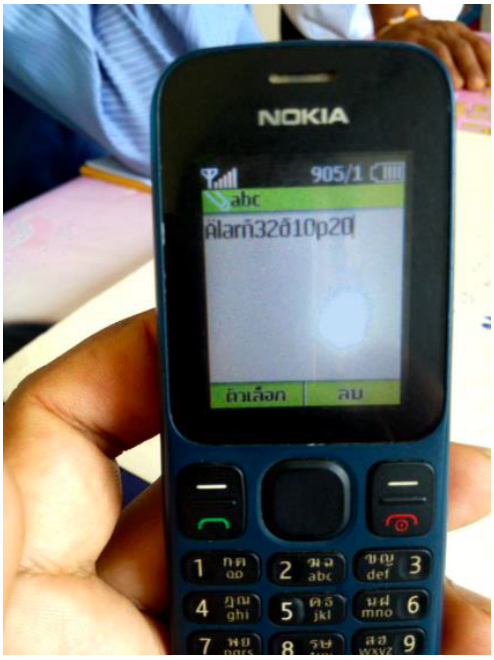

Figure 7.5: An example message showing a sample of special characters accidentally typed during a training in Laos

The component of the system that is the most difficult to update is the reporting format used by the health workers. Thus, in the Laos SIM project the full set of SMS commands have been finalized for the project before it is expanded beyond the current pilot sites. Full implementation of the commands, and the backend functionality is under development - but what is essential is that the health workers will not see any changes in their interface.

7.6.6 Design-Actuality Gaps

These systems were collaboratively designed by people familiar with each individual context, and the projects utilized technologies that have been validated in the developing world. However, there have still been challenges which could be attributed to a mismatch between assumptions made in the design, and the on-the-ground reality, or in Heeks [63] terminology, a *design-actuality gap*. The larger the gap is, the higher chance of failure. Hence a projects initial focus should be to reduce this gap as much as possible. Therefore, as described earlier,

each project had many design iterations in the early stages of roll out. Even so, there will inevitably be unforeseen circumstances that create unexpected design-actuality gaps.

As one example, PDSS unexpectedly found that, for political reasons, the executive district health officer is constantly changing. These people were supposed to be the main district level advocates for better adoption, however, most had never heard of the project before assume office. PITB is now engaging the district statistical officers who have much more employment stability.

One design-actuality gap that was larger than expected for the Laos LSIM is character sets for SMS. The Lao language has its own script, so many people have only limited familiarity with the Latin script. Sending messages using the Latin script was perceived as acceptable, since there are only a few characters to type, however and inability to distinguish between Latin characters complicates the message syntax (Fig. 7.5). Another difficulty arises with the receipt of text messages, which are important for giving acknowledgements. Since Lao is spoken by a small population, handsets generally do not support the Lao script. We had assumed that transliteration into Latin characters (or Karaoke as it is called locally) would be acceptable (as is common in many countries), but it turns out that many people are not familiar with this and would have to have help reading messages. Another work around would be to use Thai characters, which are available on handsets and close enough to Lao that they can be read. However, this was not an option, since as a government project, there was a prohibition on using Thai.

7.7 Conclusion

Issues associated with SMS reporting systems are amplified when the system is scaled to the national level. One major barrier with using SMS as a reporting medium is the complexity associated with SMS format. Considering that, making a completely automated SMS reporting system is difficult, especially at large scale. We suggest having mediators look at incorrectly formatted SMSs to fix trivial errors and follow-up on unregistered numbers. Monitoring SMS data coming from thousands of facilities is a very daunting task. Engaging

provincial and district level managers and distributing mediator responsibilities is critical in achieving acceptable reporting rates and good data quality.

It is important to reach out to health workers to determine barriers faced by them, after the deployment, and learn local adaptations deployed by them to overcome these barrier. This plays a critical role in decisive system design improvements. This could be done through district and provincial level mediators who are already engaged with these health workers for better reporting. This will ensure that the design-actuality gap is minimized quickly thereby reaching the goal of good data reporting.

National scale deployments are inherently complex due to many stakeholders with different goals for the system and roles in the implementation. The fundamental risk in accommodating these goals is complexification of the system which will have an adverse impact on training and on the health worker. Thus, it is necessary to manage stakeholder expectations and to reach common goals to ensure feasibility of the implementation. However, we expect that issues integrating a SMS system with the national HIS are surmountable and can create a more comprehensive information system.

Chapter 8

NAME RESOLUTION FOR DATA CLEANING

8.1 Summary

In this chapter, I consider datasets that are difficult to merge due to transliterated name resolution problem. The amount of time and energy that goes into matching the (location) names is immense for people working on cleaning and analytics stages of the pipeline. In turn, I explored what algorithms can be used to simplify and automate this process. I proved that a heuristic based solution that requires no ground truth and is more language independent is possible. Furthermore, I discussed how fully automating the data merge is very difficult, and, thus, proposed a solution that keeps a human in the verification loop yet reduces their workload. Most strikingly, I proposed that algorithms can be enhanced to incorporate local knowledge on language / region specific norms on name abbreviation, and emphasized the importance of first or last names for identifying an individual.

8.2 Introduction

In many instances, developing nations have digitized data with the assitance of both international and local organizations. This results in various datasets that have patient count, vaccine distribution, population distribution, immunization counts and other indicators that are very helpful in tracking the success of various interventions. It is very easy to generate individual reports, which is primary use of these datasets. There are recent efforts on creating high level regional indicators that are calculated using combination of several of these datasets. Currently, merging data from various datasets is challenging because of lack of standardization and different identifiers being used in each dataset. The best way to merge these is to use the region's name, patient's names or any other names/address that are com-

mon in every dataset. Even though that seems straightforward strategy, it is not an easy task. Other than the spelling mistakes and inconsistencies that one regularly expects, this task is specially challenging due to transliteration of foreign language names to Roman or English based characters. This results in very different spellings and even different n-grams for the same name.

Many organizations that work in international development face this problem. Several of them have staff that can write small scripts, which can ease the merging/cleaning process. While quite a bit of these organization manually clean these datasets sometimes in a basic tool like excel. Given this problem space is huge, several researchers have proposed various tools to simplify this task [69, 133]. It is important that the manual process is simplified, if not completely eliminated to reduce the time and effort spent in data cleaning and merging.

My work aims to build a system that can merge data based on matching the names or shortlist the possibilities, saving the user from countless hours of manual merging process as well as making it easier for layman user to achieve this task. To keep the system generalizable to other languages, I explored unsupervised algorithms that are easy to extend to other datasets.

8.3 Related Work

Several researchers have looked at how to generate transliterated word from one language to another in efforts to do better information retrieval [1, 158]. The idea behind this approach is to generate all possible transliterations based on the phenomes of the source language so a query in source language could be searched in transliterated content [146]. Some has done work in the space of mapping names from specific languages to English [166]. Freeman [54] even devise a modification of edit distance so all possible variation of Arabic names in English can be identified. These solutions are language specific and are not very generalizable to other languages.

Data merging based on names is a challenging task even without transliteration due to the inconsistencies in the values [88]. The current explored approaches includes approximate

Heuristic	String Matching Algorithm Combination
Variation 1	Jaro Winkler AND Edit Distance AND Double Metaphone AND NYSIIS
Variation 2	Jaro Winkler AND Edit Distance AND Double Metaphone
Variation 3	Jaro Winkler AND Double Metaphone
Variation 4	Variation 3 OR (Low Jaro Winkler AND Edit Distance AND Double Metaphone)
Variation 5	Variation 4 OR Variation 3 with removed bracketed words
Variation 6	Variation 4 OR Variation 3 with removed bracketed words and numerals

Table 8.1: List of Heuristics.

string matching or fuzzy matching algorithms [127, 147]. Some have done exploratory studies on literature of duplicate record detection, that basically covers popular string matching approaches [51].

