University of Nairobi

School of Computing and Informatics

M .Sc. Computer Science

Project Report

Name: Davis M Onsakia

Registration Number: P5 8/61467/2010

Project title: A business model for encouraging citizens to open up their personal data for anonymous statistical use

Supervisor: Dr. Wanjiku Nganga

Declaration of original work

This is my original work and to the best of my knowledge has not been submitted before to any institution and is in partial fulfilment towards the requirements for the award of MSc. Computer Science degree.

Name: Davis Onsamia Signature: Date: 20/10/2012

Supervisor

Name: Dr. W. Nganga Signature: JwNganga Date: 16/11/2012

Abstract

Governments all over the world have a lot of data about their citizens. The citizens in turn have a lot of information about themselves which the government does not have. If these two data sets are consolidated and then mined by interested parties, there is a lot of valuable knowledge that can be gleaned from it. This research has looked into these issues and proposed a model which can be used by the government to encourage its citizenry to share their personal data with it and also with other interested parties, in a legal and acceptable manner for anonymous statistical use.

Since it was not feasible to carry out a census to ascertain the perception in regard to the issues around personal data access and sharing, a sample of the Kenyan population was chosen to provide feedback about these issues.

From the research findings, most Kenyans (62%) are ready and willing to open up their personal data details for anonymous statistical analysis to the government and other interested stakeholders in the Personal Data Ecosystem (PDE). However, they indicated that they need to give their consent for any access of their personal data.

There was also correlation between gender, level of education and age and willingness to sharing of personal data with the government and other stakeholders in the Personal Data Ecosystem (PDE).

We considered the Kenya scenario and the responsibility of the Government of Kenya (GoK) and its citizens and other stakeholders in the PDE in achieving this objective.

Acknowledgements

I wish to express my sincere appreciation to the invaluable guidance I have received from my supervisor, Dr. Wanjiku Nganga. Without her encouraging words and guidance, I would not have made it this far. To her I say, 'Thank you very much!'

To my family who have endured my absence during my study period, I salute them for their understanding and patience in waiting for me even when I was delaying arriving home and 'disappearing' over weekends.

Above all, to God who makes all things possible, I say 'Thank You Father'.

Table of Contents

Tables

Table of figures

Acronyms and abbreviations

API – Application Programming Interface

CRB – Credit Reference Bureau

EASSy – Eastern African Submarine cable System

GoK – Government of Kenya

GSM – Global System for Mobile communication

HELB – Higher Education Loans Board

HOSIF – Higgins Open Source Identity Framework

ICT – Information and Communication Technology

KIAF – Kantara's Identity Assurance Framework

KNBS – Kenya National Bureau of Statistics

KRA – Kenya Revenue Authority

LDAP – Light Directory Access Protocol

NHIF – National Hospital Insurance Fund

NOFBI – National Optic Fiber Backbone Infrastructure

NSTIC – National Strategy for Trusted Identities in Cyberspace

ODI – GoK's Open Data Initiative

OECD – Organization for Economic Cooperation and Development

OITF – Open Identity Trust Framework

P3P – Platform for Privacy Preferences Project

PbD – Privacy by Design

PDE – Personal Data Ecosystem

PDEC – Personal Data Ecosystem Consortium

PDS – Personal Data Store/Service

PIN – Personal Identification Number

PSC – Public Service Commission

SAML – Security Assertion Markup Language

TEAMS – The East Africa Marine System

UMA – User Managed Access

USSD – Unstructured Supplementary Service Data

WEF – The World Economic Forum

WS -Trust – Web Services – Trust specification

XDI – XRI Data Interchange

XRI – Extensible Resource Identifier

Introduction

With the arrival of the undersea cables TEAMS, EASSy, Seacom and France Telecom's Lower Indian Ocean Network 2 (LION 2) at the Kenyan Coast, what is needed is content to pass through the dumb pipes. These undersea cables link Kenya to the rest of the world on the information highway. Content can be individual, private or government generated. However, it will require to be locally hosted for citizens to enjoy faster access speeds. The Kenya ICT Board (Kenya ICT Board, 2012) is working on how this can be achieved in the most cost effective, efficient and sustainable way through their many initiatives.

On the local ICT infrastructure, the National Optic Fiber Backbone Infrastructure (NOFBI) is expected to connect the country's 47 counties (currently 37 counties have been connected). Once all the 47 counties are connected, content will be required which will pass through these network links in addition to leasing out extra capacity to other interested parties like banks and Internet Service Providers (ISPs). This research's findings and recommendations, once implemented, are expected to help in generating the necessary content to pass through this robust ICT infrastructure that is expected to be in place.

The government's initiatives, like the NOFBI project, are in addition to what private investors like mobile and internet service providers are laying on the ground. It is expected that in the long term, every single inch of the country will be connected to some network (whether wireless or wired).

Currently, the Government of Kenya (GoK) has a lot of personal data of its citizens which it collects at different points such as during birth registration, application for a death certificate, Identification Card (ID) or a passport, among many other instances. However, we asked ourselves the question: what if the government allowed access to these data (to interested parties) for some anonymized statistical analysis? This question can be answered by another question: what if it did and infringed on the privacy and security concerns of its citizens? Basically this was the crux of this project: How can the government share its citizens' personal data without infringing on their privacy and/or security and how can citizens either be 'encouraged' or 'educated' to share or allow access to their personal data for some anonymous statistical use? The statistical analysis probably can be for planning, research, academic or marketing purposes. It should be noted that this research concentrated on anonymized personal data details of citizens rather than identifying traits or features. This was deliberately done to preserve the privacy and security of the citizens.

This project considered citizens' static data details like the date of birth, name, etc and also their dynamic data sets. Dynamic personal data details are like employment history, salary, location etc. The consideration of these two data sets was to help to get a complete picture of each individual citizen without necessarily identifying the person.

Interviews and a structured survey were used to ascertain which ways and approaches can be adopted that can encourage the common mwananchi to open up his/her data bank details for anonymous statistical analysis.

Finally, we have proposed a framework and guidelines which can be used by the government and/or relevant stakeholders to encourage the citizenry to share some of their personal details for anonymous statistical use.

In this age of empowered citizenry, people want to be in charge of their lives and this includes a desire to control all data pertaining to them. These data might be that they have willingly submitted to authorities or data that the authorities have collected and stored concerning their citizens courtesy of their oversight role in society. Considering that we generate over 2.5 quintillion bytes of data everyday globally (IBM, 2012), the question that one can ask is: what can we do with all these torrents of data that is being generated daily? This question can easily be answered by looking at business models of internet giants like Google, Facebook etc. Their business thrive on the (anonymous) collection, aggregation and monetising of personal data (sometimes they indicate that they collect personal data in small-print privacy laws which most people just click 'OK' without reading). Therefore, their lifeline is basically data that is willingly (sometimes unknowingly) submitted by many users of their websites and services.

Back home, what can be done on the data submitted by the citizens together with what the government has about them? Can the government make use of it by aggregating and monetising it, like the above-mentioned giants? If yes, under what regulatory framework? And if no, what are we losing as a country by sitting on a gold mine – if the benefits of what Google and their ilk are making from the use of our data are anything to go by? These are some of the issues which gave impetus to carry out this project.

The ways which the internet giants have managed to attract and retain users and get data from them will be compared to methods which the GoK too can use to woo citizens to provide their personal data. However, unlike these companies which might not be explaining the implications of one logging to their site and supplying their personal details, the government is obliged to explain clearly what it means and how the citizens will benefit either immediately or in the long run. The value proposition must be very clear: what needs to be in place for the citizens to provide relevant and meaningful data. This will be part of creating awareness and in the process building confidence with the citizenry on the functionality of the ecosystem.

The question of how data will be consolidated from the different government systems with what individual citizens might be willing to provide has also been tackled: How can the data currently residing in government's disparate systems, together with what the citizens will be willing to provide, be consolidated so that useful statistical analysis can be carried on it? How can the government which is normally 'closed' (although this perception is changing with the implementation of the ODI project) allow individual citizens to access their personal data and update it if necessary so that a researcher, statistician, academician etc can get a correct picture of what s/he needs once they can be allowed to access such data?

How can this be done? Can a web interface, may be called, a Government Citizens Portal, which is accessible over the web help in this? Or can the ODI portal be enhanced with these project's recommendations? These are some of the questions that this research has attempted to find answers to.

Globally, there are concerted efforts to change data from being organization-centric (or government-centric) to user-centric. This is due to the fact that it is the individual who matters and not the organization or government. Basically, this is what Credit Reference Bureaus (CRBs) are trying to do: getting all financial data related to an individual, analyzing it and then compiling this into one comprehensive report which makes an individual's credit score. However, what about other data concerning to the individual – which only the individual knows? It should be noted that the individual has no direct input to the data resting in the CRB's systems – all that data is inferred from financial transactions of an individual end-user from different financial institutions that the person deals with. This research has proposed how the individual can be brought onboard and what his role will be in regard to personal data held about him in GoK systems in terms of sharing it and/or making money out of it.

Government leading the way

The GoK 'opened' some of its data sets for public access through the Open Data Kenya (ODI) initiative (Government of Kenya, 2012) under the transparency, openness and citizen participation banner in July 2011 (Kenya ICT Board, 2011). It is expected that this might encourage wananchi to also share some of their personal information to the public. It should be noted though that the government moved ahead with this project development and launch without the necessary 'legislative, policy and legal framework, including protection for data reuse' (The World Economic Forum, 2012) in place. The same report notes that such policy frameworks are now being backfilled. Some of the legal frameworks being developed in this area are the Data Protection (Commission for the Implementation of the Constitution, 2012) and Freedom of Information (Commission for the Implementation of the Constitution , 2012) bills.

From the above, it can be appreciated that the government is at the forefront in driving the change that it needs to see taking place in the country without necessarily waiting to have the legal framework setup before embarking on rolling out an innovation that can positively change how they provide service to the citizenry. Although this might not be the best approach since the constitution empowers the citizenry to question such implementations, it is a clear indicator that the government is willing to test innovative solutions and ideas whose impact will be appreciated in the fullness of time. And the ODI is such an implementation.

Problem Statement

The government has a lot of data about its citizens and the citizens in turn have data (about themselves and their lives) that the government does neither have the capacity nor the motivation to collect and consolidate with what they have. We need a mechanism in which these two data sets can be 'merged' and incentives which can be used to entice the citizenry to allow access and/or sharing of their anonymized personal data details to interested stakeholders for statistical analysis.

We know that if we can consolidate these two data sets, there is a lot that we can learn as a country (both as citizens, public sector actors as well as private sector stakeholders) about our fellow citizens and our country. This can generate economic value by having in place better planning and marketing mechanisms, among a host of other benefits.

This research has explored and recommended how this can achieved in a cost effective, efficient and legally acceptable manner.

Research objectives

The major research objectives of this study were:

1) Find out if citizens are willing to supply their personal details to the government or to any other interested party (for free or through some incentives);

2) Propose ways in which the citizenry can be encouraged to open up access to their personal data for some anonymous statistical analysis (by giving some value propositions);

3) Propose methodologies or mechanisms which will facilitate easier access of personal data by individual citizens and sharing of the same to interested stakeholders;

4) Propose changes, if feasible, in which the current legal and/or regulatory framework might need to be amended to allow for anonymous access to individual citizens' data by authorized entities for statistical analysis.

Rationale of this study

We need to use data that is available in the country for decision making and in its current state – being in disparate silos; sources and forms will not help us to achieve this noble dream. This research has attempted to tackle the twin challenges of motivating the citizenry to allow access to their personal data for sharing with interested parties and of consolidating the available government data to one central place. For the latter part the eGovernment Directorate is at advanced stages in its implementation (eGovernment Directorate, 2012).

Currently, there are ways in which you can electronically track some government services. For example, you can send a message to some GSM short code or using USSD to get details about the processing of your passport, ID etc. What if all these services are available from a central place (or using one code)? An individual will just require to login (through some interface or send an SMS to one code) and then be able to check the status of all government services including job applications which he has made through the Public Service Commission (PSC) website and a host of other services. From this one focal portal point, an individual can also make his tax submission online to Kenya Revenue Authority (KRA) without necessarily going to the KRA website to do so. This way we will have an empowered individual who does not need to remember many usernames and passwords as well many GSM short codes – basically one central access point provides a window of access to all Government services! These are some of the benefits that have been explored in this research to ascertain whether they can be incentive enough for citizens to update their personal data before they can allow access to same.

Also, it should be possible for individuals to access the Higher Education Loans Board (HELB) site from this same portal to check on their loan application status, repayment schedule and print even a clearance certificate from the system if they have completed repaying their loans – without necessarily going to HELB offices to get the same. To some extent, this concept has been demonstrated that it works the way individuals currently login to the KRA website and print their legally accepted PIN certificates. This will not only improve service delivery from these corporations but at the same time will save a lot of resources and increase satisfaction from individual citizens' perspective. Such an automated service delivery is expected to empower individual citizens and reduce corruption at these entities by enhancing transparency and accountability.

Assumptions and limitations

The following assumptions were made prior to carrying out the research:

Assumptions

i. There will be co-operation from interviewees during the research exercise;

ii. Resources for carrying out the survey will be available;

iii. We will have peace during the scheduled research period;

iv. Interviewees and questionnaire respondents will be honest in their responses.

Limitations

Despite the above assumptions, the following were identified as limitations of the research. We have included our approach on how we mitigated their effect on the research findings:

i. Poor perception of the government by the citizens when it comes to their privacy.

Our approach: Explained the study objectives before we could solicit for feedback to dispel the notion that the government might use the data against them.

ii. Sample design: An ideal research design for this project would have been a census but due to lack of enough resources we used sampling.

Our approach: Although we used a sample population, we believe that its findings are representative of the whole population.

iii. Privacy and security concerns.

Our approach: Explained the research objectives to the prospective interviewees before we could solicit their responses and assured them that whatever they provided was going to be used for research purpose only.

Literature Review

Opening up citizen's personal data for anonymous statistical analysis has implications in relation to both their privacy and security. This is the case more so in cases where the data has indentifying attributes.

In this regard, there has been a lot of research and concern about individual information privacy and security over the web and generally in life. It is due to this concern that a myriad number of laws and regulations have been formulated as well entities formed concerning privacy of an individual. The Electronic Frontier Foundation (EFF) (Electronic Frontier Foundation, 2012), for example, has listed the laws and regulations governing individual privacy on their website.

There are also annual events scheduled to discuss about data protection and privacy issues. One such event is the International Conference of Data Protection and Privacy Commissioners (ICDPPC). The latest such event was held from 2nd - 3rd November 2011 in Mexico City (International Conference of Data Protection and Privacy Commissioners, 2011) – this was the 33rd event in the history of ICDPPC. The next event is scheduled to be held from 23rd – 24th October 2012 (International Conference of Data Protection and Privacy Commissioners , 2012).

Information privacy has been defined as an individual's claim to control the terms under which personal information – identifiable to the individual – is acquired, disclosed and used (National Information Infrastructure, 1995). This definition is quite comprehensive and its meaning is what we will be implying when talking about 'information privacy' of an individual citizen.

On the research front, Nissenbaum (1998) introduced a concept she called *privacy in public*. In this paper, she argued that individuals are entitled to their own privacy even if what is currently known about them was freely provided by the individuals themselves. She further introduced the concept of *contextual integrity,* where she indicated that disparate pieces of information collected over time about an individual, if aggregated, can form a picture which the individual cannot desire to be known about him.

Basically, she was arguing against data mining of personal data collected over a period of time from different sources and in different contexts. From a government perspective, however, it should be noted that the government might be carrying out its own data mining on the current citizens' data for planning purposes – without necessarily the consent of the individual citizens. Even if the government wants to get the citizens' consent, it might not be feasible. For the case of to-be provided more personal data, data mining might have to be carried out by interested parties rather than the government itself. It should be borne in mind too, that data without descriptive and predictive data mining might not be useful to anybody. All this should and must be done however within the acceptable legal environment and limits. Also, it is important to note that statistical analysis will be done on anonymized data rather than identifying one. The entities that might be interested in the citizen's data will be free to carry out either descriptive or predictive data mining since once they have access to the data there is no much control on what they can do with it.

In addition to the above, the World Wide Web (WWW) consortium developed and adopted the Platform for Privacy Preferences (P3P) (World Wide Web Consortium, 2002) project as a protocol for privacy protection on the web. However, it is only Microsoft that has adopted this protocol in the Internet Explorer 9 with other internet companies giving it a cold shoulder.

P3P allows websites to declare the intended use of information they collect when browsing. It was designed to give users more control of their personal information on the web. The developed government data sharing platform will ideally adhere to this requirement although it is our belief that most of what might be expected from this protocol should ideally be captured in some agreement between the government and its citizens before rollout of the project. It should be noted though that this protocol is no longer supported by W3C as it received poor support from the major browser vendors.

In 2010, having recognized the importance of users' privacy over the web, different stakeholders in the PDE came together and formed the Personal Data Ecosystem Consortium (PDEC) (Personal Data Ecosystem Consortium, 2012) to 'educate industry, facilitate cooperation, and inspire the development of an ecosystem where individuals are empowered to collect, manage, and obtain value from their own personal data.' This group has come up with some innovative policies (found on their website) in the PDE which this research has endeavoured to abide by and emulate so that citizens' personal data is safe and secure.

On a more inclusive and global scale, The World Economic Forum (WEF, 2010) recognized personal data as the 'new currency' in the global arena. It was with this recognition that WEF in June the same year, setup a multi-year project dubbed *Rethinking Personal Data* to bring together leading experts, public authorities, advocates and executives from telecom, technology, media and internet firms to:

a. Increase understanding of the different stakeholder interests in the PDE;
b. Illustrate the opportunity to be derived by leveraging on personal data;
c. Detail real-life use cases and pilot studies;
d. Identify a set of principles which could serve to establish a balanced PDE across all the stakeholder communities.

The WEF released a preliminary report in September of the same year (The World Economic Forum, 2010) where it called a PDE where users, the public sector and the private sector mutually benefit a win/win/win ecosystem for sharing of personal data. The report further indicated that despite the challenges in the personal data ecosystem, 'wait and see is no longer a viable strategy for most actors'. This therefore means that Kenya too cannot be left behind if it expects to benefit from this new economy.

The committee members released their final report on their findings in January 2011 (The World Economic Forum, 2011).The report (The World Economic Forum, 2011, p. 7) defined *personal data* as digital data (and metadata[1]) created by and about people. It further categorized personal data as volunteered data, observed data or inferred data. It should be noted that they left out what can be termed as static data – data that the government has about its citizens; which individual citizens cannot arbitrarily alter. Their definition therefore, together with the inclusion of static data, is what will be used and implied throughout this research when mentioning *personal data*.

[1] Metadata is data about data

However, it was noted that in a pre-read document which was prepared by Davis et al. (2010), these authors had included static data (indicated as Government records) as part of personal data although the final WEF report seems to have ignored this. For our research therefore, we relied more on the personal data map of these authors as shown in the figure below.

The personal data map as was presented during the WEF meeting by Davis et al. (2010).

