Examining the Challenges in Development Data Pipeline

Fahad Pervaiz
University of Washington
Seattle, Washington
fahadp@cs.washington.edu

Aditya Vashistha
University of Washington
Seattle, Washington
adityav@cs.washington.edu

Richard Anderson
University of Washington
Seattle, Washington
anderson@cs.washington.edu

ABSTRACT

The developing world has increasingly relied 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 their own efforts for frequent information gathering. Overall, the amount of data collected is tremendous, yet there are significant issues in doing useful analysis. Most of these barriers manifest in data cleaning and merging, and require a data engineer to support some parts of the analysis. In this paper, we investigate the challenges of cleaning development data through an interview based study. We conducted face to face interviews of 13 stakeholders, eight from international development organizations and five government workers from Pakistan, including both managers and data analysts. From analysis of the interviews we identified common challenges faced in processing development data including correcting open text fields, merging hierarchical data, and extracting data from textual formats such as PDF. We construct a basic taxonomy of data cleaning challenges, and identify areas where support tools can improve the process. Ultimately, the objective is to empower regular data users to easily do the necessary data cleaning and scrubbing for analysis.

CCS CONCEPTS

• **Information systems → Data cleaning**.

KEYWORDS

HCI4D, ICTD, data cleaning, data analysis, data collection, global development

ACM Reference Format:
Fahad Pervaiz, Aditya Vashistha, and Richard Anderson. 2019. Examining the Challenges in Development Data Pipeline. In *ACM SIGCAS Conference on Computing and Sustainable Societies (COMPASS) (COMPASS '19), July 3–5, 2019, Accra, Ghana.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3314344.3332496

1 INTRODUCTION

Global development and policies for health, education, and governance in developing countries increasingly rely on data analytics. Significant amounts of data is collected to evaluate and support the decision making for this development. This area is complex involving a diverse set of organizations including government agencies, non-governmental organizations (NGOs), donors that are funding development work and third-party companies.

Supporting data for development is complex and fragmented with different organizations interested in different parts of data collection that supports their area of focus like maternal care or immunization. This leads government agencies to collect a wide range of indicators that gives them an overall status while NGOs have systems in place that gather specific variables to enable more in-depth analysis on a topic.

Currently, various datasets are managed and maintained by several organizations in disconnected systems. The coverage of these datasets overlaps to a certain degree and serves a specific purpose or analysis. The utility of these datasets is enhanced many fold once datasets from different systems are merged. This is complex to achieve and requires a substantial amount of effort spent on data cleaning and establishing data accuracy.

Data cleaning in this space is very distributed due to multiple players participating in it. Different collaborators will work on different stages of this process, some in collaboration and others in isolation. The data goes through multiple stages of the processing pipeline, which is explained in detail later in the paper, allowing the process to be segregated easily. This isolation also means that people doing data cleaning might have no control over collection and may not understand the context in which the data was collected.

ICTD literature demonstrates copious amounts of research on the data collection stage of the pipeline, yet it neglects analysis of the processing that occurs after collection. Building a taxonomy for data cleaning will help identify the biggest gaps. Although some researchers have provided a dirty data taxonomy with respect to developed world data [22] where the challenges are around schema, reverse engineering or lack of constraints on data types [34], to our knowledge, there is no work on data cleaning taxonomy for developing regions.

This paper explores the existing data collection and cleaning processes in development data. In the last decade, incredible improvements have been made on the data collection side that help the processing pipeline. In reality, the collection and processing are disconnected stages in terms of people involved. This causes some information about the data to be lost during this transition, and people doing the cleaning have no control over collection. Insights on how the data goes through various processing stages for analysis will help us build better tools for data cleaning and achieve better analytics.

We attempt to understand the data cleaning processes for development through interviews with stakeholders from different types

COMPASS '19, July 3–5, 2019, Accra, Ghana

ACM ISBN 978-1-4503-6714-1/19/07…$15.00
https://doi.org/10.1145/3314344.3332496

of organizations. We interviewed 8 people from international development organizations, and 5 people who worked for the government in Pakistan. This allowed us to get both a global perspective, as the the people from development organizations had worked with projects from well over 100 different countries, and well as a more in depth perspective from a specific country. The goal is to compile the outstanding issues expressed by the practitioners and point out gaps that have the biggest impacts. This allows us to build a basic taxonomy and identify areas where support tools can help achieve better cleaning of development data.

1.1 Why data cleaning is different for development data

Development data has a set of characteristics which arise from practices in collection, limitations in infrastructure, and organizational settings. This leads to messy datasets with varying degrees of integrity, which contrasts with consistently structured datasets from conventional information systems [1]. The data cleaning challenges in the development domain can therefore be much more time consuming to resolve.