8.4 Methodology

This section explains the algorithms used for analysis and the dataset used for the evaluation.

8.4.1 Algorithm

There are two family of string matching algorithms that I considered. First is Phonetic algorithms that uses sounds of the pronunciation to match two strings. This approach will result in positive match for similar sounding strings. The second family is approximate string matching or fuzzy string matching that calculates a relative score, which basically indicates the similarity of two string. These rely on the spelling of the string rather than sounds.

Phonetic Algorithms

Soundex, invented by Russell [138], is a phonetic algorithm that indexes a string based on the sound so minor spelling mistakes does not effects the matching. The algorithm assigns identical code digits to phonetically similar groups of consonants and ignores the vowels. The encoding consist of first letter followed by numbers based on the encoding. This

algorithm's coding is based on English sounds so might not work really well with sounds of other languages.

New York State Identification and Intelligence System (NYSIIS) [154] code is a phonetic algorithm designed to better handle the surnames. It retains the position of vowels as well as replacing consonants with other phonetically similar letters. It is an improved version of soundex for better name matching and is adopted by NEW York State Criminal Justice Services too.

Double Metaphone [124] is an improved version of Metaphone [123], proposed by Philips, that uses 16 consonant sounds that covers large variety in English and Non-English words. It considers multiple encoding for a name that greatly improves the accuracy.

Approximate String Matching (ASM) Algorithms

ASM algorithms are used to measure similarity between two string and is generally used for spell checking. There are several distance based algorithms proposed that measure how many edit it needs to transforms one word into the other. Levenshtein-Distance [87] is the most commonly used edit distance string matching algorithm.

Jaro [73] is a string matching algorithm that uses length, number of common characters between the two string and the number of transpositions required to convert one string into the other. It uses transposition as an indicator for out-of-order characters. Jaro-Winkler [162] is modification of Jaro algorithm that takes into the account that most common typo happens towards the end of the string.

Heuristics for better matching

I devised several heuristics, given in table 1, based on the combination of various string matching algorithms defined in previous section. The idea behind these heuristics is to use the strengths of both ASM and Phonetic algorithms in matching while keeping the false positives to zero or manageable.

The variation 1 till 3 were defined based on the combination of the algorithms. Initial study resulted in the confirmation that Jaro-Winkler is better than Simple Jaro algorithm and that Double-Metaphone is better than Soundex. Hence I did not include those in the combination when defining the basic heuristics. Rest of the heuristics were guided by the results of the first ones therefore the detailed discussion on their formulation is in evaluation section.

8.4.2 Datasets

The dataset that I used to evaluate the heuristics is related to Niger's population. Niger is a West African country with population of 21.48 million. This dataset has two parts, one is coming from Niger's voting bureau that has the age distribution and the second dataset is from Niger's census data that has the population distribution across locations. The objective is to merge these to achieve both age and population distribution with exact locations. To merge, I have to compare location name from both datasets, which have various spellings used due to transliteration. Unfortunately, the available ground truth against was available for only 109 locations. There were community IDs given in both datasets that could reduce the options for matching but I kept the problem challenging and operated as if we do not have any other reliable attributes in our data that could help us shortlist the potential matches.

8.5 Evaluation

The evaluation was done in two phases. First, I evaluated the unsupervised string matching algorithms individually against the dataset. This is to verify how well the algorithms perform from off the shelf implementations. Second phase was focused on performance of the defined heuristics. Several heuristics were evolved in this phase, based on the results of previous phase, to achieve better performance.

Since the evaluation of the algorithms is from the perspective of data merging, I defined a metric that I call detectable false positive. One of the biggest fear that organizations have with automated cleaning/merging is incorrect matching. Even if the wrong merge happens

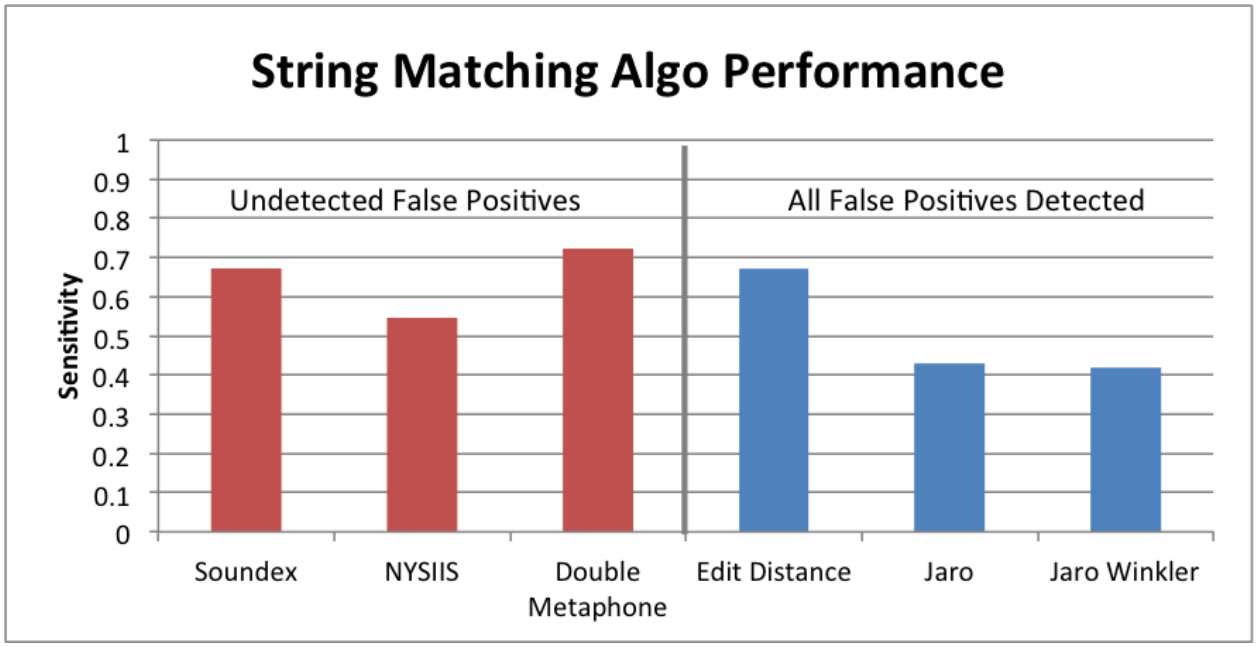

Figure 8.1: Sensitivity of string matching algorithms against the Niger transliterated locality names

to 1 percent of the data, it is not acceptable specially in health sector. The idea behind detectable false positives is that we lower our matching thresholds just enough that a false positive will always results in several matches including the true positive match. This will reduce the overall accuracy but will maintain a confidence that a single positive match for a location name is indeed a true positive.

My proposal for the system is that system should have zero undetectable false positives. That way, the system can merge all the true positives without human interventions. If there are multiple matches then it is presented to the user who can pick the right match. Even with user involved in this verification loop, the overall workload will be less with full confidence on the automated merged data.

8.5.1 String Matching Algorithms

Sensitivity of both Phonetic and ASM algorithms is given in Figure 1. Phonetic algorithms result in a encoding representation of the given string. The matching happen by comparing these encoded strings. This means that there is no way to lower the threshold of these algorithms to avoid undetectable false positives. While the ASM algorithm always present a distance or relevance score that one can draw different threshold on to optimize for less false positives.

Due to the nature of the phonetic algorithms, I was unable to change them to avoid false positives. Hence all the phonetic algorithms shows presence of undetectable false positives comparing to ASM algorithms. This also means that I can not compare the sensitivity numbers across the type of algorithms since the thresholds of ASM algorithms were change to less false positives than more true positives.

The surprising part is that Levenshtein edit distance performed significantly better than both Jaro algorithms. The 0.72 sensitivity of Double-Metaphone sets an expectation for us. This is the best I could get from off the shelf implementation. If I were able to get same or better sensitivity without undetectable false positives from our heuristics algorithms then that will be good result.

