

**UNIVERSITY OF THE WITWATERSRAND**

**FACULTY OF HEALTH SCIENCES**

**SCHOOL OF PUBLIC HEALTH**

**MSC EPIDEMIOLOGY IN RESEARCH DATABASE MANAGEMENT**

**TITLE:**

**INTEGRATED LONGITUDINAL DATA VISUALIZATION AT INDEPTH NETWORK SITES**

**TUMAINI KILIMBA**

**STUDENT ID: 969326**

**INTERNAL SUPERVISOR: GIDEON NIMAKO**

**EXTERNAL SUPERVISOR: KOBUS HERBST**

CONTENTS

[1. INTRODUCTION 3](#_Toc418577821)

[1.1. PROBLEM STATEMENT 4](#_Toc418577822)

[1.2. MOTIVATION 4](#_Toc418577823)

[2. LITERATURE REVIEW 5](#_Toc418577824)

[3. RESEARCH AIMS AND OBJECTIVES 7](#_Toc418577825)

[3.1. OVERALL AIM 7](#_Toc418577826)

[3.2. SPECIFIC OBJECTIVES 7](#_Toc418577827)

[4. RESEARCH METHODS 8](#_Toc418577828)

[4.1. SCOPE OF RESEARCH 8](#_Toc418577829)

[4.2. OVERVIEW OF THE SOLUTION 8](#_Toc418577830)

[5. EXPERIMENTAL SETUP, MODELS AND VALIDATION 12](#_Toc418577831)

[5.1. EXPERIMENTAL ENVIRONMENT 12](#_Toc418577832)

[5.2. DATA SOURCES 16](#_Toc418577833)

[5.3. SOFTWARE ENGINEERING PRACTICES AND PUBLIC ACCESSIBILITY 16](#_Toc418577834)

[6. RESULTS 17](#_Toc418577835)

[7. DISCUSSION 17](#_Toc418577836)

[8. ETHICAL CONSIDERATIONS 17](#_Toc418577837)

[9. PROJECT MANAGEMENT, RISK MANAGEMENT AND BUDGET 18](#_Toc418577838)

[10. SUMMARY 18](#_Toc418577839)

# INTRODUCTION

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 [3].

VA is referred to as “the science of analytical reasoning facilitated by interactive visual interfaces” [1]. 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 [4–7].

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

## 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 [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.

## 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 a data visualization [10].

# LITERATURE REVIEW

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 [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 [10].

Of crucial importance when dealing with large datasets is the need for the users of the data be they scientists and other stakeholders to be able to discover the relations among and between the results of data analyses and queries [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 [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 may be the breakthrough in resolving the scientific visualization bottleneck.

A lot of work has been done on spatial visualizations for public health [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 [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.

Along a similar spatial theme, [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 there is 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 [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 [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 [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[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 [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 [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 [8] by incorporating dynamic display of data and interactivity into multi-panel graphs for epidemiological research [8].

# RESEARCH AIMS AND OBJECTIVES

## OVERALL AIM

The overall aim is to increase the utilization of data in INDEPTH sites through visualization, aimed at improving hypotheses generation and increase operational awareness.

## SPECIFIC OBJECTIVES

The specific objectives of this research are:

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. To evaluate the usability of the developed platform for easy integration into the operational research cycle of the site.

# RESEARCH METHODS

## SCOPE OF RESEARCH

The scope of the intended research is limited to the integration, transformation and visualization or datasets from clinical research data, demographic surveillance data and electronic health record data (EHR).

## 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.

*Scientific Portal*

This screen will be 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 visualizations (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 [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 will be 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.

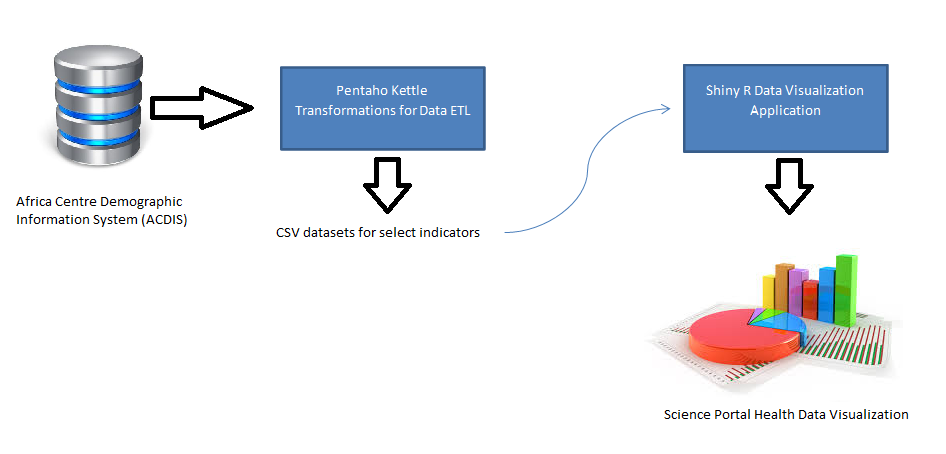


Figure : Scientific Portal Solution

*Data Operations Portal*

The Data Operations Portal screen will be 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 number of data forms allocated for a particular survey round, the number of forms at a particular stage of data management [elaborate on different stages], and archived (complete) versus un-archived (in progress) document counts of key studies. Additionally it will also estimate time to reach target status (archived status) given the historical rate of transition from non-archived to an archived state. 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 five minutes.