Data-driven decision making for international development is becoming more common. Unfortunately, many settings 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. This last mile problem is especially challenging when limited training has been provided to the people responsible for data collection, curation, and analysis. Although other constraints like power and access to computing devices are manageable, ensuring quality training is a critical and challenging task [30]. This is especially important if data is flowing up the hierarchical chain instead of directly being reported at the national level. This increases the chances of errors appearing in the data because of it getting mistyped or wrongly aggregated at any level.

Institutions, including governments, play a critical role in data collection and processing. Semi-functional governments lead to uncoordinated attempts to digitize data. Moreover, there are usually multiple organizations that are pushing for data compilation in a given region. Every stakeholder has their own agenda and priorities. This causes tension and pressure on certain aspects of data collection while other parts are ignored.

Development data poses more cleaning challenges because typically the data is entered on paper forms which clerical staff later digitize [26]. 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, for example, having values as part of the field names [38]. This is in a 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 constant change of requirements for collection. Over time, more and more indicators are added to the collection process with the addition of requirements from newer partners. This causes a huge burden for the lower level staff to fill more forms. For example, our discussions with a manager of a basic health facility in Punjab, Pakistan, revealed that the staff must fill out seven different forms, with redundant information, at weekly, fortnightly, and monthly frequency. This presents a major obstacle for staff to complete all the forms accurately or keep up with the latest guidelines to record additional indicators.

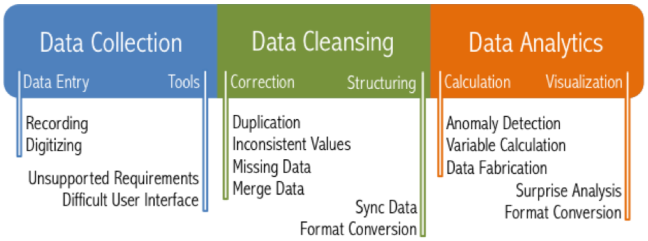


Figure 1: Various stages that play a role in data cleaning along with example lists of challenges for each stage.

All these aspects lead to inconsistent, poorly recorded, incorrectly formatted, missing, and fabricated data. Additionally, we have seen reorganization in developing countries that changes the administrative hierarchy of the region. Since all the data collection is tracked and identified through this hierarchy, this makes it harder to do temporal analysis. The changing nature of data collection environment makes developing world data very different from rest of the world.

1.2 Data Cleaning Pipeline

In this section, we describe different stages as shown in Figure 1 that play a key role in the data cleaning pipeline. We divide these stages into three main categories: data collection, data cleansing, and data analytics. These categories also represent the pipeline that data goes through. It is not necessary that data goes through all the stages or in a specific order.

Data collection provides an early opportunity to identify errors and collect clean data —but is a major entry point of mistakes and errors into the pipeline. Due to lack of resources and poor infrastructure in developing regions, large scale data collection typically happens over paper forms. For example, at rural health facilities in Pakistan all attendance records are kept in large paper based log books. Field workers complete these forms manually and send them to a local data collection center. It is then manually transcribed by data entry workers into a computer, often with poor accuracy [12]. Several researchers have designed solutions to address this gap by integrating the paper collection process with the digital collection process. It ranges from using a smartphone camera to digitally capture a paper form [11, 32] to using cloud and crowd for processing forms [8, 25].

Several mobile-based tools are designed to cater to the needs of remote digital data collection in low resource settings. Open Data Kit (ODK) [21] and CommCare [37] are mobile data entry tools that can be configured with any form to collect data. DHIS2 [21] and OpenMRS [36] are customizable information systems with dashboards and reporting tools that also have mobile data collection components. Even with progress in mobile tools, digital data collection still faces challenges [7, 31]. These tools can introduce other issues in the data, such as duplicate entries or the generation of multiple versions of databases due to some unsupported requirement. ODK 1.0, for example, currently does not allow editing an an existing form. If one wants to add a new field to an existing form, then one must create a new form altogether. This action creates a new database that will require merging with the original form's database to get one single consistent dataset.

Data cleaning is the critical and most time-consuming component of data processing. It is well known that 80% of data processing time is spent on cleaning and aligning data [10]. This ranges from removing errors, duplicate values, and inconsistency to merging the data from multiple sources. Several researchers have studied this area, laying out the scope of the problems and current approaches with respect to data warehousing [17, 22, 34]. 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 more real datasets from resource-constrained settings in the developing world where even the structure of datasets is very messy and unorganized [38].