8.5.2 Heuristic Algorithms

The results of defined heuristics are shown in Figure 2. The first two variation shows undetectable false positives due to requirement of matching from several algorithms. This trims away the true positive from the possible matches because one algorithm did not agree with it the match. Hence variation 3 performs the best among the first three predefined variations.

To clarify, the results of heuristic variation is not a simple AND of the sensitivity results from first analysis. Variation 3, for example, performs better comparing to separately calculated sensitivity of Double-Metaphone and Jaro-Winkler because there is a considerable disjoint in true positives of the individual results. The heuristic not only take benefit of that

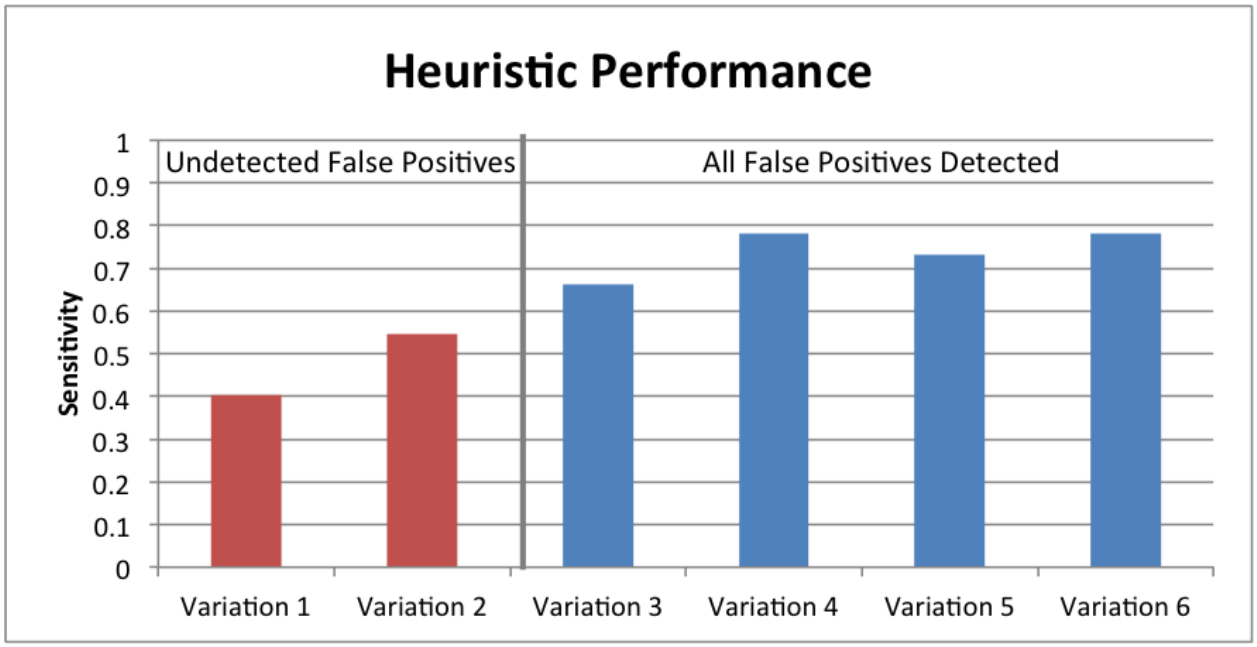

Figure 8.2: Sensitivity of different heuristics defined, using a combination of string matching algorithms, against the Niger transliterated locality names

but also able to remove undetectable false matches due to it. Also, a lower Jaro threshold was used for the heuristic algorithm since the Double-Metaphone was good is catching the failing cases by Jaro.

Variation 4 was defined based on the assumption that if I lower the threshold of Jaro from 0.88 to 0.84, I will get more true positives along with more noise. Then the Levenshtein edit distance and Double-Metaphone will be able to reduce the noise. This increased the sensitivity from 0.66 to 0.78, which is better performance than the best number I got from Phonetic algorithms. Plus, this improvement is without any undetectable false positives.

Variation 5 and 6 was to test the result of removal of certain post-fixes. The data showed that the failing matches are due to presence of numeral numbers at the end of the location names or extra detail mentioned with brackets around it. The result shows that removing numeral does not help but removing the bracketed words does have slight improvement.

Location Name 1	Location Name 2
akoual	akouel (akoyal)
guidan sabon gari(kankare)	sabon gari kankare
kassou chagolo	guidan kassou chagalo
zango algabit ii	zongon algabid
guidan neyno(moutoun naya)	guidan neino moutoun
tchilili	intchilili

Table 8.2: Examples of failed matches

8.6 Discussion

Achieving good results while keeping the undetectable false positive zero is a hard problem in the transliterated name resolution space. While standard string matching algorithm did reasonable but with undetected false matches, I was able to achieve better results while keeping the confidences that positive match are indeed right matches.

To improve these results, we need to identify prefix and postfix within names. Niger data shows that 'guidan' and 'int' are prefixes that we should trim before matching the strings. Identifying these is a hard problem since it requires user's knowledge about the language. Same prefixes can not be applied in other languages. I know that 'Muhammad' or 'M' are prefixes in Arabic, Farsi and Urdu names that one should ignore when matching names in those languages.

We could build a system that scans a dataset and find the most common prefix/postfix words that appear in all the names. The system can further use user to verify the results with the assumption that the users are knowledgeable about the language and we have datasets large enough to find these patterns. This is a collective knowledge that if we compile over time for various languages then it will improve the system in the long run.

Table 2 shows the examples that resulted in failed matches. Some of these examples clarify that one or similar rules can not result in all data matched. First example shows that we should drop bracketed word since it is a variation in spelling for the same name but in

second example we see that bracketed word is part of the place name. One solution is that we create rules such that 1-grams are handled differently versus matching n-gram names, where n is greater than 1. To prove or disprove this, we need bigger ground truth data to test the solution. Based on the discussion above, there are several heuristics that one could explore but ground truth data was the limiting factor in my analysis.

I defined heuristics based on my intuition and initial results I got from the string matching. I believe that better results could be achieved if we get weightage of individual algorithms from a model. Other parameters like n-gram, numerals, bracketed words, length of words and other variables could help model figure out an even better matching.

Transliterated names that results in multi-word location names are easy to handle if a 2-gram location name is represented in 3-gram name in other datasets. Phonetic algorithms are good in ignoring the space when matching. The problem is when one is a sub-string of the other. This makes it harder to identify that if the names are indeed different or there is a prefix or suffix missing. So far I have not seen strong motivation to split the n-grams and do more complex analysis for each word matching but we need more data to study this in more detail.

8.7 Conclusion

In this chapter, I evaluated the ASM and Phonetic algorithms for the matching of transliterated names. I defined heuristics that use the strength of both types of algorithms while keeping false positives detectable. I have shown in my results that the heuristic based, new algorithms outperforms the phonetic algorithms without undetectable incorrect matches. Beyond this I discussed other heuristics that are worth exploring with additional data or diverse ground truth data.

Chapter 9

CONCLUSION

In this thesis, I propose a taxonomy of challenges found in the development data pipeline. This taxonomy is based on extensive qualitative research that sought to understand these challenges from the viewpoint of development data workers as well as a series of case studies from various domains of health and finance. In Chapter 3, I formalized the set of challenges, frustrations, and barriers in the development data pipeline, and each subsequent chapter focused on a case study that supported and added nuance to these.

The key issues that emerged from the qualitative study, presented in Chapter 3, are extracting data from legacy textual formats, merging data between existing large data sources, and validating data accuracy. Moreover, partnerships between local governments, non-profits, non-governmental organizations, and social enterprises contribute to frustrations at each stage of the data pipeline due to misaligned objectives and expectations. These challenges are substantially difficult to address in developing world contexts due to the lack of availability of tools and resources, poor planning, and uncoordinated efforts of multiple stakeholders.