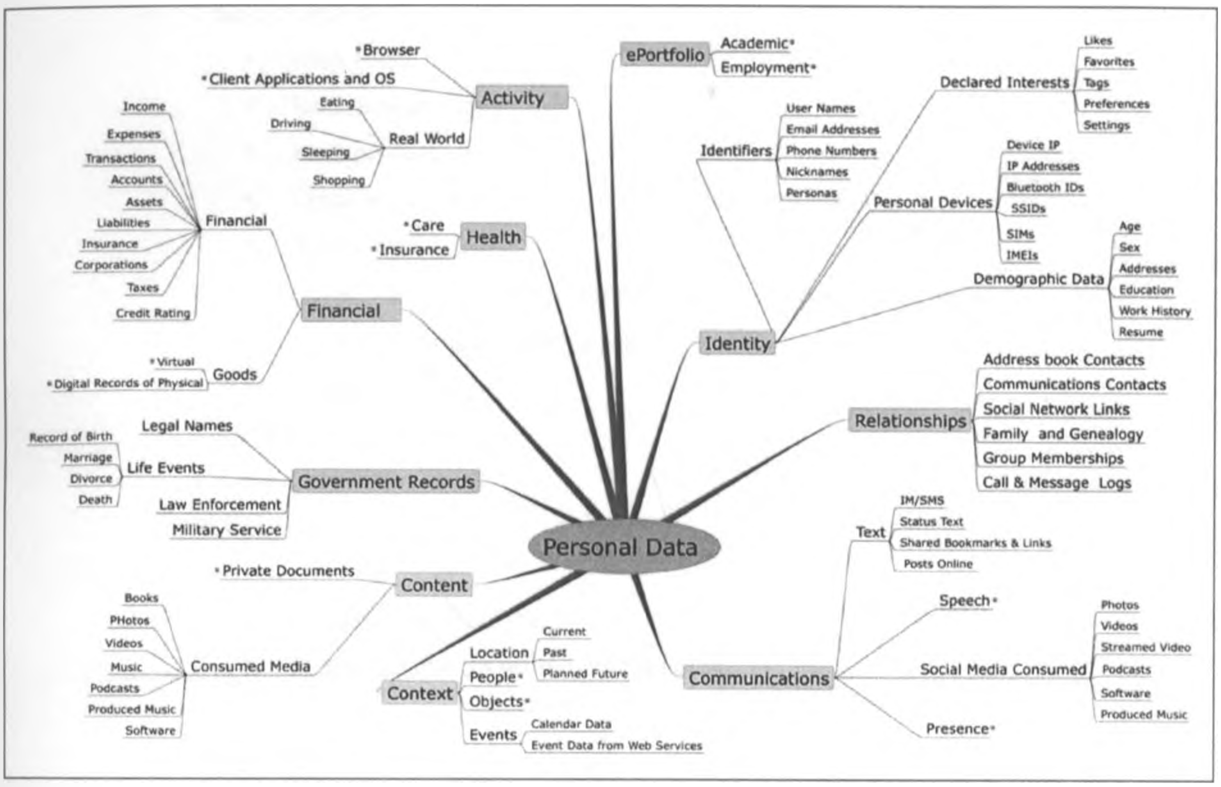


Figure 1: The WEF personal data map

However, it should be noted that some of the details of personal data envisaged from this map were considered not feasible bearing in mind the scope of this project. For example, under communications arm areas like speech, social media etc were not considered.

In this regard, personal data details that were considered in this research include the following, among others:

- Static data –provided majorly by Government systems – which cannot arbitrarily be altered by individual citizens:
  - Personal details:
    - Full Name
    - Date of birth
    - PIN number – KRA generated and issued
    - ID Number
    - NSSF Number – from NSSF systems
    - NHIF Number – from NHIF systems
    - Passport number (if ever applied for one)

  - Demographic data
    - County of birth
  - Criminal records – which the individual can request to be changed (only on presentation of evidence). Alteration of an individual's personal details can be changed as per rights bestowed on citizens as per Constitution Article 35 (2) (Government of Kenya, 2010).
- Dynamic data sets – to be supplied by the individual citizen – can be changed at will:
  - Financial details:
    - Bank Name
    - Bank Branch name
    - Bank account number
    - Credit scores (from Credit Reference Bureaus (CRBs))
    - Assets owned:
      - List of properties owned, location and their tentative value
      - Approximate net worth
    - Liabilities (if any)
  - Postal address
  - Current county of residence
  - Physical address
  - Telephone number
    - Fixed telephone number
    - Mobile number (to list if more than one)
  - Political stand
  - Physical description
  - Religion
  - Parents details – whether alive or dead – some of these to be picked from the GoK systems
    - Parents' names
      - Father's full name
      - Mother's full name
    - Dates of birth – for both mother and father
    - Death's details – if dead
      - Date(s) of death
      - Cause(s) of death
    - Occupation of both parents
    - Employer details of both parents – or the last employment they had
  - Marital status
    - If married:
      - Name of spouse
      - Number of children (currently) and their gender
      - 'Ideal' number of children planning to sire
    - If single:

    - Whether planning to get married or not and reasons thereof
  - If widowed
    - Cause of death of spouse
    - Date of death
    - Number of children (currently) and their gender
    - Whether planning to remarry or not and reasons thereof
  - If divorced
    - Date of divorce
    - Number of children (currently) and their gender
    - Duration in marriage
    - Reasons for divorce
    - Whether planning to remarry or not and reasons thereof
  - If separated
    - Date of separation
    - Reason(s) for separation
    - Duration in marriage
    - Number of children (currently) and their gender
    - Custody of children – if any
    - Whether planning to remarry or not and reasons thereof
- Email addresses (official and personal)
- Health data (medical records)
  - Weight
  - Height
  - Medical history etc
- Academic and professional qualifications
- Occupation
- Employment history
- Travel history (and probable purposes – leisure/vacation, business etc)
- Travel plans (and probable purposes – leisure/vacation, business etc)
- Hobbies – a listing
- Ethnicity
- Web profile(s) – social media usernames, preferred search engines, etc

All the above will be required so that we can get a complete picture of who the individual is: who they know, where they are, where they have been and probably where they plan to go, WEF (2011, p. 5)! Although the data will be anonymized, this will be important to give a complete picture of a citizen's profile.

The WEF in its 2011 report (The World Economic Forum, 2011) made the following observations:

- Mining and analysing personal data of individuals 'will give us the ability to understand and even predict where humans focus their attention and activity at the individual, group and global level'.

It should be noted however, that the report did not propose ways of analysing and mining this information for use by interested stakeholders; to get the nuggets from the gold. This is what this research has tried to achieve: ways in which individuals can be encouraged to open up their personal data for mining and analysis so that we can be able to understand the human behaviour and predict where they focus their attention and this will be extrapolated to the group or even global level.

- Personal data is generating a new wave of opportunity for economic and societal value creation. However, it is noted, that this observation fails to elaborate on the economic and societal value creation that can be derived from personal data. It is possible that it might not be feasible to enumerate the benefits that might be derived from personal data till individuals allow access to their data for mining and analysis purposes, but it is possible to give a rough breakdown of 'what is in it for me?' This is what this research has endeavoured to bring out.

The question that this research has tried to tackle is: bearing in mind all these required (private) data from citizens, how can they be motivated to provide personal details about themselves? These data, once provided, will be a goldmine not only to researchers but also to the government. But this then begs the question: since individual personal data is not so critical to the functioning of the government and bearing in mind the previous negative perception the populace has towards the government in regard to their personal privacy, how can they be 'encouraged' to freely submit their data and how can this be accomplished? And how can they be assured that the provided data will be secure and safe?

Will personal data be the oil, a new asset class touching all aspects of our life, as claimed in WEF (2011, p.5) report? These are some of the questions that this research has sought to answer, from a Kenyan perspective.

From an infrastructure perspective, consolidating all the required 39 million plus (Kenya Government, 2009) Kenyans' personal data to one central place will require a highly reliable, secure and available ICT infrastructure at its core and robust innovation at its edge. Does the government have the capacity to host all its citizens' personal data in one central place? Alternatively, do we need to consolidate all the data to one central place or can we use mash-ups and/or real-time APIs[2] to get the necessary data from the disparate government systems dynamically, as and when required? And what needs to be put in place to implement any chosen solution? This research has tried to evaluate the current government ICT infrastructure to try and answer these questions. This is very key to the success of the project of sharing personal data of the citizens for without it there is no way citizens can access the system to update their details or even researchers access data, from the system. Additionally, it must be borne in mind that the government is currently able to handle its citizen's data needs since most of it is in different systems which are not in any way linked or communicate.

The WEF (2012) have come up with some model which can help to understand the dynamics of the PDE. The model is show below:

[2] Application Programming Interface (API) is a source-code based specification intended to be used as an interface by software components to communicate with each other – Wikipedia.org.

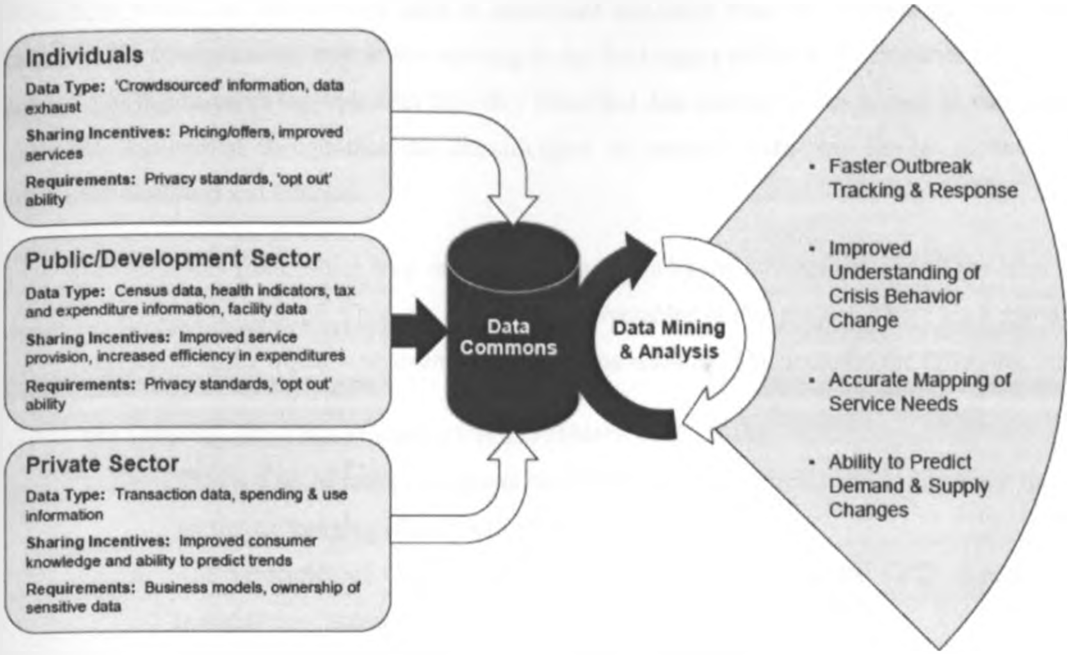
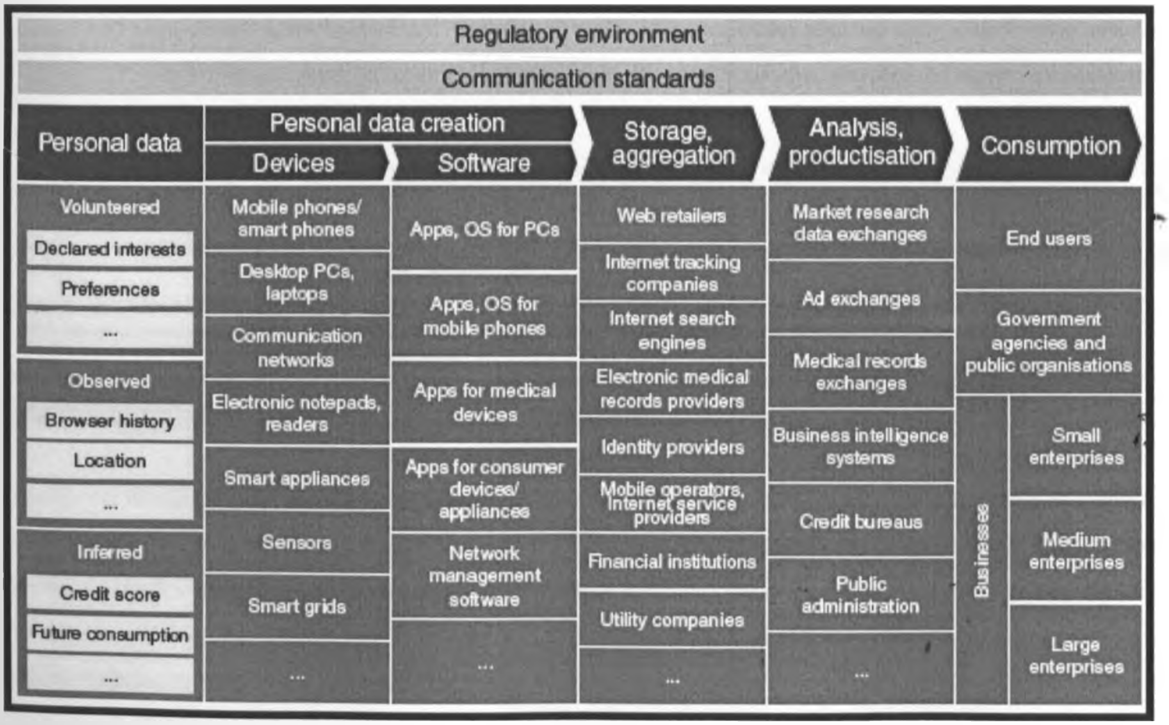

Individuals
Data Type: 'Crowdsourced' information, data exhaust
Sharing Incentives: Pricing/offers, improved services
Requirements: Privacy standards, 'opt out' ability
Public/Development Sector
Data Type: Census data, health indicators, tax and expenditure information, facility data
Sharing Incentives: Improved service provision, increased efficiency in expenditures
Requirements: Privacy standards, 'opt out' ability
Private Sector
Data Type: Transaction data, spending & use information
Sharing Incentives: Improved consumer knowledge and ability to predict trends
Requirements: Business models, ownership of sensitive data
Data Commons
Data Mining & Analysis
• Faster Outbreak Tracking & Response
• Improved Understanding of Crisis Behavior Change
• Accurate Mapping of Service Needs
• Ability to Predict Demand & Supply Changes

Figure 2: Dynamics of a PDE

This model proposes collection of data from the different stakeholders in the PDE to a 'Data Commons', what technically can be called a data warehouse and then using data mining technologies in extracting information which can be used for various purposes by the stakeholders. It should be noted that this is just one of the outcomes of this research.

Bain & Company (2010) came up with a PDE model which they used as a guide in carrying out their research in relation to personal data. The model had details about the PDE from data creation to data consumption:



Regulatory environment

Communication standards

Personal data	Personal data creation		Storage, aggregation	Analysis, productisation	Consumption	
	Devices	Software				
Volunteered	Mobile phones/ smart phones	Apps, OS for PCs	Web retailers	Market research data exchanges	End users	
Declared interests	Desktop PCs, laptops	Apps, OS for mobile phones	Internet tracking companies	Ad exchanges	Government agencies and public organisations	
Preferences	Communication networks	Apps for medical devices	Internet search engines	Medical records exchanges	Businesses	Small enterprises
...	Electronic notepads, readers	Apps for consumer devices/ appliances	Electronic medical records providers	Business intelligence systems		Medium enterprises
Observed	Smart appliances	Network management software	Identity providers	Credit bureaus		Large enterprises
Browser history	Sensors	...	Mobile operators, Internet service providers	Public administration		
Location	Smart grids		Financial institutions	...		
...	...		Utility companies			
Inferred			...			
Credit score						
Future consumption						
...						

Figure 3: The Bain & Company PDE model

It should be noted that the personal data of individual end-users from government agencies (static data) is missing in this ecosystem the way it was missing in the final report of the WEF. However, static data has been considered in this research together with the other identified data sources in this model, as was indicated earlier on. It was appreciated though that the classification of personal data was similar to that of the WEF: volunteered, observed and inferred.

On the interoperability front, which trust model can be adopted by the government and all the other stakeholders in the PDE was also considered. This was one of key deliverables of this project. There are a myriad number of existing trust frameworks which were examined during this research. This includes the following, among others:

- The Open Identity Trust Framework (OITF) (Mary, et al., 2010)
    - This is a set of technical, operational and legal requirements and enforcement mechanisms for parties exchanging identity information.
    - The Principles of Openness are touted as the strength of the OITF model as they afford transparency, accountability and open competition.
- Higgins Open Source Identity Framework (HOSIF) (Eclipse, 2011)
    - This is an open source Internet identity framework designed to integrate identity, profile and social relationship information across multiple sites, applications and devices.
    - However, HOSIF is not a protocol; it is software infrastructure to support a consistent user experience that works with all popular digital identity protocols, including WS-Trust, OpenID, SAML, XDI, LDAP, etc.
- Kantara's Identity Assurance Framework (KIAF) (Kantara Initiative, 2010)
    - This framework introduces the concept of User-Managed Access (UMA). UMA lets web application creators easily craft systems that give control of data back to the individuals. It offers them centralized security, privacy and control for sharing data with friends and family, business associates and organizations. It reinforces the concept of user-centricity in web access.
- Polis Personal Data Management Framework (Algorithms and Privacy Research Unit, Democritus University of Thrace, 2012)
    - This framework has been implemented on the principle that 'all personal data is considered private'.

It should be noted that no truly large-scale implementation of any of the above-mentioned trust frameworks has yet been rolled out. So trials on these trust frameworks will be necessary for a start before a complete rollout.

Privacy by design (PbD) (Privacy By Design, 2011) principle was incorporated in this research. This is one principle which incorporates privacy throughout the life cycle of any large scale project like this one where the privacy of the individual is very important. This principle has been researched and advocated widely by Dr. Ann Cavoukian (Wikipedia, 2011), and even has a dedicated website where this principle has been explored by various experts in the privacy industry.

It should be borne in mind also that one of the objectives of the WEF committee in the *Rethinking Personal Data* project was to find out the principles involved in the PDE. Dr. Ann Cavoukian has come up with seven (foundational) principles (Cavoukian, 2011) that govern personal data sharing under the PbD banner – somehow fulfilling one of the objectives of the WEF committee – and this project has endeavoured to adhere to all of them.

On the other hand, Mydex CIC (Mydex , n.d.) proposed a Personal Data Store (PDS) for individual data management. According to Mydex, a PDS allows the individual to determine who can access what about them. This proposal about personal data management is in tandem with the recommendations put forth by Nissenbaum (1998). However, this approach might not be feasible in this project since the individual citizen is not expected to store any data concerning themselves. However, some portion of their proposals has been adopted for this project since user centricity is paramount to the success of such a project.

It should be noted though that other than Mydex, there are as well a number of other Personal Data Store (PDS) providers who have tested and developed proven PDS technologies which have been explored in this project and evaluated to determine what they can contribute.

According to Mydex report, there are four groups of organizations that are involved in the collection, management, use and sharing of personal data:

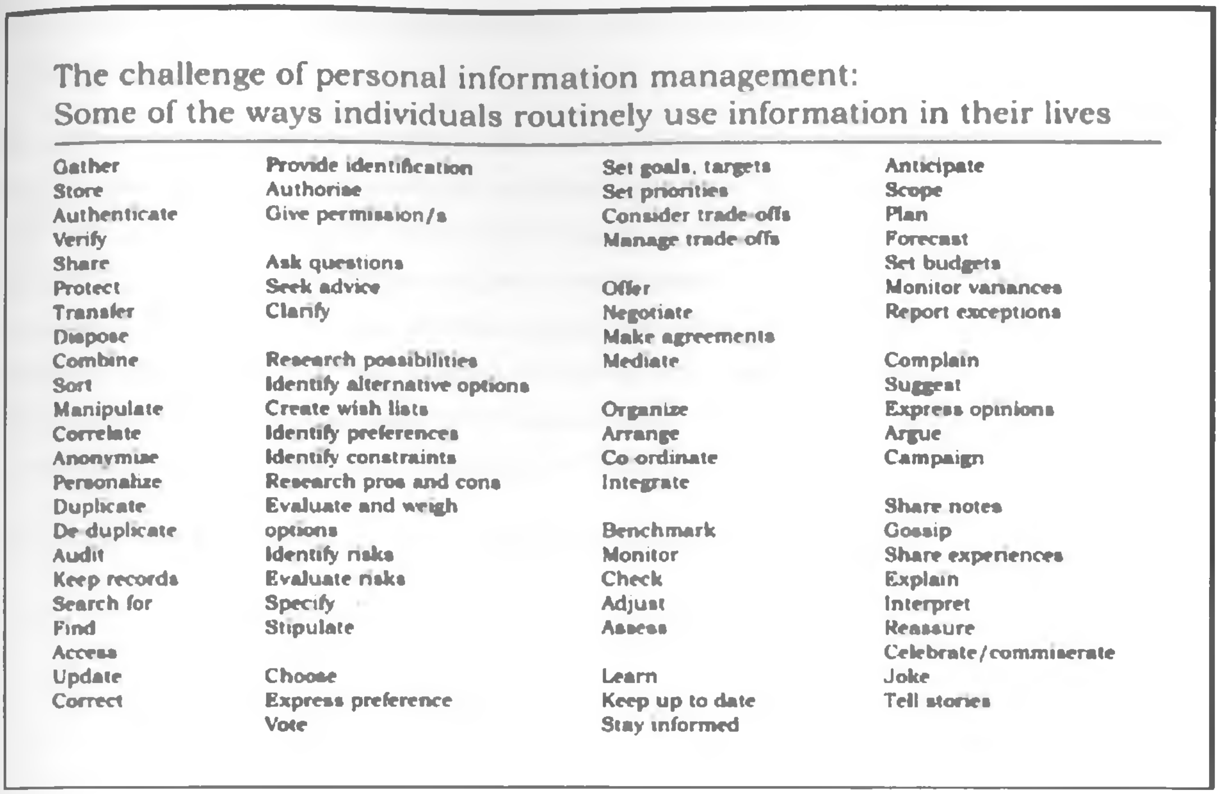
i. Organizations who transact with customers;
ii. Public sector bodies and government;
iii. Third party bodies that collect, process and sell personal data;
iv. Individuals.