Several tools have been built to simplify the cleaning process. These tools use pattern detection, from user's interaction for data transformation, and heuristics to infer the intended structure and data domain [20, 35]. Wrangler and Potter's wheel are interactive tools that show users the transformations as they progress. Their goal is to engage regular users into the cleaning process without making it too complicated for them [18].

The presence of several tools, which are intended for different stages of the pipeline, poses structural and formatting issues in the space of data processing [19]. Information systems like DHIS2 and OpenMRS solve some of these problems by providing more complete solutions, but 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 so that it is easy to import data for further analysis [27, 38].

There have been a range of systems that aid data exploration through a visual interface [16, 20, 35]. These tools support what is called the sensemaking model [4]. This helps the lay user to better understand the data for cleaning, integration and other calculations [5]. In our own experiences, government staff in Nigeria were using the interactive visualization in DHIS2 to find the anomalies in the data and follow up on it to remove any errors.

2 RELATED WORK

In this section, we present stages of data cleaning with reference to previous work done in respective areas.

2.1 Data Correction

There are several aspects of data quality that need to be fixed before any analysis. These include dirty data, duplicated data, and the merging of different data sources. Dirty data means missing values, wrong or inconsistent values and unexpected formats [22]. Duplication is an issue in which the same entity is present in different parts of the data and could be represented in various formats [9]. Merging datasets is a challenge when there is no standard common identifier.

Data errors can creep in at any step of data processing including data entry, measurement, processing, or integration [17]. Fixing wrong or inconsistent values requires domain knowledge, making it hard to automate the process unless the tool is very domain specific. Rahm et al. [34] have proposed standard ways to handle dirty data with limitations on conflict resolution. Modern cleaning tools like Wrangler [20] and Google Refine [24] include the user in the cleaning process and rely on learning from example and display only a snapshot of the transformations to the user for verification.

Duplicate data is a common problem especially when integrating multiple datasets. Duplicate data can also arise if a system does not support updating records. For example, Open Data Kit (ODK) treats each filled form as a single record and does not provide support to update previous entries. This allows multiple entries for the same entity to be created. Several works have approached this problem, mostly by defining a similarity function to find clusters of matching entities [9, 15, 23, 29]. It is a difficult problem since an entity might have different representations in different datasets, as a name might be "John Doe" in one dataset, and "Doe, John" in another.

The lack of standard identifiers makes it harder to merge datasets from different sources. The foremost strategy is to use a common attribute from both datasets. Usually this is the name of a location or person. This approach is challenging since the probability of spelling mistakes is high and local names transliterated into Roman characters can result in multiple spellings [13]. A second strategy, known as probabilistic record linkage, takes a wider range of possible identifiers and calculates a probability of the two records being same [2, 28].

2.2 Data Collection and Structuring

Several tools have been proposed to reduce error rates at the collection stage. Shreddr [8] is a tool that takes an image of a paper form and use the crowd to digitize it. Errors in the transcription step are reduced by having multiple people perform the task. Usher [6] targets data collection done on a mobile device in the field. Instead of having the user enter the same field every time, it updates the form and gives the user options from previous entries. This reduces the chance of making a spelling mistake because the user can select rather than enter the same word repeatedly.

The data processing pipeline consists of various tools that expect data to be in a certain format and structure. Hence the organization of data is very critical for the pipeline to operate seamlessly and is a common problem in development data. There are several standard formats that are proposed for different purposes [27, 33]. Most tools in the development domain rely on the comma-separated values format with each column representing an indicator and the first row as the name for the indicators [38]. This problem is addressed by industry through the additional support for new formats in standard tools. However, the support is generally added as the need grows, causing periods when support does not exist. Organizations handle this by writing custom scripts for data conversion.

Table 1: Summary of Participants by Organization Types

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

3 METHODOLOGY

We base our findings on face-to-face interviews that we conducted. The participants of these interviews belonged to multiple departments of the Punjab provincial government of Pakistan and three international non-governmental organizations (NGOs). These NGOs are all head-quartered in the United States. Two of these NGOs are large international global health organizations that have worked in well over 100 countries on a multitude of projects. The third is a global health research institute that specializes in building global datasets.

Some participants, as illustrated in Table 1, are managers in some capacity for various projects while others are individuals working on specific data collection projects. This allowed us to get a well rounded perspective as certain details like issues around partnerships are dealt by managers while basic cleaning tasks are handled by individuals on the project. The interviewees were selected based on connections we have established over the past years while working with them in different capacities. Interviews in Pakistan were conducted in Urdu, the common local language. The first author of this paper is from that region and fluent in Urdu. The interview questions and analysis was informed by the authors collective experience working with large development datasets.