I determined barriers in scaling an information system and defining a data standard by developing a vaccine cold chain reporting system and a general use data standards. In this process, I found that the evolution of data collection and its formatting is inevitable even when all stakeholders agree to establish the collection format. It is important that we extend the existing formats to evolve instead of replacing them completely. This will keep the newer data compatible with older data as the information system changes and new needs arise. Moreover, establishing a data standard that is implemented by diverse tools supporting the domain and is accepted by all those involved is a difficult task since every stakeholder will

have slightly different needs. Nonetheless, if the domain is well defined and limited in scope, then achieving a consensus on a standard is possible as I illustrated for the cold chain.

Manually reviewing data is unavoidable in some cases as I experienced in processing unstructured crowd sourced SMS data for fraud analysis, which I presented in Chapter 6. I reviewed SMS texts to look for trends. This is similar to having open ended questions in surveys that then must be processed. Although I determined potential keywords to extract from texts, I encountered the issue of various spellings for the same keyword due to transliteration variations. This is the same issue as name resolution that hinders data merge, which I discussed in Chapter 3. I determined that building a string match for transliterated words will allow us to address several issues in data processing.

I noted the barriers to and issues with data collected over SMS, arguing that potential solutions to these issues could be found by studying large scale national collection systems. In talking to people involved in both systems, I discovered that sending a feedback message about what is reported, what data is not yet received, and guidelines for the format if unrecognized syntax is encountered contributes to the workers feeling recognized and appreciated for their efforts. The usability of the SMS format was evaluated differently by new users versus expert users for the same reporting type. While new users appreciated a self exploratory format, the expert users wanted a format that is quick to construct without too much effort needed to look up codes. From this I ascertained that the SMS format should be formatted in a way that the system can easily identify if an error is made, while at the same time able to rely on minimum codes so expert reporters can format from memory. Finally, in large scale training, the ability to standardize and structure the training content with specific worksheets and flyers for the trainers is very important for a new SMS reporting system, or any other system for that matter. This is because of the variable quality in training due to a large variety of trainers that result in variable quality in data being reported by the new trainees.

I prove that the name resolution problem can be addressed with a heuristic that considers the output of string approximation algorithms as well as phonetic algorithms. This approach

is language independent as it is an unsupervised algorithm and does not need language specific training. Second, I presented an approach that aims for high sensitivity at the cost of lower specificity since some stakeholders explained that perfect matching is important to them. This strategy was to prove that we can build algorithms that involve human in the loop, match significant portion of the data, and leave the rest for manual matching with potential options listed. Overall, this reduces the manual labor required to match such datasets while keeping the precision 100%. Moreover, one can easily change the thresholds for needs when 100% precision is not important, so the algorithm would aim for overall high accuracy that includes both positive and negative matches, as well as match the majority of the data with some errors that will go undetected.

Development data needs support on two fronts: streamlining the future data pipelines and patch-up of the existing pipelines. Streamlining the future pipelines means standardizing the training processes for better quality data collection, having clear collection forms, setting up universal information systems that could scale as need changes, creating data standards that all applications within a domain follow and having proper documentation about the data collection and processing so future use of data is well informed. Ideally, all these efforts would lead to a perfect system, but, realistically, it only helps future projects. Significant development efforts are dealing with legacy data and systems, which require tools that will help with issues in current pipelines that are not going away anytime soon. These include extracting data from legacy reports like PDFs, establishing the accuracy of the data and matching transliterated names of location or people for data merging or processing open text fields in a form. These tools should be flexible to different types of inputs and outputs so they can be easily adapted to different systems and needs.

The development data pipeline is complex with different tools and datasets for the same domain due to limited scope of each development project led by different organizations. Each project should consider scalability to other regions, adaptability to work with other systems, and the extension or repurposing of data for other analyses. This could be better achieved by reaching out to other organizations in the intended space to get data schema

that others follow, national data (e.g., facility list or population distribution), and keep well organized public documentation on system architecture, data collection, data attributes, and processing done on the data. While this documentation may be about a private access system, it would allow archiving of the documentation for future use as well as a guide for other stakeholders in the domain. Open sourcing the the documentation could allow for consensus among stakeholders with regard to standards for data systems.

In reality, it is a challenge to guide various stakeholder to consensus. Their motivations and goals are often mitigated by a number of perceived and stated goals, and are dependent on the guidance of funding organizations and agencies. Likewise, competition between various non-profits both for projects and funding, may not provide the necessary motivation to work together on developing new systems that seek to overcome development data challenges. It may be that approaching these issues at the source of the funding would have more impact if donors could work toward establishing data standards, information systems, reusable data, and in general, a robust data pipeline for a variety of stakeholders. Funding agencies might consider broadening their scope to target missing tools in the data pipeline and fund projects to build such tools for general use by any organization. This would allow researchers and policy makers to more quickly analyze data without the need for extensive data cleaning.

BIBLIOGRAPHY

[1] Nasreen Abdul Jaleel and Leah S Larkey. Statistical transliteration for English-Arabic cross language information retrieval. In *Proceedings of the twelfth international conference on Information and knowledge management*, pages 139–146. ACM, 2003.

[2] Frank W. Agbola, Angelito Acupan, and Amir Mahmood. Does microfinance reduce poverty? New evidence from Northeastern Mindanao, the Philippines. *Journal of Rural Studies*, 50:159–171, February 2017.

[3] Elske Ammenwerth, Stefan Gräber, Gabriele Herrmann, Thomas Bürkle, and Jochem König. Evaluation of health information systemsproblems and challenges. *International journal of medical informatics*, 71(2-3):125–135, 2003.

[4] David S Anderson, Chris Fleizach, Stefan Savage, and Geoffrey M Voelker. *Spamscatter: Characterizing internet scam hosting infrastructure*. PhD thesis, University of California, San Diego, 2007.

[5] Richard Anderson, John Lloyd, and Sophie Newland. Software for national level vaccine cold chain management. In *Proceedings of the Fifth International Conference on Information and Communication Technologies and Development*, pages 190–199. ACM, 2012.

[6] Richard Anderson, Trevor Perrier, Fahad Pervaiz, Norasingh Sisouveth, Bharath Kumar, Sompasong Phongphila, Ataur Rahman, Ranjit Dhiman, and Sophie Newland. Supporting immunization programs with improved vaccine cold chain information systems. In *Global Humanitarian Technology Conference (GHTC), 2014 IEEE*, pages 215–222. IEEE, 2014.

[7] Caroline Asiimwe, David Gelvin, Evan Lee, Yanis Ben Amor, Ebony Quinto, Charles Katureebe, Lakshmi Sundaram, David Bell, and Matt Berg. Use of an innovative, affordable, and open-source short message service–based tool to monitor malaria in remote areas of Uganda. *The American journal of tropical medicine and hygiene*, 85(1):26–33, 2011.

[8] David Avison and Guy Fitzgerald. *Information systems development: methodologies, techniques and tools*. McGraw Hill, 2003.

[9] Jim Barrington, Olympia Wereko-Brobby, Peter Ward, Winfred Mwafongo, and Seif Kungulwe. SMS for Life: a pilot project to improve anti-malarial drug supply management in rural Tanzania using standard technology. *Malaria journal*, 9(1):298, 2010.

[10] Yahel Ben-David, Shaddi Hasan, Joyojeet Pal, Matthias Vallentin, Saurabh Panjwani, Philipp Gutheim, Jay Chen, and Eric A Brewer. Computing security in the developing world: A case for multidisciplinary research. In *Proceedings of the 5th ACM workshop on Networked systems for developing regions*, pages 39–44. ACM, 2011.

[11] M. Berg, J. Wariero, and V. Modi. *Every Child Counts: The Use of SMS in Kenya to Support the Community Based Management of Acute Malnutrition and Malaria in Children Under Five*. Columbia university. Earth institute. ChildCount : with UNICEF Innovation group, 2009.