This listing mirrors closely to the Bain & Company (2010) PDE model depicted in an earlier figure (fig. 2).

To empower the individual, Mydex propose the deployment of PDSs. The proposed PDS empowers the user to gather, store, access, update and change their personal details and also it empowers them to share their personal information in ways in which they can control – enabling them to choose what information they wish to share with who for what purpose(s). This is the user centricity that we have endeavoured, as well, to maintain in this project.

However, there is a serious weakness in deploying PDS for managing personal data – this is due to the fact that the data is only accessible to the individual and the people who can be allowed access to it. This weakness leaves the government out of the PDE since there are some personal data which does not belong to the individual like criminal records. Therefore PDS as recommended in the Mydex report have not been adopted fully in this research – there has been modifications to include the government's stake and interest.

The report lists ways in which an individual routinely use information in their lives as:



The challenge of personal information management:
Some of the ways individuals routinely use information in their lives

Gather
Store
Authenticate
Verify
Share
Protect
Transfer
Dispose
Combine
Sort
Manipulate
Correlate
Anonymise
Personalize
Duplicate
De-duplicate
Audit
Keep records
Search for
Find
Access
Update
Correct

Provide identification
Authorise
Give permission/s

Ask questions
Seek advice
Clarify

Research possibilities
Identify alternative options
Create wish lists
Identify preferences
Identify constraints
Research pros and cons
Evaluate and weigh options
Identify risks
Evaluate risks
Specify
Stipulate

Choose
Express preference
Vote

Set goals, targets
Set priorities
Consider trade-offs
Manage trade-offs

Offer
Negotiate
Make agreements
Mediate

Organize
Arrange
Co-ordinate
Integrate

Benchmark
Monitor
Check
Adjust
Assess

Learn
Keep up to date
Stay informed

Anticipate
Scope
Plan
Forecast
Set budgets
Monitor variances
Report exceptions

Complain
Suggest
Express opinions
Argue
Campaign

Share notes
Gossip
Share experiences
Explain
Interpret
Reassure
Celebrate/commiserate
Joke
Tell stories

Figure 4: Mydex: Ways in which individuals routinely use information in their lives

The report further indicates that PDS help in information sharing under what is called *Select Disclosure. Select Disclosure* functions like an Information Sharing Agreement: the individual specifies what information they wish to share with which organizations or individuals, for what purposes under what terms and conditions. It works in two ways: *bespoke and automatic*. Bespoke information sharing happens on a one-to-one basis – it is negotiated individually for each new situation. In this case the individual might allow access to his personal information for free or charge for it under his terms.

The second type of *Select Disclosure* is an automatic 'subscribe to me' service. Here organizations subscribe to updates from specific fields within the individual's PDS. To gain access they have to sign the individual's terms and conditions. The individual can choose which organizations he or she wishes to accept or reject as a subscriber. Once the subscription is in place, every time the individual changes the relevant field in his data store, the subscribing organization is alerted to this fact. This is true user centricity – the individual is in charge and has complete control of his personal data.

With this research, *bespoke Select Disclosure* is what was considered since it is about interested stakeholders who want to access anonymized personal data of individuals. Automatic disclosure might be relevant to institutions like banks which want to keep track of their clients.

What about privacy of the personal data in a PDS setup? The report proposes that this issue be handled by the user himself. He has authority over whom to give access to what and even in what context. So the individual is empowered in privacy settings; this is a principle which Mydex calls *privacy as a personal setting*. This is something that has been adopted when dealing with dynamic sets of data for the individual citizen and even some static data sets.

With the current ubiquitous use of social media, it should be possible for the portal/system where users access their personal data, held by the Government, to link with social sites like Facebook, Twitter, LinkedIn, etc to be able to scour these sites for users' dynamic data so that the same can be integrated to the portal and makes up the user's dynamic data. This will help to give a complete picture of a user's web profile as well as save the citizen the hustle of keying in same information again. This can be achieved through use of APIs or real-time mash-ups which must give the user an option to allow access to their personal data on the social site or even options on what data the tool can access from the site. A classic example of this is Facebook Connect (Facebook Limited, 2011) which was launched by Facebook on 9th May 2008.

Mydex identifies the following as being important to successful personal data sharing:

i. Exemplary data security, both in storage and sharing;
ii. Absolute ease of use – the user interface should be user friendly;
iii. Easy population of the data store, with equally easy access and use – to correct, update, link, share etc;
iv. Easy-to-use and understand data sharing agreements, protocols and processes;
v. The development of technical, legal and other standards to support data sharing and data sharing agreements;
vi. The ability of data fields in PDS to talk to data fields in organizations' databases without confusion or error. This will require the development of sophisticated data architectures.
vii. The ability to gather and share bespoke bundles of data from the data store.

This research has tried to explore on how the above conditions can be met in a Kenyan setup in regard to personal data resting on some government server. However, it should be noted that some of these issues were indicated as areas of tension by the WEF Report (2011) as well as by the NSTIC (The WhiteHouse, 2011) which were mentioned earlier on. And hence this underscores their importance in the PDE.

According to Mydex, some of the key attributes of information sharing agreements should be:

i. They are practically oriented and specific, focussing on a specific problem or information sharing need;
ii. They release a genuinely new class of information – Volunteered Personal Information (VPI) that previously only the individual knew, could see, or had access to;
iii. They give individuals the confidence to share information they would have previously withheld – because now they know appropriate safeguards are in place;
iv. They are – have to be – user-friendly;
v. They are machine-readable so their generation and consumption can be automated, including their comparison to a baseline set that are pre-approved by the individual;
vi. They operate above the level of all global privacy regulations offering individuals and organizations a release from country-to-country regulation differences, arbitrage etc;

vii. The deployment of these agreements within a broader trust framework will create a secure, efficient and workable foundation for rich, mass scale information sharing between individuals and organizations.

It should be noted that these information sharing agreements requirements fall under interoperability requirements in the PDE as was identified by the NSTIC (The WhiteHouse, 2011). They should be watertight for effective data sharing.

Further, it will be noted that later in this chapter, the success cases for effective data sharing that have happened so far have closely adhered to the above attributes of data sharing. It has been noted from research findings that this is not be far from the truth.

The benefit of development of value-adding services has been demonstrated as a possibility when you see the number of applications which have been developed based on the open data available on the government Open Data Portal (Government of Kenya, 2012).

Keele (Keele, 2009) introduced the concept of *privacy by deletion*: where personal data is deleted when it is deemed that it is no longer necessary for the purpose for which it was collected. This concept although sounds very practical and realistic, it might not apply for static data but might need to be considered for dynamic aspects of an individual's personal data. Then the question is: When can dynamic sets of an individual's personal data be deemed 'no longer necessary'? This is a subject that has been explored in this study to find an acceptable timeline within which this can be done and whether it applies to dynamic personal data or static data as well. This thought has also been explored to see where it is applicable (or whether it is applicable at all): at the PDS or during the data sharing phase (to be included in the information sharing agreement?)?

The issue raised by Keele (Keele, 2009) can also be handled by the privacy standards to be adopted or legislated like the Data Protection Bill, 2012 (Commission for the Implementation of the Constitution, 2012). In fact, this bill lists one of the key principles to be followed when dealing with personal data as: '*information is not kept for any longer than is necessary for achieving the purpose for which it was collected*'. If this Act gets legislated into law, it will definitely meet this requirement as was envisaged by Keele.

Success cases

Some of the encouraging cases where personal and public data is being shared by an individual (or government) to another third party include the following:

a. ***The Blue Button*** (The White House, 2011) project by the US Government: This is a web-based system through which patients easily download their health information and share it with health care providers, caregivers and others they trust. It is a partnership between the Department of Veterans Affairs (VA) and the Centres for Medicare & Medicaid Services (CMS) in the Department of Health and Human Services (HHS). It was launched in August 2010 and so far has recorded great success.

The sharing of data between the veterans and the medicare providers is done using a simple ASCII text file. This sorts out the technical interoperability between these stakeholders in this project on data sharing.

The benefits: Having ready access to personal health information from Medicare claims helps beneficiaries understand their medical history and partner more effectively with providers. With the advent of the Blue Button project, Medicare beneficiaries are able to view their claims and self-entered information—and be able to export that data onto their own computer.

The Blue Button system is accessible to about one million Veterans as well as 47 million Medicare enrolees.

b. ***Citizens @ The Centre: B.C Government 2.0*** (British Columbia Public Service, n.d.): This is a strategy of the British Columbia (BC) Public Service. Their number one principle is to empower citizens to create value from open government data. They want the citizen to get value of what the government has publicly availed to them. This is important since the government has opened up its data but not yet the individuals'. The strategy recognizes that 'the growing movement towards sharing of data also has great potential benefits as citizens adapt and combine government data in creative new ways.'

The major challenge that was highlighted in this strategy is on identity management: how to confirm that the citizen is who he claims he is. They were proposing an electronic system in which citizens will be issued with electronic credentials which uniquely identifies each citizen of the BC. This way they can be sure that the person claiming to be him is indeed him.

We thought this is a good approach which can also be adopted in this project like using the online PIN issuance process of KRA: each citizen applies online for his access credentials and once he logs into the system it can be confirmed that it is indeed him. May be this might need to be enhanced to include features like a photograph and other biological features like fingerprint, iris scan etc.

c. ***The Kenya Open Data Initiative (ODI)*** (Government of Kenya, 2012) : The government has shared some of its data to the public. Then is it not obvious that the next stage should the individual citizen too to share his data?

d. ***The Moldova Open Data Initiative*** (Republic of Moldova, 2011): Implemented in similar fashion with the Kenya's ODI project.

e. ***The India's Magic Project – the Unique Identity Project*** (The Economist, 2012)*:* This project involves scanning of poor Indian's citizens' irises and fingerprints for tracking purposes for the distribution of relief food. Although there has been opposition to the project due to issues about privacy and confidentiality, the citizens themselves are fast embracing the idea because they are directly benefiting from it and in this way preventing corruption in the relief food distribution chain.

This project mirrors closely with our research's objectives: citizens sharing their personal data details with commensurate direct benefits. However, there is a difference since this project collects identifying information (which is key to them) unlike ours where identification of individual citizens is not anticipated.

Summary

From the researches that have been carried out in the PDE from both legal and technical aspects, it has been noted that most of it and debate is mostly centred on dynamic aspects of an individual's personal data. This project however is expected to incorporate static data into the PDE so that we can harness its value and have a complete picture of the citizenry.

Therefore, it should be noted that the expected list of personal data details to be captured from the citizens have incorporated static data. This means that most of the proposed static data included on the to-be captured data is ours.

Also, we will be looking into a business model which can help the different stakeholders (individuals, government and businesses) in this ecosystem to make money from the citizens' personal data. This is one area which has not been fully explored in other researches in regard to personal data. Although there was mention of 'value proposition' by reports such as that of the WEF Report (2011), there were no precise information on incentives that can be preferred by individuals in the PDE for them to submit or share their personal details. This has been covered in this research.

Study design

Methodology

The following research techniques were used to carry out this study:

i. Interviews with relevant stakeholders in the PDE;

ii. Review of what has been done by the GoK and researchers in this area including:

   a. Analysis of current and proposed legislation and regulations governing privacy;

   b. Analysis of existing trust frameworks and their implementation in identity solutions (trial / pilot or live deployments);

   c. Analysis of the current government ICT infrastructure.

iii. Administration of a structured questionnaire;

iv. Development of a system prototype to implement the survey findings.

Research Framework

The following steps were followed in carrying out this research:

- ***Problem statement*** – to understand what the research was going to find out. In this particular case it was about a business model to encourage citizens to open up their personal data for anonymous statistical analysis ;
- ***Background study*** – a study of currently what is happening in the personal data landscape and how this research was going to add to that body of knowledge;
- ***Questionnaire survey*** – came up with a raft of questions that helped us to solicit responses from the citizenry about their perception in regard to sharing their personal data;
- ***Data collection*** – collected responses from the citizenry using interviews and a structured questionnaire;
- ***Data pre-processing*** – formatted the collected raw responses to a format which was easy to analyze by available data processing programs;
- ***Data analysis*** – analyzed the received data to get the perception in regard to personal data management from the citizens;
- ***Study findings and conclusions*** – from the data analysis exercise came up with a list of findings in regard to the objectives of the study;
- ***Prototype development*** – developed a prototype that can help in personal data sharing in the PDE.

Diagrammatically, the research framework below was used in carrying out this study.

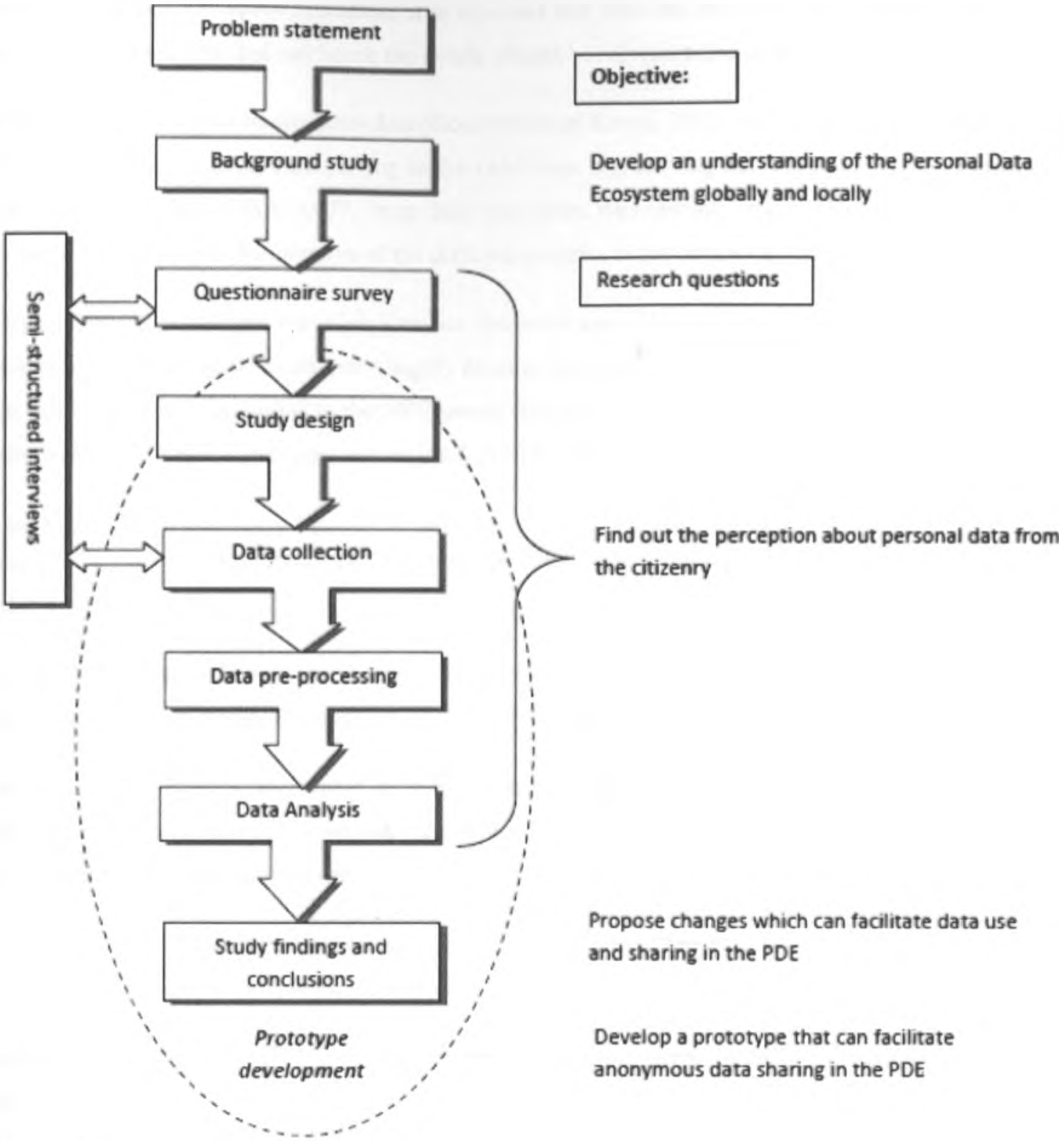

Problem statement
Objective:
Background study
Develop an understanding of the Personal Data Ecosystem globally and locally
Semi-structured interviews
Questionnaire survey
Research questions
Study design
Data collection
Find out the perception about personal data from the citizenry
Data pre-processing
Data Analysis
Study findings and conclusions
Propose changes which can facilitate data use and sharing in the PDE
Prototype development
Develop a prototype that can facilitate anonymous data sharing in the PDE

Figure 5: Research Framework

Sample design

The research method that was used to carry out this study was *quota sampling* (Kothari, 2004). This was due to the fact that we wanted views of anonymous statistical use of personal data both from the urban and rural areas – these two areas make up two strata as per this research method. From these two strata, we picked the citizens who participated in this research using the *simple random sampling technique* (Kothari, 2004). This method was chosen so that every Kenyan (in the two strata), has an equal of chance of being selected to participate in this survey.

It should be noted though that an ideal research design for this exercise would have been a census[3]; where every Kenyan would have given his/her opinion on the subject matter. However, due to the limitation on 'expenditure of effort, time and money' according to Kothari (2004), this was considered not feasible and hence the chosen

[3] Census is a complete enumeration of all items (citizens) in a population (Kothari, 2004, p. 14)

methodology. Despite the above limitation, it is expected that with the chosen research method, the error rate will considerably be minimized and hence the results should be representative of the perception on the ground.

According to the population census 2009 data (Government of Kenya, 2012), we had 12,487,375 Kenyans living in urban areas and 26,122,722 others living in the rural areas representing 32.30% and 67.70% respectively of Kenya's total population of 38,610,097. From these two strata, we randomly selected Kenyans who participated in this survey based on the classification of the different counties in the country as per this census exercise.

The targeted sample population was adult Kenyans above the age of 18 years old. This is due to the fact that it is only people above this age who can make legally binding decisions according to the law of the land. Out of the 38 plus million Kenyans, according to the 2009 census (Government of Kenya, 2012), Kenyans above the age of 18 years were 19,414,893: Urban population was 7,219,183 while rural population was 12,195,710.

Sample size

The expected sample size (quota) for each stratum was determined using the Survey Systems website (Creative Research Systems, 2012).

The confidence was taken as 95% (so standard deviation =1.96 – as per table of area under normal curve of 95% confidence level) and acceptable error as ±5.

At confidence level of 95%, acceptable error of ±5 and population of 19,414,893, using the Survey Systems (Creative Research Systems, 2012) website the required total sample size is *384*. This figure was segregated to required samples from the rural and urban areas as follows:

- Sample size required from urban areas is:32.30% X 384 = 124.032 ≈ 124
- Sample size required from rural areas is: 67.70% X 384 = 259.968 ≈ 260

The formula used for the Survey System sample size calculator (Creative Research Systems, 2012) is as shown below:

$$SS = \frac{Z^{2*}(p)^{*}(1-p)}{c^{2}}$$

Where:

Z = Z value (e.g. 1.96 for 95% confidence level)

p = percentage picking a choice, expressed as decimal (.5 used for sample size needed)

c = confidence interval, expressed as decimal (e.g., .04 = ±4)

Correction for finite population formula:

$$\text{new ss} = \frac{\text{ss}}{1 + \frac{\text{ss}-1}{\text{pop}}}$$

Where: pop = population

Data collection

A survey was at the heart of this research. The survey covered both rural and urban Kenya. This was to help us get balanced and unbiased views from the population's perception on personal data.