This portal shall not be externally accessible as it is purely for operational monitoring.

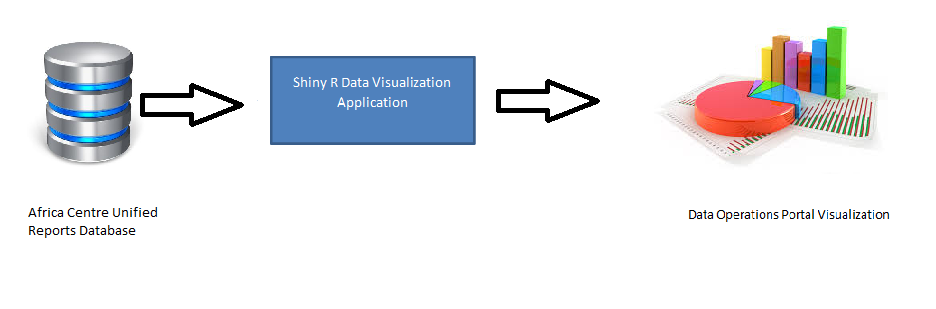


Figure : Data Operations Portal

*Community Engagement Portal*

The third and final screen will be 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 will be externally accessible over the internet for the community at large to access.

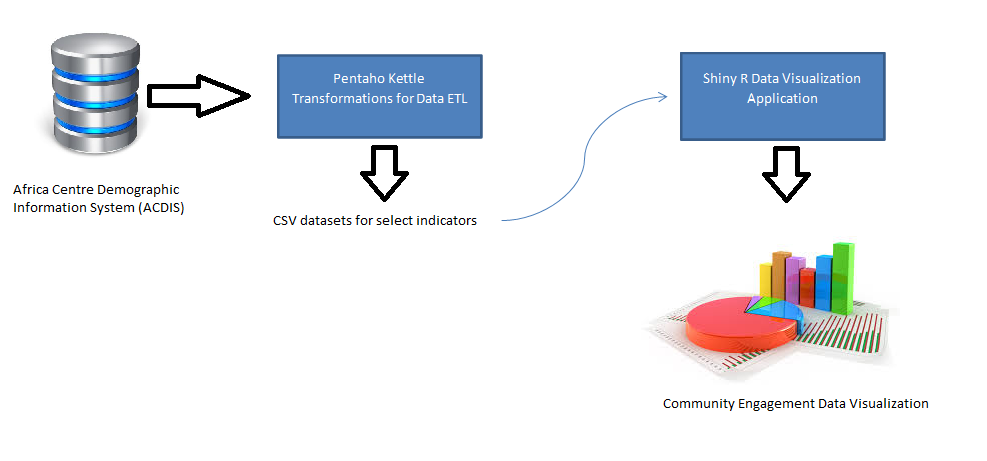


Figure : Community Engagement Portal

The Extract-Transform-Load steps will be automated to allow streaming of data from the various data sources into the visualization database. This will ensure that the visualizations stay current and in the case of the Data Operations Portal, will facilitate real time monitoring.

Individual HDSS sites have their data cleaning and validation procedures which handle the issues of dirty and missing data. 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 local 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.

# EXPERIMENTAL SETUP, MODELS AND VALIDATION

In this section we look at how we set up our proposed solution, so that it can be replicated, as well as how we validate the framework developed.

## EXPERIMENTAL ENVIRONMENT

At the beginning of the project, indicators were identified with the help of domain experts on site which informed the initial backbone of the Science Portal, the Operations Portal and the Community Engagement data visualization modules.

### Scientific Portal: Data

The first stage of developing the Scientific Portal involved developing the ETL data transformations which generate the datasets for the indicators that were identified initially as proof of concept indicators. These are Population Structure (Residents), Population Structure (Residents and Non-Residents), All-Cause Mortality Rates (Residents), All-Cause Mortality Rates (Residents and Non-Residents), and HIV Prevalence Rates. This was done using Pentaho Data Integration 5.3 (Pentaho Kettle) and it was chosen as it has been proven to support all aspects of Business Intelligence [15,16]. This is ideal as it creates opportunities for future expansion into other types of data analytics, such as data mining and predictive analysis, which may interest the INDEPTH sites in the future. Pentaho is also the tool used to generate the datasets used in INDEPTH’s iShare data repository [17], and data managers from 17 HDSS sites have currently already been trained on using the tool with more expected to be trained in the future.

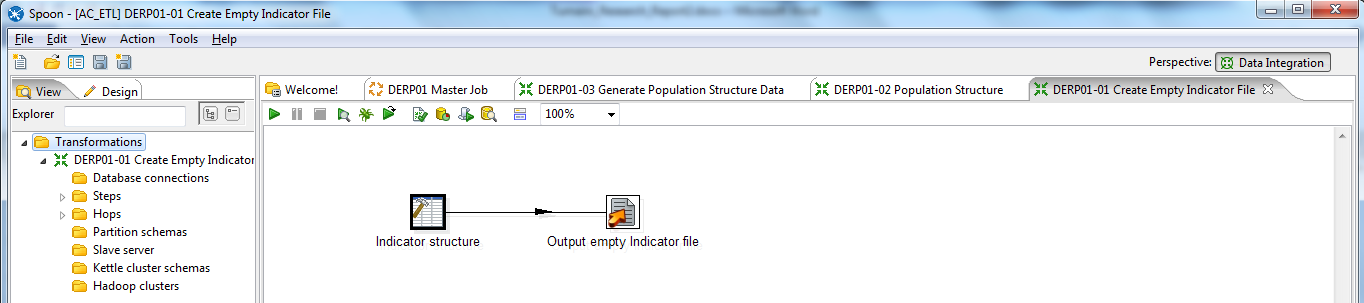


Figure : Pentaho Kettle transformation which creates an empty indicator mapping file. It will eventually have entries which maps an indicator to the URI of a dataset file on a file system

Figure 4 shows the first Pentaho Kettle transformation we create which is a simple two-step process which when run, results in the creation of an empty indicator mapping with the fields label, file, rate, multiplier and description. These field names are conventions and cannot be renamed as the application specifically looks for fields with those names (a functional requirement).

