Integrated Longitudinal Data Visualization Platform for INDEPTH HDSS Sites

Tumaini Kilimba

A Dissertation submitted to the Faculty Of Health Sciences for the partial fulfilment of the requirements for the degree of

Master of Science

Supervisors:
Gideon Nimako[1], Kobus Herbts[2]

[1]School of Public Health
Faculty of Health Sciences
University of the Witwatersrand

[2]Africa Centre for Health and Population Studies
Mtubatuba
Kwa-Zulu Natal

DECLARATION

I declare that this dissertation is my own, unaided work under the supervision of Gideon Nimako and Kobus Herbts. It is being submitted for the Degree of Masters of Science at the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination at any other University.

_____

22nd July, 2015

ABSTRACT

A lot of work has been done on spatial visualizations for public health. The feasibility of creating poverty maps for Indonesia at various administrative levels to help with the implementation of programs which target the poor was investigated in [6]. Their focus however was on a new methodology for imputing per capita consumption for each household in the population based on data collected from household surveys and data collected from population censuses. Their report however did not focus much on how the visualization platform was to be built. The final product though, after all the computations is the visualization, the poverty map of the country.

ACKNOWLEDGEMENTS

XXX

Contents

Dedicated to

XXX

AAA and
BBB

1

Introduction

The first chapter introduces the subject of visual analytics in general and then goes on to introduce visual analytics in the context of public health and epidemiology. We will delve into the literature around this subject and then we will espouse our own motivations in creating a generalizable public health data visual platform

Contents

Visual analytics tools are used to synthesize information and glean insight from massive, dynamic, ambiguous and often conflicting data. Visual Analytics (VA) is often referred to as a means for dealing with complex, large information sources that require human judgment to identify the expected and discover the unexpected. It is a multidisciplinary field whose core areas are analytical reasoning techniques, visual representation and interaction techniques, data representations and transformations as well as production, presentation and dissemination [1].

Analytic reasoning techniques enable users to get deep insights which support assessment, planning and decision making. It also allows users to assimilate large amounts of information at once [2]. In a healthcare setting expected outcomes include more efficient and effective clinical performance monitoring and improvement as well as improved modelling of patient flow and management. Additionally, one can expect increased quality of care, improved safety and efficiency and better support for clinical costing and resource coordination. It can also lead to better planned growth and competitive advantage [cite 3].

VA is referred to as "the science of analytical reasoning facilitated by interactive visual interfaces". Map based community health visualizations have provided a comprehensive and powerful interface for scientists and policy makers to visualize health care quality, public health outcomes and access to care. This has helped in making evidence-based decisions about improving healthcare [cite 4-7]

Multi-panel graphs have also been used to good effect in a graphical tool for epidemiologicalstudies to reveal the distribution of an outcome by time and age simultaneously [cite 8].

The growth of surveillance systems in both quantity of data and variety of outcomes is likely to necessitate constant innovations in data processing, synthesis, and communication [cite 8] Therefore, techniques to support production, presentation and dissemination of analytic results will allow us to communicate the information in the appropriate context to a variety of audiences.

1.1 Problem Statement

One critical requirement for successful public health surveillance is the ability to analyse and present data so that it is understandable to a range of public health stakeholders. In public health, VA can be viewed as the bridge between the availability of surveillance data in database architectures and useful information derived from this available data [cite 9].

INDEPTH Health and Demographic Surveillance Sites (HDSS's) deal with complex longitudinal data and, as a result, knowledge transfer to stakeholders is challenging. Better visualisation of this data is therefore required in order for potential scientific users to maximise exploratory data analysis and hypothesis generation. It would also aid decision-makers and the society at large to visualise this information in terms understandable by them. Such a visualization tool will also improve field work research activities by providing summary data of operational progress, e.g. fieldwork data collection progress, data entry progress, or other parameters such as data quality. This will serve to improve operational decision making and data quality. However, datasets at HDSS sites are normally under-visualised. These HDSS sites currently have no generalizable framework for implementing a data visualization platform to be used at these sites.