3.1 Interviews

All the interviews were semi-structured and were conducted in person. Each interview took about 40 minutes. Prior to the interviews, we asked participants from development organizations permission to audio record their interviews. Among the eight participants, only one declined our request and was not recorded. No attempt was made to record the interviews with participants who were Punjab, Pakistan government employees. This was done to alleviate their fears of saying something that higher officials might not like while being recorded. We considered recording as an obstacle to getting their honest feedback and did not want it to be a potential concern. We generated thorough field notes during the interviews when voice recording was not used.

The interview questions focused on the the regular processes they followed for data processing and any frustrations they had during this process. We had a set of questions that we asked and followed up on particular topics that came up. The questions targeted various scenarios that may cause annoyance with the data. That way we were able to uncover difficulties that the participant might consider as part of their regular job. The interview instrument and process was reviewed and approved by the institutional review board at our university.

3.2 Analysis

We used an iterative approach to generating interview questions and qualitative analysis. After several interviews we discussed results and updated questions based on current findings. We transcribed all the interviews for analysis along with the detailed field notes. To process the data, we did thematic analysis [3]. We had regular discussions over the themes coming from the analysis to do further iterations and make sure we were thorough. We then used affinity diagrams [14] to identify themes of challenges from the interviews.

4 FINDINGS

Several specific challenges were identified from the interviews. In Figure 2, we have grouped the individual issues into the three categories that were discussed earlier in the paper. Issues that were mentioned by three or more participants are in bold. The rest of the section describes these issues.

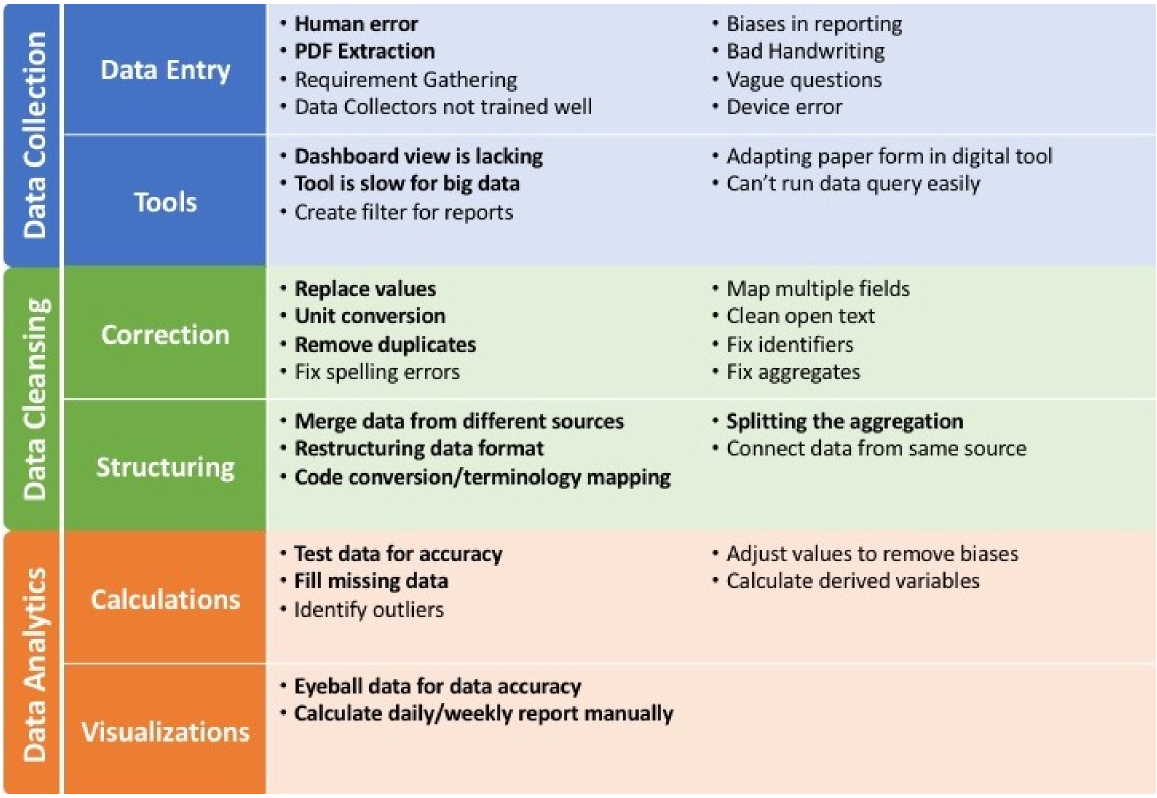
4.1 Data Collection

Data collection is generally the most time-consuming stage. The process of data collection has shifted in the last decade from digitizing paper forms to using direct entry with digital tools. We split the data collection issues into two categories, one related to the data gathering process and the other being issues with tools.