The label field stores the identifying name of the indicator.

The file field gives the URI of the indicators corresponding dataset.

The rate field is a Boolean field which indicates whether a specific indicator is a rate or not (Y for rate, N for not a rate).

The multiplier field gives the constant with which to multiply the result of the division between the numerator and the denominator (only in the case of rate indicators).

The description field stores a descriptive summary of the indicator which can include definitions, data collection schedules, data collection methods, data collection start and end dates or any other information which the data manager feels would aid in the overall understanding of the visualization of a particular indicator. The contents of this field are pulled into the data visualization interface as part of an information panel in the multipanel interface.

With the empty indicator mapping file created, we need to create a transformation to add an entry for each indicator we want to visualize to the mapping file. Figure 5 shows the data transformation which adds a new entry for the Population Structure indicator to the indicator mapping file. Figure shows the output of this transformation.

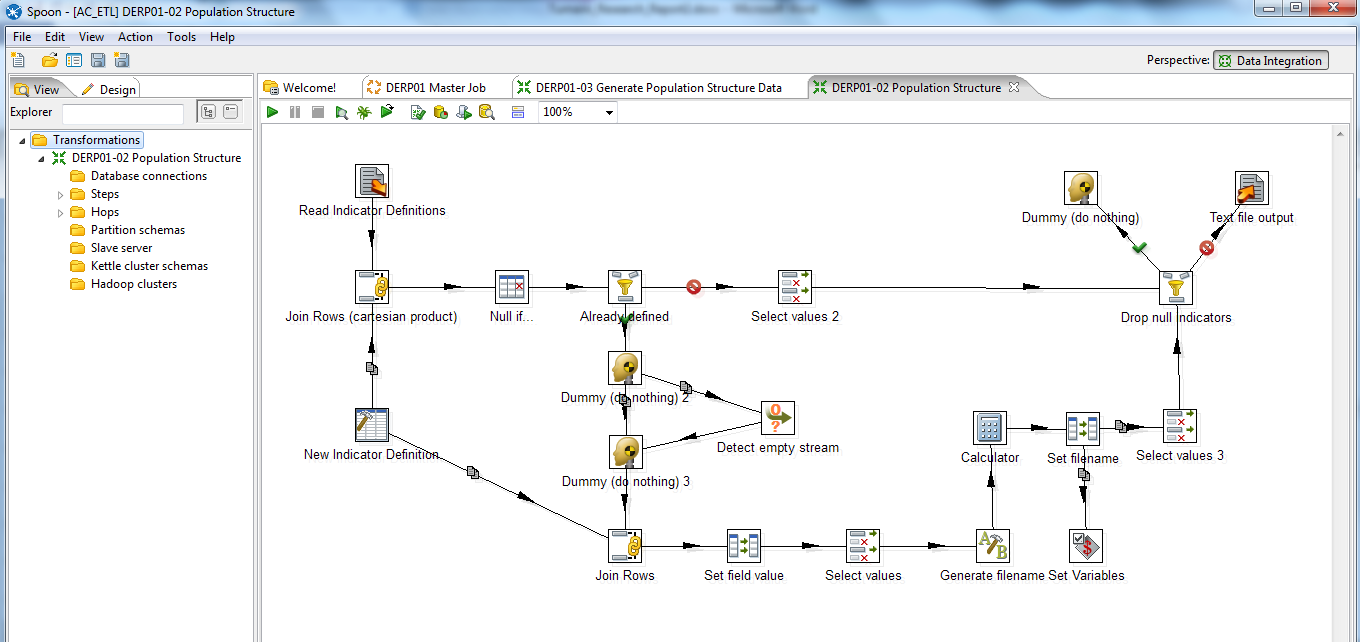


Figure : Pentaho Kettle Transformation which adds a new row (Population Structure) to the indicator mapping file.