The survey was carried out by circulation of questionnaires with specific questions in relation to personal data as well reading of the same to illiterate or semi-illiterate citizens for interpretations and filling in their responses. Some of the questions required quantitative as well as qualitative feedback and hence data analysis was both quantitative as well as qualitative.

Also, the web was used to seek feedback from Kenyans in the diaspora (as well as locally) on their perception about personal data and anonymous statistical analysis of the same. However, the final analysis is based on the Kenyan population whether living locally or abroad. This is due to the fact that an online survey is expected to attract Kenyans as well as non-Kenyans. And indeed we did get some few non-Kenyans filling the online questionnaire.

Google Docs was used for online survey due to the fact that is free and allows the researcher the option to download the raw data for any further analysis. Also, it does not have a limitation on the duration for which you should have accessed and downloaded the data. Most online survey tools have limitations of being expensive as well as the time frame within which you should download your data from their servers.

Pilot run of the survey tool

A trial run of the survey was carried out between 4th and 9th April 2012. From the results of the pilot, it necessitated some few modifications on the tool. The real survey was carried out between 12th April 2012 to 4th June 2012. Although the survey targeted people of 18 years and above, it should be noted though that there was a provision for people below 18 years to fill the questionnaire (for online surveying) but their feedback was expected to be discarded – not to be considered for data analysis. However, at the end of the data collection exercise it was noted that there was no citizen below 18 years who filled the questionnaire, whether the one available online or the paper-based one.

For the paper-based survey, the researcher had to move from one area to another assisted by assistants whom he had trained on how to administer the questionnaire. Questions which were not clear were expounded by the researcher and his assistants to interviewees.

Since it was expected that the online questionnaire might reach a wider outreach (including non-Kenyans) it was modified to include this class of people as well as Kenyans in the diaspora.

Kenyans in the diaspora were considered as belonging to the urban class in the data analysis.

Data pre-processing

From the collected data for the paper-based questionnaires, the responses were keyed in to Microsoft Excel in the format which the online survey had for easy analysis. The data was as well cleaned up of any typographical errors ready for analysis stage.

Data analysis

Data analysis was carried out using Microsoft Excel. The analysis was carried out by us. Also data sourced from government systems and publications was used in comparison with data which was collected from this research.

As was noted above, data analysis was both quantitative as well as qualitative.

Research findings

Gender representation

Out of the received responses, 36% were from the female respondents while the remainder (64%) were from their male counterpart – both online and on paper.

Sharing of anonymized personal data with an interested party

There was a higher perception that citizens are willing to share their personal data (62%) with the government or an interested party compared to those who were not willing (38%).

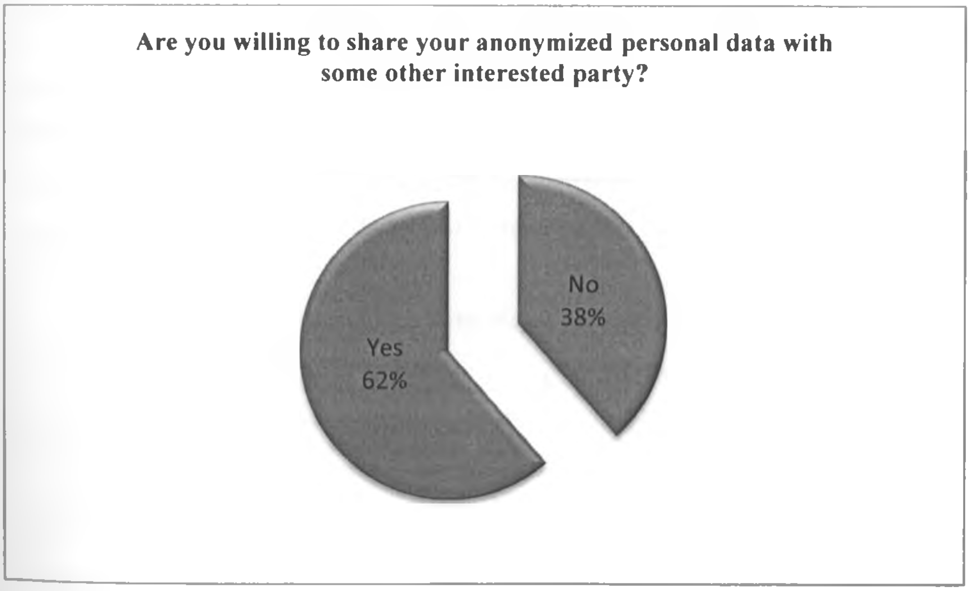


Figure 6: Willingness to share data

This finding mirrors closely with the survey which was carried out by the Office of the Privacy Commissioner of Canada (Office of the Privacy Commissioner of Canada, 2011) which found that 44% of Canadians did not approve sharing of their personal data with the United States (US). However, it should be noted that their research was about sharing of identifying data rather than anonymized data.

Demographics

However, for the sampled population, 30% of the female population was NOT willing to share their personal data while 70% were willing to share their personal data. On the male front, 43% were NOT willing to share their personal data while the remainder was willing to share their personal data. From this, it can be concluded that *women are more willing to share their personal data with the government and/or other interested party than their male counterpart.*

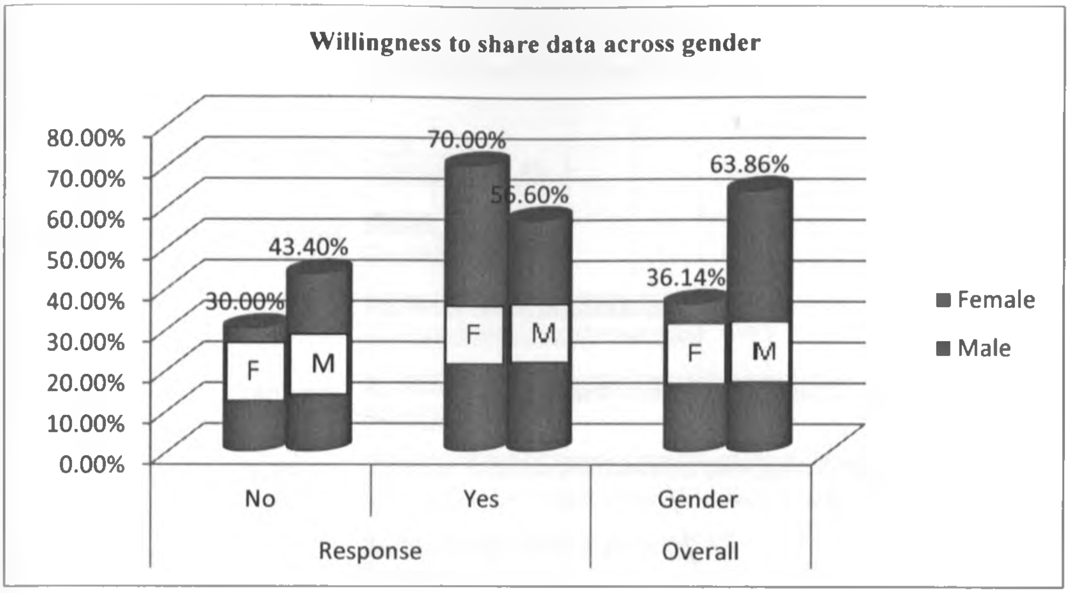


Figure 7: Willingness to share data across gender

Willingness to share personal data across the different age groups

On the issue of age, there was general willingness to share their data save for age groups 26 – 35 and 46 – 55 where more of the sampled population were not willing to share their data compared to those who were willing. However, it should be noted that among the sampled population, *the young people (18 – 25) were more willing to share their data compared to their older compatriots.*

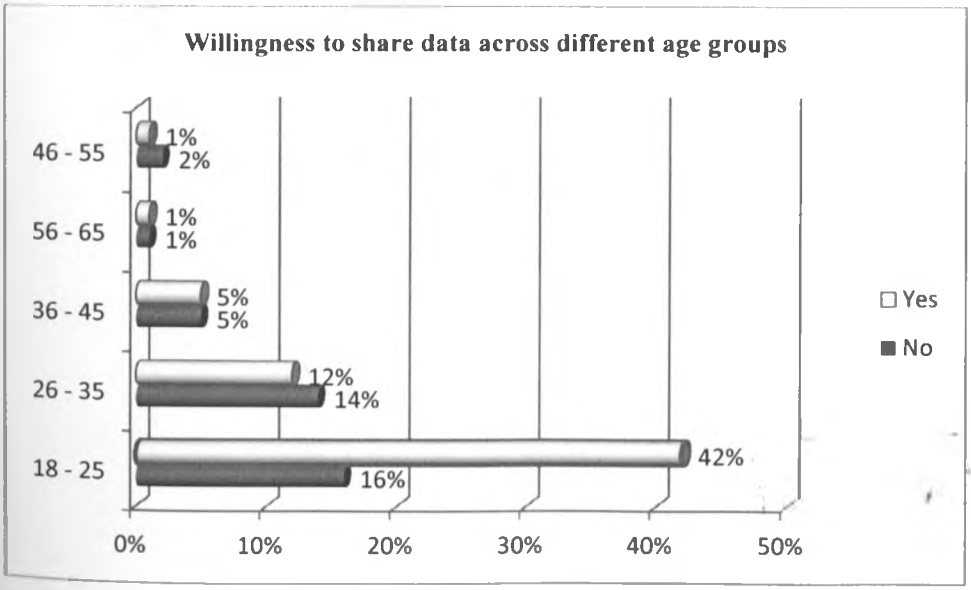


Figure 8: Willingness to share data across age groups

Educational level and willingness to share data

From the survey, it was clear that *there was close correlation between educational level and willingness or non-willingness to share personal data*. Citizens with less education were reluctant to share their data compared to those with higher education. It was noted for example that citizens with no formal schooling were not at all willing to share their personal details.

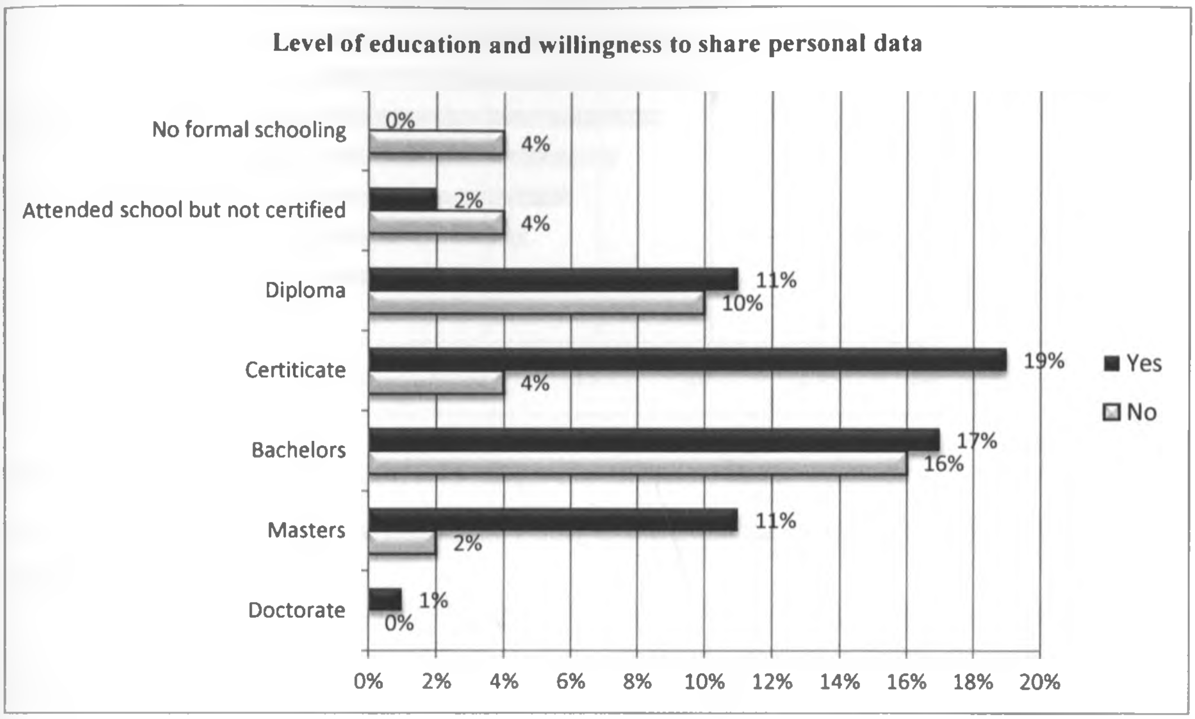


Figure 9: Level of education and willingness to share data

Personal data details that the citizens are willing to share

Below are the details citizens are willing to share frequency chart:

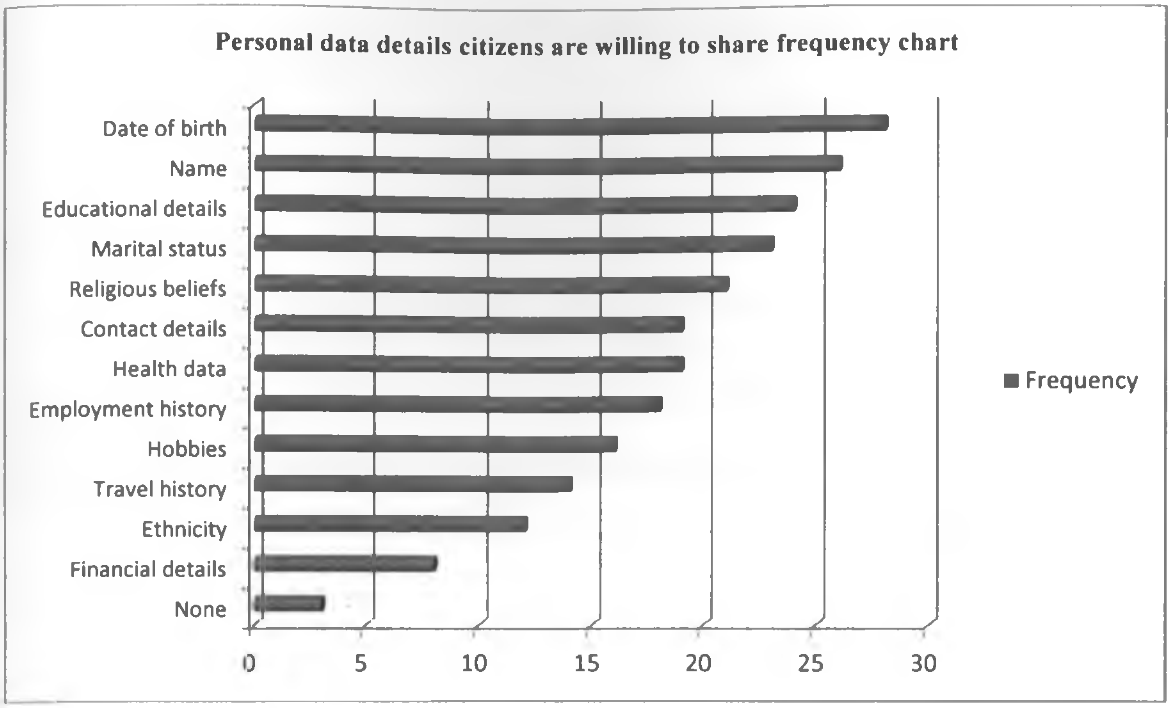


Figure 10: Personal data details citizens are willing to share frequency chart

It is clear from the above chart that citizens are willing to share with the government and other interested parties the following details:

- Name;
- Date of birth;
- Marital status;
- Hobbies;
- Educational details;
- Employment history.

It should be noted that this is quite contrary to the research expectations since we were trying to find out if citizens were willing to share their non-identifying personal data details. It is clear that sharing your name will definitely and largely help to unmask your identity!

Common reasons for non-willingness to share personal data across the citizenry

Some of the reasons given by respondents for not willing to share their personal data included – from both gender:

- Privacy concerns in regard to identity theft;
- Enjoying/liking their privacy;
- No trust of people/institutions which will handle the data;
- A feeling that their security will be compromised;
- A feeling that personal data is private and hence only for him/her alone;
- A fear that their data might be used for wrong purposes;

- Fear of who might access their data – they indicated that they needed to know the person/entity first before sharing their data;
- Need to know the usage of data they are providing before providing it;
- Due to the high insecurity in the country.

Reasons for non-willingness to share data and level of education

The following were some of the reasons which were advanced by the respondents on why they were not willing to share their personal data, across the different levels of education.

Are you willing to share your anonymized personal data with some other interested party?	**Category of concern/reason**
Level of education and reasons for not wanting to share personal data	
Bachelors	
Because I am not sure who is accessing it. And if they do for what purpose?	Privacy and security
Because my data is private	Privacy
For me to disclose any data, I have to be in a position to know or at least guess the intentions of whoever will acquire that data, that is, what does he/she intend to do with the data?	Privacy and security – data sharing objectives
I cannot trust other people with my data	Privacy and security
I do not fully understand the motive behind the interested party wanting my information / data.	Privacy and security – data sharing objectives
I do not trust most of the people/institutions handling my data	Privacy and security
I would need to know the use of the information I am providing to enable me to provide this data even if it is anonymous.	Privacy and security – data sharing objectives
I would not be comfortable with my personal data being shared unless I know the purpose it is intended for	Privacy and security – data sharing objectives
It is important to know the identity of other person you are in communication with.	Privacy and security – identity theft
Personal pertains to AN INDIVIDUAL, it is private.	Privacy
Privacy concerns, stolen identity e.t.c.	Privacy and security – identity theft
It might compromise on my security	Personal security
My data is private and hence confidential	Privacy
Certificate	
Because my data is private	Privacy
I fear that they might use my data for wrong purposes.	Privacy and security – identity theft
I would like to keep my personal information private.	Privacy
Diploma	
Because my data is private	Privacy
Due to high level of insecurity in the country.	Personal security
I enjoy/like my privacy.	Privacy
I like my privacy	Privacy
It will not help	Ignorance
Doctorate	
None – all surveyed indicated willingness to share their data	
Masters	
Some data is very personal to me and hence I'm not willing to share.	Privacy
Unless I know the interested party and fully aware of the purpose of the data	Privacy and security – data sharing objectives and identity theft
No formal schooling	
Because it cannot assist me	Ignorance
Because my data is private	Privacy
Attended school but not certified	
Because my data is private	Privacy
It is personal property	Privacy

Table 1: Reasons for non-willingness to share data and level of education

It was noted that most citizens were not willing to share their data with the government or other interested stakeholders in the PDE due to privacy and security concerns. This means that if the government educates them on the benefits of data sharing and guarantees them security (that their data cannot be accessed without their permission) there is a high probability that they will soften their stance to personal data sharing. These concerns were replicated across the different educational levels. However, it was noted that the higher the educational qualification a respondent had, they demanded more security of their personal data. Most respondents with lower level education were more concerned with their privacy as compared to their counterpart with higher education, who demanded both – privacy and security.

Personal data details that the citizens are NOT willing to share – whether the data is anonymized or not

The following are some of the personal data details that citizens were NOT willing to share with any other party:

- Financial data like salary, monthly earnings from business ventures, bank account etc;
- Health data like HIV status etc;
- Contact details;
- Ethnicity;
- Religious beliefs.

Incentives for sharing personal data

Citizens indicated the following as key motivators for sharing their anonymized personal data details, arranged in order of priority:

- If somebody explains to me the data sharing objectives;
- If my data sharing will help other Kenyans;
- If I receive payment in kind, cash or through electronic means like Mpesa, Paypal etc;
- If I'm promised that the research findings will be shared with me later;
- I don't need to be informed or paid but my data can be used.

These are the major incentives that citizens indicated for data sharing. Hence any exercise that will require their consent for data sharing will have to meet these expectations.

However, there were some few cases where the citizens did not care if they will be paid or not – they were still willing to share their data. Also, there were cases where citizens were willing to be paid in kind (for example by being granted discounts to access various government services).

