Data Migration for Cold Chain Inventory Systems

Fahad Pervaiz, Richard Anderson

ABSTRACT

Developing countries face the challenge of migrating their health systems data from disjoint, standalone information sources to modern web-based databases. We study this issue in the domain of immunization logistics by looking at the problem of maintaining a national vaccine cold chain inventory. We present modular components that can be composed to create a migration path from existing spreadsheet inventories to web-based databases and other software tools. Specifically we (1) introduce the Cold Chain Equipment Inventory (CCEI), a new data standard for cold chain inventories that was developed in collaboration with UNICEF, WHO, and NGO cold chain experts, (2) present Data Extractor (Dextra), a tool that facilitates the translation of existing inventories into CCEI, and (3) evaluate CCEI and Dextra by converting multiple real health equipment inventories to a format that can be used by modern database systems. Our findings show that CCEI and Dextra together are capable of transforming a variety of health inventories into formats compatible with existing data management systems such as Open Data Kit (ODK) Tables and District Health Information System (DHIS).

Categories and Subject Descriptors

J.3 [**Computer Applications**]: Health - *Medical information systems*

General Terms

Management

Keywords

Health Information Systems, Vaccine Cold Chain, CCEI, Dextra, ODK Tables, DHIS.

1. INTRODUCTION

Health information systems are critical for strengthening public health in developing countries [2][20]. 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. A health information system reports data up through the health hierarchy and is maintained in a central database. Reports and analysis can be generated either centrally or in a distributed manner. A health information system often supports multiple health domains, where each health domain is a programmatic area around a disease, intervention or service.

However, in many countries, health information systems are still primitive with reports on paper forms, and the resulting information is stored in standalone databases or spreadsheets. Nonetheless, infrastructure is improving rapidly in many countries, making networked solutions feasible in locations where they were impossible just a few years ago. Network bandwidth is increasing with the deployment of fiber and expanding wireless networks. Costs for data communication have reduced substantially, and computing infrastructure has improved with increased availability of PCs and access to hosted solutions.

Still, infrastructure alone does not make it easy to introduce a modern health information system [13][14][24]. 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 will 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 often be highly political [23].

Recognizing the complexity of existing information systems and organizations, how does a country successfully implement a modern health information system? In this paper we explore this research question and present a proof of concept demonstration that generalizable solutions to this problem are possible, and that the resultant systems may be more powerful than existing tools.

Specifically, our work focuses on the management of vaccine cold chain information. Vaccines are one of the most successful public health interventions in history and successful vaccination programs are integral to the public health efforts of any nation. However, in order for vaccines to remain viable, they must be stored at appropriate temperatures from the time they leave the factory to the time they are delivered to the recipient. This network of warehouses, health facilities, and vehicles that keep the vaccines at safe temperatures are collectively referred to as the vaccine cold chain.

National-level administration of the cold chain is an immense logistical challenge. A basic aspect of cold chain administration is the cold chain inventory that represents the current state of the cold chain within a country. This inventory consists of two parts: the data model and the implementation. The data model defines the information that is collected about the cold chain—e.g. the nature of the facilities and the number and type of refrigerators. The model must be designed to be lightweight enough to be practical, but powerful enough to answer questions about the adequacy and status of the cold chain. The implementation refers to the physical means of storing the data model, be it on paper, in a set of spreadsheets, or in a modern web database.

In most countries, cold chain data is not accessible by the healthcare system, and, at best, is stored using spreadsheet tools or isolated applications available to a few managers. However, increasingly the technology and infrastructure exist to make this information available across the entire health system.

In this paper we present a data model and implementation solutions that are both powerful and feasible. Specifically, our main research contributions are:

(1) We introduce the Cold Chain Equipment Inventory (CCEI), a new data model developed in collaboration with UNICEF, WHO, and NGO cold chain experts. By implementing CCEI, countries will have access to a powerful and flexible data model that

permits accurate querying, predictive modeling, and regular updates.

(2) We present Data Extractor (Dextra), a new data-mapping program that transforms data from multiple standalone sources to CCEI.

(3) We demonstrate that a cold chain inventory can be integrated with an existing web-based health information system by implementing the functionality of the standalone Cold Chain Equipment Management (CCEM) inside the District Health Information System 2 (DHIS2). DHIS2 is a modern database system that employs a client-server model, allowing online administration and access of the cold chain inventory [11].

(4) We demonstrate that CCEI presents data in a way that can be easily consumed by external applications by importing it into Open Data Kit (ODK) Tables, a data management system for Android.

Our findings show that Dextra and CCEI together are capable of generating a fully functional cold chain inventory system, supporting online visualization and access as well as distributed management on mobile devices.

The rest of this paper is organized as follows: we begin by discussing in detail the problems faced by cold chain logisticians and health officials. Then we discuss current cold chain solutions and why they are insufficient. We then we provide a high-level outline of our technical contributions before describing them in detail our main contributions: the CCEI data model, Dextra for transforming data, the implementation of a cold chain inventory inside of DHIS, and finally, the use of ODK Tables for managing a cold chain inventory through CCEI. Together these demonstrate the basic steps in the migration path from isolated data sources to a modern health information system.