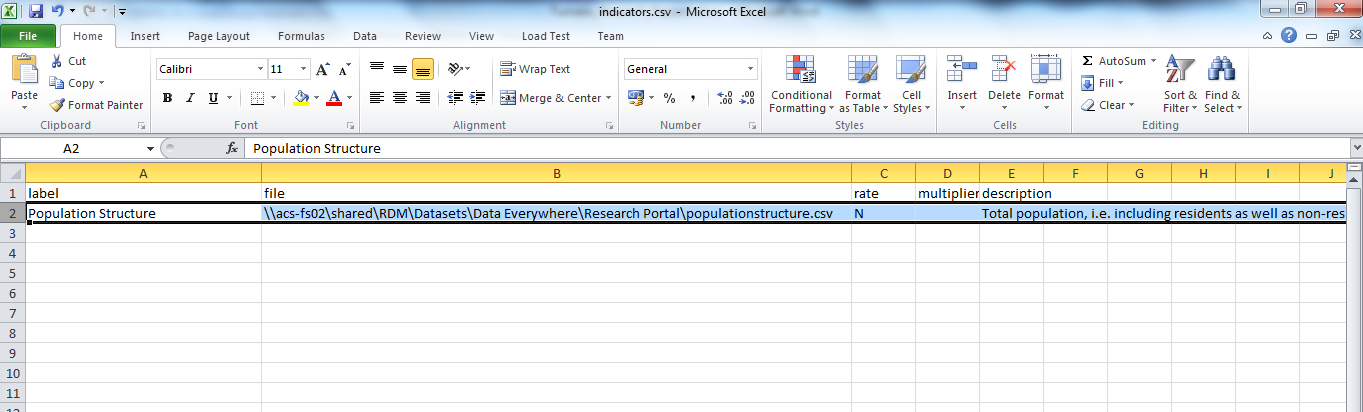


Figure : Mapping file which maps an indicator to its respective dataset URI on a file system

With the mapping file having an entry for the Population Structure indicator, all which is left is to create the actual dataset. This is done with yet another transformation (Figure 7). This transformation extracts the data relevant to the indicator of interest from the ACDIS database, transforms it to fit a predefined structural convention, and stores the dataset at the location specified in the mapping file entry created in the previous step. The structural convention is that the dataset must contain the following columns; year (observation year), agegrp (five year age group), sex (gender), denominator and numerator. For indicators which were not flagged as rates in the indicator mapping file, the numerator field will contain no values and the denominator field will store the absolute counts.

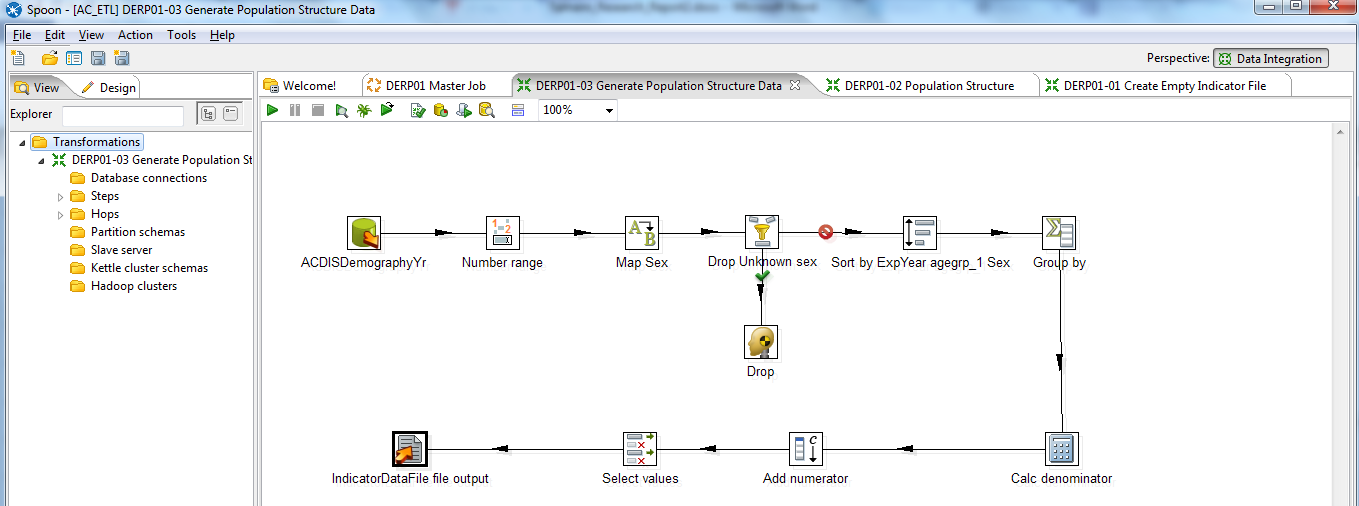


Figure : Transformation which creates the dataset for the Population Structure indicator

In summary, we have the following steps:

1. Create empty mapping file (CSV file) with the fields label, file, rate, multiplier and description.
2. Create a Pentaho Kettle transformation which adds an entry to that mapping file for any indicator of choice.
3. Create a Pentaho Kettle transformation which creates a dataset with the fields year, agegrp, sex, denominator and numerator, storing it in the file location indicated in the file field in the mapping file in step 2.

For each new indicator to be added to the visualization platform, repeat steps 2 and 3. In our case, we have five indicators which were identified initially to create a proof of concept, Population Structure (Residents), Population Structure (Residents and Non-Residents), All-Cause Mortality Rates (Residents), All-Cause Mortality Rates (Residents and Non-Residents), and HIV Prevalence Rates. By repeating steps 2 and 3, we now have five transformations to add entries for each indicator on the mapping file and five transformations which create datasets for these indicators. Therefore, our mapping file has five entries, and our dataset folder has five datasets (see Figure 8 and 9).