[12] Nita Bhalla. Poor nations need help to use big data to tackle disease, poverty: expert, July 2017.

[13] Moez Bhatti. QKSMS. https://github.com/moezbhatti/qksms, 2019. Accessed March 2019.

[14] Tony Blakely and Clare Salmond. Probabilistic record linkage and a method to calculate the positive predictive value. *International journal of epidemiology*, 31(6):1246–1252, 2002.

[15] Jørn Braa, Ole Hanseth, Arthur Heywood, Woinshet Mohammed, and Vincent Shaw. Developing health information systems in developing countries: the flexible standards strategy. *MIS Quarterly*, pages 381–402, 2007.

[16] Jørn Braa and Calle Hedberg. The struggle for district-based health information systems in South Africa. *The information society*, 18(2):113–127, 2002.

[17] Jørn Braa, Eric Monteiro, and Sundeep Sahay. Networks of action: sustainable health information systems across developing countries. *MIS quarterly*, pages 337–362, 2004.

[18] Kristin Braa, Petter Nielsen, and Ola Titlestad. Innovation for Health in Developing Countries. In *Medical Technology–Meeting Tomorrows Health Care Challenges*, page 21.

[19] Kristin Braa and T Sanner. Making mHealth happen for health information systems in low resource contexts. In *Proceedings of the 11th International Conference on Social Implications of Computers in Developing Countries*, pages 530–541, 2011.

[20] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101, 2006.

[21] Waylon Brunette, Samuel Sudar, Mitchell Sundt, Clarice Larson, Jeffrey Beorse, and Richard Anderson. Open Data Kit 2.0: A services-based application framework for disconnected data management. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*, pages 440–452. ACM, 2017.

[22] Waylon Brunette, Samuel Sudar, Nicholas Worden, Dylan Price, Richard Anderson, and Gaetano Borriello. ODK Tables: building easily customizable information applications on Android devices. In *Proceedings of the 3rd ACM Symposium on Computing for Development*, page 12. ACM, 2013.

[23] Waylon Brunette, Mitchell Sundt, Nicola Dell, Rohit Chaudhri, Nathan Breit, and Gaetano Borriello. Open Data Kit 2.0: expanding and refining information services for developing regions. In *Proceedings of the 14th workshop on mobile computing systems and applications*, page 10. ACM, 2013.

[24] Jenna Burrell. Problematic Empowerment: West African Internet Scams as Strategic Misrepresentation. *Information Technologies and International Development*, 4(4):15–30, 2008.

[25] Stuart K Card, Jock D Mackinlay, and Ben Shneiderman. Using vision to think. In *Readings in information visualization*, pages 579–581. Morgan Kaufmann Publishers Inc., 1999.

[26] Sam Castle, Fahad Pervaiz, Galen Weld, Franziska Roesner, and Richard Anderson. Let's Talk Money: Evaluating the Security Challenges of Mobile Money in the Developing World. In *Proceedings of the 7th Annual Symposium on Computing for Development*, ACM DEV '16, pages 4:1–4:10, New York, NY, USA, 2016. ACM.

[27] CHAI. Clinton Health Access Initiative, 2019.

[28] Rohit Chaudhri, Gaetano Borriello, and Richard Anderson. Pervasive computing technologies to monitor vaccine cold chains in developing countries. *IEEE Pervasive Computing. Special issue on Information and Communication Technologies for Development*, 2012.

[29] Rohit Chaudhri, Eleanor O'Rourke, Shawn McGuire, Gaetano Borriello, and Richard Anderson. FoneAstra: enabling remote monitoring of vaccine cold-chains using commodity mobile phones. In *Proceedings of the First ACM Symposium on Computing for Development*, page 14. ACM, 2010.

[30] Kuang Chen, Emma Brunskill, Jonathan Dick, and Prabhjot Dhadialla. Learning to Identify Locally Actionable Health Anomalies. In *AAAI Spring Symposium: Artificial Intelligence for Development*, 2010.

[31] Kuang Chen, Harr Chen, Neil Conway, Joseph M Hellerstein, and Tapan S Parikh. Usher: Improving data quality with dynamic forms. *IEEE Transactions on Knowledge and Data Engineering*, 23(8):1138–1153, 2011.

[32] Kuang Chen, Joseph M Hellerstein, and Tapan S Parikh. Data in the First Mile. In *CIDR*, pages 203–206. Citeseer, 2011.

[33] Kuang Chen, Akshay Kannan, Yoriyasu Yano, Joseph M Hellerstein, and Tapan S Parikh. Shreddr: pipelined paper digitization for low-resource organizations. In *Proceedings of the 2nd ACM Symposium on Computing for Development*. ACM, 2012.

[34] Peter Christen. *Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection*. Springer Science & Business Media, 2012.

[35] Camille Cobb, Samuel Sudar, Nicholas Reiter, Richard Anderson, Franziska Roesner, and Tadayoshi Kohno. Computer security for data collection technologies. In *Proceedings of the Eighth International Conference on Information and Communication Technologies and Development*. ACM, 2016.

[36] Gordon V Cormack et al. Email spam filtering: A systematic review. *Foundations and Trends® in Information Retrieval*, 1(4):335–455, 2008.

[37] Henry Corrigan-Gibbs and Jay Chen. Flashpatch: spreading software updates over flash drives in under-connected regions. In *Proceedings of the Fifth ACM Symposium on Computing for Development*, pages 1–10. ACM, 2014.

[38] Catalina M Danis, Jason B Ellis, Wendy A Kellogg, Hajo van Beijma, Bas Hoefman, Steven D Daniels, and Jan-Willem Loggers. Mobile phones for health education in the developing world: SMS as a user interface. In *Proceedings of the First ACM Symposium on Computing for Development*, page 13. ACM, 2010.

[39] Tamraparni Dasu and Theodore Johnson. *Exploratory data mining and data cleaning*, volume 479. John Wiley & Sons, 2003.

[40] Paul A David. Understanding the economics of QWERTY: The necessity of history. *Economic History and the modern economics*, pages 30–49, 1986.

[41] Paul A David and Mark Shurmer. Formal standards-setting for global telecommunications and information services. Towards an institutional regime transformation? *Telecommunications policy*, 20(10):789–815, 1996.

[42] DAWN. FIA finds the educated "equally vulnerable" to online bank fraud, 2018.

[43] Sarah Jane Delany, Mark Buckley, and Derek Greene. SMS spam filtering: methods and data. *Expert Systems with Applications*, 39(10):9899–9908, 2012.

[44] Nicola Dell, Nathan Breit, Jacob O Wobbrock, and Gaetano Borriello. Improving form-based data entry with image snippets. In *Proceedings of Graphics Interface 2013*, pages 157–164. Canadian Information Processing Society, 2013.

[45] Nicola Dell, Trevor Perrier, Neha Kumar, Mitchell Lee, Rachel Powers, and Gaetano Borriello. Paper-digital workflows in global development organizations. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 1659–1669. ACM, 2015.

[46] Brian DeRenzi, Nicola Dell, Jeremy Wacksman, Scott Lee, and Neal Lesh. Supporting community health workers in India through voice-and web-based feedback. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 2770–2781. ACM, 2017.

[47] Brian DeRenzi, Leah Findlater, Jonathan Payne, Benjamin Birnbaum, Joachim Mangilima, Tapan Parikh, Gaetano Borriello, and Neal Lesh. Improving community health worker performance through automated SMS. In *Proceedings of the Fifth International Conference on Information and Communication Technologies and Development*, pages 25–34. ACM, 2012.

[48] Leslie L Dodson, S Sterling, and John K Bennett. Considering failure: eight years of ITID research. In *Proceedings of the Fifth International Conference on Information and Communication Technologies and Development*, pages 56–64. ACM, 2012.

[49] Ficawoyi Donou-Adonsou and Kevin Sylwester. Financial development and poverty reduction in developing countries: New evidence from banks and microfinance institutions. *Review of Development Finance*, 6(1):82–90, June 2016.

