A Framework for Adaptation of Virtual Data Enumeration for Enhancing Census – Tanzania Case Study

Ramadhani A. Duma
St. John's University of Tanzania, P.O.Box 47, Dodoma, Tanzania
Email: radsiffi@yahoo.com

Abstract—Population census is an enormous and challenging national exercise with many stakeholders whose participation is required at all levels of government or public administration. The problem of high cost in conducting the traditional census process imposes extra and unaffordable cost in most of the developing countries which resulted into ten-year defaulting for census enumerations. This challenge compels nations to seek for assistance mostly from various donors nations in every census enumerations exercise. Virtual Census enumerations play a vital role in demographic data enumerations since it does not require physical involvement in Enumeration Area as in traditional enumerations approach. In this paper the main focus is on data integration from different heterogeneous sources, addressing cleansing challenge for data integrated from data sources with no common key for integrations, building virtual data integration framework for enhancing virtual censuses enumeration process. The developed framework and algorithms can be used to guide design of any other data integration system that need to address similar challenges in related aspects. The outcome of this work is suitable and cheaper technique of demographic data enumeration as compared to traditional technique which involves a lot of manual works and processes.

Index Terms—Census, Virtual enumeration, Data integration, Heterogeneous data, Data cleansing, Data mining

I. INTRODUCTION

Census is an official state count of people carried out at specific time intervals mostly in years (example after every 5, 10, or 15 depending on the state plans) where data or information about the individual members of a given population is obtained. The data collected in census contain information about the number and characteristics of the population such as age structure, birth rate, death rate, population dispersion or spatial distribution. The enumeration exercise involves different activities such as data collection, analyzing, and publishing to the public [1].

The practical usefulness of census cannot be undermined, however, the practicality of taking census have been scared into doubt due to high costs of the operation; and this problem is reflected in widespread defaulting on the ten-year periodicity by a number of developing countries. An assessment of various censuses carried out in Africa has identified the following recurrent weaknesses in the process; the high cost of the census as a data collection method, dependence on external funding, long delays in the publication of the results, poor quality of the results in many instances, high cost /utility ratio [2].

In the majority of developing countries censuses are financed in part by their governments while the remaining part is provided with assistance from the international communities. Literature suggests that: sometimes in developing countries, governments tend to allocate unrealistic low budget for censuses with the expectation that international assistance will provide the remaining balance. There are other claims that many countries organize donors' meetings for this purpose, and the outcome is not always successful as per their expectations. Meanwhile, the population census schedule continues to be moved forward from it's defined period of time, as it happens in Tanzania. Delaying the census exercise can compromise the overall census objectives. It can also affect the decision making processes and negatively impact the quality of census operations. It is strongly believed that in order to get the best results, the allocations of the necessary resources for a census should be made available in advance of the census moment [2].

The population censuses in Tanzania dates back to 1910. However, the first scientific census took place in 1958. Five censuses have been conducted since Tanzanian independence, in 1967, 1978, 1988, 2002 and 2012. These dates indicates that Tanzania has not been able to maintain the census taking every ten years, the delay could be due to a number of reasons, the main factor being economic [3].

Most of Southern sub-Sahara African countries have included Information and Communication Technologies (ICTs) in the national development plan. Currently in Tanzania, ICT has been used in facilitating health services, agriculture, education and governance. The information warehouses could be used as source of

information for the implementation of Virtual census. In Tanzania for example, data from Tanzania Revenue Authority (TRA), National Identification Authority (NIDA) and Registration Insolvency and Trusteeship Agency (RITA) can be integrated for census purposes. Virtual census approach has been used in some European countries, especially Netherland and Finland. This approach relies on information collected from different data sources as shown in Fig. 1.
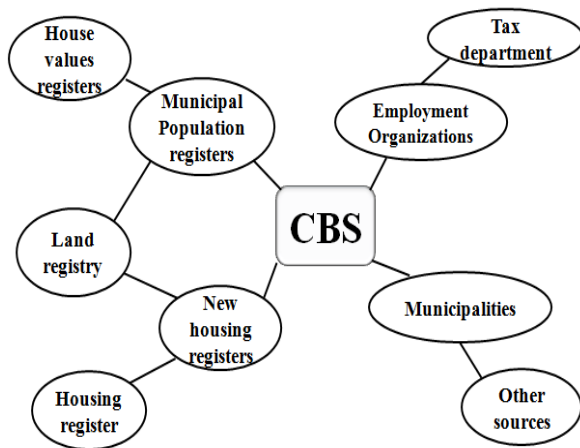

Fig.1. Central Bureau of Statistics (CBS): The Dutch Virtual Census Model. Source: [4].

Data integration provides additional and reliable results since they are from various data sources, these sources can be either independent or dependent from other sources. The data integration from multiple database systems has gained the attention of many researchers in recent, they attempt to logically integrate data from several independent distributed DBMSs, while allowing the local DBMSs to maintain complete control of their systems (i.e. all existing databases are autonomous) [5].