1.2 Motivation

The current under-visualization of HDSS datasets shows little promise of improvement in a harmonised way (across multiple sites) unless specific research efforts are directed towards finding a generalizable solution for delivering interactive visualizations, supporting exploratory analyses and real-time displays of operational progress. Furthermore, there is a paucity of research specifically on the technologies and tools which can be used to create such a data visualization [cite 10].

2

Background and Related Works

This chapter gives the background and related works that have been done in the field of public health data visualization

The topic of visualization of public health data was identified in 2009 by the CDC as one of six major concerns which must be addressed by the public health community in order to advance public health surveillance in the 21st century [cite 9]. However, very little has been researched in terms of standardization of the workflow and linking technologies for heterogeneous data sources needed specifically for visualizations in public health science [cite 10].

Of crucial importance when dealing with large datasets is the need for the users of the data, be they scientists or any other stakeholders to be able to discover the relations among and between the results of data analyses and queries [cite 10]. However, due to bottlenecks resulting from resource cost and lack of required skills, data visualization becomes an end product of scientific analysis rather than an exploration tool which facilitates scientists to generate better hypotheses in the continually more data-intensive scientific process [cite 10]. The use of such visualizations are usually utilised in business analytics, open government data systems, and media infographics but have generally not been used in public health. However, the capabilities currently being seen by web-based tools and technologies may be the breakthrough in resolving the scientific visualization bottleneck.

A lot of work has been done on spatial visualizations for public health [cite 4,6,7]. The feasibility of creating poverty maps for Indonesia at various administrative levels to help with the implementation of programs which target the poor was investigated in [cite 6]. Their focus however was on a new methodology for imputing per capita consumption for each household in the population based on data collected from household surveys and data collected from population censuses. Additionally, their report did not focus much on how the visualization platform was to be built. The final product though, after all the computations is the visualization, the poverty map of the country.

Along a similar spatial theme, [cite 7] looked at how the agents of parasitic diseases are spatially distributed using map visualizations. The tools of interest used in their research were the two closely linked Google products, Google Earth and Google Maps. Though they provide a little

more implementation details, the mix of tools used are not all open-source.

Other research on implementing a data visualization platform for community health assessment (CHA) used open source technologies but was limited to a desktop application and could not be accessed online [cite 11]. Web based tools have been seen as the preferred platform of choice for public health researchers as they permit distributed access, reduced software implementation costs and wider exposure of public health information for public dissemination [cite 4,12,13]. The report also gives scant details of the actual open source tools used and how they were put together in a way which would allow recreation of the steps.

The use of multi-panel graphs to illustrate trends and anomalies which would otherwise be obscured by traditional epidemiological visualization techniques such as pyramids and time-series plots was explored in [cite 8]. Under future developments in their report however, they acknowledge that two other features if incorporated to these graphs would enhance their impact, namely the dynamic display of data and interactivity.

Existing tools offer a range of features and functions to allow for exploration, analysis and visualization of public health users data, but the tools are often for siloed applications incapable of reciprocal operation with other, related information systems. They are isolated to the jurisdictions and organisations which developed them limiting their widespread adoption by other agencies or organisations[cite 13]. Interoperability of the visualization tools has been identified in a systematic literature review of infectious disease visualization tools as a prominent theme, due to challenges associated with increasingly collaborative and interdisciplinary nature of disease surveillance, control and prevention [cite 13].

The CDC's inclusion of VA as one of its six areas of focused research in public health lends credence to the importance of this research project. Furthermore there is little that has been researched in terms of standardisation of the workflow, tools and linking technologies for visualizations in the public health domain, as well as few existing web based tools for health related data visualization [cite 4,10]. This research will add to the body of literature broadly

in the subject of tools and technologies for data visualization in general, and specifically for implementing a data visualization platform which is generalizable for all INDEPTH HDSS sites with the aim of allowing it to become a central piece of the scientific process. It will also augment on work already done [cite 8] by incorporating dynamic display of data and interactivity into multi-panel graphs for epidemiological research [cite 8].