[50] eHealth Africa. VaxTrac System, 2019.

[51] Ahmed K Elmagarmid, Panagiotis G Ipeirotis, and Vassilios S Verykios. Duplicate record detection: A survey. *IEEE Transactions on knowledge and data engineering*, 19(1), 2007.

[52] Polio Eradication. Global Polio Eradication Initiative - Data and monitoring - Polio this week, 2019.

[53] S Thomas Foster and Kunal K Ganguly. *Managing quality: Integrating the supply chain*. Pearson Prentice Hall Upper Saddle River, New Jersey, 2007.

[54] Andrew T Freeman, Sherri L Condon, and Christopher M Ackerman. Cross linguistic name matching in English and Arabic: a one to many mapping extension of the Levenshtein edit distance algorithm. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 471–478. Association for Computational Linguistics, 2006.

[55] Lise Getoor and Ashwin Machanavajjhala. Entity resolution: theory, practice & open challenges. *Proceedings of the VLDB Endowment*, 5(12):2018–2019, 2012.

[56] José María Gómez Hidalgo, Guillermo Cajigas Bringas, Enrique Puertas Sánz, and Francisco Carrero García. Content based SMS spam filtering. In *Proceedings of the 2006 ACM symposium on Document engineering*, pages 107–114. ACM, 2006.

[57] Joshua Goodman, Gordon V Cormack, and David Heckerman. Spam and the ongoing battle for the inbox. *Communications of the ACM*, 50(2):24–33, 2007.

[58] George D. Greenwade. The Comprehensive Tex Archive Network (CTAN). *TUGBoat*, 14(3):342–351, 1993.

[59] Zoltan Gyongyi and Hector Garcia-Molina. Web spam taxonomy. In *First international workshop on adversarial information retrieval on the web (AIRWeb 2005)*, 2005.

[60] Pat Hanrahan. Tableau software white paper-visual thinking for business intelligence. *Tableau Software, Seattle, WA*, 2003.

[61] Carl Hartung, Adam Lerer, Yaw Anokwa, Clint Tseng, Waylon Brunette, and Gaetano Borriello. Open Data Kit: tools to build information services for developing regions. In *Proceedings of the 4th ACM/IEEE international conference on information and communication technologies and development*, page 18. ACM, 2010.

[62] Niall Hayes and Chris Westrup. Consultants as intermediaries and mediators in the construction of information and communication technologies for development. *Information Technologies & International Development*, 10(2):pp–19, 2014.

[63] Richard Heeks. Information systems and developing countries: Failure, success, and local improvisations. *The Information Society*, 18(2):101–112, 2002.

[64] Richard Heeks. Health information systems: Failure, success and improvisation. *International Journal of Medical Informatics*, 75(2):125–137, 2006.

[65] Jeffrey Heer and Ben Shneiderman. Interactive dynamics for visual analysis. *Queue*, 10(2):30, 2012.

[66] Joseph M Hellerstein. Quantitative data cleaning for large databases. *United Nations Economic Commission for Europe (UNECE)*, 2008.

[67] Joseph M Hellerstein. People, Computers, and The Hot Mess of Real Data. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 7–7. ACM, 2016.

[68] Justin J Henriques, Benjamin T Foster, William G Schnorr, and Reed Barton. Implementation of a Mobile Vaccine Refrigerator with Parallel Photovoltaic Power Systems. In *2012 IEEE Global Humanitarian Technology Conference*, pages 128–131. IEEE, 2012.

[69] Mauricio A Hernández and Salvatore J Stolfo. Real-world data is dirty: Data cleansing and the merge/purge problem. *Data mining and knowledge discovery*, 2(1):9–37, 1998.

[70] Samia Ibtasam, Hamid Mehmood, Lubna Razaq, Jennifer Webster, Sarah Yu, and Richard Anderson. An Exploration of Smartphone Based Mobile Money Applications in Pakistan. In *Proceedings of the Ninth International Conference on Information and Communication Technologies and Development*, ICTD '17, pages 1:1–1:11, New York, NY, USA, 2017. ACM.

[71] Jelena Isacenkova, Olivier Thonnard, Andrei Costin, Aurélien Francillon, and David Balzarotti. Inside the scam jungle: A closer look at 419 scam email operations. *EURASIP Journal on Information Security*, 2014(1):4, 2014.

[72] HV Jagadish, Johannes Gehrke, Alexandros Labrinidis, Yannis Papakonstantinou, Jignesh M Patel, Raghu Ramakrishnan, and Cyrus Shahabi. Big data and its technical challenges. *Communications of the ACM*, 57(7):86–94, 2014.

[73] Matthew A Jaro. Probabilistic linkage of large public health data files. *Statistics in medicine*, 14(5-7):491–498, 1995.

[74] Nan Jiang, Yu Jin, Ann Skudlark, and Zhi-Li Zhang. Greystar: Fast and Accurate Detection of SMS Spam Numbers in Large Cellular Networks Using Gray Phone Space. In *The 22nd USENIX Security Symposium (USENIX Security 13)*, pages 1–16, 2013.

[75] Sean Kandel, Andreas Paepcke, Joseph Hellerstein, and Jeffrey Heer. Wrangler: Interactive visual specification of data transformation scripts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3363–3372. ACM, 2011.

[76] Amir Karami and Lina Zhou. Improving static SMS spam detection by using new content-based features. *Americas Conference on Information Systems 2014 Proceedings*, 2014.

[77] Dean Karlan, Jake Kendall, Rebecca Mann, Rohini Pande, Tavneet Suri, and Jonathan Zinman. Research and impacts of digital financial services. Technical report, National Bureau of Economic Research, 2016.

[78] Amy K Karlson, AJ Brush, and Stuart Schechter. Can I borrow your phone?: understanding concerns when sharing mobile phones. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1647–1650. ACM, 2009.

[79] Josephine Karuri, Peter Waiganjo, ORWA Daniel, and Ayub MANYA. DHIS2: The tool to improve health data demand and use in Kenya. *Journal of Health Informatics in Developing Countries*, 8(1), 2014.

[80] Anas Katib, Deepthi Rao, Praveen Rao, and Karen Williams. Jeev: a low-cost cell phone application for tracking the vaccination coverage of children in rural communities. In *2013 IEEE International Conference on Healthcare Informatics*, pages 115–120. IEEE, 2013.

[81] Won Kim, Byoung-Ju Choi, Eui-Kyeong Hong, Soo-Kyung Kim, and Doheon Lee. A taxonomy of dirty data. *Data mining and knowledge discovery*, 7(1):81–99, 2003.

[82] Jon A Krosnick, Sowmya Narayan, and Wendy R Smith. Satisficing in surveys: Initial evidence. *New directions for evaluation*, 1996(70):29–44, 1996.

[83] A. D. Langmuir. Evolution of the concept of surveillance in the United States. *Proceedings of the Royal Society of Medicine*, 64(6):681–684, June 1971.

[84] Bruce Y Lee, Tina-Marie Assi, Korngamon Rookkapan, Angela R Wateska, Jayant Rajgopal, Vorasith Sornsrivichai, Sheng-I Chen, Shawn T Brown, Joel Welling, Bryan A Norman, et al. Maintaining vaccine delivery following the introduction of the rotavirus and pneumococcal vaccines in Thailand. *PloS one*, 6(9):e24673, 2011.

[85] Mong Li Lee, Hongjun Lu, Tok Wang Ling, and Yee Teng Ko. Cleansing data for mining and warehousing. In *International Conference on Database and Expert Systems Applications*, pages 751–760. Springer, 1999.

[86] Adam Lerer, Molly Ward, and Saman Amarasinghe. Evaluation of IVR data collection UIs for untrained rural users. In *Proceedings of the first ACM symposium on computing for development*, page 2. ACM, 2010.