II. RELATED WORK

There are various works related to this study, in this section therefore those works are explained briefly in order to have a clear picture of what they achieved. Authors in [6] proposed a framework to integrate heterogeneous relational databases using Global View in which through view client can update remote sites by applying SQL query on global view. Integration methodology and creation of global view relies on mapping attributes such as unique keys for the relations. Their main focus was to provide a transparent access and manage distributed database systems on relational databases. However, the approach is not effective to be applied for the case of Tanzania since not all data sources have relational databases. In other data sources, data were stored in flat files and sometimes in hardcopies, taking example Local Government Authorities (LGA) as observed in Tanzania. Similar framework for data integration systems has been introduced in the paper [7] where data integration from multiple local component

databases using a proxy table technique in the multi-database system. The approach is good since it relies on mapping of attributes (common keys) between proxy table created and remote local table but it does not suite to Tanzanian data source participating during integration processes. This is because there is no common identification (ID) to identify every person in the country.

Sabri and Rahal [8] proposed a platform for service-oriented data mining to build a system that enables data extraction through predictive rules which produces a flexible and scalable system to aid in decision-making. The paper covers data integration from different sources where integrated data were stored in the global source; all relevant data were copied in the warehouse. Moreover authors explained data selection and cleaning processes, *data selection* by applying filters to select a subset of rows or columns and *cleaning* to process missing data.

For data integration systems prototype to be interoperable into different environments as observed during data collection phase; author found out that most of the data sources in Tanzania were operating on different platforms. Author of this work recommend the use of extensible markup language (XML) simply because it supports a number of key standards for character encoding. Also, XML is free of format in identifying information, data storage and its transfer in disparate data sources, which is the key behavior in any data integration framework. Authors in [9] have addressed well the issues of integrating and querying disparate heterogeneous information using unified XML views, but the paper does not explain anything on how to cleanse the data redundancy especially when participating systems (organs) has no common key for integrations.

To integrate data involves various components, the task of the author is to identify important components such *Data Sources*, *Processes* involved, *Data Exchange Design*, *Information Sharing Protocol*, *Access Mode* and *Access Technologies*. The most important part for data integration systems is a Data Source. Data sources are mostly developed for a particular function. In Tanzania for example, each organ (data source) stores information in a database with a different data model. It is a common practice that these kinds of data sources are accessed through different interfaces, protocols and languages. These data sources are autonomous (i.e. organs are independent, self-regulating and managing). It is the will of the organ to decide how contained information is stored. Moreover, they are independently deciding which other systems/ organs are allowed to communicate with them depending on their goal. Furthermore, every source does not always exist in the same host. They are on different hardware platforms and operating systems and can only be accessed through certain network protocols. During data collection phase, it was found that different potential data sources organs which deal with the personal data and some organs keep the same personal data which leads to data redundancy and duplicates when integrated.

Absence of common key to identify person between the participating organs /data sources was identified.

Every organ uses its' own key which means nothing in another organ. For example NIDA uses NIDA_ID as their unique key. RITA uses BNo as their unique key while ISD uses Passport_No. as their unique key [9]. This work also identify a dissimilarity in database schemas. In RITA schema for example all names (i.e. first name, surname and middle name) are in one column or field called "Names", but in other databases for example the NIDA every record is in its own field (i.e. first name, surname and middle name were separated). Not only these but also, it was noted that date format do not match, in NIDA they uses DD/MM/YYYY date format while in RITA they use DDDD, MMMM DD, YYYY and heterogeneity in DBMS.

Processing and functionalities: in practice data sources are either structured or unstructured. After the extraction process, the data should be transformed to change it to the appropriate structure. Transformation also involves combining data from several sources, generating aggregates, generating surrogate keys, data sorting, obtaining new calculated data values, and sometimes applying data validation constraints. The transformed data is now read to be loaded into the intended database for cleansing process as explained in Fig. 4. There must be connections to different data sources to make them accessible from a common logical data access point in which the results are combined from across multiple data sources systems. Finally, publish the result sets as views and /or data services executed by client application or users when requested [10].

Data Communication protocols are formal descriptions of rules which governs data /information exchange in participating data source system. Author recommends the most popular network communications protocol which is TCP/IP simply because I finds it best suite with the proposed framework. TCP/IP also has other protocols which are often packaged together with it such as File Transfer Program (FTP), Telnet for remote log-in capability which are also very suitable for the proposed framework [11].

Data exchange involves the process of taking data structured under a source schema and actually transforming it into data structured under a target schema, so that the target data is an accurate representation of the source data. During data integration data is restructured in data exchange. For this framework, data live at the source e.g. Registration, Insolvency and Trusteeship Agency (RITA), National Identification Authority (NIDA), Tanzania Revenue Authority (TRA), etc,. [12].