The graph below shows the level of rating given to each incentive:

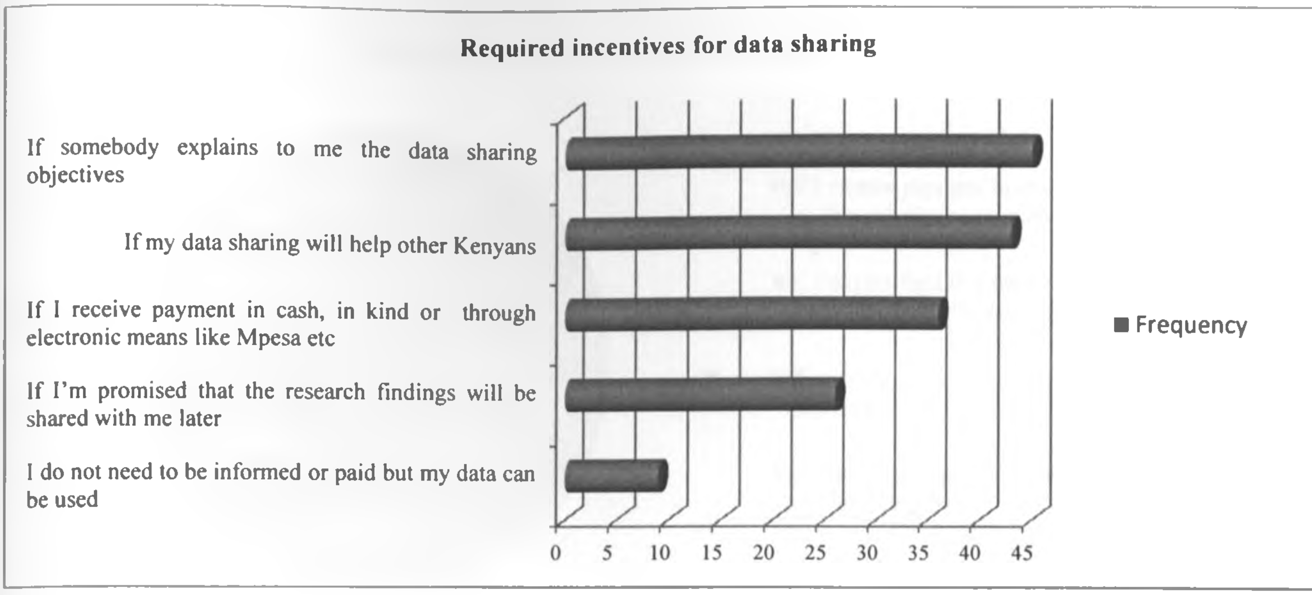


Figure 11: Incentives for data sharing in the PDE

The frequency table for the above graph is as shown below:

Required incentives for personal data sharing	***Frequency***
I do not need to be informed or paid but my data can be used	9
If I'm promised that the research findings will be shared with me later	26
If I receive payment in cash, in kind or through electronic means like Mpesa etc	36
If my data sharing will help other Kenyans	43
If somebody explains to me the data sharing objectives	45

Table 2: Frequency table for incentives for data sharing in the PDE

Educational level and required incentives for data sharing

The following are the findings as per level of education and the required incentives for data sharing:

a. No formal schooling

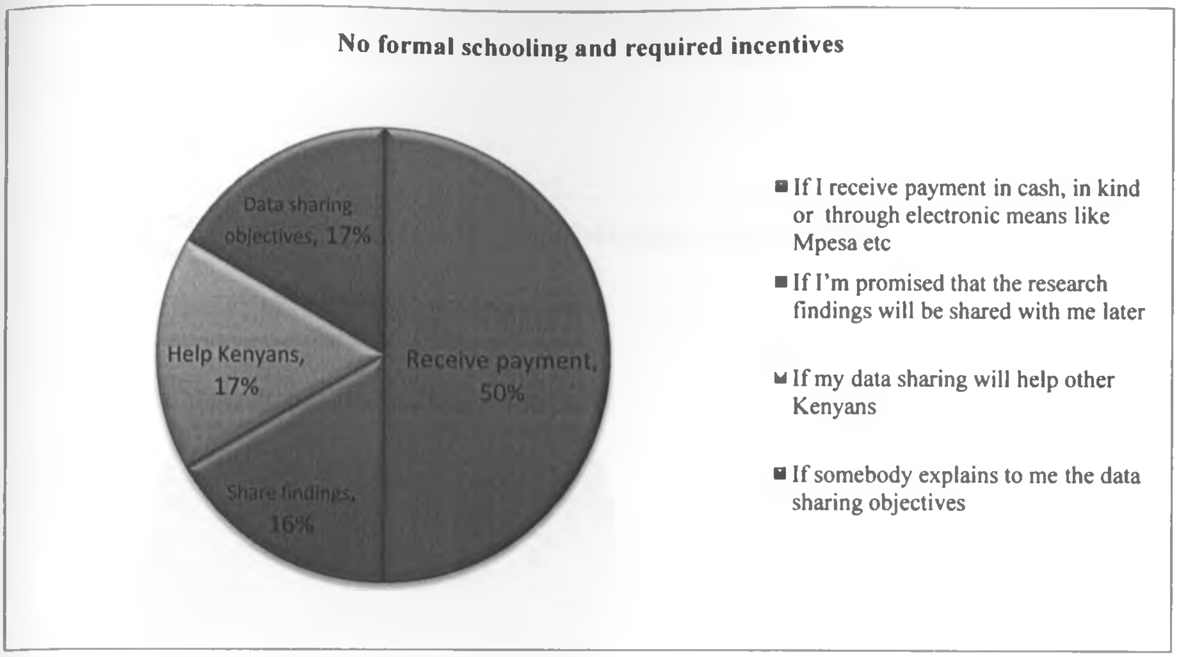


Figure 12: Required incentives for citizens with no formal schooling

Clearly, their major incentive is payment of some form being made to them before access to their personal data is granted.

b. *Attended school but not certified*

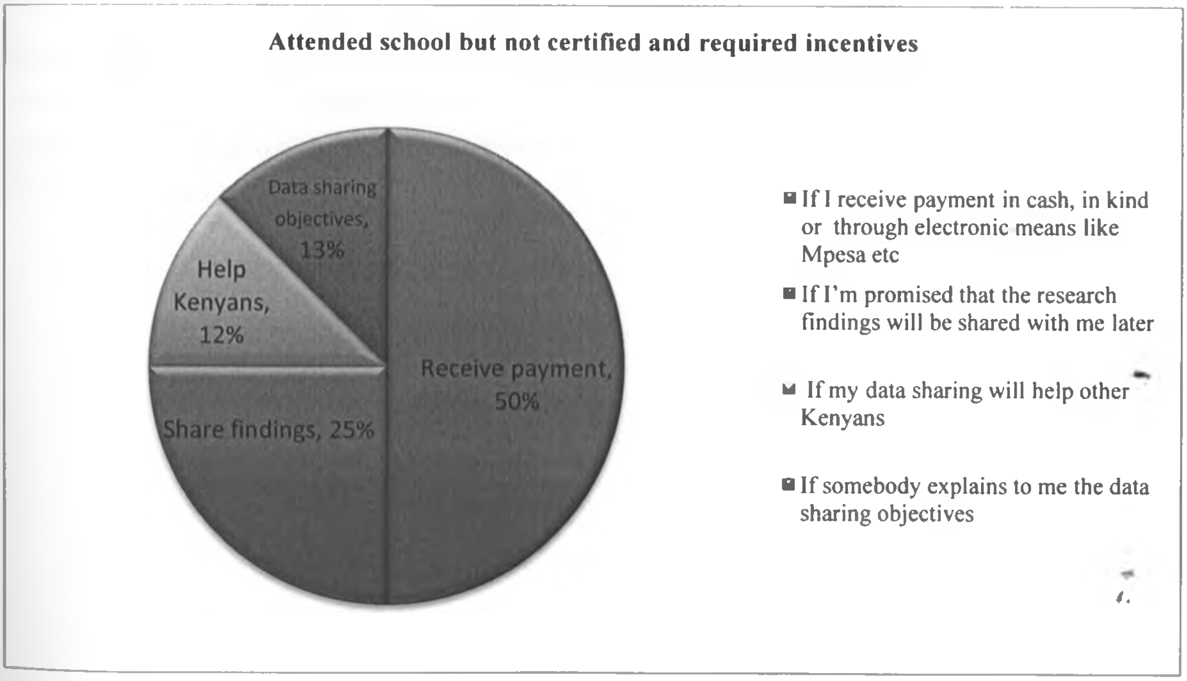


Figure 13: Required incentives at Attended school but not certified

Clearly, also, their major incentive is payment of some form being made to them before access to their personal data is granted. Although it should be noted that with this group, the desire to be informed of the research findings for access granted to their personal data has increased from 16% in the former group to 25%. This might imply that they want to know of what became of their data sharing exercise and what might they learn from others in the PDE.

c. Certificate

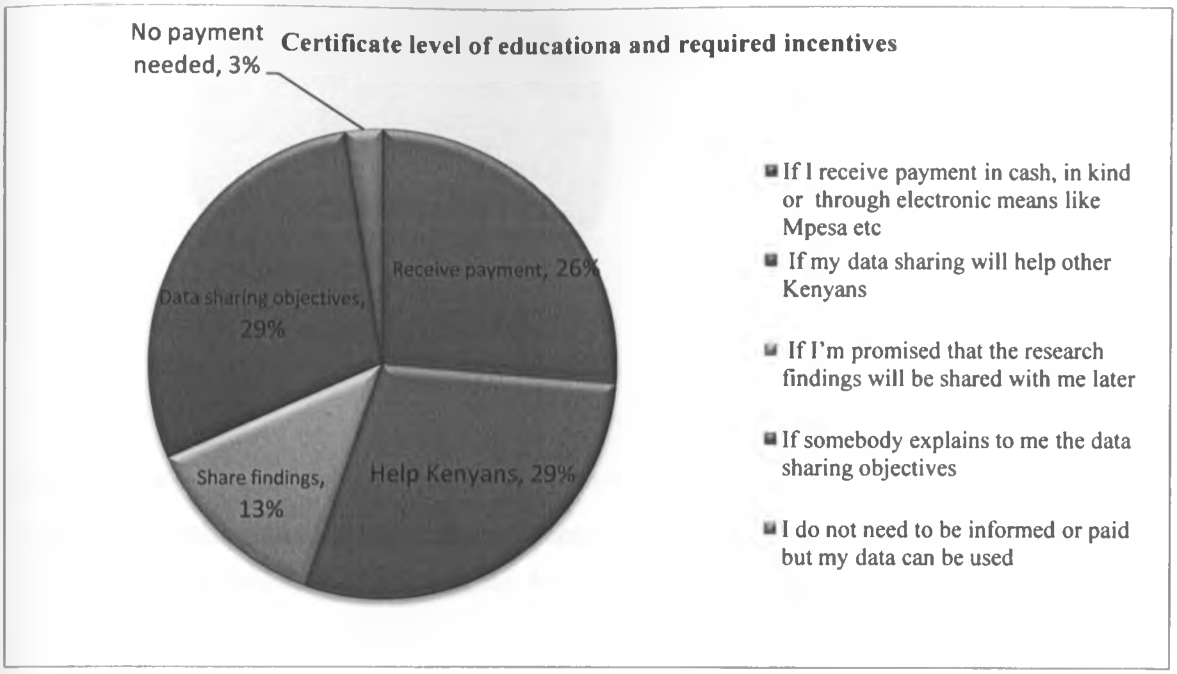


Figure 14: Required incentives at Certificate level of education

It was noted that this group considered the incentives of knowing the data sharing objectives and willingness to help other Kenyans more important than other incentives in data sharing. The incentive to receive payment shrunk to 26% compared with the other two former cases where this incentive was considered important by half of the sampled population.

d. Diploma level

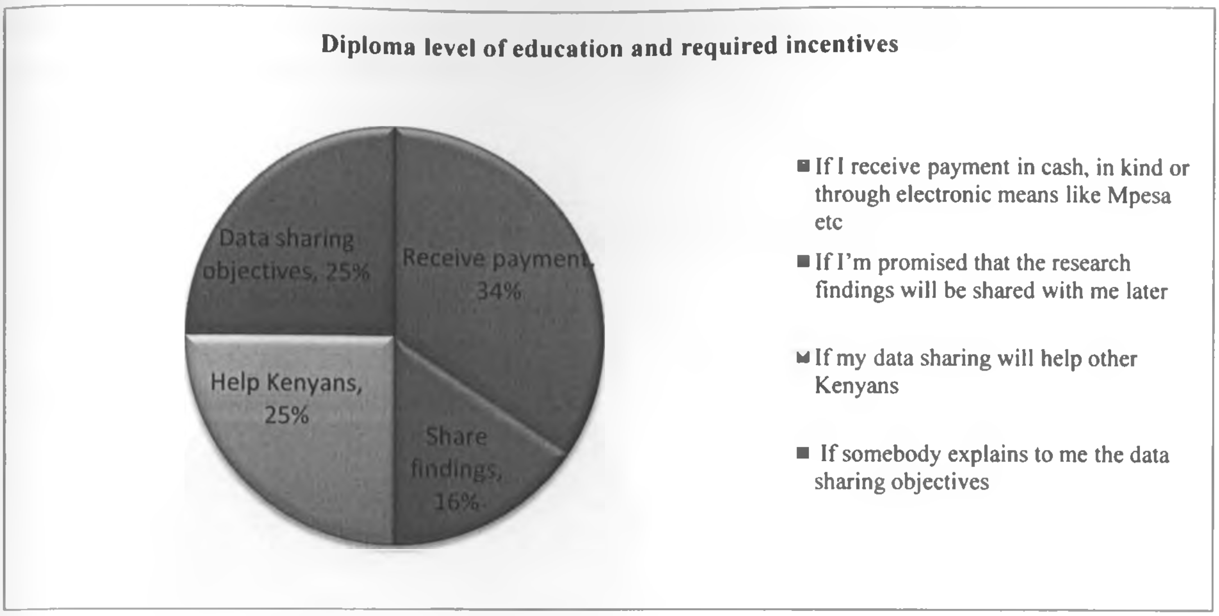


Figure 15: Required incentives at Diploma level of education

It was clear that this group had a higher desire to receive payment for them to share their data. However, it was noted that the other incentives of knowing the data sharing objectives and willingness to help other Kenyans were close second ; considered important each by a quarter of the sampled population.

e. Bachelors

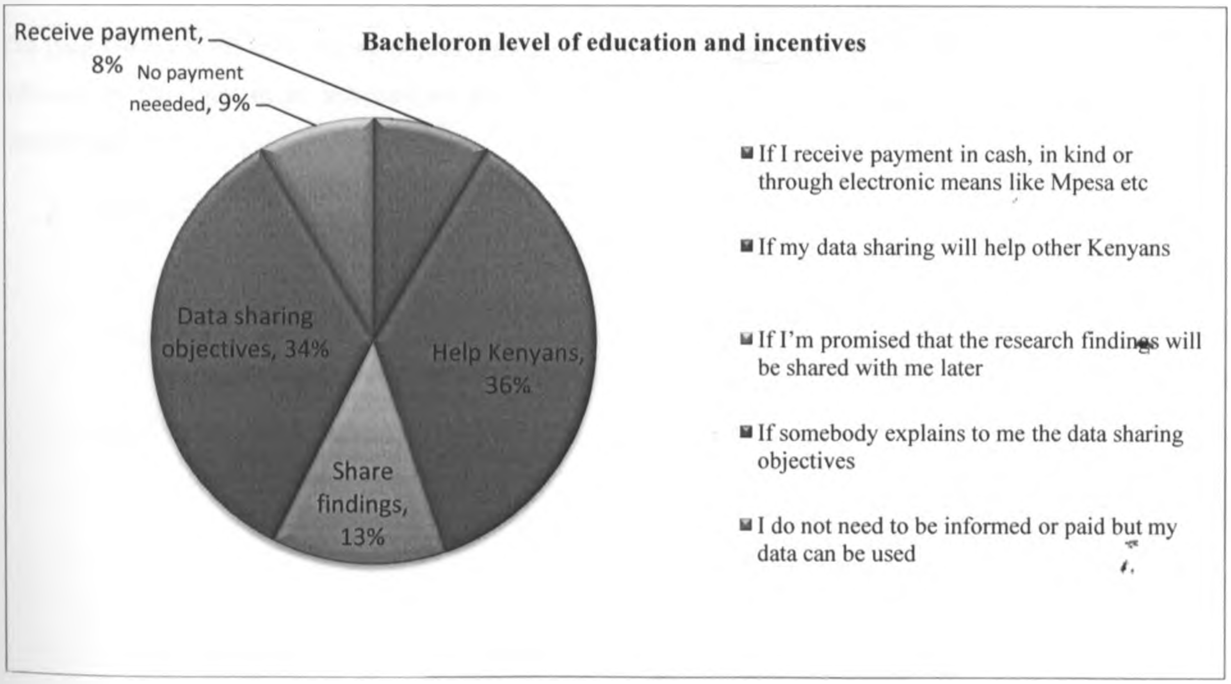


Figure 16: Required incentives at Bachelor level of education

For this group, it was noted that the desire to help other Kenyans and knowing the data sharing objectives were more important compared to the other motivators. The incentive of being paid to allow access to their personal data was noted to have considerably shrunk to a paltry 8% - almost the same number as to those who did not care to be paid to grant access to their anonymized personal data!

f. Masters

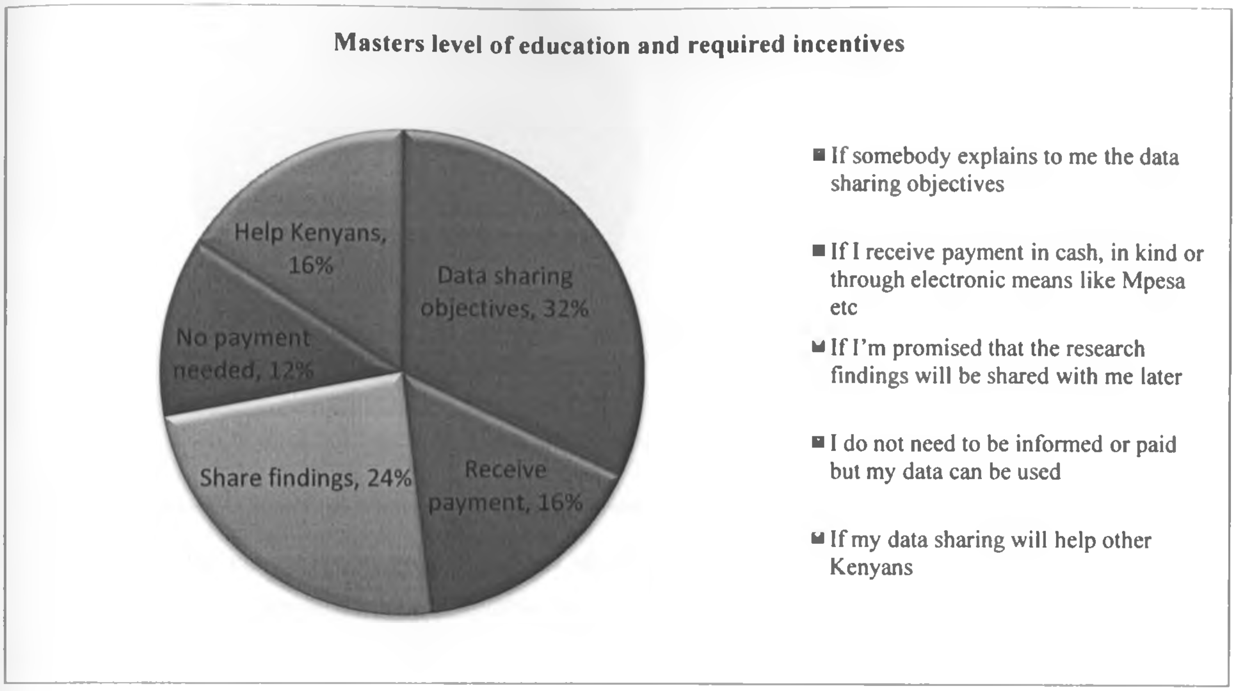


Figure 17: Required incentives at Masters Level of education

The desire to know the data sharing objectives was the major overriding incentive for this group. It was closely followed by the quest to be informed of what became the data sharing exercise; by sharing with them the research findings.