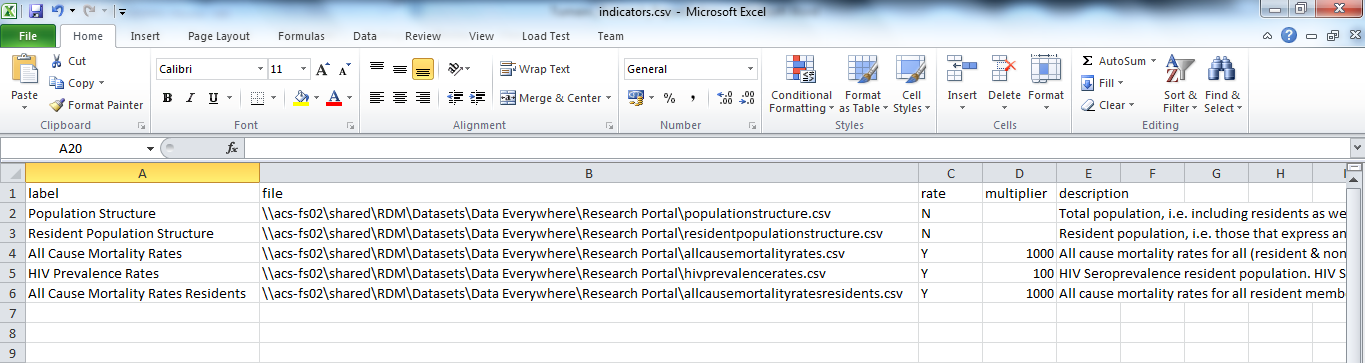


Figure : Mapping file, indicators.csv with five entries for indicators to be visualised

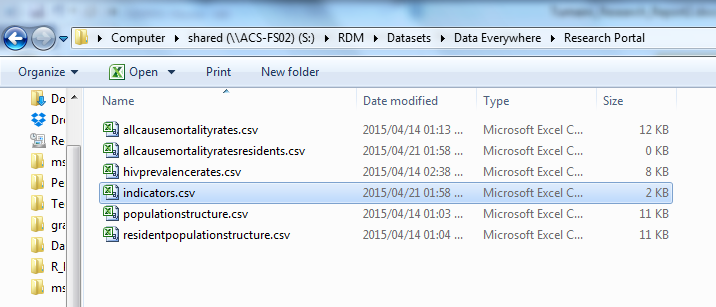


Figure : Datasets to visualize, all conforming to the same structural and field naming convention

Now that we have a mapping file with an entry for Population Structure (which gives us the label of the indicator, its location on a file system, its type, its multiplier and its description), and a corresponding data set all that is left is the application which renders the visualization.

### Scientific Portal: Application

Once the integration procedures are completed, the queries which were ran before integration on the original data sources are re-ran on the integrated data to verify that the pre-integration results and post integration results are the same. Additionally, the numbers being displayed for any particular outcome on the visualizations shall be verified independently by a domain expert for accuracy. By doing so, we shall assert data integrity and validity of the integration framework and the rendered visuals.

The visualization database will be on MySQL 5.0 [18] with the visualization rendering implemented in one of the following technologies: Pentaho Dashboard [19], D3.js [20], or Bokeh [21].

The proof of concept will run on a Lenovo T420 with 8 GB of RAM, an Intel Core i5 CPU at 2.5GHz, on a 64 bit operating system (OS). The OS will be Ubuntu 12.04 LTS (an open source OS).

Code development is a laborious process, and without a good revision control environment code changes can break the system and debugging can eat into development time. We will be using Git, an open source revision control and collaborative programming tool [22].

The onsite implementation set up will use the current available servers. The candidate server will have a minimum 16 GB of RAM and will run the Ubuntu 12.04 OS.

The data integration process requires validation. In order to reduce risk and guarantee that data has been transformed and integrated properly, it is important to have a quality assurance and validation strategy. Methods employed for data migration validation testing include sampling [23] and writing MINUS queries [24]. In sampling, the assumption is made that errors are uniformly distributed (which is not necessarily true). When sampling, a smaller subset from the larger dataset is selected and the information gathered from the smaller dataset will allow inferences to be made on the larger datasets data quality and validity. However, the sampling technique is highly inefficient, error prone and the final results are not entirely reliable [24].

Writing MINUS queries is a process which will allow us to ascertain what records are in the source dataset and have *not* been migrated to the target dataset. The tool which performs the ETL process will have validation rules which filter out invalid data from the source and write these error rows to a separate file. This means that when writing the minus queries we should be aware that some records will not be migrated as they break some validation rules, and as such cannot be considered a failure of the integration framework.

We will go one step further and run analytical statistics on the pre-integrated data sets, e.g. calculation of person-years in the DSS site between two defined periods. The same analyses will also be run post-integration to ascertain that the integration is not affecting key analyses. These analytical queries will be identified in the planning stage.

## DATA SOURCES

The crux of this research project is reliant on data, and lots of it. The scope of this research is limited to patient records (EMR data), longitudinal demographic surveillance data and clinical research/clinical trial data. Thus, in developing the proof of concept, copies of the databases which will be found at the INDEPTH site of implementation will be availed to us.

## SOFTWARE ENGINEERING PRACTICES AND PUBLIC ACCESSIBILITY

The development of software needs to be engineered and managed. As such, we will develop the project software using the Agile Methodology. The emphasis will be on delivering on small features in short iterations (weekly) with continual integration. Defensive programming will be employed in the form of Test Driven Development which has been shown to reduce production code defects by 40% without impacting on the team productivity [25].

Furthermore, the code base will be available on a Git server hosted at the University of the Witwatersrand. This will enable collaborative programming as well as a version control environment.