The most common frustration that participants brought up was human error in collecting data from the field. Human errors are due to multiple reasons, ranging from spelling mistakes to entering incorrect data. A non-government participant mentioned: *"We have signs to pick ID to each facility or segment or household and it always happen that they don't pay attention that they input the wrong thing then it's a challenge to fix it and we look at the start time of the survey and then try to match up the linked surveys or sending email to supervisor there asking what's going on so a lot of human error is our challenge."*

Extracting past data from PDF reports was another common source of frustration during the data cleaning stage. This extraction is required for many different reasons. It is common for external organizations to not have access to the raw structured data, but only get prepared reports – even recently created ones. While others are pulling legacy data from reports to build longer time series data for the region. An NGO participant said: *"I think the most extraordinarily challenging dataset was our India vital registration from a particular series of reports that were all PDF reports. . . Even just the extraction was terribly difficult because they are the report per state of India per year that was very long and in a slightly different format sometimes and had page breaks so to format you get an army of people to do PDF extractions and those might not all be right."*

Several other issues related to collection were brought up by fewer participants. This included unclear handwriting for the projects that are still using paper-based forms. One participant complained about the questions being vague for locals, which leads to inconsistent answers by various collectors. The collectors themselves are often not well trained due to limited time and resources for training. An NGO interviewee said: *"A challenge from my perspective will be translating paper based data into an electronic tool. The number one problem is training of the data collectors by technical people who*



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 Cleansing	Correction	• **Replace values** • **Unit conversion** • **Remove duplicates** • Fix spelling errors	• Map multiple fields • Clean open text • Fix identifiers • Fix 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 2: Summary of specific issues grouped into categories. Issues in bold text were mentioned by three or more participants.

have been well trained or already know. The data collector that are bothered to collect the serial numbers has a huge implication on the data cleaning. People's handwriting is a huge factor... This all could be easy if we had more data available and data collector are well trained to ask the right questions in the right way."

One participant mentioned how sometimes GPS devices are erroneous, which results in incorrect entries. Moreover, two people talked about how the respondents of their surveys are biased for personal reasons, such as when reporting their weight or have incentive to lie to get more funding for their center. As one participant from an NGO says *"Numbers can be inflated for political reasons or deflated."*

Digital tools are well integrated in most current collection practices. The major frustration with these tools has to do with the data entry aspects and not with the data cleaning operations. Three participants explained that the tool they were using slows down when they process large datasets. One government participant mentioned that: *"Field Note: The participant explained that data is too big so he cannot use excel because it takes a long time to load and is slow to use. He uses R script every time to remove duplicates from the data."* Many also expressed that the dashboard they were using, which was built for their application, was lacking and does not provide easy access to filters or statistics. Hence, they have to export the data to excel to create the intended reports.

A few people brought up other issues regarding tools. One explained that adapting a paper-based form exactly to digital form is challenging due to the different manner of data entry. The other issue is running specific queries conveniently as an NGO participant explains: *"I can't run these queries easily so I have to export this into excel but everything else related to the facility goes away so I have to work with MS Access [tool]. My approach to data cleaning is limited by my abilities in the tool. I can sort refrigerators by age and pull it out [in excel] and look at it in the form that makes sense."*

4.2 Data Cleaning

Data correction is part of the cleaning process that focuses on fixing basic issues in the data. Common tasks that participants mentioned are replacing values, unit conversion, and removing duplicates. Government participants talked about manually following up on human errors and replacing values with real ones while others talked about replacing missing data with *not-a-number* values that are better supported by data analysis tools. For example, one NGO participant stated that: *"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 government interviewees touched upon the issue of removing duplicates from the incoming data. Moreover, unit conversion was commonly discussed as one NGO worker stated: *"Oh and sometimes we get data in rate based, so the numbers are actually in rate and you have to know that from the documentation that they're not numbers. You have to* *multiply that rate by a population to get numbers so like some form of metric conversion is really common."*

Correction issues discussed by fewer participants includes fixing the identifier that appears due to human error while filling forms. The more difficult issue is to clean and extract values out of open text form fields. While open text brings more flexibility in the data entry process, it ends up being a hectic cleaning process: *"They could be recording things in milliliters or different units and different short-hands 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 it is an overly big pain. It's doable but it's one of our biggest problem in general especially in terms of any sort of human error that we are collecting."* Open but limited fields even with predetermined codes or statuses result in misspelling that wa mentioned by two people. One of them said: *"The biggest one [problem] is when we have excel based data source that didn't come from a system that was enforcing data integrity. . . One of them was related to breastfeeding and in this dataset, which was big but not humongous, they spelled breastfeeding 47 different ways."* Lastly, one participant explained that they had to map values from multiple fields into one to compute the value for the intended field.