[87] Vladimir I Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. In *Soviet physics doklady*, volume 10, page 707, 1966.

[88] Sergio Luján-Mora and Manuel Palomar. Reducing inconsistency in integrating data from different sources. In *Database Engineering and Applications, 2001 International Symposium on.*, pages 209–218. IEEE, 2001.

[89] Hong Ma. Google Refine–http://code. google. com/p/google-refine. *Technical Services Quarterly*, 29(3):242–243, 2012.

[90] Robert MacIntosh and Dmitri Vinokurov. Detection and mitigation of spam in IP telephony networks using signaling protocol analysis. In *IEEE/Sarnoff Symposium on Advances in Wired and Wireless Communication, 2005.*, pages 49–52. IEEE, 2005.

[91] Sriganesh Madhvanath, Geetha Manjunath, Suryaprakash Kompalli, Serene Banerjee, Sitaram Ramachandrula, and Srinivasu Godavari. PaperWeb: paper-triggered web interactions. In *Proceedings of the 3rd ACM Symposium on Computing for Development*, page 43. ACM, 2013.

[92] Kedar S Mate, Brandon Bennett, Wendy Mphatswe, Pierre Barker, and Nigel Rollins. Challenges for routine health system data management in a large public programme to prevent mother-to-child HIV transmission in South Africa. *PloS one*, 4(5):e5483, 2009.

[93] John M Maurice and Sheila Davey. *State of the World's Vaccines and Immunization.* World Health Organization, 2009.

[94] Kate McKee, Michelle Kaffenberger, and Jamie Zimmerman. Doing Digital Finance Right, June 2015.

[95] Wes McKinney et al. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, volume 445, pages 51–56. van der Voort S, Millman J, 2010.

[96] Nora Méray, Johannes B Reitsma, Anita CJ Ravelli, and Gouke J Bonsel. Probabilistic record linkage is a valid and transparent tool to combine databases without a patient identification number. *Journal of clinical epidemiology*, 60(9):883–e1, 2007.

[97] Najmeh Miramirkhani, Oleksii Starov, and Nick Nikiforakis. Dial one for scam: A large-scale analysis of technical support scams. *arXiv preprint arXiv:1607.06891*, 2016.

[98] Gianluca Miscione. Scalability as Institutionalization-Practicing District Health Information System in an Indian State Health Organization. 2007.

[99] Alvaro E. Monge. Matching algorithms within a duplicate detection system. *IEEE Data Eng. Bull.*, 23(4):14–20, 2000.

[100] Joseck Luminzu Mudiri. Fraud in mobile financial services. *Rapport technique, MicroSave*, page 30, 2013.

[101] Henry Mwanyika, David Lubinski, Richard Anderson, Kelley Chester, Mohamed Makame, Matt Steele, and Don de Savigny. Rational systems design for health information systems in low-income countries: An enterprise architecture approach. *Journal of Enterprise Architecture*, 7(4):60–69, 2011.

[102] Akshay Narayan and Prateek Saxena. The curse of 140 characters: evaluating the efficacy of SMS spam detection on android. In *Proceedings of the Third ACM workshop on Security and privacy in smartphones & mobile devices*, pages 33–42. ACM, 2013.

[103] The Nation. Internet banking frauds on the rise, 2018.

[104] Carib Nelson, Paulo Froes, Anne Mie Van Dyck, Jeaneth Chavarría, Enrique Boda, Alberto Coca, Gladys Crespo, and Heinz Lima. Monitoring temperatures in the vaccine cold chain in Bolivia. *Vaccine*, 25(3):433–437, 2007.

[105] Pharmalink Newsletter. Pharmalink Newsletter, Vol 11, Issue 1, November 2011. Effective Pharmaceutical Supply Chains - On the Road in Low Income Countries, 2011.

[106] Nexleaf. Cold Chain Monitor, 2018.

[107] Matthew J O'Brien, Allison P Squires, Rebecca A Bixby, and Steven C Larson. Role development of community health workers: an examination of selection and training processes in the intervention literature. *American journal of preventive medicine*, 37(6):S262–S269, 2009.

[108] Pakistan Bureau of Statistics. Administrative Units, 2017.

[109] Government of The Punjab. District Health Information System (DHIS), 2018.

[110] WHO — World Health Organization. The WHO Performance, Quality and Safety (PQS), 2019.

[111] World Health Organization. Master facility list resource package: Guidance for countries wanting to strengthen their MFL, March 2018.

[112] PAHO. Vaccination Supplies Stock Management (VSSM), 2010.

[113] Md Jahid Hossain PANIR. Role of ICTs in the health sector in developing countries: a critical review of literature. *Journal of Health Informatics in Developing Countries*, 5(1), 2011.

[114] Tapan S Parikh. Engineering rural development. *Communications of the ACM*, 52(1):54–63, 2009.

[115] Tapan S Parikh, Paul Javid, Kaushik Ghosh, Kentaro Toyama, et al. Mobile phones and paper documents: evaluating a new approach for capturing microfinance data in rural India. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 551–560. ACM, 2006.

[116] PATH. Cold Chain Equipment Manager (CCEM), January 2012.

[117] PATH. PATH Seattle, 2014.

[118] Somani Patnaik, Emma Brunskill, and William Thies. Evaluating the accuracy of data collection on mobile phones: A study of forms, SMS, and voice. In *Information and Communication Technologies and Development (ICTD), 2009 International Conference on*, pages 74–84. IEEE, 2009.

[119] Fahad Pervaiz, Richard Anderson, and Sophie Newland. Data Specification for Information Systems for the Immunization Cold Chain. *Proceedings of 2015 Technet Conference*, 2015.

[120] Fahad Pervaiz, Rai Shah Nawaz, Muhammad Umer Ramzan, Maryem Zafar Usmani, Shrirang Mare, Kurtis Heimerl, Faisal Kamiran, Richard Anderson, and Lubna Razaq. An assessment of sms fraud in pakistan. In *Proceedings of the 2nd ACM SIGCAS Conference on Computing and Sustainable Societies*. ACM, 2019.

[121] Fahad Pervaiz, Trevor Perrier, Sompasong Phongphila, and Richard Anderson. User errors in SMS based reporting systems. In *Proceedings of the Seventh International Conference on Information and Communication Technologies and Development*, page 55. ACM, 2015.

[122] Fahad Pervaiz, Aditya Vashistha, and Richard Anderson. Examining the challenges in development data pipeline. In *Proceedings of the 2nd ACM SIGCAS Conference on Computing and Sustainable Societies*. ACM, 2019.

[123] Lawrence Philips. Hanging on the metaphone. *Computer Language*, 7(12 (December)), 1990.

[124] Lawrence Philips. The double metaphone search algorithm. *C/C++ users journal*, 18(6):38–43, 2000.

[125] Rowan Phipps, Shrirang Mare, Peter Ney, Jennifer Webster, and Kurtis Heimerl. ThinSIM-based Attacks on Mobile Money Systems. In *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies*, page 23. ACM, 2018.

[126] Rowan Phipps, Shrirang Mare, Peter Ney, Jennifer Webster, and Kurtis Heimerl. ThinSIM-based Attacks on Mobile Money Systems. In *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies*, COMPASS '18, pages 23:1–23:11, New York, NY, USA, 2018. ACM.

[127] Ari Pirkola, Jarmo Toivonen, Heikki Keskustalo, Kari Visala, and Kalervo Järvelin. Fuzzy translation of cross-lingual spelling variants. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 345–352. ACM, 2003.

[128] Bruno Pouliquen, Ralf Steinberger, Camelia Ignat, Irina Temnikova, and Anna Widiger. Multilingual person name recognition and transliteration. *Corela. Cognition, représentation, langage*, 2005.

[129] Satish K Puri, Sundeep Sahay, and John Lewis. Building participatory HIS networks: A case study from Kerala, India. *Information and Organization*, 19(2):63–83, 2009.