To establish an environment for strategic intelligence for this framework the following processes are necessary; setting up a strategic intelligence foundation which concerned with the establishment of a centralized data repository containing integrated data from distributed organs. Creating reports helps users /stakeholders to understand, interpret and monitor the information for further decision such as data analyst that uses data in the reports to identify the reason a particular condition occurred.

Data Access technologies: there is different standardized data access technologies developed to serve as interfaces between different database systems. Examples of the technologies are SQL, ODBC, JDBC, XQJ, ADO.NET, XML, XQuery, and the like. For this study data were compiled from different heterogeneous data source systems need to be accessed and later published on the web, reporting results to the different government authorities, calculating total number of populations, or creating internal reports for planning purposes. Author is interested in different elements of the data and expects it to be formatted and transformed according to the needs. [13]XQuery is a query language designed to address these needs. It allows you to select the XML data elements of interest, restructure them and return the results in a structure of your choice. Apart from the fact that XQuery helps in combining different databases and plain file documents, it is also flexible in querying and transforming collections of either structured or unstructured data. XQuery API for Java (XQJ) enables Java programmers to execute XQuery against an XML data source while eliminating vendor **lock in** since it is not a proprietary technology [14].

The proposed system is a web based system; it uses Internet web access technologies to deliver information and services to users or other information systems /applications. Web browser was typically used as the front-end with the database as the back-end. Server-side scripting is used to send queries to the database and translate answers to HTML. The main reason for choosing this mode is its ability to be accessed via a variety of platforms, in which web based applications are far more compatible across different platforms than traditional installed software. One of the main requirements for web based systems is a web browser which enables users to access web based applications. Some of these includes: Internet Explorer, Firefox and Netscape. These web browsers are available to a number of operating systems (i.e., Windows, Linux, Mac OS, and so on) [15].

III. PROPOSED APPROACH

In order to solve the mentioned challenges, author surveyed various potential data sources and literatures. The information obtained was used to design a framework that was later used to develop a Virtual Census Framework as illustrated in Fig. 2 to represent the designed theoretical framework. The components in the framework include data sources, processes (including aggregation algorithms such as algorithm 1 and 2), data exchange design, information sharing and communication protocol, access mode and technologies, network topology.
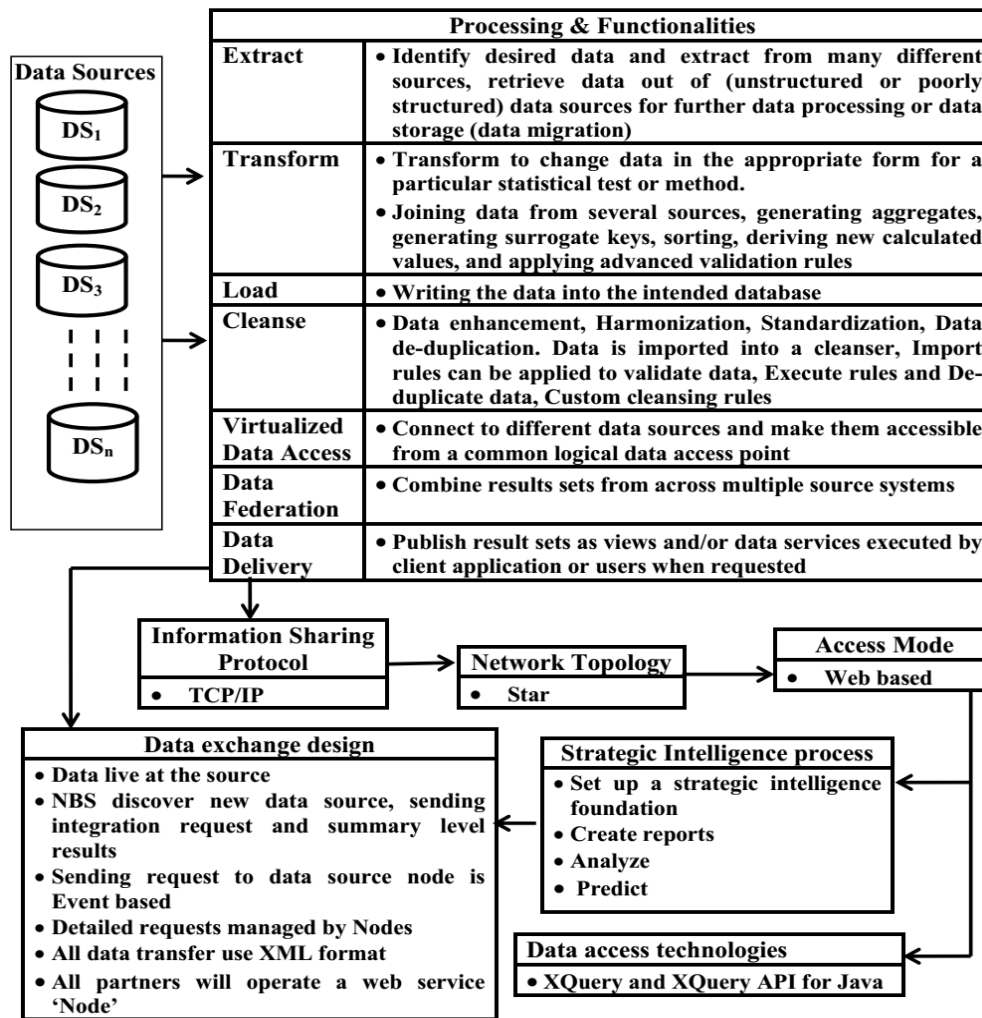
Data Sources	Processing & Functionalities	
DS₁ DS₂ DS₃ ||| DSₙ	Extract	• Identify desired data and extract from many different sources, retrieve data out of (unstructured or poorly structured) data sources for further data processing or data storage (data migration)
	Transform	• Transform to change data in the appropriate form for a particular statistical test or method. • Joining data from several sources, generating aggregates, generating surrogate keys, sorting, deriving new calculated values, and applying advanced validation rules
	Load	• Writing the data into the intended database
	Cleanse	• Data enhancement, Harmonization, Standardization, Data de-duplication. Data is imported into a cleanser, Import rules can be applied to validate data, Execute rules and De-duplicate data, Custom cleansing rules
	Virtualized Data Access	• Connect to different data sources and make them accessible from a common logical data access point
	Data Federation	• Combine results sets from across multiple source systems
	Data Delivery	• Publish result sets as views and/or data services executed by client application or users when requested