2. IMMUNIZATION LOGISTICS

This paper focuses on software systems for managing the equipment in the vaccine cold chain. Although we are interested in addressing the broad data management challenges that exist across health systems in developing countries, we have chosen to focus this paper on the smaller use case of software systems for managing the equipment in the vaccine cold chain.

2.1 Vaccine cold chain

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 [22]. There are robust global organizations supporting 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 separate department inside the Ministry of Health. 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 are 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 use. This is done with refrigerators and freezers at storage locations, and refrigerated trucks and cold boxes for transit, which are collectively referred to as the "vaccine cold chain."

2.2 Cold Chain Inventories

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 will manage 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 vaccine cold chain is whether or not it has sufficient capacity to store the country's 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 even more important with the introduction of new vaccines such as the new rotavirus and pneumococcal vaccine. These new vaccines take up more space, and are more expensive and more sensitive to heat than older vaccines, which add to 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.

2.3 Challenges

There are a number of challenges associated with cold chain inventories. These can be divided into problems with the data model and problems with the implementation. The data model specifies both the types of information that are collected as well as how they are organized. If key attributes are not included in the inventory, then its functionality is limited because some queries cannot be performed. The specific way that attributes are represented will also impact how the information can be used and how the collected information aligns with international standards.

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 used to provide feedback to people involved in the immunization system. Finding a 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.

To the best of our knowledge, few countries maintain a list of all health facilities, or Master Facilities List[1]. Many countries conduct periodic assessments of their vaccine cold chains, which involve conducting a partial or complete cold chain inventory. This has traditionally been done through facility visits by a trained team, which is very expensive[2]. 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 their inventories up to date 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.

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.

3. EXISTING SYSTEMS

A wide range of systems is used to manage cold chain inventories. The context differs between countries, so it is natural to see a range of 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.

3.1 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 on Excel. The 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 in Section 5).

3.2 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 [3], 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 multiyear 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 most significant contributions of CCEM.

3.3 Web-Based Databases

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.

[1] A notable exception is Kenya [17], which has its Master Facilities List on line http://www.ehealth.or.ke/facilities/

[2] As a data point, Kenya, with a population of 42 million, has roughly 5,000 health facilities with vaccine storage. Travel to remote sites can be very slow.

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.

An alternative to building a custom inventory system is to build on top of a more general platform. We describe in Section 8 how DHIS2 has been extended to support cold chain inventories. 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 [5][6].

3.4 Other applications

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. Two examples are the logistics management system OpenLMIS [28] and the vaccine stock management system VSSM [19]. 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.

4. Our approach

When a country initiates a cold chain inventory, they frequently must design a system from the ground up, devising their own data model, and implementing it with tools on hand that do not necessarily provide the features expected of a modern inventory system. As a result, the collected data is often insufficient or difficult to manage, greatly limiting the impact that the inventory can have on the administration of the cold chain. This situation can be improved based on knowledge gained from the successes and failures of previous efforts. In this paper, we draw on extensive experience with country-level cold chain inventories and present a proof of concept demonstration that with a minimal set of tools, countries supporting diverse inventory infrastructures can transition to a modern system.

We have designed and developed several modular components that can be combined to provide a migration path from existing spreadsheet inventories to a web-based database and other software tools. At the center of our work is a robust new data model for cold chain inventories, the Cold Chain Equipment Inventory or CCEI. To support the conversion of data from existing sources into CCEI, we developed a tool called Data Extractor (Dextra). In addition, we extend a widespread web-based health information system, DHIS2, to support a cold chain inventory. We will now discuss each of these components in greater detail.

4.1 Data Model: CCEI

CCEI serves two distinct purposes. First, for cold chain logisticians it is a definition of a cold chain inventory. It specifies the information that must be collected, as well as the details of how information is recorded. Second, the data model gives a standard for implementation of cold chain inventory tools including the data types and organization, as well as a specification of input and output formats.

CCEI was developed in collaboration with UNICEF, WHO and NGO cold chain experts and informed by over a dozen existing cold chain inventories systems. By implementing inventories based on CCEI, countries will have access to powerful and flexible tools that support accurate querying, predictive modeling, and regular updates.

It is important to emphasize that the data model is separate from the implementation. Choices of data model can limit the utility of the data. For example, if assets do not have a unique identifier, it is not possible to track performance of refrigerators over time. As another example, if the data model does not contain the immunization population associated with facilities, it is not possible to evaluate whether or not storage capacity is adequate.

4.2 Migration: Dextra

Although CCEI is a powerful data model, it must be implemented in order to be useful. Migrating existing systems to a new data model is a challenging task. To aid in this process, we have developed Dextra, a software tool that takes existing implementations of cold chain inventories—be they in modern databases or in disconnected spreadsheets—and facilitates transformation into the CCEI standard.

Dextra provides an extensible set of transforms for converting data between different models. It maintains a copy of the original data as well as the data as it is transformed into CCEI. There is thus no risk of data loss or corruption, and administrators are always able to view their original data. Dextra supports a number of visualizations and additional utilities that facilitate data cleaning, because the process of data integration is tightly connected with cleaning.