g. Doctorate

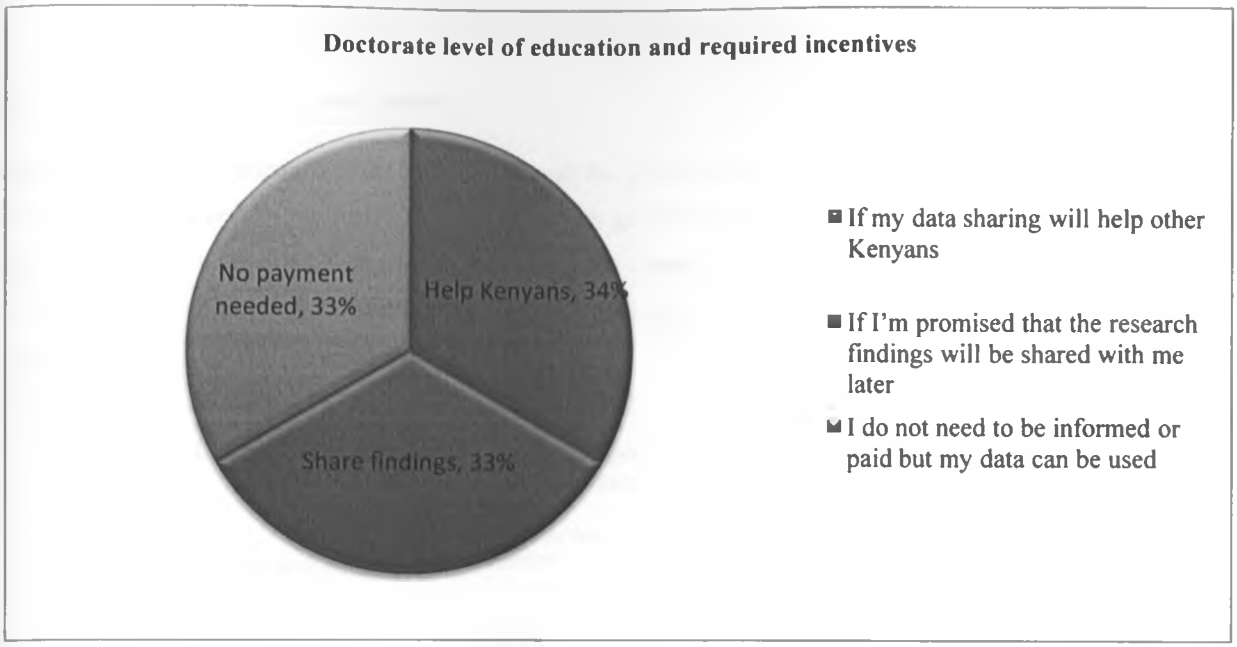


Figure 18: Required incentives at Doctorate level of education

At this level, it was noted that the three incentives identified in the above chart were equally important: none was considered more superior to the other, other than the marginal higher mark awarded to the desire to help other Kenyans (at 34%) compared to the others which were at 33% each.

Summary of the required incentives at the different levels of educational qualifications:

- Incentives required by citizens with no formal schooling or with formal education with no certification were almost similar.
  - i. Half of the sampled population from these two classes indicated a desire to receive some payment to allow access to their personal data.
  - ii. However, it was noted that a higher number of the populace with some education but no certification wanted the research findings to be shared with them later as compared to their counterpart with no formal schooling, as an incentive for data sharing.
- The desire to receive payment to allow personal data access reduces considerably as a citizen acquires higher education qualifications.
  - i. It was noted that among the surveyed citizenry, people with Doctorates did not need any payment at all to allow access to their personal data.
  - ii. This might mean that if the government wants citizens to open up their personal data for anonymous statistical analysis, they might have to institute mechanisms which can help and encourage citizens to further their education.
- Overall, the following were noted to be key incentives across the different educational levels for opening up their personal data for anonymous statistical analysis:
  - i. If citizens can receive some payment;
  - ii. If the data sharing can help in assisting other Kenyans;
  - iii. If the data sharing objectives can clearly be explained to them;

iv. If they are promised that the data sharing research findings will be shared with them later.

In charge of personal data management

Majority of the respondents wanted themselves and the government to manage their personal data (71%). A small number (1% and 3%) wanted a third party and the government respectively to manage their personal data. Also another 1% wanted their data to be managed by a commission to be setup after legislation of the Data Protection and Freedom of Information Bills into law. It should be noted though that a substantial number (24%) wanted to manage the data themselves.

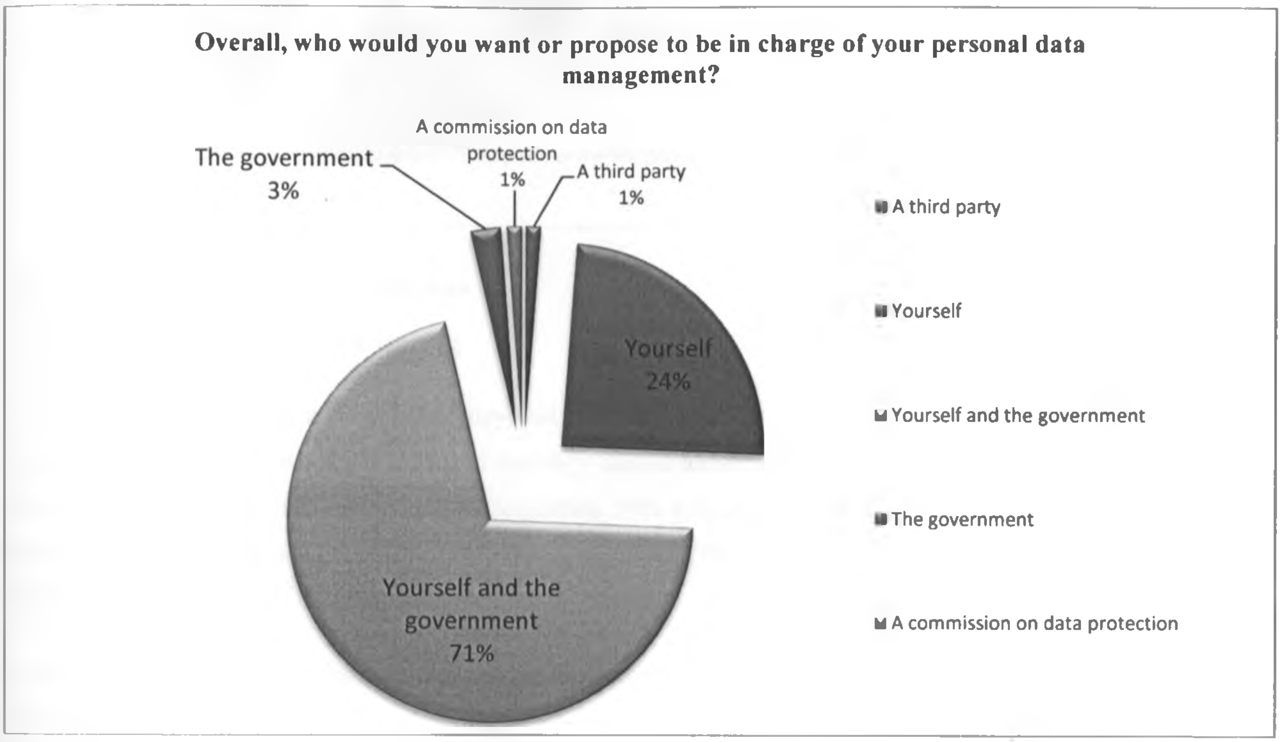


Figure 19: In charge of personal data management

How they would like to access their personal data

38% of the polled citizens indicated that they would like to access their personal data using a mobile device while 41% indicated that they would like to access their data using a computer (can be a desktop PC or a laptop). 19% wanted to be able to access the data either using a mobile device or a computer – they did not prefer one to the other. There is a surge in the need to access the data using a mobile device than the traditional method of accessing government services and systems using a computer. Therefore this means that any application which should be developed to allow management of personal data should be accessible using mobile devices like a mobile phone, an iPad, etc.

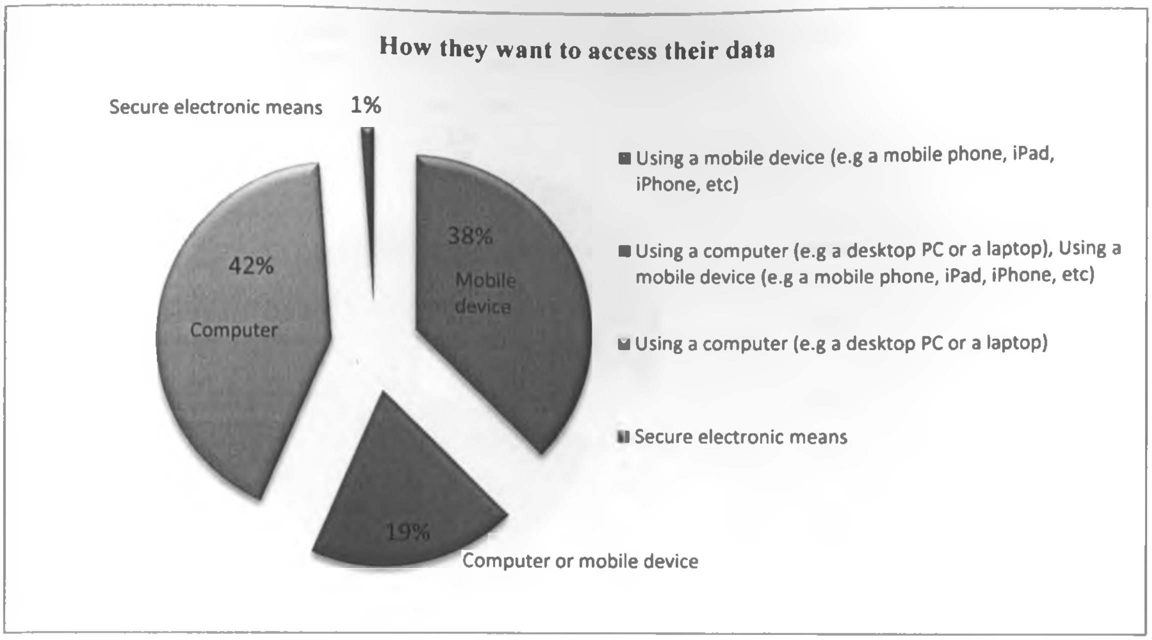


Figure 20: How they want to access their data

Storage of personal data in GoK systems

Most of the polled citizens indicated that they had a problem (78%) with the GoK storing their personal data details forever while the remainder indicated that they had no issues with such an arrangement. For the people who had issues with the GoK storing their personal data, 59% indicated that the government will need to archive the data after either their demise or if not considered necessary. This was closely followed by 23% who wanted the GoK to delete their records completely from their systems.

In fact one of the proposed principles in the Data Protection Bill 2012 outlaws keeping information for longer than necessary.

Infringement of rights by the government for sharing data without consent

Majority of the sampled citizens (71%) indicated that they would consider it an infringement of their rights if their data was shared without their knowledge or consent, to any party. However, there was a small minority who did not care whether their data was shared without their knowledge, consent or permission and hence expected no payment for access.

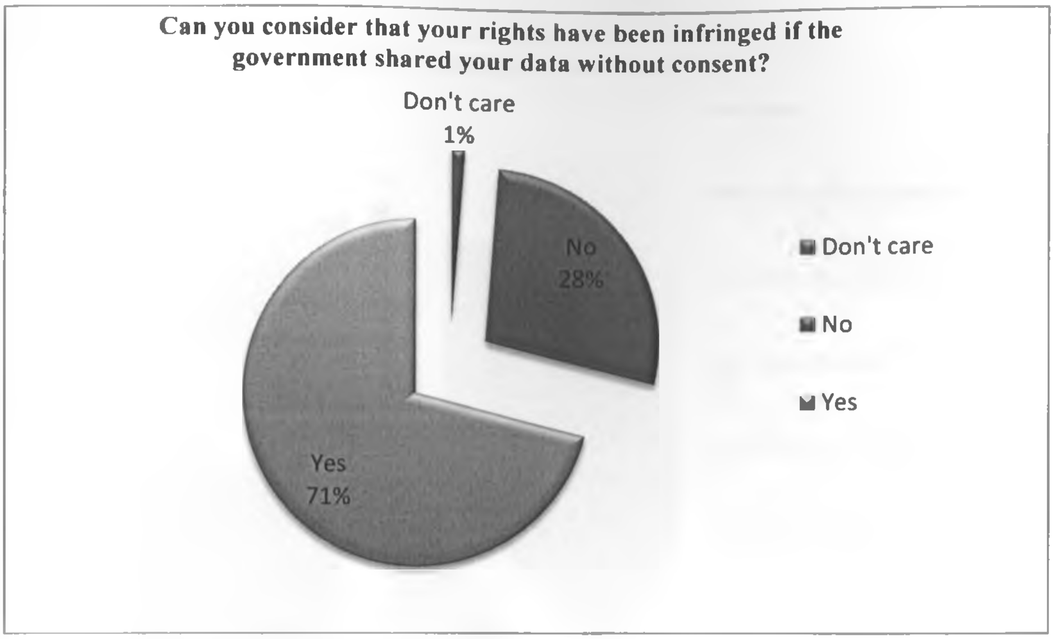


Figure 21: Infringement of rights for unauthorized data sharing

Actions citizens to take for unauthorized personal data sharing

Over a third of the sampled population (39%) indicated that they will sue the government and another 35% indicated that they will petition the government to change its privacy laws to better protect their personal data. 15% will sue the government and at the same time will petition it to change its privacy laws.

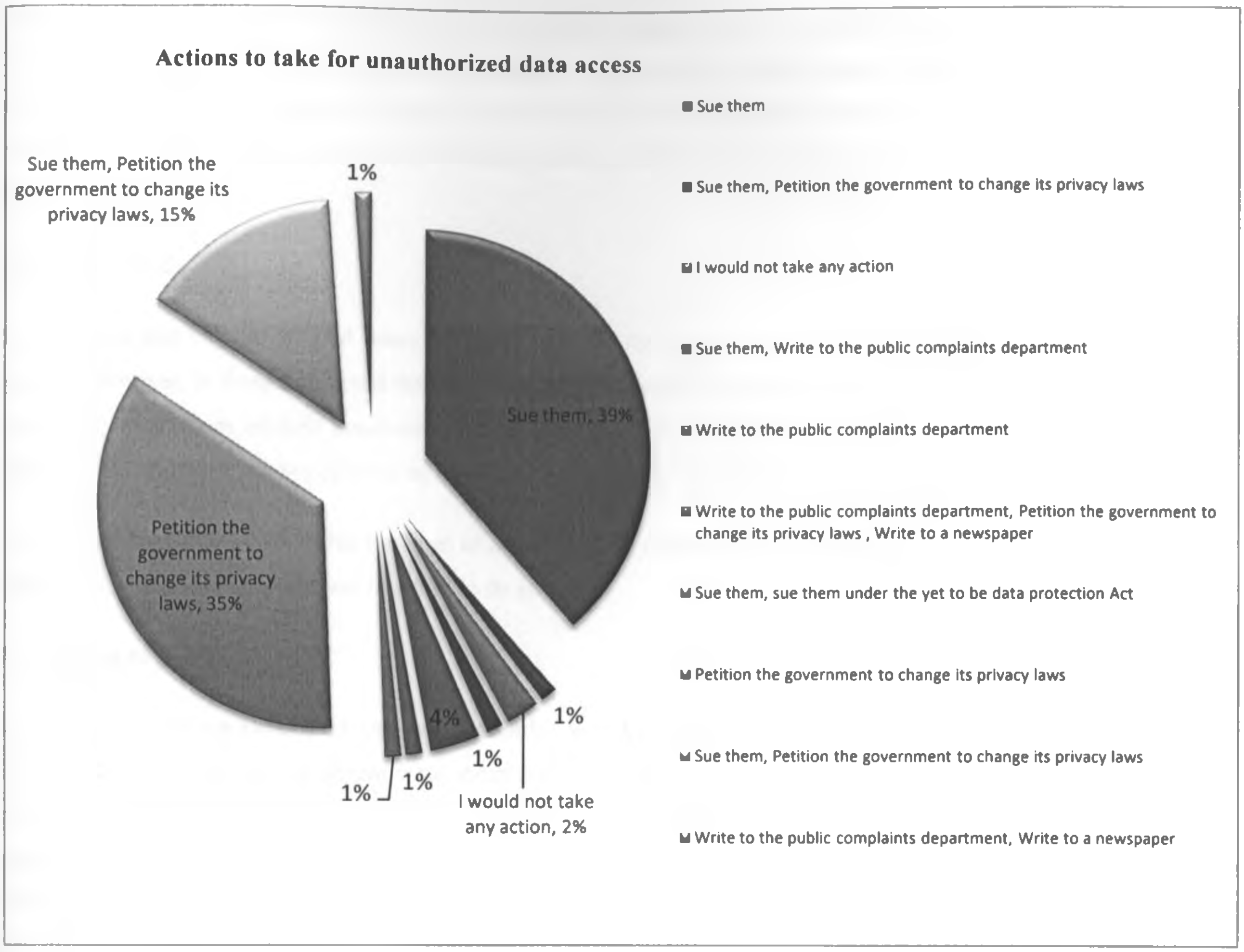


Figure 22: Citizen's actions for any unauthorized data access

This means that citizens perceive that they have a right to their privacy and hence any unauthorized access of their personal data by any entity will not be taken lightly.

Discussions on the research findings

Data sharing

From the foregoing research findings, it is clear that citizens will prefer bespoke select disclosure (Mydex, 2009) data sharing – they would like to know who wants to access what about them – not an automated data sharing mechanism.

Also, from these findings, it is clear that the time when organizations (or government) managed individual information and individuals just accessed it is coming to an end – users want control of their personal data. Also, it is worth to note that a good percentage of the sampled population did indicate that they wanted a partnership where themselves and the government can manage their personal data. This might be the case due to the fact that the government has some data about its citizens which citizens themselves cannot alter arbitrarily – and hence a partnership kind of arrangement looks more feasible and acceptable to them.

User centricity

Mydex report indicated that individuals are effectively disempowered since they have no control of their personal data. It proposed that this need to be changed – users should be at the centre of their personal data. This research has indeed confirmed this to be the case, from a Kenyan perspective. Bodies like the Internet Society (ISOC) have too proposed user centricity (Internet Society, 2008) as being important in empowering individuals when using the internet.

Incentives for data sharing

It was noted that citizens wanted some motivating factors for them to allow access to their data by interested parties. However, it should be noted that the incentives needed by the citizenry were not necessarily monetary rather they were more of their involvement in the data sharing exercise as well as a willingness to share with them any findings that might come out of a research which might involve their personal data.

It was clear from this research that the level of education of a citizen was a key determinant to the willingness to share data as well as the incentives required to do so.

Not willing to share

It was noted that some 38% of the sampled population were not willing to share their anonymized personal data. Most of the reasons advanced therein were about fear of compromised personal security and privacy. Therefore for anonymized personal data sharing to be a success, the government and other stakeholders in the PDE will need to address some of these concerns. It seems like in most instances citizens understood that it was their identifying personal data that will be shared. Even with explanation that it was their anonymized personal data details they were not fully convinced. They advanced arguments such as: access to my bank account, PIN number or ID number whether it will be accompanied by names or not will mean that you can know the citizen if there is enough motivation for doing so. This means that the idea about anonymized personal data sharing will take off once some of these concerns have been adequately handled by the concerned parties.

It is worth noting and encouraging though that the majority supported the idea of anonymized personal data sharing.

System implementation

Requirements analysis

As per the research findings enumerated in the preceding chapter, it is clear that as much as users want to be in charge of their data in partnership with the government, they want to be informed (to give their consent) for any access to their data whether it is anonymized or not.

Also, they wanted a system which can allow access to some of their personal data fields – granting access does not mean complete access to who they are.