3

Research Aims and Objectives

In this chapter we will elaborate on the problem we are addressing. We shall also discuss how this problem is relevant to Health and Demographic Surveillance Sites (HDSS's) and our motivation to address it

Contents

3.1 Problem Statement

One critical requirement for successful public health surveillance is the ability to analyse and present data so that it is understandable to a range of public health stakeholders. In public health, VA can be viewed as the bridge between the availability of surveillance data in database architectures and useful information derived from this available data [cite 9].

INDEPTH Health and Demographic Surveillance Sites (HDSS's) deal with complex longitudinal data and, as a result, knowledge transfer to stakeholders is challenging. Better visualisation of this data is therefore required in order for potential scientific users to maximise exploratory data analysis and hypothesis generation. It would also aid decision-makers and the society at large to visualise this information in terms understandable by them. Such a visualization tool will also improve field work research activities by providing summary data of operational progress, e.g. fieldwork data collection progress, data entry progress, data archiving progress or other parameters such as data quality. This will serve to improve operational decision making and data quality. However, datasets at HDSS sites are normally under-visualised. These HDSS sites currently have no generalizable framework for implementing a data visualization platform to be used at these sites.

3.2 Motivation

The current under-visualization of HDSS datasets shows little promise of improvement in a harmonised way (across multiple sites) unless specific research efforts are directed towards finding a generalizable solution for delivering interactive visualizations, supporting exploratory analyses and real-time displays of operational progress. Furthermore, there is a paucity of research specifically on the technologies and tools which can be used to create such a data visualization [cite 10].

3.3 Overall Aim

The overall aim of the project was to increase the utilization of data in INDEPTH sites through interactive data visualization. This was aimed at improving hypotheses generation at these research sites as well as increase operational awareness.

3.4 Specific Objectives

The specific objectives of this research were:

1. To design a data visualization platform for the Africa Centre for Health and Population Studies (ACHPS).
2. To build a data visualization platform for ACHPS in order to increase data utilization and hypotheses generation at the site
3. To create a developer manual for data visualization so that the process for building the platform can be reproduced.
4. iv. To evaluate the usability of the developed platform for easy integration into the operational research cycle of the site

4

HDSS Sites Visualization Platform

This chapter gives a description of the the scope of the research, and overview of the solution, including a discussion on the three themes of Scientific Portal, Data Operations Portal and Community Engagement Portal. It goes on to describe in detail how the the application which serve the three themes were constructed, in terms of the tools, programming platforms and languages as well as relevant algorithms to replicate key artefacts.

Contents

4.1 Scope of Research

The scope of this research is limited to the integration, transformation and visualization of datasets from demographic surveillance data, verbal autopsy data [cite for VA] and socio-economic status data.

4.2 Overview of the Solution

The Africa Centre has embarked on a concept project dubbed Data Everywhere. Its aim is to increase the comprehension, access and utility of data collected through the use of a data visualization platform with 3 themes; Scientific Portal, Data Operations Portal and Community Engagement Portal. On site, this will be realised through the placement of three 52 inch touch screens in three strategic positions depending on the target audience. These touch screens will allow for users to visually interact with data on demand, selecting, filtering and visual feedback on being touched (active assimilation), as well as animate visualizations to show temporal trends when on standby mode, allowing for passive assimilation of potential insights. Additionally, the Scientific Portal and Community Engagement Portal will be hosted on a web server which allows access to their respective visualizations remotely from a browser.

4.2.1 Scientific Portal

This screen is placed strategically in the "Science Lounge" at the Africa Centre, a lounge area where Africa Centre scientists congregate for informal discussions and coffee breaks. It will allow scientists to either passively glean insights from the wall mounted 52 inch touch-screen as the animations show trends through time, or engage with the visualizations by directly interacting with the visualizations through selections, filtering and dynamic visual feedback.