4.3 A Modern Database: DHIS2

An important component to having a cold chain inventory that is kept up to date and utilized for management of the cold chain is to make it accessible on the web. A modern database system will support a client-server model, hosting the data in the cloud and serving it through a browser with data, application, and presentation layers. For deployment of a web-based cold chain inventory, there are tremendous advantages if it can be tied to existing health information systems. When this is possible, it does not require the introduction of a new software system to a country, with the associated costs of maintenance and training.

We demonstrate the possibility of using an existing health information system for a cold chain inventory by adding the functionality of CCEM, an existing cold chain inventory system to DHIS2, a widely-used system for reporting health indicators and visualizing health data [11][7]. This integration required an extension of the underlying data model of DHIS2 to accommodate reporting information associated with fixed assets such as refrigerators. This extension has enormous potential to make the cold chain inventory immediately more useful, taking the information out of the hands of a few administrators and providing access to healthcare officials, policy makers, and statisticians, all of whom will be able to help strengthen the country's vaccine cold chain.

4.4 External Applications: ODK Tables

Finally, we show that the CCEI data model is easily consumed by external applications, which provides additional value to the inventory. We demonstrate this by importing CCEI data for several countries into ODK Tables [8]. These cold chain inventories started in very different formats and were transformed to the CCEI format using Dextra.

ODK Tables is a data management system for Android. Using ODK Tables we were able to define custom interfaces to the data and sync changes between devices. Employing ODK Tables and the CCEI data model, administrators could download the database

to a fleet of low-cost Android devices. This would give them the ability to update and curate the data in a distributed fashion, functioning in the absence of internet connectivity and enforcing data cleanliness through electronic data entry techniques.

5. CCEI—A Uniform Data Model

The first part of our solution is to develop a data standard for the cold chain equipment inventory. This schema is referred to as CCEI. In the health domain, alignment of reporting standards is particularly important when assessing health systems and the impact of health programs. Developing standards introduces a number of challenges. One is balancing simplicity with the number of indicators being reported. Another is balancing the ideal indicators with what is feasible to collect.

In practice, CCEI makes three important contributions:

1) It is a definition of cold chain equipment inventories to which cold chain logisticians can agree.
2) It is a data standard that allows data to be shared between applications, and permits applications to interoperate.
3) It allows the inventory to be represented in a structured manner, increasing the quality of the data.

5.1 CCEI data model and definitions

The CCEI data model is designed to store the core data of a cold chain inventory, as well as to allow for a set of extensions that include related information. One of the contributions of the CCEI data definitions is to support applications built on top of inventories, such as cold chain temperature monitoring or logistics planning, so that stakeholders may realize added benefits from these additional components.

The CCEI model has been developed in a collaborative manner. An initial set of inventory definitions was put together based on several existing cold chain tools, and then edited by the UNICEF Cold Chain Logistics group as a Google document. 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 [21] and CHAI [9], who provided very detailed feedback. The current draft is now under a more formal review. The project has 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 components of a CCEI relate directly to its intended use. The approach taken in this effort was to focus on a minimal core with the possibilities of adding additional modules later. This approach was effective in allowing the effort to move forward and attracting interest in developing extensions to the model. Examples of planned extensions include temperature monitoring, transportation and equipment maintenance.

The core of CCEI is the model of facilities and their associated equipment. One of the requirements for CCEI was that it should 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. Since many of the assets in the cold chain are standard equipment, the inventory also includes an official catalog of models of equipment available.

The basic 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 country's 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.

Figures 1 and 2 show simplified versions of the facility, refrigerator and equipment catalog components of the data definitions. 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. Although we do not have space in this paper to present the full model, we encourage the reader to explore the full model online.[3]

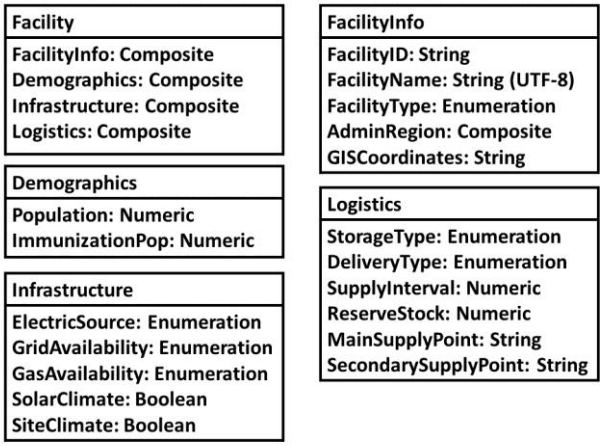

Figure 1 Facility data definitions (simplified)