Information Sharing Protocol
- TCP/IP

Network Topology
- Star

Access Mode
- Web based

Data exchange design
- Data live at the source
- NBS discover new data source, sending integration request and summary level results
- Sending request to data source node is Event based
- Detailed requests managed by Nodes
- All data transfer use XML format
- All partners will operate a web service 'Node'

Strategic Intelligence process
- Set up a strategic intelligence foundation
- Create reports
- Analyze
- Predict

Data access technologies
- XQuery and XQuery API for Java

Fig.2. Virtual Census Framework

Table 1. Tanzanian Virtual Census Framework Components for Designed Framework

Component	Description
Data sources	Heterogeneous and autonomous data storage centers or organs. Found primary data sources for Tanzanian Virtual Census enumerations are NBS, NIDA, RITA, Local Government Authority (LGA), NGOs, Public Services Management (PSM), Private organizations and Immigration Service Department (ISD).

Processing and Functionalities	Main processes involved are data extraction, transformation, loading, cleansing, virtualization of data access, data federation and delivery.
	    Data extraction and transformation into the appropriate structure for a particular arithmetical test.
	    Loading the transformed data into the intended database.
	    Data is then imported into a cleanser, import rules can be applied for validation.
	    Finally, publishing processed result set.

Data exchange design	Data live in the source NBS identify suitable source and sends integration requests Request is event based and details for requests is managed by Nodes All data transfer use XML format
Information sharing protocol	TCP/IP to remote servers
Topology	Simple Star Topology single hop Communication. It includes the data source organ and destination organ (NBS) for data storage and processing
Access mode	Web based
Strategic intelligence process	Data analysis and reports
Data access technologies	XQuery and SDO

This framework (i.e Virtual Census Framework), involves different defined steps proposed as the result of the findings. The following are the main steps in order to meet the required goal; Identification of the required data source is the first thing. In Tanzania the organ responsible for census activity is the National Bureau of Statistics (NBS), the researcher recommends NBS to act as an identifier of the new data sources, retrieving data out of data sources for further processing and storage. The organ NBS should discover or suggest new data source to participate during the integration process and sending integration request to any discovered data source. Request to data source node is event based (i.e. the data will be received only when requested). All data transfers use XML format as described in algorithm 2. The summary of the proposed Virtual Census Framework's components identified is shown in the Table 1.

By using the designed framework, the Virtual Census enumeration model was developed. In the development of the system to represent the framework, the prototyping method was used. This method was used in order to produce high quality system. Key advantage of the prototyping model was to assure the clients that the development team was working on the system and the user can be well informed about the unrealistic requirements. Fig. 3 shows the basic steps in the prototyping model employed by the author.

Initially, author collected requirements of the proposed tool as per expectations in which these requirements were subject to change as new ideas appear. Collected data should be analyzed to meet objectives where all conflicting requirements were addressed. It includes requirement documentation and examining if the identified target or needs were met. Upon successful requirement collections and analysis followed the system design which was the step before starting to write the codes, the author finds it necessary to have a complete understanding of the system to be developed through designing the architecture. For this purpose the system design was developed according to the requirement specification document. Coding and Implementation, System implementation uses the structure created during architectural design to meet the stakeholders' requirements and tool requirements developed in the early phases.