From an analysis of personal data management systems which are currently being used (Personal Data Ecosystem Consortium, 2012); it was clear that there is no fully matured system that can meet these requirements. However, out of the systems sampled, the Polis Personal Data Management Framework (P.S, et al., 2012) was meeting most of these requirements albeit with a lot of modifications.

Also, citizens indicated that they wanted to access their data and manage it using mobile devices as well as computers. This means that a system which can meet their expectations when it comes to personal data management, should work in a decentralized setup and accessible over mobile devices as well computers (laptops and desktop computers).

System design

Systems architecture

From the above analysis, it was decided that the Polis Personal Data Management Framework be adopted in developing the system that can meet citizens' requirements and expectations in regard to personal data management.

The resultant system's architecture is as shown below:

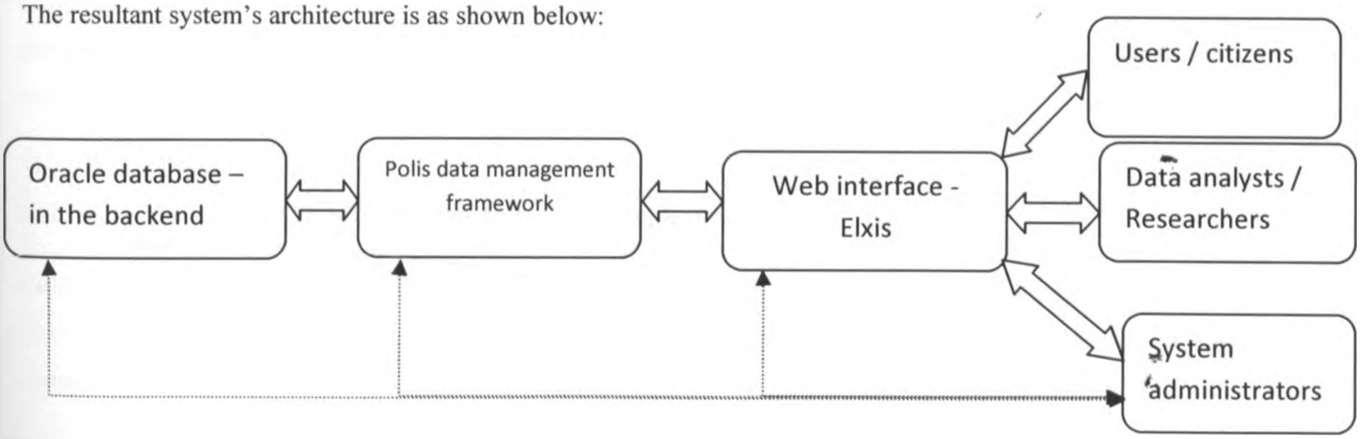

Users / citizens
Oracle database – in the backend
Polis data management framework
Web interface - Elxis
Data analysts / Researchers
System administrators

Figure 23: System architecture

The Polis Personal Data Management architecture is represented below:

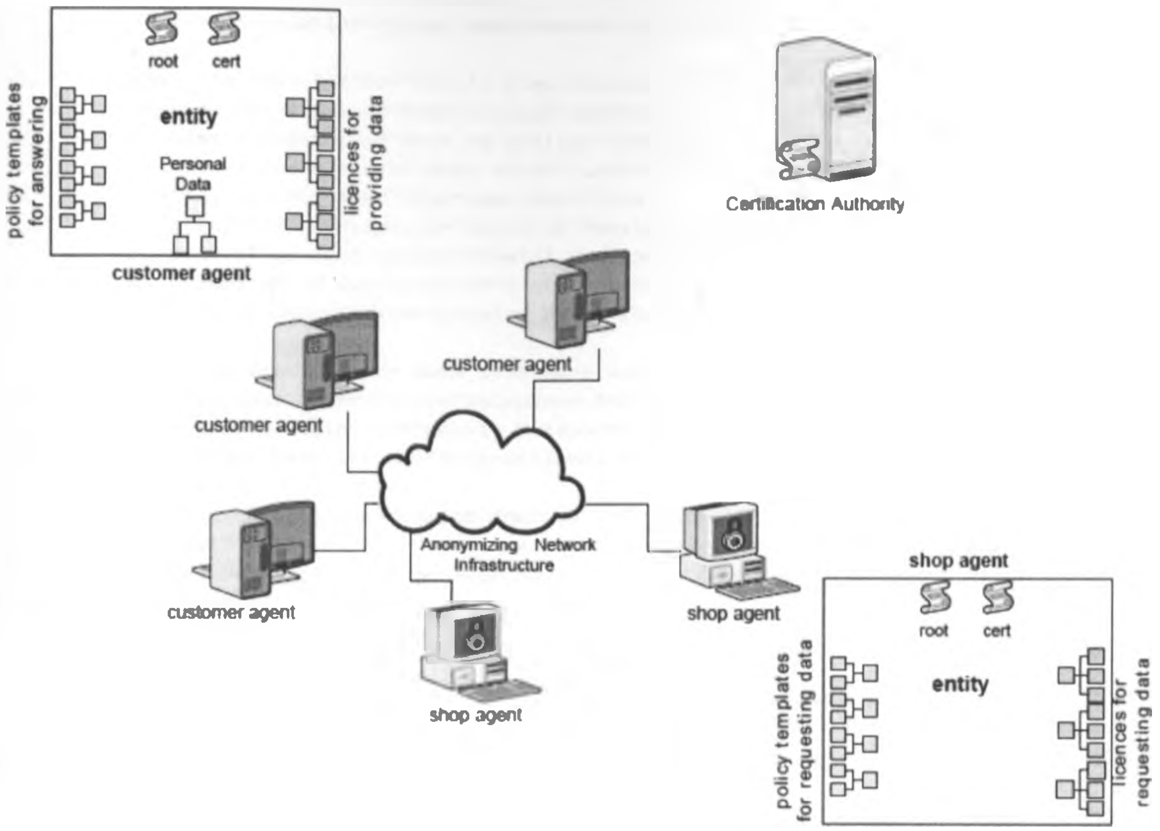

root
cert
policy templates for answering
entity
Personal Data
licences for providing data
customer agent
Certification Authority
customer agent
customer agent
Anonymizing Network Infrastructure
customer agent
shop agent
shop agent
shop agent
root
cert
policy templates for requesting data
entity
licences for requesting data

Figure 24: The polis architecture: source - (Efraimidis, et al., 2009)

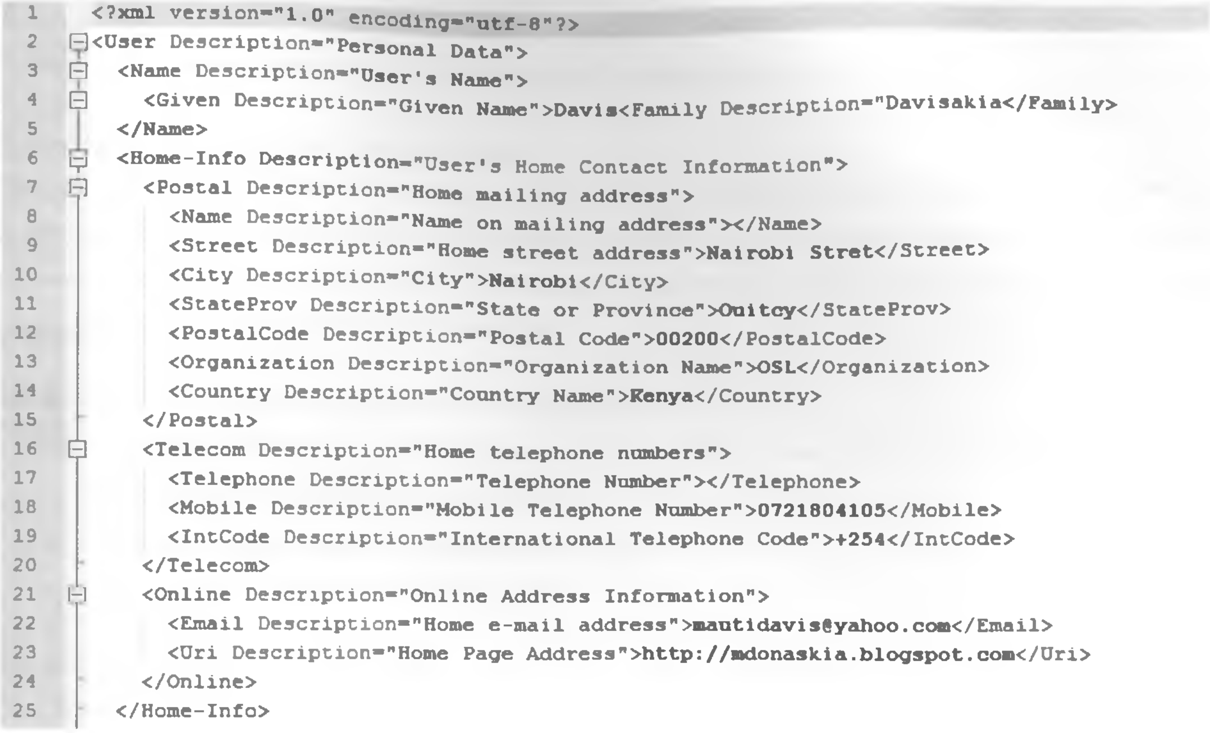
In this architecture, every citizen is represented by a dedicated entity known as a customer agent. This entity is used to instantiate a corresponding polis-agent, which is the main architectural component of Polis. The customer agent is used to manage the personal data of the entity and provide controlled access to it. Potential personal data users use the shop agent to retrieve data from the entity. The shop agent requests for data from a citizen using predefined templates.

There might be a third party acting as certification authority to confirm that indeed the citizen is whom he says he is. However, it should be noted that this role was not implemented for this research.

Citizens' personal data were organized into eight categories as follows: Personal details, Home-info, Social-media, Financial-info, Academic-professional-qualifications, Employment-history, Travel-history and Travel-plans. Each one of them had one or more sub-categories.

Storage of personal data

Each entity stores its personal data in an XML document. An example of an entity personal data stored in an XML file is shown below:

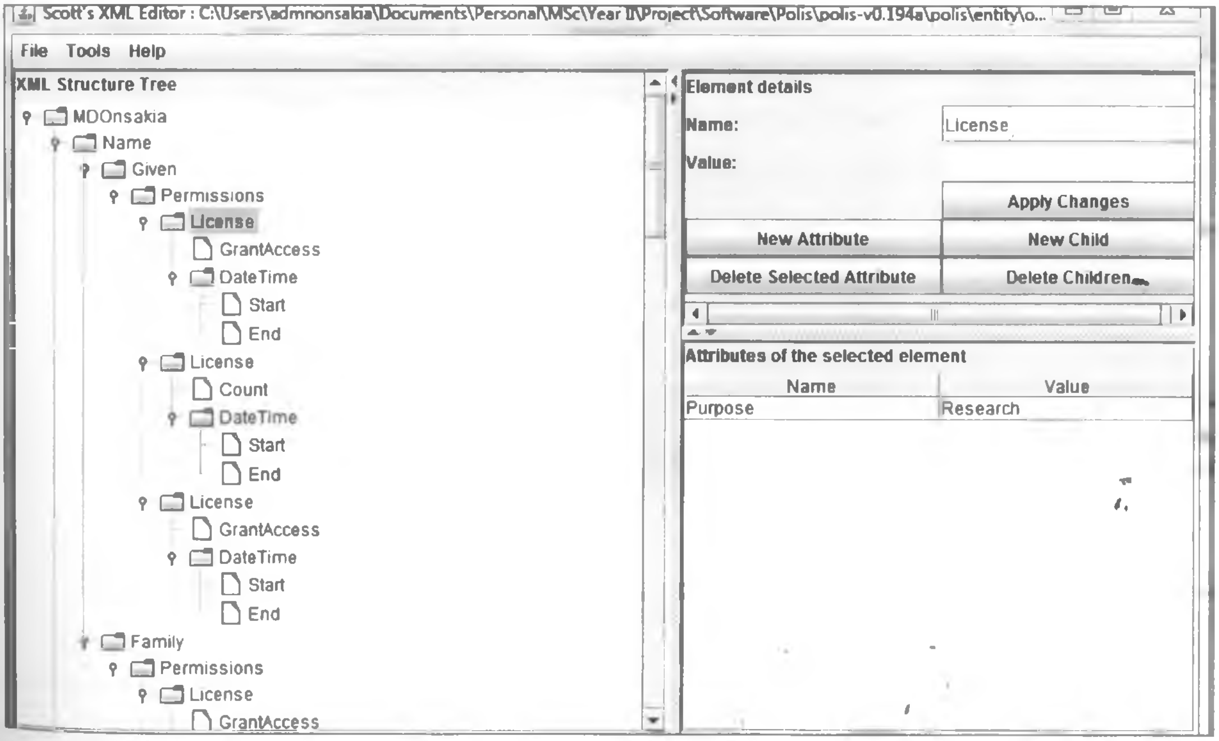


```xml
<?xml version="1.0" encoding="utf-8"?>
<User Description="Personal Data">
  <Name Description="User's Name">
    <Given Description="Given Name">Davis<Family Description="Davisakia</Family>
  </Name>
  <Home-Info Description="User's Home Contact Information">
    <Postal Description="Home mailing address">
      <Name Description="Name on mailing address"></Name>
      <Street Description="Home street address">Nairobi Stret</Street>
      <City Description="City">Nairobi</City>
      <StateProv Description="State or Province">Ouitcy</StateProv>
      <PostalCode Description="Postal Code">00200</PostalCode>
      <Organization Description="Organization Name">OSL</Organization>
      <Country Description="Country Name">Kenya</Country>
    </Postal>
    <Telecom Description="Home telephone numbers">
      <Telephone Description="Telephone Number"></Telephone>
      <Mobile Description="Mobile Telephone Number">0721804105</Mobile>
      <IntCode Description="International Telephone Code">+254</IntCode>
    </Telecom>
    <Online Description="Online Address Information">
      <Email Description="Home e-mail address">mautidavis@yahoo.com</Email>
      <Uri Description="Home Page Address">http://mdonaskia.blogspot.com</Uri>
    </Online>
  </Home-Info>
```

Figure 25: Personal data XML file

Request of data from citizens

This is done using predefined templates. An example of a policy template for requesting data from the individual customer agents is as shown below:



Scott's XML Editor : C:\Users\admnonsakia\Documents\Personal\MSc\Year II\Project\Software\Polis\polis-v0.194a\polis\entity\o...

XML Structure Tree

```
MDOnsakia
  Name
    Given
      Permissions
        License
          GrantAccess
          DateTime
            Start
            End
        License
          Count
          DateTime
            Start
            End
        License
          GrantAccess
          DateTime
            Start
            End
    Family
      Permissions
        License
          GrantAccess
```

Element details

Name: License

Value:

Attributes of the selected element

Name	Value
Purpose	Research

Figure 26: Policy template for requesting data

It should be noted that request for data is at field level rather than citizen level: for each field, the citizen would have indicated whether he wants his data to be accessed, for what purpose and for how long.

Software development methodology

The prototype was developed by following the agile software development methodology (Beck, et al., 2001) . Agile methodology was chosen since it is based on iterative and incremental development which is very necessary in a case of such a project whose requirements might change midway. It is also flexible to change and hence very responsive. The methodology values are:

- **Individuals and interactions** over processes and tools;
- **Working software** over comprehensive documentation – this shortcoming was overcome by documenting every step that was being taken;
- **Customer collaboration** over contract negotiation;
- **Responding to change** over following a plan.

The methodology proponents indicate that *'..while there is value in the items on the right, we value the items on the left more.'* (Beck, et al., 2001)

The methodology is graphically shown below:

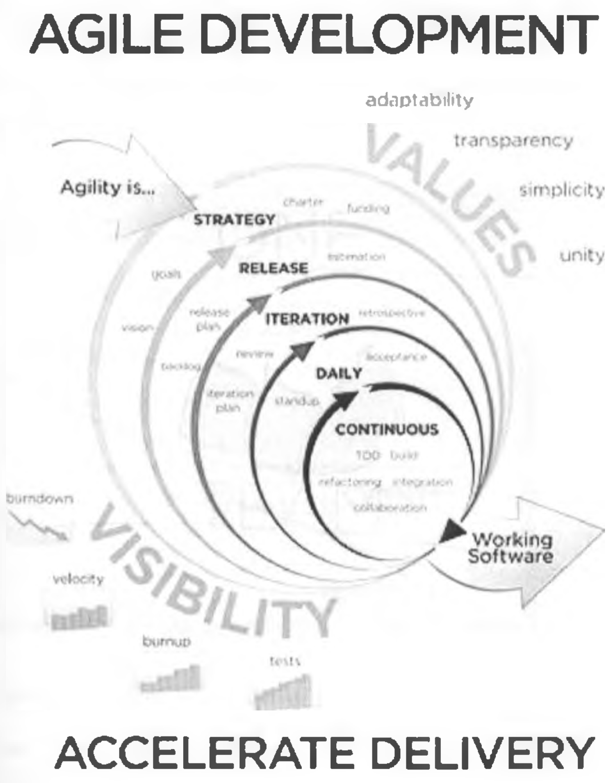


Figure 27: The Agile software methodology

Tools used in system implementation:

- The back-end database was Oracle;

  - Oracle database was chosen over other relational database management systems since it supports Oracle Stored Procedures (OSPs) which the agents use to communicate to each other.
- The front-end was a web interface running on Elxis Content Management System (CMS). Elxis is an open source CMS – Elxis was chosen since it supports Oracle database unlike most open source CMSs;
- The agents communicate using OSPs;
- The system is based on the Polis Personal Data Management Framework.

Prototype implementation

The prototype was majorly been implemented using open source software and technologies save for the Oracle Database.

The front-end is an interface that allows querying of data from the system by prospective data analysts. The prospective data users will only be able to access what citizens have set as visible to other parties.

The data from the system is extracted in a simple text file which can be used for analysis later in any data processing system.

Application architecture

The authentication in the system has been implemented using either a citizen's National Identification Document (ID) number, a Personal Identification Number (PIN) or a passport.