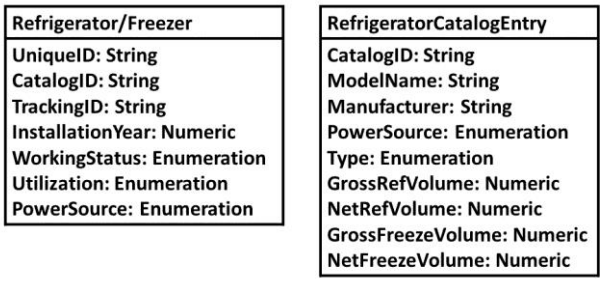

Figure 2 Refrigerator and Refrigerator Catalog data definitions (simplified)

The process of defining specific indicators or attributes presents a set of well-known challenges. This includes balancing completeness of information with parsimony of data collection. Another challenge is in designing indicators that can be collected

[3] A current draft of the model is available at scribd.com/doc/152853455/CCEI

in an accurate manner and yield information that is useful. To show the flavor of the challenges faced, we provide two brief examples that generated substantial discussion:

- *Population*. Determining the population covered by a health facility is often difficult. Two values are recorded, the total number of people served by the facility, and the total number of people receiving immunization services at the facility.
- *Grid Power Availability*. Some measure of the quality of grid power is needed to identify if a facility can use certain types of equipment. A measure of hours-per-day is used, with the values broken into four buckets: 0, under 8, 8 to 16, and more than 16. Beyond the question on the choice of each bucket is the definition of hours per day.

Despite these challenges, CCEI has proved to be capable of modeling data in both a standardized and practical manner. We will describe some of the specific implementations of CCEI in sections 7 and 8.

6. Data Extractor (Dextra)

Dextra is a tool built to migrate datasets from a variety of diverse sources into a well-structured database. It was designed keeping in mind the requirements of transforming current cold chain inventory datasets into the new CCEI standard. The challenges of this design process include importing dirty datasets, keeping the CCEI data up-to-date without disrupting the current data maintenance process, and dealing with intermittent access to the original datasets. A goal of Dextra is to make the migration process easy by adding transformation rules that are determined from different domain-specific datasets, thereby reducing the setup overhead in subsequent deployments.

6.1 System

To migrate a data source, Dextra first imports all source datasets in their original format. A local copy of these datasets is maintained so that intermittent availability of the data does not affect the system's processes. These datasets are then maintained by Dextra and can be aligned and transformed into any given data definition. The system can be configured to let subsets of this data be modified from within Dextra while the rest remains immutable. The data can only be updated by providing the latest copy of the source dataset. The most challenging part of the migration process is to understand the transformations that are required to align source datasets into the required destination format.

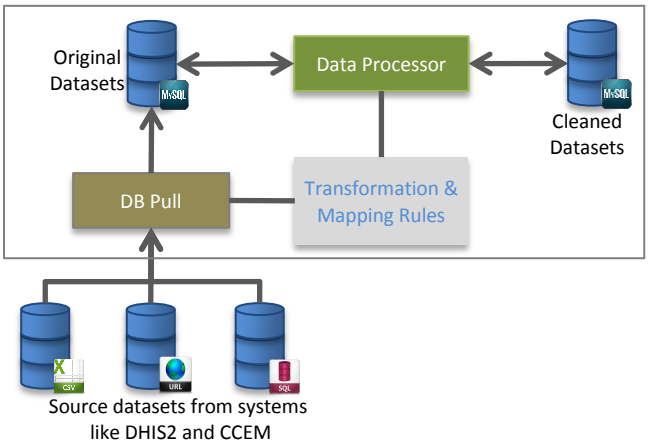

Figure 3 Data flow and architecture of Dextra

Although several these features are already in some other tools [15][18][25], Dextra provides the flexibility of a custom system, which means that users can try out different transformation models based on what they learn from their domain-specific datasets and adjust it according to their needs. Currently, Dextra provides the following features:

- Extract data from Excel files, MySQL and an ODBC connection. Most of the cold chain datasets are maintained only in Excel files.
- Suggest a data schema with connections between columns based on the column names. Cold chain datasets tend to have similar column headings for a particular field across all tables.
- Set a validity period for all datasets so that if the local copy expires, subsequent merges or transformations will not be allowed.
- Map fields from different source tables to the given data definition:
  - Select rows based on equality or inequality conditions
  - Join tables based on common attributes between tables
  - Split a column into different fields
  - Merge values from multiple columns into one column
- Provide transformations:
  - Conversion between basic data types
  - Translating code values to more descriptive String variables
  - Convert GIS coordinates

To setup Dextra, a user needs to populate the configuration database that contains all the mapping and transformation rules. Currently, this is done through a very basic interface that allows the user to edit the suggested schema and add mapping rules. Based on the mapping rules, Dextra adds some transformation rules itself that can then be modified by user. These suggestions are based on rules we uncovered while manually going through the transformation process for cold chain data from Kenya, Benin, Malawi, Tanzania and the Indian States of Bihar, Gujarat and Kerala.

The ideas for Dextra draw on Extract-Transform-Load (ETL) tools [18]. However, Dextra is specialized for the cold chain domain, making it easier to use by cold chain experts, logisticians and managers. Examples of existing ETL tools are Apatar, CloverETL, Microsoft SSIS and many more [18]. These tools can be configured to transform datasets from one definition to another [27]. However, they may be challenging to setup if one is not familiar with the data issues and does not know the required transformations for the given dataset. Google Refine [15] and Data Wrangler [25] are powerful tools that do not share this shortcoming. These tools help users to make sense of dirty datasets and apply appropriate transformations. However, none of these tools provide the possibility to learn domain-specific transformation requirements from the data and they do not make it easy to transform other datasets in the same domain [1].