The second category for cleaning is about data structuring. The biggest and most discussed challenge among all categories was merging data from different sources. The merge includes syncing data from various current systems as well as connecting it to past data for analysis. This is challenging due to names not matching for locations, referred as the name resolution problem, changes in geographic boundaries for a unit area, and errors in identifiers. As one NGO participant described the situation: *"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 the admin unit has even been spelled differently."*

Even if the data merges at a higher level, there is still the common problem of mapping various codes and terminologies to standard ones. This sometimes requires manual follow up and metadata lookup outside the dataset to determine a correct mapping. Along with that, standardizing the data format itself is a regular issue. As it is mentioned by an NGO participant: *"Data is hierarchical and hierarchies are different. We have to read the documentation to understand their division and codes used in the data. At some level, we use string matching to match the datasets."*

There are other structuring issues that were infrequently discussed. One is splitting the aggregate to get more specific local numbers when the data was summed up at a regional level, like district or provincial level. The strategy mostly used was to split it proportionally to fill in the missing data. Another issue mentioned was connecting data from different parts of the form that belong to the same survey. This is largely due to the digital survey applications like ODK or surveyCTO that treat each form as a separate dataset, making it cumbersome to connect the dataset all together once the survey is done.

4.3 Data Analytics

Calculation is a category of data analytics where some processing is done to achieve cleaned data. The most discussed strategy was to test data for accuracy. This is mostly done by using other datasets to calculate the approximate range of values to verify the data. The verification is based on either a simple distribution or a complex model. One NGO participant said: *"So [for verification] we are relying on using other data that we have, either from the same country or same region or same age, sex group and time period. So we can see if the numbers fall under the same realm of plausibility. So we are checking if something looks really crazy and double checking only that."* A similar strategy is used to fill the missing data, which is another challenge that is frequently discussed: *"We had all sorts of other problems because entire state-years were missing in random ways so there is a sparse matrix of state-years available but we had a full matrix of national level data by year so we tried to conform state level to the national level and fill in those missing cells in the matrix using a model to do that. It was a very complicated model."*

Less mentioned frustrations were to calculate derived variables like calculating age from date of birth. Some mentioned that they calculate the outliers in the data to clean any obvious errors. While some go as far as evaluating the biases in the data that could be under or over reported and adjust the values to fix it. These biases range from reporting more patients to get more funding for vaccines, self-reporting less weight to avoid societal shaming, or, in general, having different reporting practices for cause of death on death certificate. As one participant explained: *"There are analytic challenges in how we adjust the self reported data [on weight], quantify the bias and remove it. And we have leverage measure data that we have from similar countries hopefully in the same time period and if we have measured data source and a self-report dataset in the same country, so we match those up controlling for change across time and see what is the difference and often times the self-reported is under measured."* Another NGO participant mentions: *"A big issue is also the ICD code, the way people map diseases. . . If something is wrong because they coded in different ways. So, we try it adjust for those biases."*

The final category for analytics is visualization that is used as visual aid for cleaning data. Several participants referred to using graphs, either on dashboard or self-generated, to eyeball the data for any anomalies. As one participant said: *"And then once we have the data and cleared it, specifically I work with spatial data, I tend to do a visualization first so look at and make a sanity check. If it passes the sanity check,. . . because if there are points out in the middle of the ocean for health facility then there is something wrong with the data"*. Another interviewee explained that: *"We plot the data on a time series to look for outliers and mark that this looks weird and this looks weird and then verify it."* This is a quick manual way to verify data accuracy though one participant talked about how graphs help to view the results from verification models in order to spot errors in the data quickly. Another frustration that government interviewees discussed was that they must send data forward in the form of visualizations. They feel annoyed that they create the same report every week or month, requiring countless man hours, and lament that this process could be automated.

5 DISCUSSION

Several cleaning challenges arise once development data has been collected. As we have seen in our findings, some are trivial, while others require sophisticated solutions like building machine learning prediction models. These solutions demand varying degrees of person hours, which in some cases could be avoided through automation, yet in others cases remain inescapable. Based on detailed comments from our interviews, we present our synthesis in this section.

In our opinion, these are the three key areas that have emerged as the source of major frustration for development data. This is based on what participants point out as most time consuming or challenging part of data cleaning for them.