The production ready code base as well as all accompanying documentation will be hosted on GitHub, as an open source project for collaborative improvement of the integration framework.

Black box testing will be done with selected users and by ourselves with each new functionality added. White box testing will be accomplished by using Test-Driven Development to continuously and exhaustively test the code-base whenever a new functionality is integrated.

Dissemination of the project outcomes will be done mainly through publishing in peer reviewed journals, presentations at INDEPTH conferences as well as other relevant conferences.

# RESULTS

The platforms usability was assessed using the USE Questionnaire. This questionnaire measures usability in 3 dimensions, Usefulness, Satisfaction and Ease of Use using Likert scales per question. For internal systems it is suggested to split Ease of Use into Ease of Use and Ease of Learning [14]. In our case we measure responses in the 4 dimensions (see Questionnaire in Appendix C).

Results are described by calculating the median scores to give us a measure of central tendency for each of the 4 dimensions. An overall median score is also calculated. Non-parametric statistical tests for paired data (Wilcoxon signed rank test) was used to assess the data for changes from first use to last use.

# DISCUSSION

# ETHICAL CONSIDERATIONS

Permission will be obtained from the Africa Centre for Health and Population Studies to implement the system. All individual level data which is going to be used during the research, design and implementation process will be de-identified before being made available to us, giving us no access to identifiers. Additionally, the data visualizations themselves will be on aggregate level data and no individual level data is exposed to the end user.

# PROJECT MANAGEMENT, RISK MANAGEMENT AND BUDGET

In software projects without a comprehensive risk management plan, there is a heightened risk of failure, running over budget, or running off schedule. The table in Appendix A shows potential risks as well as the corresponding mitigation strategies. Appendix B shows the project timelines.

In terms of budget, the entire project uses open source tools. Furthermore, all the necessary hardware and software is already available for developing the proof of concept. For site implementation, the infrastructure and servers already exist. As such the project requires no additional budget.

# SUMMARY

Crucial to the success of public health surveillance is the ability to analyse and present data so that it is understandable to a variety of public health stakeholders. In 2009 the CDC declared the visualization of public health data as one of its six areas of focus for the improvement of public health surveillance in the 21st century. The current under-visualization of HDSS data needs to be addressed with targeted efforts and this research will form the foundation towards the use of data visualizations in these sites as part of their scientific process rather than as an end-product. The work done by Chui *et al* will be improved upon by the addition of interactivity to the visualizations as well as dynamism and spatial dimension.

**BIBLIOGRAPHY**

1. Thomas J, Cook K. A visual analytics agenda. Comput Graph Appl … [Internet]. 2006 [cited 2014 Sep 28];(February):10–3. Available from: http://ieeexplore.ieee.org/xpls/abs\_all.jsp?arnumber=1573625

2. Thomas J, Cook K. Illuminating the path: The research and development agenda for visual analytics. 2005 [cited 2014 Sep 28]; Available from: http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Illuminating+the+Path.+The+Research+and+Development+Agenda+for+Visual+Analytics#0

3. Bellucci E, Boo YL, Nguyyen L. Data Visualisation in Healthcare Analytics | Visual Analytics [Internet]. [cited 2014 Oct 3]. Available from: http://visanalytics.org/info/projects/data-visualisation-in-health/

4. Sopan A, Noh AS-I, Karol S, Rosenfeld P, Lee G, Shneiderman B. Community Health Map: A geospatial and multivariate data visualization tool for public health datasets. Gov Inf Q [Internet]. Elsevier Inc.; 2012 Apr [cited 2014 Oct 3];29(2):223–34. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0740624X12000159

5. Scotch M, Parmanto B, Gadd CS, Sharma RK. Exploring the role of GIS during community health assessment problem solving : experiences of public health professionals. 2006;10:1–10.

6. Suryahadi A (SMERU), Widyanti W (SMERU), Perwira D (SMERU), Sumarto S (SMERU), Elbers C (Vrije U, Menno P (World B. Developing a Poverty map for Indonesia: an initiatory work in three provinces [Internet]. 2003 [cited 2014 Sep 28]. Available from: http://siteresources.worldbank.org/ASEM/Resources/Poverty+Map+Indonesia+I.pdf

7. Stensgaard A, Saarnak C. Virtual globes and geospatial health: the potential of new tools in the management and contol of vector-borne diseases. Geospatial … [Internet]. 2009 [cited 2014 Sep 28];3(2):127–41. Available from: http://www.geospatialhealth.unina.it/articles/v3i2/gh-v3i2-02-stensgaard.pdf

8. Chui K, Wenger J, Cohen S, Naumova E. Visual analytics for epidemiologists: understanding the interactions between age, time, and disease with multi-panel graphs. PLoS One [Internet]. 2011 Jan [cited 2014 Oct 3];6(2):e14683. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3039641&tool=pmcentrez&rendertype=abstract

9. Report MW. CDC’s Vision for Public Health Surveillance in the 21st Century. Morb Mortal Wkly Rep. 2012;61.

10. Fox P, Hendler J. Changing the equation on scientific data visualization. Science(Washington) [Internet]. 2011 [cited 2014 Sep 28];705(2011). Available from: http://data2discovery.org/dev/wp-content/uploads/2013/05/Fox-and-Hendler\_Visualization\_Science-2011-Fox-705-8.pdf