6.2 Implementation of CCEI

One of the first applications of Dextra was to convert instances of Microsoft Access-based inventories in CCEM to CCEI. The key data tables from CCEM were exported as Excel files and run through Dextra to create the CCEI representation. Dextra was able to extract the current structure of CCEM based on the naming convention used in header rows. The exact CCEM to CCEI field mapping was then added manually. After this configuration was provided, Dextra converted all the CCEM datasets to CCEI standards. The instances of CCEM for the different countries had minor differences in the tables that were handled automatically by

Dextra. This converted data was stored in a MySQL database so that other applications can be built on top of and consume the data.

The Tanzania cold chain inventory data, maintained in excel files, was the second test case for Dextra. This dataset was different from CCEM as it came from a WHO spreadsheet tool. It contained some extra information on the maintenance and replacement cost of refrigerators that is not part of the CCEI definition. In addition, the data was lacking key CCEI information, such as details on power source availability at facilities, vaccine schedules, and the GIS coordinates of health centers. There were also several blank fields in the CCEI representation of this data after the Dextra transformation process. Though this representation was as complete as possible given the original dataset, the blank fields indicated that more details should be collected for complete reporting.

The third dataset used to test the capabilities of Dextra was data from a recently completed cold chain inventory survey, an effort led by INCLEN [16]. This was paper-based survey data digitized into Excel files and had a large number of blank fields that represent no data. For example, if a standard survey contains fields for seven refrigerators but the given facility has only six. Blank fields in other data sources represent unknown values for those fields. Therefore, extra rules were added to handle the blank fields in this source data differently. This dataset also required additional mappings to associate the facility IDs between asset and facility information, and to extract multiple pieces of equipment.

It took one to two hours to manually add mapping rules, which must only be done once per source system. Finalizing mapping rules takes time, as the user has to match each field in CCEI (destination data model) to one or multiple fields in the source data model. This mapping was at times challenging due to lack of understanding of the source schema. Once these rules were added, transformation of the data took several minutes. The authors performed these transformations, and the system has yet to be tested on a layperson user.

6.3 Data Integration Challenges

We experienced several data integration challenges that made it difficult to integrate data from different spreadsheets, tables and/or sources. Given below are some of the issues that appeared in our datasets:

- Different spellings for the same value in a field make it harder to join data from various tables based on that field, such as joining facility and population served tables on the facility name field (learned from Tanzania dataset).
- Various formats used to represent the same value hinder the task of producing data in standard format. For example, one dataset had multiple formats for GIS coordinates in the same column.
- A lack of identifiers in source data makes it harder to split a table into multiple tables. It is difficult to track which fields in the new data definition are coming from same row of source data. In the Tanzania dataset, for example, there are no unique identifiers assigned to refrigerators and facilities.
- Duplicates in source data make it hard to extract distinct information. For example, in the Tanzania dataset, a health facility with its detail is duplicated against each refrigerator located in that facility. This requires extra effort to extract a list of health facilities from the data.
- Mapping multiple fields to one field requires rules to select a field or combine multiple fields based on certain conditions. For example, CCEM has a separate field for each reason a refrigerator may be dysfunctional, while CCEI has only a single field.
- Unique tuples of the same entity, like refrigerator, are placed in datasets as shown in Table 1, which represents nine distinct refrigerators in a table of three rows. This requires transposition of parts of each row before loading it into the new format, which complicates the mapping rules (learned from INCLEN's dataset).

Table 1. Schema for refrigerator storage against each facility

Facility	Ref_ID	Ref_Name	Ref_ID	Ref_Name	Ref_ID	Ref_Name
10201	1	Ref 1	2	Ref 2	3	Ref 3
10202	4	Ref 4	5	Ref 5	6	Ref 6
10203	7	Ref 7	8	Ref 8	9	Ref 9

- Translating code values like 0, 1 or -1 to its actual representation may vary, as it is possible that it might not be consistent even across the same table. In one field, 0 may mean YES and -1 may mean NO, while in another, 0 may mean NO and 1 may mean YES (learned from the CCEM dataset).

We are also continuing to learn from the difficulties in managing these messy datasets, and we are adding more and more default transformation rules to Dextra in order to handle the dirty data. This makes it easier to configure Dextra for a new cold chain inventory dataset because it removes the redundancy of adding the same transformation rules again and again for every new dataset. Obviously, these rules are system-suggested so any expert user can edit them.

Currently, the system is mature enough to handle existing cold chain inventories without suffering any performance issues. The configuration process has been simplified enough that one only needs to understand the mapping between two formats, not the issues within the data, in order to setup Dextra. In our experience, this takes days worth of workload off the user.