- Merging data between existing large data sources.
- Validating data accuracy.
- Extracting data from PDFs reports.

Merging data is the most frustrating process due to several factors. The most common reason is the name resolution problem, where the names of locations do not match up in datasets of the same region. This is often because those names are from a local language that has been transliterated to either English or French. Moreover, different people will transliterate differently, resulting in various spellings for the same location. This could be avoided by building a master location list as a supporting dataset that systems can use. However, even with some push for this, it continues to be an issue. The best way to solve this is by building NLP based algorithms that can accurately match names with all sorts of spellings. This could further be used for matching patient names, which is also a problem, especially for migrating populations with no other identifiers.

The second reason for merging issues is the mapping of codes and terminologies within datasets that differ. This is mainly due to lack of standardization in this space that continues to be problematic. For instance, one participant explained that different codes are used to denote the cause of death in different countries. Another participant expressed concerns about the use of different numeric codes for equipment functionality statuses in various tracking systems. This is not just limited to terminology but also the units being used or the format in which the dates is written. This increases the complexity of merging the datasets and requires a lot of hours to untangle the differences. A solution for this is to use a machine learning model that can predict the unit or variable based on the distribution in the dataset. Indeed, this is not simple due to changing reporting norms that alter the distribution over time. The more deterministic way to do this is to look up this information via a side channel like reading through the documentation or directly asking the source partner if that is possible.

Data validation is a major concern in the development space, and is generally the focus once a digital data collection system is in place. This was an especially hectic process for the Punjab government participants who have to go through the data and manually follow up on all the errors that they find. They recounted how they eyeball the data using the tabular form as well as visualizations. Then they have to call the source facility to validate any suspicious entries. This is a tedious process for them, and could be made easier with validation algorithms based on the statistics of past entries to highlight the data that needs attention. In this way, the manual follow-up could still be part of the system, while at the same time reduce the cognitive load of finding the errors.

Some international development organizations handle this problem differently. They build a sophisticated model that tries to fit the data to find any values that are outside the norms for a given facility or data point. One organization described how the data was extracted by hand from PDFs, and, even with trained personal, they would discover errors. They would then build models to predict the values, so they could go back and verify if an error was made during extraction. Sometimes the error was made in the original PDF report itself, which is harder to fix given the lack of raw data. In turn, they rely on predicted values.

Extracting data from PDF files is important in development analytics due to the large amount of historical and even recent data available only in the form of PDF reports. This is challenging not only because extracting numbers is difficult, but also because the numbers are aggregated in different ways with the details tangled in the text of the reports. Even if you write a tailored script for a specific set of reports, putting the numbers in the right perspective deterministically is tough and results in countless person hours spent to verify the extraction.

Even though there are several other commonly mentioned issues, these three areas account for the most frustration due to lack of good supporting tools. This results in significant amount of manual labor to extract, merge and validate data in a complete work flow. There is a need for more tailored solutions addressing these problems that can help users build a complex model for automatic processing, thereby significantly reducing the burden of data cleaning. On the other hand, common data correction problems such as replacing values, unit conversion and removing duplicates are well researched and can be handled by tool like Wrangler [20] and Google Refine [24]. These tools make it easy by letting user show the intended task through an example conversion.

Over time, data collection processes and the structure of data have evolved. As a result, legacy data issues around merging or generating time series trends exist. In part this is due to the fact that the indicators being collected have changed either because there is a lack of data standards or the standard was ignored in the collection process. Moreover, data entry norms are changing. For instance, diseases that were not being reported are now being reported more commonly. One participant from a non-government organization explained this situation as follows: *"But there are some things that are very difficult to be backward compatible because it's not just the code you used to report a disease changed but the entire reporting culture around a disease is evolving"*. Lastly, boundaries of districts and sub-districts are changed over time, making it impossible to do one to one comparison of past data with current one. This requires splitting the aggregate or predicting the numbers based on comparable datasets.

Partnership is the keystone for development where international development organizations work together with local governments as well as non-government organizations. This is especially helpful to establish a local context for the project as explained by one NGO participant: *"For example, one country changed from an oral polio vaccine to an intravenous polio vaccine and they updated their cards and it just became a problem because we were not aware . . . and that's what is great about having this working relationship with the Ministry of Health, but if you don't have that and you don't know how to adapt your survey, then you will be missing a lot of your data. I think context* *is the most valuable thing you could ever have when you are working on a project like this."*

While partnerships are very important, they are also the source of some frustrations around data collection and processing. There are several datasets from different partners with varying credibility that complicates data processing. Obtaining these datasets presents a challenge to different stakeholders, who complain about lack of easy access and limited permissions. Additionally, this is complicated with misaligned objectives between partners. One NGO participant said: *"Part of the problem is that this [data cleaning] is all very fragmented [among partners]. It's all very chaotic and you are just trying to do it and I think it's sort of sad situation. There is not that network of all of the things that needs to go in [data processing]."* This struggle results with some partners not working well with each other, unwilling to use the tools created by others and sometimes nobody taking responsibility to steward the data once the project is over.