11. Parmanto B, Paramita M. Spatial and multidimensional visualization of Indonesia’s village health statistics. … J Heal … [Internet]. 2008 [cited 2014 Sep 28];14:1–14. Available from: http://www.biomedcentral.com/1476-072X/7/30

12. Verdegem P, Verleye G. User-centered E-Government in practice: A comprehensive model for measuring user satisfaction. Gov Inf Q [Internet]. Elsevier Inc.; 2009 [cited 2014 Oct 3];26(3):487–97. Available from: http://dx.doi.org/10.1016/j.giq.2009.03.005

13. Carroll LN, Au AP, Detwiler LT, Fu TC, Painter IS, Abernethy NF. Visualization and analytics tools for infectious disease epidemiology: A systematic review. J Biomed Inform [Internet]. Elsevier Inc.; 2014;51:287–98. Available from: http://dx.doi.org/10.1016/j.jbi.2014.04.006

14. Lund AM. Measuring Usability with the USE Questionnaire [Internet]. Usability and User Experience. 2001 [cited 2014 Oct 30]. Available from: http://www.stcsig.org/usability/newsletter/0110\_measuring\_with\_use.html

15. Tuncer O, Berg J van den. Implementing BI concepts with Pentaho, an evaluation. Delft Univ Technol Netherlands [Internet]. 2010 [cited 2014 Sep 28]; Available from: http://www.enixe.nl/upload/productvergelijk/pentahoevaluation.pdf

16. Pentaho Community [Internet]. [cited 2014 May 11]. Available from: http://community.pentaho.com/

17. iSHARE Repository [Internet]. [cited 2014 Oct 8]. Available from: http://www.indepth-ishare.org/index.php/home

18. MySQL :: The world’s most popular open source database [Internet]. [cited 2014 Sep 28]. Available from: http://www.mysql.com/

19. Community Dashboard Framework | Webdetails • Business Analytics [Internet]. [cited 2014 Sep 28]. Available from: http://www.webdetails.pt/ctools/cdf.html

20. D3.js - Data-Driven Documents [Internet]. [cited 2014 Sep 28]. Available from: http://d3js.org/

21. Welcome to Bokeh — Bokeh 0.6.1 documentation [Internet]. [cited 2014 Sep 28]. Available from: http://bokeh.pydata.org/

22. Git [Internet]. [cited 2014 May 10]. Available from: http://git-scm.com/

23. Manjunath T. A study on sampling techniques for data testing. Int J Mod Eng Res [Internet]. 2012 [cited 2014 May 11];3(1):13–6. Available from: http://csjournals.com/IJCSC/PDF3-1/Article\_3.pdf

24. Paygude P, Devale PR. Automated Data Validation Testing Tool for Data Migration Quality Assurance. Int J Mod Eng Res. 2013;3(1):599–603.

25. Williams L. Test-driven development as a defect-reduction practice. Proc 14th Int Symp Softw Reliab Eng [Internet]. 2003 [cited 2014 May 11];1–12. Available from: http://ieeexplore.ieee.org/xpls/abs\_all.jsp?arnumber=1251029

APPENDIX A

| **Risk** | **Risk Level**  **Low(L)/Medium(M)/High(H)** | **Likelihood of Event** | **Mitigation Strategy** |
| --- | --- | --- | --- |
| Limitations in technical skills in the project team | **M:** Project assumes that technical skills are in place | Somewhat likely | Relevant skills to be acquired in the 2 month period prior to project start |
| Project running off schedule | **M:** Project is tied to a hard deadline | Somewhat likely | Using Agile methodology to break project artifacts into features to be delivered on weekly schedules. A one month buffer window will be incorporated to cover any unforseen schedule overruns. Monthly face to face with the primary supervisor will also ensure oversight. |
| Non availability of data sources | **M:** assumes no derailment | Somewhat Unlikely | The criteria for choosing the implementation site includes availability of the required data for integration |
| Change of supervision | **L:** Well-defined | Unlikely | N/A |
| Acquiring cartographic boundaries (for map based visualizations) from the relevant authorities may not be straightforward or timely | **M:** Assumes cartographic data already on site | Somewhat likely | Communicate with site supervisor early to know whether this data is available. |

APPENDIX B

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Activities  2014-2015 | October | November | December | January | February | March | April | May | June |
| Protocol submission to Wits Ethics Committee |  |  |  |  |  |  |  |  |  |
| Receive Standard Datasets from INDEPTH site |  |  |  |  |  |  |  |  |  |
| Creation of analytical queries for pre and post integration for validation purposes |  |  |  |  |  |  |  |  |  |
| Creation of Data Cleaning scripts to be used during ETL processes |  |  |  |  |  |  |  |  |  |
| Set up ETL solution |  |  |  |  |  |  |  |  |  |
| Writing data integrity programs for post integration validation |  |  |  |  |  |  |  |  |  |
| Create visualization engine |  |  |  |  |  |  |  |  |  |
| Document generalizable data visualization implementation processes |  |  |  |  |  |  |  |  |  |
| Write research report |  |  |  |  |  |  |  |  |  |
| Submission of research report |  |  |  |  |  |  |  |  |  |
| Prepare articles for submission to peer reviewed journals |  |  |  |  |  |  |  |  |  |