[130] Supply Chain Quarterly. Supply Chain Management Newsletter, 2013.

[131] Erhard Rahm and Hong Hai Do. Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull.*, 23(4):3–13, 2000.

[132] Soatiana Rajatonirina, Jean-Michel Heraud, Laurence Randrianasolo, Arnaud Orelle, Norosoa Harline Razanajatovo, Yolande Nirina Raoelina, Lisette Ravolomanana, Fanjasoa Rakotomanana, Robinson Ramanjato, Armand Eugène Randrianarivo-Solofoniaina, et al. Short message service sentinel surveillance of influenza-like illness in Madagascar, 2008-2012. *Bulletin of the World Health Organization*, 90:385–389, 2012.

[133] Vijayshankar Raman and Joseph M Hellerstein. Potter's wheel: An interactive data cleaning system. In *VLDB*, volume 1, pages 381–390, 2001.

[134] Bradley Reaves, Logan Blue, Dave Tian, Patrick Traynor, and Kevin RB Butler. Detecting SMS spam in the age of legitimate bulk messaging. In *Proceedings of the 9th ACM Conference on Security & Privacy in Wireless and Mobile Networks*, pages 165–170. ACM, 2016.

[135] Bradley Reaves, Nolen Scaife, Adam Bates, Patrick Traynor, and Kevin R. B. Butler. Mo(bile) Money, Mo(bile) Problems: Analysis of Branchless Banking Applications in the Developing World. In *USENIX Security*, pages 17–32, 2015.

[136] Bradley Reaves, Nolen Scaife, Dave Tian, Logan Blue, Patrick Traynor, and Kevin RB Butler. Sending out an SMS: Characterizing the Security of the SMS Ecosystem with Public Gateways. In *2016 IEEE Symposium on Security and Privacy (SP)*, pages 339–356. IEEE, 2016.

[137] Johan Rewilak. The role of financial development in poverty reduction. *Review of Development Finance*, 7(2):169–176, December 2017.

[138] R Russell and M Odell. Soundex. *US Patent*, 1, 1918.

[139] Johan Ivar Sæbø, Edem Kwame Kossi, Ola Hodne Titlestad, Romain Rolland Tohouri, and Jørn Braa. Comparing strategies to integrate health information systems following a data warehouse approach in four countries. *Information Technology for Development*, 17(1):42–60, 2011.

[140] Sundeep Sahay and John Lewis. Strengthening metis around routine health information systems in developing countries. *Information Technologies & International Development*, 6(3):pp–67, 2010.

[141] Sundeep Sahay, Eric Monteiro, and Margunn Aanestad. Configurable politics and asymmetric integration: Health e-infrastructures in India. *Journal of the Association for Information Systems*, 10(5):4, 2009.

[142] Sundeep Sahay and Geoff Walsham. Scaling of health information systems in India: Challenges and approaches. *Information Technology for development*, 12(3):185–200, 2006.

[143] Terje Aksel Sanner and Johan Ivar Sæbø. Paying per diems for ICT4D project participation: a sustainability challenge. *Information Technologies & International Development*, 10(2):pp–33, 2014.

[144] Scribd. Cold Chain Equipment Inventory Data Model, 2013.

[145] Christopher J Seebregts, Burke W Mamlin, Paul G Biondich, Hamish SF Fraser, Benjamin A Wolfe, Darius Jazayeri, Christian Allen, Justin Miranda, Elaine Baker, Nicholas Musinguzi, et al. The OpenMRS implementers network. *International journal of medical informatics*, 78(11):711–720, 2009.

[146] Kazuhiro Seki, Hiroyuki Hattori, and Kuniaki Uehara. Generating diverse katakana variants based on phonemic mapping. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 793–794. ACM, 2008.

[147] Muniba Shaikh, Nasrullah Memon, and Uffe Kock Wiil. Extended approximate string matching algorithms to detect name aliases. In *Intelligence and Security Informatics (ISI), 2011 IEEE International Conference on*, pages 216–219. IEEE, 2011.

[148] Steven Shapiro, Barnaby Richards, Michael Rinow, and Timothy Schoechle. Hybrid standards setting solutions for today's convergent telecommunications market. In *SIIT*, 2001.

[149] Alex Shovlin, Mike Ghen, Peter Simpson, and Khanjan Mehta. Challenges facing medical data digitization in low-resource contexts. *2013 IEEE Global Humanitarian Technology Conference (GHTC)*, pages 365–371, 2013.

[150] Pritpal Singh, Ruth McDermott-Levy, Elizabeth Keech, Bette Mariani, James Klingler, and Maria Virginia Moncada. Challenges and successes in making health care more accessible to rural communities in Waslala, Nicaragua using low-cost telecommunications. *2013 IEEE Global Humanitarian Technology Conference (GHTC)*, pages 309–313, 2011.

[151] RW Snow, D Zurovac, G Jagoe, J Barrington, S Githinji, S Kigen, AM Mbiti, D Memusi, AN Muturi, A Nyandigisi, et al. Reducing Stock-Outs of Life Saving Malaria Commodities Using Mobile Phone Text-Messaging: SMS for Life Study in Kenya. 2013.

[152] Frank Stajano and Paul Wilson. Understanding scam victims: seven principles for systems security. Technical report, University of Cambridge, Computer Laboratory, 2009.

[153] T Svoronos, P Mjungu, R Dhadialla, R Luk, C Zue, J Jackson, and N Lesh. CommCare: Automated quality improvement to strengthen community-based health. *Weston: D-Tree International*, 2010.

[154] Robert L Taft. *Name search techniques*. Bureau of Systems Development, New York State Identification and Intelligence System, 1970.

[155] Pakistan Today. Fraudsters continue scamming unsuspecting citizens in BISP's name, 2018.

[156] Aditya Vashistha, Richard Anderson, and Shrirang Mare. Examining Security and Privacy Research in Developing Regions. In *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies*, COMPASS '18, pages 25:1–25:14, New York, NY, USA, 2018. ACM.

[157] VillageReach. The Framework for OpenLMIS, 2012.

[158] Paola Virga and Sanjeev Khudanpur. Transliteration of proper names in cross-lingual information retrieval. In *Proceedings of the ACL 2003 workshop on Multilingual and mixed-language named entity recognition-Volume 15*, pages 57–64. Association for Computational Linguistics, 2003.

[159] Lu Wei-Chih, Matt Tierney, Jay Chen, Faiz Kazi, Alfredo Hubard, Jesus Garcia Pasquel, Lakshminarayanan Subramanian, and Bharat Rao. UjU: SMS-based applications made easy. In *Proceedings of the First ACM Symposium on Computing for Development*, page 16. ACM, 2010.

[160] Hadley Wickham et al. Tidy data. *Journal of Statistical Software*, 59(10):1–23, 2014.

[161] Wikipedia. Lottery scam - Wikipedia, 2019.

[162] William E Winkler. String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. *Proceedings of the Section on Survey Research Methods*, 1990.

[163] Beyond Wireless. Cold chain temperature monitoring, 2019.

[164] Theo Wirkas, Steven Toikilik, Nan Miller, Chris Morgan, and C John Clements. A vaccine cold chain freezing study in PNG highlights technology needs for hot climate countries. *Vaccine*, 25(4):691–697, 2007.

[165] Qian Xu, Evan Wei Xiang, Qiang Yang, Jiachun Du, and Jieping Zhong. SMS spam detection using noncontent features. *IEEE Intelligent Systems*, 27(6):44–51, 2012.

[166] Ahmed H Yousef. Cross-language personal name mapping. *arXiv preprint arXiv:1405.6293*, 2014.

[167] Michel Zaffran, Jos Vandelaer, Debra Kristensen, Bjørn Melgaard, Prashant Yadav, KO Antwi-Agyei, and Heidi Lasher. The imperative for stronger vaccine supply and logistics systems. *Vaccine*, 31:B73–B80, 2013.