Furthermore, there are several key differences for challenges faced by international non-governmental organizations as compared to a local government. A government has more control over the incoming data and has easy access to the data entry source. This results in their ability to manually follow up to fix the errors and easy availability of any metadata about the data. This convenience is commonly not available to an NGO. This is the case, even though NGOs often have better resources to build complex verification processes while governments rely on simpler tools to achieve the same thing. Moreover, a low level cleaning task like removing duplicates and replacing values is more likely to be handled by local workers during the data entry stage. Governments also face difficulties in explaining complex analysis due to analytics illiteracy.

6 CONCLUSION AND FUTURE WORK

The development data area is complex with changing data structures, multiple organizations, and disconnected efforts to process data. There is a push to standardize the tools in order to streamline processing. On the contrary, our interviews demonstrate how different projects have different needs, and one generic tool cannot solve everything. As this area evolves, there is need for tools for specific problems. Care needs to be taken to allow these tools to be deployed in tailored combinations based on the demands of projects.

As more tools saturate this area, data challenges have also evolved. Partly this is due to new tools introducing data models that complicate merging and analysis. This is especially problematic when dealing with multiple datasets from different systems or time periods. One of our participants expressed that: *"If I am just working with one dataset, there is rarely a challenge that I see is too big."*. However, sometimes these tools introduce new issues as exemplified in this participant's statement: *". . . , they put that all in one row. So you get these datasets that has more than 1000 columns . . . When I looked at it first, I thought this is really a bad practice, now what I have seen that it's the easiest way for them to get data entered or that a tool like Redcap supports it, so now I am likely okay, that just means I have more work in the data preparation part of it."* So new tools that simplify one problem might introduce further issues at later stages of processing pipeline.

This paper is a preliminary effort to compile the complex issues with development data. There are several limitations to this work that need to be further expanded. We were restricted in our participant selection because we used only our own contacts and interviewed everyone face to face. We speculate that less tech-savvy organizations will have different or additional issues. The Punjab government has adapted technologically friendly policy in the past ten years, so has achieved a higher degree of adoption of digital tools. There are other provincial and country level governments that are very new to this digital world and they might be facing different challenges. The same is true for NGOs who with vary degrees of international experience and exposure might have more or less experience in data handling and processing. Hence, it is important that local organizations are also looped into this discussion for building a detailed taxonomy of data cleaning.

7 ACKNOWLEDGEMENT

We would like to acknowledge all the participants who volunteered time to talk with us and discuses data issues. We would also like to thank the partners who helped us recruit participants for our interviews. Lastly, we would like to thank the members of our research team who helped review and edit this paper.

REFERENCES

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

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

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

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

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

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

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

[8] Kuang Chen, Akshay Kannan, Yoriyasu Yano, Joseph M Hellerstein, and Tapan S Parikh. 2012. Shreddr: pipelined paper digitization for low-resource organizations. In *Proceedings of the 2nd ACM Symposium on Computing for Development.* ACM, 3.

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

[10] Tamraparni Dasu and Theodore Johnson. 2003. *Exploratory data mining and data cleaning.* Vol. 479. John Wiley & Sons.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[26] Kedar S Mate, Brandon Bennett, Wendy Mphatswe, Pierre Barker, and Nigel Rollins. 2009. 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 (2009), e5483.

[27] Wes McKinney and others. 2010. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, Vol. 445. van der Voort S, Millman J, 51–56.

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

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

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

[31] Tapan S Parikh. 2009. Engineering rural development. *Commun. ACM* 52, 1 (2009), 54–63.

[32] Tapan S Parikh, Paul Javid, Kaushik Ghosh, Kentaro Toyama, and others. 2006. 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*. ACM, 551–560.

[33] Fahad Pervaiz, Richard Anderson, and Sophie Newland. Data Specification for Information Systems for the Immunization Cold Chain. (????).

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

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

[36] 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, and others. 2009. The OpenMRS implementers network. *International journal of medical informatics* 78, 11 (2009), 711–720.

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

[38] Hadley Wickham and others. 2014. Tidy data. *Journal of Statistical Software* 59, 10 (2014), 1–23.