These system elements are then integrated to form intermediate aggregates, and finally the complete Virtual Census Enumeration Tool was ready to be tested by clients. In the Testing phase, the tool was given to various stakeholders and experts for testing how the system works as a whole and whether it meets the requirements. During user testing, there were two options which are YES or NO options (as shown in Fig. 3): If satisfied with design and functionalities the system is complete BUT, if not, they should give comments. Users' comments or suggestions act as the new requirements. Complete Build: This phase was similar to implementation phase; the only difference is that it includes new requirements obtained after user interaction with the system. Lastly, the Enhanced Working Virtual Census Enumeration Tool is

delivered after the final testing which was done by the researcher specifically the objectives and functionalities.
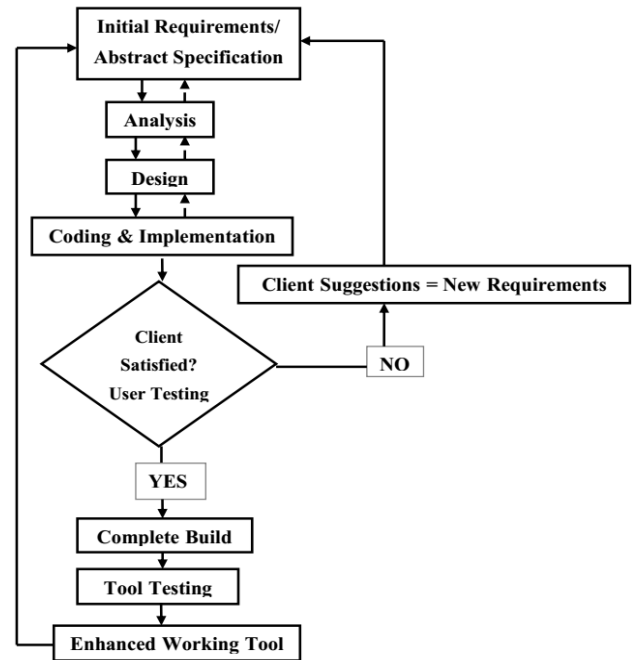

Fig.3. Prototyping Model

The prototype has been developed to reflect and collect similar information to what current traditional censuses collect. The developed system is designed to overcome the observed data integration challenges which are duplication of entries and disparities in DBMSs between the data sources.
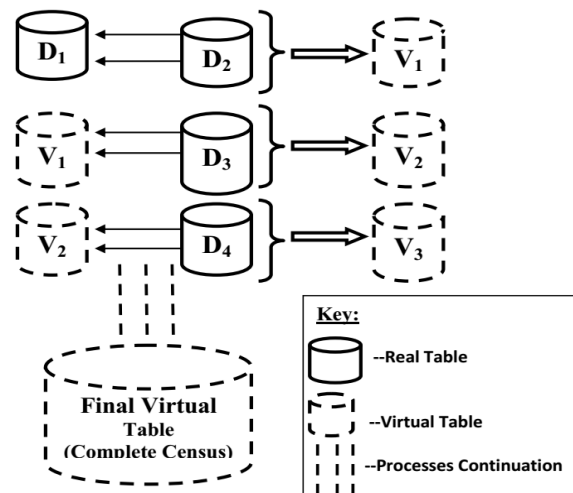

Fig.4. Steps of Cleansing Data for Database Integration Process

Duplicated entries – during data integration, it has been observed that a single person appears in more than one organ such as RITA, NIDA, etc. This resulted into producing a very large figure due to repetition of the entries since one person can be counted more than once. The duplication of information problem has been solved through cleaning processes as illustrated in Fig. 4, which show how data quality problems are addressed by data cleaning and provide steps for cleaning data. Data

cleaning is especially required when integrating heterogeneous data sources and should be addressed together with schema related data transformations.

The data cleansing approach was achieved as follows; at a very beginning the two real tables were combined together to produce a single virtual table, let say data from D_1 (i.e. data from RITA) combined with data from D_2 (i.e. data from NIDA) to produce virtual table V_1 (i.e. V1=real table$_1$ + real table$_2$). Moreover, V_1 is combined with the next real table to produce V_2 which is the same as V_1+D_3, $V_3= V_2 + D_4$, etc…

This section provides the algorithm and pseudocode employed on the data integration process. Algorithm/ pseudocode 1 describe data cleansing during the integration process and Algorithm 2 describes how to overcome the problem of dissimilarities in DBMS.