The aim of this portal is to facilitate scientific discourse, insight generation and hypothesis formulation either serendipitously (passive) or through deliberate interaction with the visu-

alizations (active), amongst Africa Centre scientists using the lounge. Furthermore, as the Scientific Portal is web hosted, it allows for external scientists who potentially want to run or collaborate on studies at the Africa Centre to get a quick feel of not only what kind of data the Africa Centre currently has but also what the data is saying in a well packaged and easily accessible manner.

The solution builds on work already done by creating multi-panel graphs with dynamic display of data and interactive capabilities [cite 8]. These graphs weave together temporality and demographics and they use time-series plots, image plots and outcome pyramids.

This portal relies on Africa Centre's demographic surveillance data. The first step in developing the visualizations for this portal was to create Extract, Transform and Load (ETL) transformations using Pentaho Kettle. These create and store the dataset for each indicator by pulling data from a Microsoft SQL Server database. An additional transformation was produced for creating/updating a lookup file which links indicators to their respective datasets via a Uniform Resource Identifier (URI). These transformations are then integrated into a single Kettle Job. The motivation for this is flexibility, as any new indicators to be visualised in the future simply need a transformation for creating the appropriate dataset and a new entry in the lookup file linking the new indicator to its dataset. This ensures that no additional programming is required on the data visualization application with each new indicator to be visualised as long as the datasets stick to predefined structural and naming conventions.

4.2.2 Data Operations Portal

The Data Operations Portal screen is placed within the Data Centre. This is the office which deals with data collection, data entry, data cleaning, data quality assurance and data archiving. This will be a dashboard which aims to give real-time feedback on the progress of data collection activities, data entry activities and data archiving activities measured against the total number of data forms allocated for a particular survey round of three studies; Household, Individual

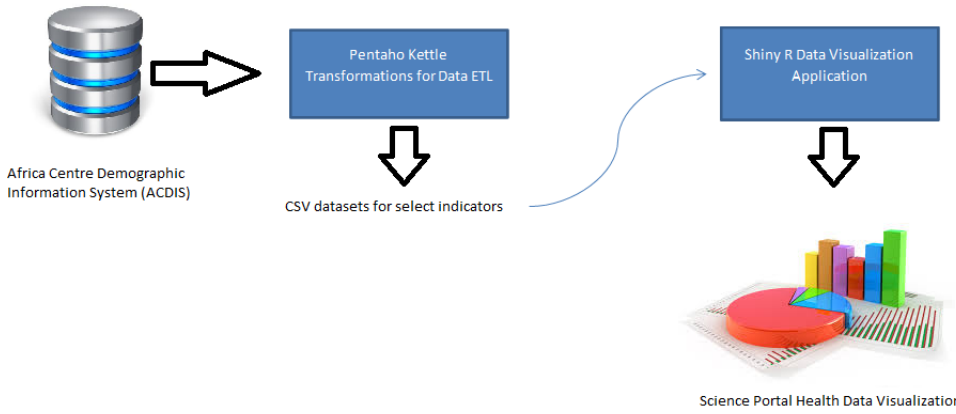

Figure 4.2.1: Overview of Scientific Portal solution

and Verbal Autopsy. This is in order to keep the Data Centre team abreast of their progress and operational bottlenecks in a transparent and accessible manner.

For this portal, data will be pulled directly from a Microsoft SQL Server via polling for changes in the underlying data every hour. This portal is not be externally accessible as it is purely for operational monitoring.