APPENDIX C – USE Questionnaire

**Usefulness**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1. It helps me be more effective | | | | | | | | |
| strongly disagree | 1 | 2 | 3 | 4 | 5 | 6 | 7 | strongly agree |
| □ | □ | □ | □ | □ | □ | □ |
| 2. It helps me be more productive | | | | | | | | |
| strongly disagree | 1 | 2 | 3 | 4 | 5 | 6 | 7 | strongly agree |
| □ | □ | □ | □ | □ | □ | □ |
| 3. It is useful | | | | | | | | |
| strongly disagree | 1 | 2 | 3 | 4 | 5 | 6 | 7 | strongly agree |
| □ | □ | □ | □ | □ | □ | □ |
| 4. It makes the things I want to accomplish easier to get done | | | | | | | | |
| strongly disagree | 1 | 2 | 3 | 4 | 5 | 6 | 7 | strongly agree |
| □ | □ | □ | □ | □ | □ | □ |
| 5. It saves me time when I use it | | | | | | | | |
| strongly disagree | 1 | 2 | 3 | 4 | 5 | 6 | 7 | strongly agree |
| 6. It meets my needs | | | | | | | | |
| strongly disagree | 1 | 2 | 3 | 4 | 5 | 6 | 7 | strongly agree |
| 7. It does everything I would expect it to do | | | | | | | | |
| strongly disagree | 1 | 2 | 3 | 4 | 5 | 6 | 7 | strongly agree |

**Ease of Use**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 8. The system is easy to use | | | | | | | | |
| strongly disagree | 1 | 2 | 3 | 4 | 5 | 6 | 7 | strongly agree |
| □ | □ | □ | □ | □ | □ | □ |
| 9. It is simple to use | | | | | | | | |
| strongly disagree | 1 | 2 | 3 | 4 | 5 | 6 | 7 | strongly agree |
| □ | □ | □ | □ | □ | □ | □ |
| 10. It is user friendly | | | | | | | | |
| strongly disagree | 1 | 2 | 3 | 4 | 5 | 6 | 7 | strongly agree |
| □ | □ | □ | □ | □ | □ | □ |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 11. It requires the fewest steps possible to accomplish what I want to do with it | | | | | | | | |
| strongly disagree | 1 | 2 | 3 | 4 | 5 | 6 | 7 | strongly agree |
| □ | □ | □ | □ | □ | □ | □ |
| 12. It is flexible | | | | | | | | |
| strongly disagree | 1 | 2 | 3 | 4 | 5 | 6 | 7 | strongly agree |
| □ | □ | □ | □ | □ | □ | □ |
| 13. Using it is effortless | | | | | | | | |
| strongly disagree | 1 | 2 | 3 | 4 | 5 | 6 | 7 | strongly agree |
| □ | □ | □ | □ | □ | □ | □ |
| 14. I feel comfortable using this system | | | | | | | | |
| strongly disagree | 1 | 2 | 3 | 4 | 5 | 6 | 7 | strongly agree |
| □ | □ | □ | □ | □ | □ | □ |
| 15. The organization of information on the system screen is clear | | | | | | | | |
| strongly disagree | 1 | 2 | 3 | 4 | 5 | 6 | 7 | strongly agree |
| □ | □ | □ | □ | □ | □ | □ |
| 16. Overall, I am satisfied with how easy it is to use this system | | | | | | | | |
| strongly disagree | 1 | 2 | 3 | 4 | 5 | 6 | 7 | strongly agree |
| □ | □ | □ | □ | □ | □ | □ |
| 17. I am able to efficiently complete tasks using this system | | | | | | | | |
| strongly disagree | 1 | 2 | 3 | 4 | 5 | 6 | 7 | strongly agree |
| □ | □ | □ | □ | □ | □ | □ |

**Ease of Learning**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 18. I learned to use it quickly | | | | | | | | |
| strongly disagree | 1 | 2 | 3 | 4 | 5 | 6 | 7 | strongly agree |
| □ | □ | □ | □ | □ | □ | □ |
| 19. I easily remember how to use it | | | | | | | | |
| strongly disagree | 1 | 2 | 3 | 4 | 5 | 6 | 7 | strongly agree |
| □ | □ | □ | □ | □ | □ | □ |
| 20. It is easy to learn to use it | | | | | | | | |
| strongly disagree | 1 | 2 | 3 | 4 | 5 | 6 | 7 | strongly agree |
| □ | □ | □ | □ | □ | □ | □ |

**Satisfaction**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 21. I am satisfied with it | | | | | | | | |
| strongly disagree | 1 | 2 | 3 | 4 | 5 | 6 | 7 | strongly agree |
| □ | □ | □ | □ | □ | □ | □ |
| 22. It works the way I want it to work | | | | | | | | |
| strongly disagree | 1 | 2 | 3 | 4 | 5 | 6 | 7 | strongly agree |
| □ | □ | □ | □ | □ | □ | □ |
| 23. It is pleasant to use | | | | | | | | |
| strongly disagree | 1 | 2 | 3 | 4 | 5 | 6 | 7 | strongly agree |
| □ | □ | □ | □ | □ | □ | □ |
| 24. I feel I need to have it | | | | | | | | |
| strongly disagree | 1 | 2 | 3 | 4 | 5 | 6 | 7 | strongly agree |
| □ | □ | □ | □ | □ | □ | □ |
| 25. Overall, I am satisfied with this system | | | | | | | | |
| strongly disagree | 1 | 2 | 3 | 4 | 5 | 6 | 7 | strongly agree |
| □ | □ | □ | □ | □ | □ | □ |

26. What would you add/remove from to the system in order to improve it?

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_