Algorithm/ pseudocode 1:

```
1.  Initial parameters are D₁ & D₂
2.  D₁=D₁{a₁,a₂,a₃,…….aₙ},
3.   where: a₁,a₂,a₃,…….aₙ,are
4.    the table
5.    column of the first table.

6.  D₂=D₂ {b₁,b₂,b₃,…..bₙ},
7.     where: b₁,b₂,b₃,…..bₙ, are
8.     table
9.       column of the second
10.      table.
11. Step1: Compare data of table
12.       D₁ & D₂ and store the result
13         into virtual table V₁
14.       V₁=Comparison (D₁, D₂);
15. Step 2: Then compare data of
16.        Virtual table V₁ against
17.         the next real table
18.        example;
19.         D₃ and store the result
20.        into virtual table V₂
21.       V₂= Comparison (V₁, D₃);
22. Step 3: Repeat step2 until all
23.       tables real tables are
24.         compared
25.       Vₙ= Comparison (Vₙ, Dₙ₋₁);
26.        where Vₙ=Final virtual
27.     table for this case it is a
28.     complete census.
29. Function:
30.     Comparison (Table₁,Table₂) { 31.
// It takes two tables as
32.       arguments and return
33.     the total cleansed virtual
34.       table.
35. while (table available) {
36. Traverse to the whole table
37.   until all column are visited
38. THEN
39. if (the person available)
40.       DROP the data
41.   else if(the person not
42.          available)
```

```
43.    ADD the data in the nᵗʰ
44.       virtual table
45. repeat procedure 18-23 until
46. the end
47.              }
48.            }
```

Algorithm 2:

Overcoming differences in DBMS; Consider Fig. 5 which shows data integration process from various data sources with different DBMS. Exchange of data in various databases with different DBMS, the data should be transparent from representation and should be platform independent. XML documents are self-describing, thus XML provides a platform independent means to describe data and therefore, can transport data from one platform to another. XML is also used when one or more organizations merge. When organizations merge, interoperability among documents is necessary which can be achieved using XML integration.

1. Establish connection between
2. The source database and server system
3. Log in to the source system
4. Fetch data from the source
5. Generate automatic XML files
6. Read the XML data and load into the server /census system
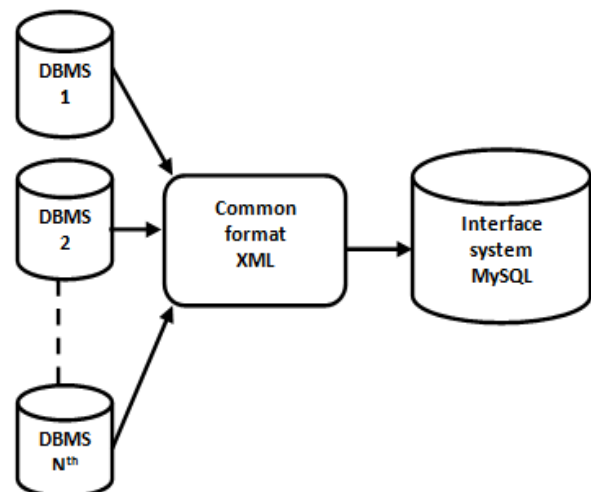7. Repeat ALL steps as in the algorithm 1

Fig.5. Data Integration from Various Data Sources with Different DBMS

IV. Results

This part describes the survey findings and hypothesis test results. In this work, the Virtual Census enumeration system was developed and tested on two major challenges which are duplicates of the entries and differences in DBMS.

A. Differences in DBMS

During this test the researcher designed the experiment setup that, there are four different computers with different databases with different DBMSs for example the first computer was installed Oracle (it contains NIDA database), the second was installed MySQL (it contains RITA database), the third computer was installed Microsoft Access (it contains LGA database) and the fourth computer is NBS which acts as the controller in which all the operations were done, they are all connected to the network. When you click Fetch Data button (as shown in Fig. 6) the system is able to extract the data from specified computer let say NIDA by providing the IP address of that computer, other required information in order to specify a required computer is user name and password.
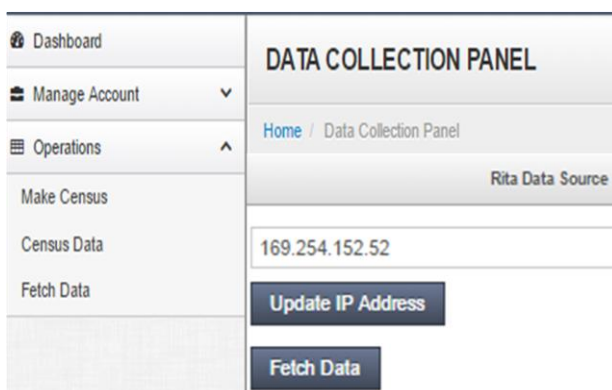
Fig.6. Data collection panel

B. Duplicate of the entries

During this test the same data was added in more than one database (i.e. the person with the same particulars such as first name, surname, date of birth, sex, place of birth, birth certificate ID). For example, the person named Mnyampaa Siffi Duma was added in NIDA, RITA and LGA databases. When you click make census button it removes all duplicates and the system output after this operation is as shown in Fig. 7. It shows the total number of population and their specific categories for population such as children, youths, adults, women, men, foreigners, teachers etc

C. Current Tanzanian Virtual Census Data Sources

After testing the system authors were confident that; Tanzanian census can be completely enumerated through integration of data from various existing sources instead of the current practice of traditional /physical data enumerations. Table 2 describes generally the census data elements and their associated data sources.