7. DHIS2

The appropriate architecture for a cold chain inventory system is a real database with a web interface supporting remote updates of the inventory and a set of cold chain specific reporting and analysis tools. It is also important to link the inventory with the national health information system because this allows the system to be tied to other program data and reporting. More importantly, centralized administration by the Ministry of Health provides resources for maintaining the system. One significant problem faced by country health programs is a proliferation of software systems—thus it is advantageous to use an existing system for a new function rather than to introduce and maintain a new system.

To demonstrate that it is possible to integrate a cold chain inventory into a national health information system, we took the functionality of the standalone cold chain tool CCEM and implemented it inside DHIS2. We produced a fully functional version of the inventory and report components of CCEM in DHIS2 for the Kenya dataset.

As of submission, we have not implemented a direct import of CCEI into DHIS2. This is planned, and is a natural extension of our tools. Importing the CCEM data required a significant amount of transformation to conform to the DHIS2 data model. This process served as a case study that motivated the development of a standard inventory data model (CCEI) and a migration tool

(Dextra). In its current format, our extension of DHIS2 has a module that converts the CCEM data format to the DHIS2 data model. The scope if this module is limited and specific to this application and dataset. Its function will be subsumed by the more general data transformation capabilities of Dextra.

DHIS2 is a health indicator reporting system used in about 30 countries and is also the health data reporting system used in about half of the states of India. The motivation behind developing DHIS2 was to improve the quality of health data and use information for action based on different tools. 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.

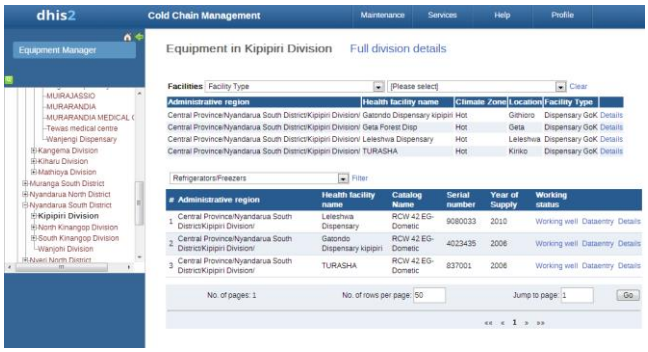

Figure 4 Screenshot of DHIS2 Cold Chain module showing a group of facilities and their associated refrigerators.

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 [10].

8. Integrating CCEI with ODK Tables

Cold chain inventories converted to use CCEI have the added benefit that they will be readily consumable by external applications. These applications will add additional value to the data. For example, Hermes [4] is a tool used for modeling country-level immunization systems. It is often used to influence policy and could easily incorporate the data presented in a cold chain inventory, facilitating more informed policymaking. OpenLMIS [28] and the vaccine stock management system VSSM [19] are two additional systems used at various levels in cold chain administration that could benefit by incorporating data in the standardized CCEI format.

To demonstrate the integration of CCEI with external applications, we took a collection of basic inventories and installed them on an additional external system: ODK Tables. ODK Tables [8] is an Android application designed to support management of a relational database on a mobile device. Using ODK Tables, workers in the field equipped with mobile devices would be able to access and update the CCEI-based inventory.

ODK Tables offers a number of features that make it a strong candidate for this use case. First, it makes no demands on the structure of the database. This means ODK Tables can be used with any database schema, including the CCEI schema described in Section 5, without requiring any transformations of the data tables. Second, it is highly customizable, using simple HTML and JavaScript files. This means that each country could use ODK Tables to manage their inventories without having to agree on how the data should be presented to users. Administrators could define their own layout with a small number of files, decoupling deployments between countries and providing a high level of control over their data. This is advantageous, as it minimizes the consensus that must be reached between countries and provides a high level of data ownership. Finally, ODK Tables provides synchronization and conflict resolution to web servers running privately or in the cloud, enabling data to be easily shared and kept up to date.

To show that country-scale datasets can indeed be effectively managed using ODK Tables, we converted a Benin cold chain inventory dataset to the CCEI schema using Dextra and loaded it onto mobile devices running ODK Tables. We defined four custom views using HTML and JavaScript that were used to display summary and detailed information about rows in the refrigerator and facility tables. An example of the facility summary can be seen in Figure 5. This resulted in a cold chain inventory management Android application, backed by the CCEI schema, where we were able to display, edit, and sync data between devices without ever having to deal with the vagaries of database management.

Using ODK Tables, we were also able to visualize the data in a variety of ways that could prove useful to cold chain administrators. This included graphing the data and showing the facilities on a map. Figure 5 shows the facilities based on their GPS coordinates. The markers have been colored based on the underlying data, where facilities with and without reliable electricity are green and red respectively.