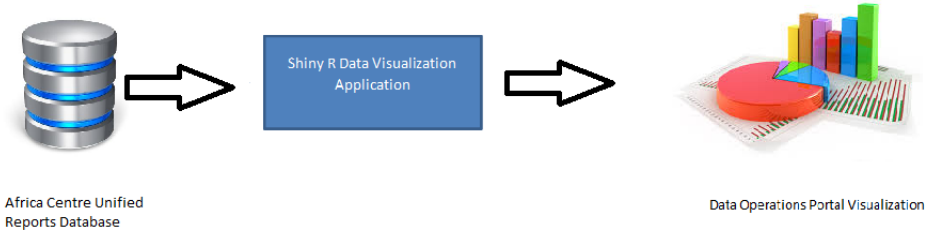

Figure 4.2.2: Overview of Data Operations Portal solution

4.2.3 Community Engagement Portal

The third and final screen is strategically placed in the Africa Centre foyer, visible and accessible to both staff and visitors. Its main focus is to package data which is of interest to the community which Africa Centre's research serves into visual representations that are easy

to interpret. This portal can be externally accessible over the internet for the community at large to access.

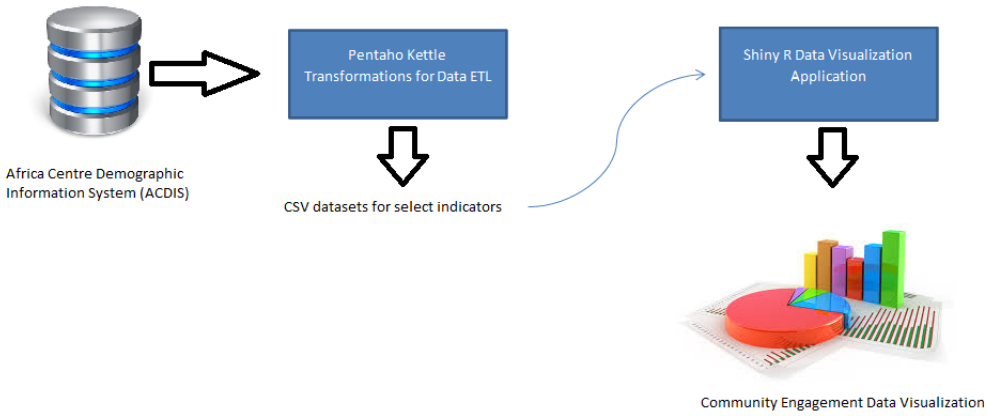

Figure 4.2.3: Overview of Community Engagement Portal solution

This research is about creating a generalizable data visualization solution for INDEPTH HDSS sites, and the output will be a software application which can visualize data at any of these sites. Different sites have differing data cleaning and data manipulation strategies and as such, the onus of ensuring that the data fed into the visualization platform is clean and has handled missing data falls on the data manager at the site. The only constant in the provided solution will be the software application; the ETL transformations and jobs at each INDEPTH site will have to be developed by the site data managers to handle each HDSS sites' database idiosyncrasies which are best known by the on site data manager. In order to do this, they will be informed by certain dataset structures and conventions which we have documented and are elaborated on in chapter 5

5

Chapter Title

This chapter illustrates xxx

Contents

5.1 Sub Title

6

Chapter Title

This chapter describes xx

Contents

6.1 Sub Title 1

6.2 Sub Title 2

7

Conclusion and Future Directions

This chapter concludes the thesis and gives roadmap for future work

Contents

A

Appendix

A.1 xxx

References

[1] J. J. Thomas, K. Cook *et al.*, "A visual analytics agenda," *Computer Graphics and Applications, IEEE*, vol. 26, no. 1, pp. 10–13, 2006.

[2] K. A. Cook and J. J. Thomas, "Illuminating the path: The research and development agenda for visual analytics," Pacific Northwest National Laboratory (PNNL), Richland, WA (US), Tech. Rep., 2005.

[3] A. Sopan, A. S.-I. Noh, S. Karol, P. Rosenfeld, G. Lee, and B. Shneiderman, "Community health map: A geospatial and multivariate data visualization tool for public health datasets," *Government Information Quarterly*, vol. 29, no. 2, pp. 223–234, 2012.