Apart from displaying the total number of population and the associated statistics, the system by itself is able to do simple data analysis and provide the enumerated information graphically as shown in Fig. 8 which shows the demographic pie chart. There is no need to count manually to get the total population since the system itself is able to do basic statistical analysis, it is not necessary to employ other statistical packages.

This study is of great importance to the country and will ensure that the census information /data can be collected at minimum cost. Virtual Census enumerations will help to eradicate critical traditional challenges by having all the required information ready in hand once required. For example, during 2012 census exercise conducted in Tanzania, several challenges arose with one major challenge being the boycott by both census enumerators and some people who neglected the census exercise. This challenge will remain a history upon successful adaptation of Virtual Enumerations.

Table 2. Census Data Elements and Associated Data Sources

| Type of census information/data	Data sources
Marital status (married, single, divorced, widowed)	RITA
Citizenship	ISD, NIDA, RITA
Life, Deaths and Disabilities	RITA, LGA, Hospitals and Ministry of health
Education (Level of education, Schools, Universities etc)	NECTA, HESLB, Ministry of education, TCU, NACTE.
Employments	PSM, NGOs, Private organizations, Social security funds
Housing and ownership of equipment	LGA, TRA, Ministry of Lands, Housing and Human Settlements
Agriculture and livestock	Ministry of Livestock Development and Fisheries, Ministry of Agriculture, Food Security and Cooperatives
Tanzanian living abroad	ISD
Availability of other social services (example Hospitals /Dispensaries, Pharmacies, Natural Medicine Shops, Schools, Colleges /Universities, Source and kind of drinking water, Churches and mosques)	LGA, Ministry of health, Ministry of education etc.
Disasters (e.g. Fire outbreak, Massive livestock theft, Killings, Accidents, floods, etc)	LGA
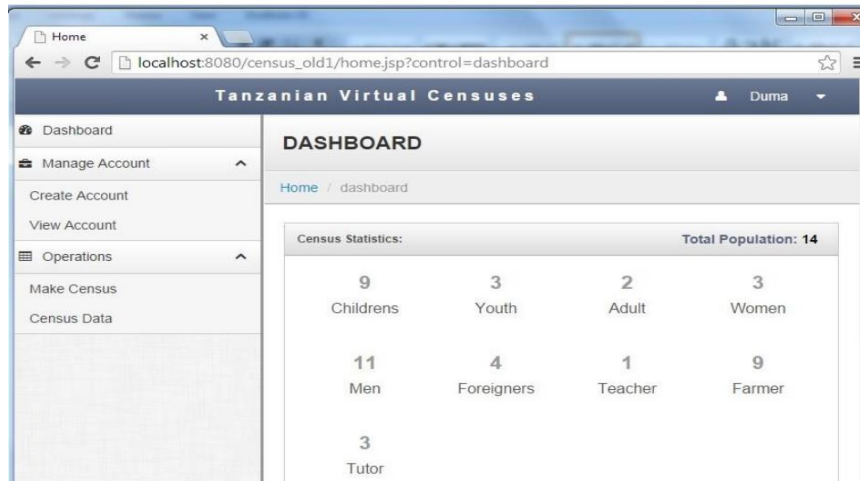
Fig.7. System Operation Interface/Dashboard
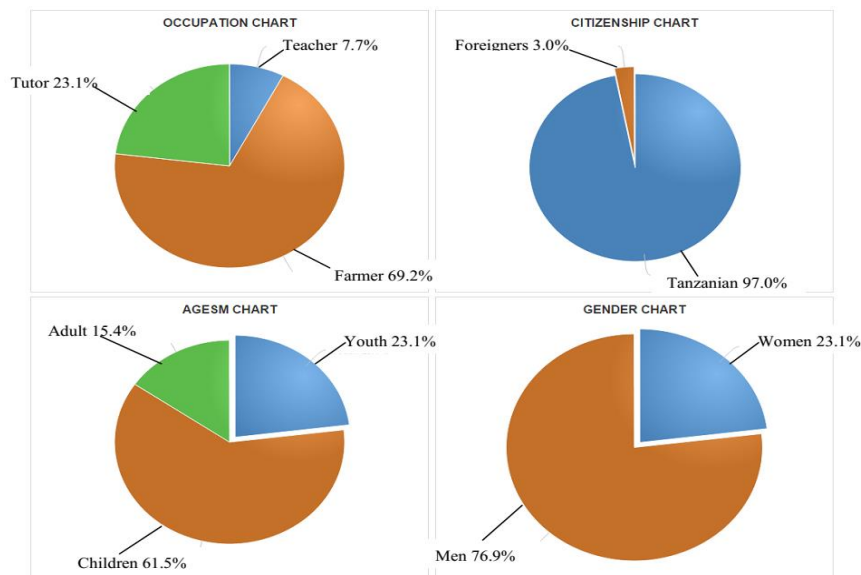
Fig.8. The Demographic Pie Chart