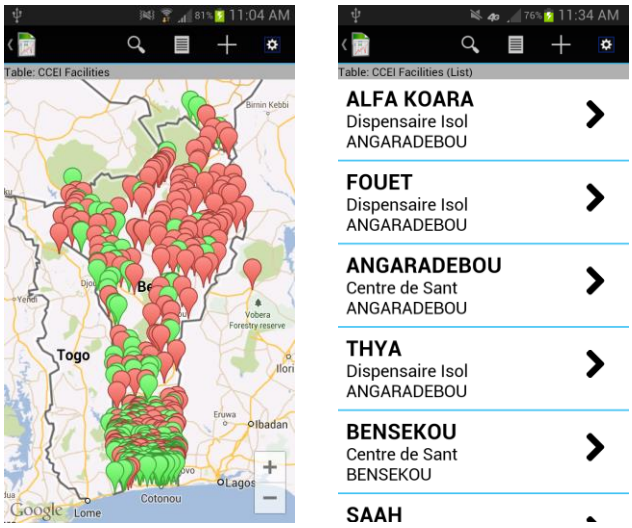

Figure 5. Benin cold chain facilities represented on a map (left) and in a summarized list view (right)

We were thus able to synchronize CCEI data between individual Android devices and with a server in the cloud. An obvious and useful extension of this functionality would be to support communication between a server capable of speaking the ODK Tables synchronization protocol and a cold chain database system such as DHIS2. Countries would then be able to add Tables to their existing infrastructure and begin incorporating real time updates to their inventory system. The issues faced in implementing such a system are beyond the scope of our current work.

9. DISCUSSION

9.1 Lessons Learned

We encountered a number of surprising discoveries in this work that bear emphasis. The first is the inextricable link between data cleaning and data characterization. The datasets used in this work are currently used in their respective countries. However, as we developed migration tools capable of visualization, we discovered discrepancies and errors in the data that had existed for lengths of time in their original format. For example, when mapping facilities according to their recorded GPS coordinates, many facilities appeared in the middle of the ocean or several countries away. While this would be quite readily discovered when administrators used the coordinates, more insidious errors were also discovered. Supposedly unique identifiers of facilities and refrigerators were frequently misspelled or entered inconsistently. Queries could then fail to find the complete and accurate list of refrigerators at a facility, resulting in incorrect assessments of capacity.

It was also interesting to discover that a number of seemingly static resources can in fact change over time, meaning that any successful system must be explicitly implemented to retain a high degree of flexibility and make a minimal set of assumptions. A notable example of this was the fact that administrative hierarchies were not fixed, but in fact fluctuated over time. In Uganda, for instance, individual administrative units were eventually split, resulting in a new hierarchy. A larger scale example is India, where in recent years a number of new states have been formed. This sort of large-scale change also occurred in Kenya, where provinces have been split up into counties, and the old districts are now sub-counties. As a result, it was necessary that CCEI and similarly Dextra be capable of handling this sort of sophisticated expression.

Finally, when developing data standards, a conscious effort must be made to provide a core functionality that is extensible and permits country-specific customizations. While CCEI has been designed to represent the data deemed necessary for an effective cold chain inventory, some countries currently monitor additional information in their spreadsheet-based systems that is not present in the core CCEI schema. For example, one of the country's inventory datasets used in this paper included information about replacement and repair costs for individual refrigerators. This information is not recorded in CCEI. However, if migrating to a new standard meant that this information would be lost or unavailable, it would be an impediment to adoption. As such it was necessary to design CCEI to represent a core set of information that would easily support customization to meet country-specific needs.

9.2 Future work

This work is a proof of concept. We have shown that with the appropriate tools, a number of disparate inventory data models and systems can be transformed to meet data standards and enjoy the benefits of being stored in a modern database. To move from a proof of concept to a complete and robust implementation, several additional steps must be taken.

First, CCEI must be finalized and released. It is currently in the late stages of review by UNICEF, WHO and NGO cold chain experts. Due to the abstraction of the migration and transformation of datasets into Dextra, any changes will be easily incorporated.

Second, Dextra itself is a work in progress. As discussed above, it was capable of converting several formats into the CCEI standard. However, there is still room for improvement. The following features will be added in the system to support more native transformations:

- Adding the Metaphone phonetic algorithm and Levenshtein distance for better string matching to handle misspelling case in an effective way [26].
- Introducing an intuitive user interface, like that of Google Refine [15], so that the user can view the system suggested transformations more easily and edit them in a user-friendly manner.
- Use an artificial intelligence engine for transformation suggestions instead of relying on a rule based suggestion system.
- Use database reverse engineering techniques to find connections between the schemas of different sources of data [12].
- Support a wider variety of source data file formats.
- Add support to update web based source systems for better integration with existing processes.

As this work progresses, the process of migrating and maintaining a connected cold chain inventory will become even simpler. The existing proof of concept shows that the technical capabilities of cold chain management have reached the point where modern, standardized systems in developing countries are within reach.

9.3 Conclusion

The main contribution of this work was to present a proof of concept demonstration of how to transition from basic spreadsheet-level health information systems to contemporary software systems. This work was done in the in the context of

immunization logistics, was based on real world country-level data sets, and targets software systems already in use in ministries of health around the world. With the appropriate tools, this process was completed in several distinct steps.

First, we introduced the new standard data model for cold chain inventories: CCEI. We demonstrated that with mapping tools such as Dextra, existing rudimentary inventory systems could be transformed into CCEI, cleaning and typing the data in the process. This format can then be added to existing health database systems and ultimately consumed by a number of external applications, lending increased value to the data. Together these demonstrate the basic steps in the migration path from isolated data sources to a modern health information system.