V. CONCLUSIONS

A conclusion is based on what has been observed in the field during the data collection phase. During testing phase many census stakeholders interacted with the prototype and they were optimistic regarding the approach, others pointed out that it will reduce a lot of tedious manual work, one should also take into account that many people are now not willing to participate and others tend to hide certain personal details, which leads to incorrectness of the census. From National Electoral Commissions (NEC) it was reported that there would be no need for investing a lot of resources (i.e. money, time and human resources) in updating what is famously known in Swahili as "Daftari la wapiga kura" or "Voters registration book" in English rather, they will be using the available /already existing data. Not only that, they agreed with the researcher that this approach will also help in eradicating conflicts currently raised about the exact number of voters. When the researcher discussed it with ISD they agreed that this approach would help in tracing the illegal immigrants who cause a lot of problems in Tanzania.

VI. FUTURE WORK

Despite the challenges encountered, there are also opportunities found in this study which need to be cultivated. It is the view of the researcher that future researchers should look on security issues i.e. that they should study invasion of privacy during data integration and sharing of resources, and what security measures can be developed for preventing the disclosure of sensitive information. It should be taken into account that data mining and integration requirements may carry conflicting goals, for example, the handling of different types of data is confined to relational and transactional data.

REFERENCES

[1] The United Republic of Tanzania.: Population Distribution by Age and Sex. Dar es Salaam, Tanzania (2013).

[2] UNFPA.: Counting the People Constraining Census Costs and Assessing Alternative Approaches. UNFPA (2003).

[3] Ngallaba, S., Kapiga, S. H., Ruyobya, I., & Boerma, J. T.: Tanzania Demographic and Health Survey. Bureau of Statistics, Planning Commission, Dar es Salaam, Tanzania and Macro International Inc., Columbia, MD, USA (1993).

[4] Gerards, E.: Dutch Virtual Census: Presentation at the International Seminar on Population and Housing Censuses Beyond the 2010 Round, The Netherlands [Video clip], retrieved on 20th January, 2013 from http://youtu.be/SLpDkcyenf0 (2012).

[5] Nordholt, E. S.: The Dutch virtual Census 2001, A new approach by combining different sources. Statistical Journal of the United Nations Economic Commission for Europe, Vol. 22, No. 1, 25-37 (2005).

[6] Ghulam, A. M., & Ghulam, M. M.: A Framework for Creating Global Schema Using Global Views from Distributed Heterogeneous Relational Databases in Multidatabase System. Global Journal of Computer Science and Technology, Vol. 10, Issue.1, pp. 31-35 (2010).

[7] Ali, M. G.: A framework for creating Global Schema using Proxy Tables from Distributed Heterogeneous Relational Databases in Multidatabase system. IJCEE, Singapore (2010).

[8] Sabri, M., & Rahal, S. A. APESS-A Service-Oriented Data Mining Platform: Application for Medical Sciences (2016).

[9] Rochlani, Y. R., & Itkikar, A. R.: Integrating Heterogeneous Data Sources Using XML Mediator. International Journal of Computer Science and Network, Vol.1, Issue 3 (2012).

[10] Mohammed, R. O., & Talab, S. A.: Clinical Data Warehouse Issues and Challenges. International Journal of u-and e-Service, Science and Technology, 7(5), 251-262 (2014).

[11] Fall, K. R. & Stevens, W. R.: TCP/IP Illustrated: The protocols. Vol 1 addison-Wesley (2011).

[12] Portier, B., & Budinsky, F.: Introduction to Service Data Objects. International Business Machines Corporation (2004).

[13] Digplanet.: XQuery API for Java, Retrieved December, 2015, from http://www.digplanet.com/ wiki/XQuery_ API_for_Java#cite_ note-QueryDataModel-1

[14] Melton, J., & Buxton, S.: Querying XML: XQuery, XPath, and SQL/XML in context. Morgan Kaufmann (2011).

[15] Shamin, O.: Development of a Web-based Demonstrator for an Approach to prevent Insider Attacks on DBMS (2014).

Authors' Profiles

Ramadhani A. Duma is an assistant Lecturer at the Centre of Information and Communications Technology of St. John's University of Tanzania. He obtained his Master of Science Degree in Computer Science from the University of Dar es Salaam in 2015. He is currently teaching computing related courses. His research interests include ICT4D, Databases, Data warehousing and Web Services.