It is important to highlight the modularity of this process. We chose to integrate the data into DHIS2 and ODK Tables, which provide valuable capabilities like web-based access and distributed data management. However, these tools could readily be replaced by additional server systems or external applications. This is a natural consequence of a powerful data model implemented in a modern system, and demonstrates that with the correct tools such a situation could (hopefully) soon become the norm.

10. References

[1] Agrawal, H., Chafle, G., Goyal, S., Mittal, S., Mukherjea, S.: "An Enhanced Extract-Transform-Load System for Migrating Data in Telecom Billing", Int. Conf. on Data Engineering (ICDE), 2008.

[2] Ammenwerth E, S Gräber, G Herrmann, T Bürkle, and J König. 2003. "Evaluation of health information systems-problems and challenges". *International Journal of Medical Informatics.* 71 (2-3): 2-3.

[3] Richard Anderson, John Lloyd, and Sophie Newland. 2012. Software for national level vaccine cold chain management. In *Proceedings of the Fifth International Conference on Information and Communication Technologies and Development* (ICTD '12). ACM, New York, NY, USA, 190-199.

[4] Assi TM, Rookkapan K, Rajgopal J, Sornsrivichai V, Brown ST, Welling JS, Norman BA, Connor DL, Chen SI, Slayton RB, Laosiritaworn Y, Wateska AR, Wisniewski SR, Lee BY. How influenza vaccination policy may affect vaccine logistics. *Vaccine*. 2012 June 22;30(30): 4517-23.

[5] Braa J, Hedberg C. The struggle for developing district health information systems in South Africa. *Information Society*. 2002;18(3):113–127

[6] Braa J, Monteiro E, Sahay S (2004): Networks of action: sustainable health information systems across developing countries. MIS Quarterly, 28(3), 337-362

[7] Braa, J., and Sahay, S., *Integrated Health Information Archtitecture: Power to the Users, Matrix Publishers*, New Delhi, 2012.

[8] W. Brunette, S. Sudar, N. Worden, D. Price, R. Anderson, and G. Borriello. "ODK tables: building easily customizable information applications on Android devices." In Proceedings of the 3rd ACM Symposium on Computing for Development (DEV 2013). Jan 2013.

[9] CHAI – Clinton Health Access Initiative - http://www.clintonhealthaccess.org/

[10] Chaudhri, R., O'Rourke, E., Borriello, G., Anderson, R., and McGuire, S., FoneAstra: Enabling Remote Monitoring of Vaccine Cold-Chains Using Commodity Mobile Phones, The 1st Annual Symposium on Computing for Development, 2010.

[11] DHIS2: http://www.dhis2.org

[12] J. -L. Hainaut, C. Tonneau, M. Joris, M. Chandelon. Transformation-based database reverse engineering. In Lecture Notes in Computer Science Volume 823, 1994, pp 364-375

[13] Heeks, R. "Information systems and developing countries: failure, success, and local improvisations". *The Information Society, 18:101-112,2002.*

[14] Heeks, R., Health information systems: Failure, success and improvisation, International Journal of Medical Informatics, 75, pp, 125-137, 2006.

[15] D. Huynh and S. Mazzocchi. Google Refine - http://code.google.com/p/google-refine/

[16] INCLEN Trust International – http://www.inclen.org

[17] Kenya, Ministry of Medical Services. Master Facility List Implementation Guide, February 2010.

[18] List of ETL tools - http://www.etltool.com/list-of-etl-tools/

[19] PAHO. Vaccination Supplies Stock Management [VSSM] In Immunization Newsletter Volume 32, No 6, Dec 2010, pp 7

[20] Panir, Role of ICTs in the Health Sector in developing countries: a critical review of literature, Journal of Health Informatics in Developing Countries, Vol 5, No 1, pp 197-208, 2011.

[21] PATH – www.path.org

[22] Polio: Global eradication initiative. http://www.polioeradication.org/Dataandmonitoring/Poliothisweek.aspx

[23] S. Sahay, M. Asnestad, and E. Monteiro, Configurable Politics and Asymmetric Integration: Health –Infrastructure in India, Journal of the Association for Information Systems, Vol 10: 5, pp. 399-414, May 2009.

[24] Sahay, S. and Walsham, G. (2006) Scaling of health information systems in India: challenges and approaches, Journal for IT and Development. 12, 3, 185-200.

[25] 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* (CHI '11). ACM, New York, NY, USA, 3363-3372.

[26] Chakkrit Snae. 2007. A Comparison and Analysis of Name Matching Algorithms In *International Journal of Applied Science. Engineering and Technology*.

[27] Vassiliadis, Panos. "A Survey of Extract–Transform–Load Technology." *IJDWM* 5.3 (2009): 1-27. Web. 8 Jul. 2013. doi:10.4018/jdwm.2009070101.

[28] VillageReach. "The framework for OpenLMIS white paper" (2012). http://openlmis.hingx.org/Share/Details/312