There will be an agent running which is accessible by logging to a screen like the one shown below (by running the command: >java -jar polis.jar:

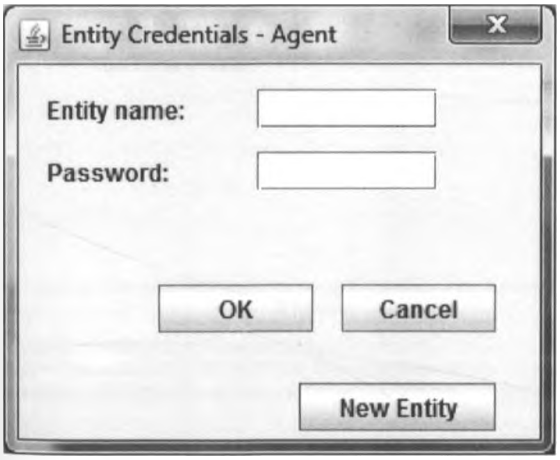
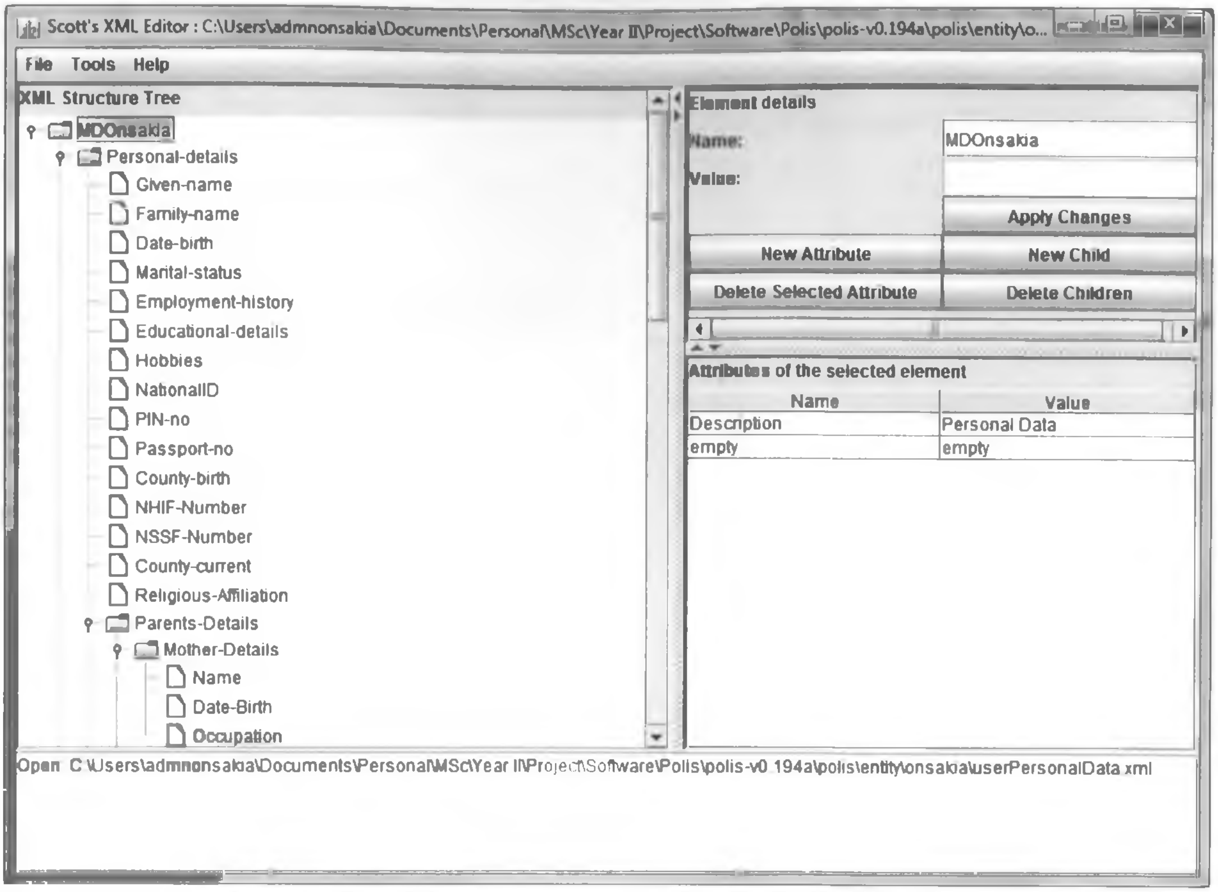


Figure 28: System login window

It should be noted that it is possible to create a new entity for managing data in the system by clicking on 'New Entity' on the above window.

Once you login you will get options like:

- You can edit your personal data – this is the window for a citizen to enter their personal data:



Scott's XML Editor : C:\Users\admnonsakia\Documents\Personal\MSc\Year II\Project\Software\Polis\polis-v0.194a\polis\entity\o...

```
MDOnsakia
    Personal-details
        Given-name
        Family-name
        Date-birth
        Marital-status
        Employment-history
        Educational-details
        Hobbies
        NationalID
        PIN-no
        Passport-no
        County-birth
        NHIF-Number
        NSSF-Number
        County-current
        Religious-Affiliation
        Parents-Details
            Mother-Details
                Name
                Date-Birth
                Occupation
```

Name: MDOnsakia

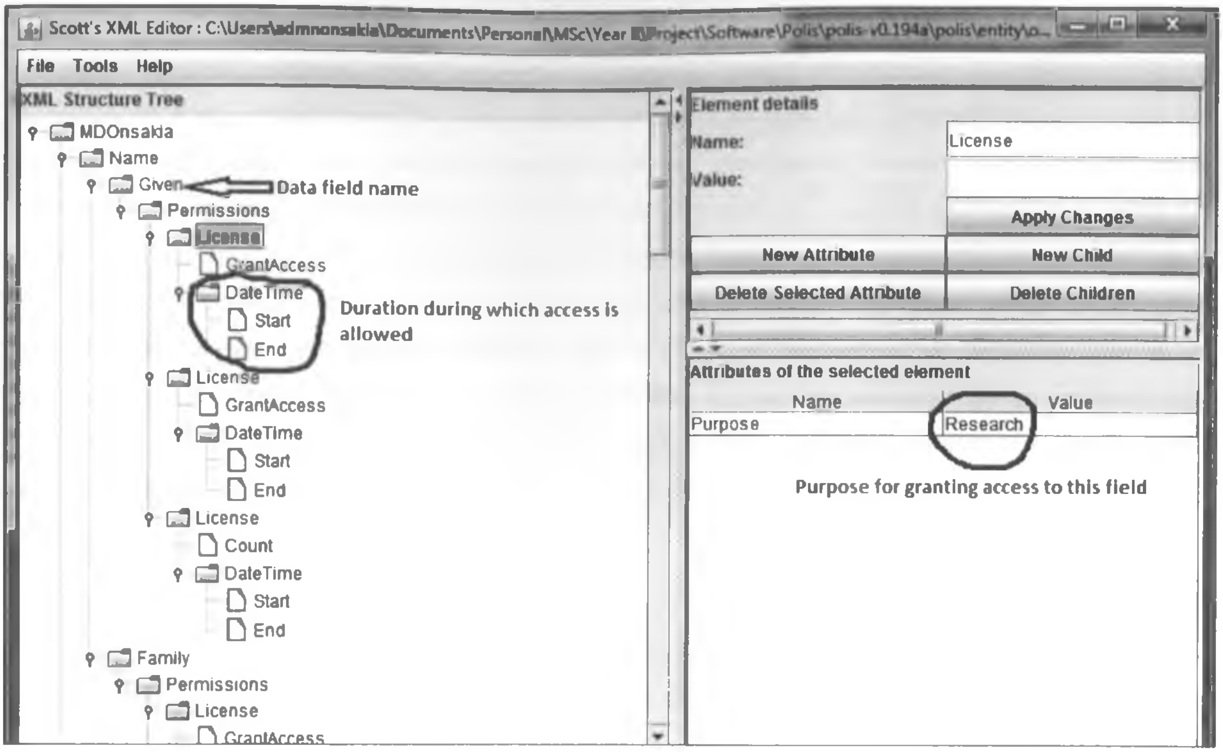
Value:

Name	Value
Description	Personal Data
empty	empty

Open C:\Users\admnonsakia\Documents\Personal\MSc\Year II\Project\Software\Polis\polis-v0 194a\polis\entity\onsakia\userPersonalData.xml

Figure 29: Citizen's personal data management window

- Access control – it is possible for a citizen to specify which field he can allow to be accessed for data analysis, for what purpose and for which duration, as shown in the snapshot below – access control for his First Name:


Scott's XML Editor : C:\Users\admnonsakia\Documents\Personal\MSc\Year I\Project\Software\Polis\polis-v0.194a\polis\entity\o...
File
Tools
Help
XML Structure Tree
Element details
Apply Changes
New Attribute
New Child
Delete Selected Attribute
Delete Children
Attributes of the selected element

```
MDOnsakia
  Name
    Given
      Permissions
        License
          GrantAccess
          DateTime
            Start
            End
        License
          GrantAccess
          DateTime
            Start
            End
        License
          Count
          DateTime
            Start
            End
    Family
      Permissions
        License
          GrantAccess
```

Name: License

Value:

Name	Value
Purpose	Research

Data field name

Duration during which access is allowed

Purpose for granting access to this field

Figure 30: Granting permissions window

It should be noted that it is possible to grant different access durations to the same data field for different purposes or reasons.

This is possible by editing your data access policy. This policy is unique to each entity (citizen).

Discussions

Data cannot be useful unless it is shared among the relevant stakeholders in the PDE, as was noted in the WEF Report (2011). To be able to share data, there needs to be interfaces (between and among systems) which are interoperable. According to the US National Strategy for Trusted Identities in Cyberspace (NSTIC) (2011, pp.8-9), there are three types of interoperability for identity solutions: technical, semantic and legal interoperability. All these types of interoperability were examined in the Kenyan context and it is proposed that sharing of data be accomplished through a simple ASCII text file (Blue Button Project way). Also, people wanted legal mechanisms put in place to enable acceptable data sharing; otherwise they were ready to sue and petition the GoK to change its privacy laws for any unauthorized access of their data. In this regard, it is noted with appreciation that the government is working on the Data Protection Bill, 2012 and Freedom of Information Bill, 2012. These bills once enacted into law will act as legal mechanisms for data protection as well as sharing of information.

The WEF Report (2011, p.15) identified five points of tension in the PDE namely: privacy, global governance, personal data ownership, transparency and value distribution. This project considered the above-mentioned five areas. However, it should be noted that global governance was above the purview of this research since we neither have the mandate nor the resources to influence issues globally. Hence this research was restricted to the Kenyan scenario although the global picture was maintained throughout. On the value distribution point, some citizens wanted to be paid (in cash or using electronic means) as an incentive to allow access to their data. Also, another majority wanted to be informed of the objectives of data sharing before they can allow access to their data. This means that this group expects to be informed about what is happening concerning their data without necessarily being paid for any access. This is a group which researchers can target in their research activities if they need access to their personal data.

The WEF Report (2011, p.19) also indicated that the solution to a balanced PDE lies in 'developing policies, incentives and rewards that motivates all stakeholders – private firms, policy makers, end users – to participate in the creation, protection, sharing and value generation from personal data'. This issue has been handled in this research and it has come out quite clearly that citizens need policies to protect their personal data and by extension their privacy and security. This was the reason why some citizens indicated their desire for protection of their data to be established under the current debated bills of Data Protection and Freedom of Information.

The WEF Report (2011, p. 10) claimed that a person's data is equivalent to their money; this has partially been proven to be true from this research since some users have indicated that they need to be paid (in cash or electronically) to allow access to their data. It should be noted that even from the WEF model (fig. 2), some of the incentives expected by individual users are better prices, offers as well as improved services.

One of the questions that the research also was attempting to answer was: how will the end-user, the individual citizen have control over his data – which they are willing to submit and some of which the government currently has? This question was addressed across the four principles identified in WEF Report (2011, p.10) namely: transparency, trust, control and value.

The issue of how users wanted to update their personal data also was explored in this age of anytime anywhere connectivity – since this is one of the key tenets of the user being charge of his data. Users clearly indicated that they wanted to access the data (to update or just check it) using either a mobile device or a computer.

When dealing with personal data, there will be issues about identity assurance: how to be sure that we will be dealing with who the citizen says he is – when accessing data from a government system bearing in mind that we will be dealing with a virtual being. Approaches like OpenID[4], Information cards[5] (I-cards) are being promoted so that a specific individual can be uniquely identified in the global internet space. This is very important since it will lend credence to the integrity of the data that is being stored of an individual citizen. In addition, this will also improve the level of reliability that will be placed by the interested stakeholders on the stored citizens' data. We explored how this will be handled in a Kenyan setting. It should be noted though that some services like Gmail[6] are already OpenID compliant and hence this might not be an entirely new concept to some citizens.

On this issue, we proposed that it might be prudent to use ID card numbers as unique identifiers of citizens since it is not possible to have duplicated ID numbers. Also, ID numbers can be used in conjunction with PIN (Personal Identification Numbers) to uniquely identify each citizen in the system. It should also be possible to have two levels of authentications: the first level is the ID number and the second one is the PIN. This way, it will be possible to be sure that the citizen is who he says he is. Although we propose usage of ID and PIN numbers for authentication, this information will not be shared with another person or entity requesting access to the citizen's data and hence the identity of the citizen will not be disclosed.

From the research findings, for the citizenry to be convinced to provide their personal details, the government should be able to prove to them that the to-be provided personal data will be securely stored – that it has put in place mechanisms that will ensure that their data will not be abused in any way. This guarantee is important since without it confidence can neither be built for the populace to volunteer their personal information nor can the information gleaned from it be considered authentic. This objective can be boosted if the government enacts the Data Protection and Freedom of Information bills into law.

This research therefore, hopes to form the basis on which the government can have an idea of the perception of its citizens in regard to their personal data. It is our hope that its findings will be incorporated into the yet-to-be gazetted Data Protection and Freedom of Information bills (CIC, 2012).

Mydex listed some of the benefits of citizens opening up their personal data for anonymous statistical use as:

i.) They can gain from it – they can indicate that for any individual or organization to be allowed to access their specific selected personal data, they get paid (in cash or in kind) for that;

[4] OpenID is an open standard that describes how users can be authenticated in a decentralized manner, eliminating the need for service providers to provide their own ad hoc systems and allowing users to consolidate their digital identities.

[5] Information cards are the digital version of cards you carry in your purse or wallet today for identification purposes over the web.

[6] Gmail is a free Google mail service.

ii.) Reinventing marketing – by sellers having access to personal data of citizens, they can only send to them what is relevant and hence not spamming which is not only intrusive but very annoying;

iii.) Their data can be used to make future decisions and hence this helps in the country's social and economic development;

iv.) Allowing access to their personal data can spark innovation and economic growth via the development of new value-adding services.

These are some of the benefits that the research has confirmed that citizens are looking in the personal data sharing initiative. Save for benefit (ii) this research indeed confirms this to be the case. On the benefit of targeted advertising, it might not be a feasible incentive in this research since access will be to anonymized personal data – not identifying in any manner.

Limitations of the study

As was indicated in the research design, an ideal research method for such a study would have been a census. However, due to lack of enough resources, sampling was done. Despite this, it is expected that the research findings mirrors the perception on the ground regarding personal data management.

Recommendations for future work

It might be necessary that other than just allowing access to their personal data, to consider mechanisms which citizens will be rewarded by updating information about themselves, as was proposed by the WEF Report (WEF, 2011). This might help in having the most update data about citizens and hence any interested party willing to reward citizens who allow access to their data to retrieve the most current data.

The issue of how the data which has been granted to an entity will be dealt with later is not very straight forward to handle. Since it is very difficult to know or to control what one does with the data he has been granted access to, it is recommended that this issue be dealt with in a legal way: incorporate the requirement that an entity which has been granted access should destroy data once the function for which it was granted for has been met. However, this might not be easy to determine, enforce or ascertain. It is important to note that this requirement has been incorporated in the proposed Data Protection Act, 2012 although its implementation might not be easy.

It will be important for the government to consolidate all the data about its citizens to a central place, before it can let the citizens start accessing it and updating the same. This can be a very important starting point since it will give them a glimpse of what the government has about them and hence making the exercise of updating their details easier. The consolidation of these citizen's data can be done using data warehousing technologies or real-time APIs or mashups.

Conclusion

From the research findings, it is clear that most citizens are willing to share their anonymized personal data details as long as their consent is sought before any access.

Proposed business model

As per the research findings and the reviewed literature, the following is being proposed as a sustainable business model for managing and linking the different stakeholders in the PDE:

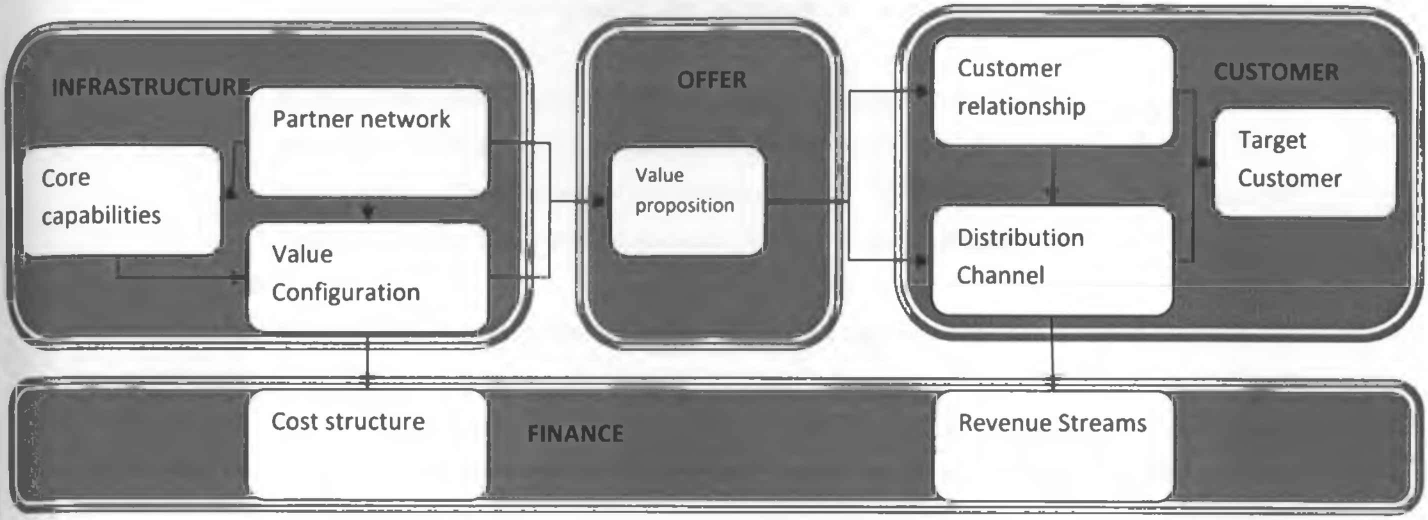


Figure 31: Proposed business model

Explanation of the different components of the business model:

- **Infrastructure**:
  - ***Core Capabilities*** : the capabilities required to run the proposed system which includes, among others:
    - Operating system required – proposed operating system to be any Linux flavour;
    - Hosting server technical specifications;
    - Backend database required;
    - System accessibility requirements etc.
  - ***Partner network***: the network of cooperation between the stakeholders in the PDE but majorly to be led by the GoK. This will deal with how the different stakeholders in the PDE will interact with the system. The stakeholders targeted include citizens, the GoK (and public sector entities) as well as the private sector actors – this includes potential data users like research institutions, universities etc.
  - ***Value configuration***: describes the arrangement of activities and resources to get value out of the model. This entails activities like how the different stakeholders will benefit from the anonymized personal data stored in the system.
- **Offe*r:***
  - ***Value Proposition***: gives an overall view of the project's benefits to its customers – citizens and potential anonymized personal data users. This will deal with the benefits of anonymized

personal data to the different stakeholders in the PDE. This includes an example of payment for accessing citizens' data as per their preference.

- **Customer**:
  - ***Customer Relationship***: explains the relationships among the different stakeholders in the PDE and their respective roles and responsibilities: Citizens supply their dynamic data while the government provides the static data for each citizen and the private sector players (which include potential data users) access data from this pool. The access is as per what citizens have defined as accessible to the public with the respective motivational factors which must be met before access is granted.
  - ***Distribution channel***: describes the channels to communicate and get in touch with the citizens and other interested parties (and distribute revenue – if any);
  - ***Customer***: describes the parties involved in the PDE – this includes citizens, government and public and private sector entities who are interested in personal data;
- **Finance**:
  - ***Revenue Streams***: describes the revenue streams through which money is earned like payment for data access from the system. Any revenue earned from the system is to be distributed among the actors in the PDE through *Distribution Channels* identified above;
  - ***Cost Structure***: sums up the monetary consequences to run this system. The initial cost and advocacy for the project is expected to be borne by the government and hence the government is expected to play a pivotal role for the success of this system.

This business model describes how the different stakeholders in the PDE will benefit from personal data consolidation and ultimately anonymous statistical analysis of the same. It also shows the capabilities and required partnership in the PDE to derive value from personal data and also as well includes expected revenue streams.

It should be noted that the role of the government in the above model might be carried by the proposed Freedom of Information and Data Protection Commission of Kenya, as per the Data Protection Bill, 2012 and Freedom of Information Bill, 2012.

Were research objectives met?

The research objectives for this study were met as demonstrated below:

1) Find out if citizens are willing to supply their personal details to the government or to any other interested party (for free or through some incentives);

   Findings: Citizens are willing (62%) to share their personal data details to the government or an interested party. However, the majority indicated a desire to be informed or their consent sought to allow access to their personal data as well as be rewarded monetarily.

2) Propose ways in which the citizenry can be encouraged to open up access to their personal data for some anonymous statistical analysis (by giving some value propositions);

Findings: Citizens indicated that they need incentives to share their personal data. However, it should be noted that the greatest incentive was not monetary but rather a desire to be informed of the data sharing objectives

3) Propose methodologies or mechanisms which will facilitate easier access of personal data by individual citizens and sharing of the same to interested stakeholders;

   Findings: The Polis Personal Data Management Framework has been proposed as the ideal platform upon which data sharing can be done. Sharing of data to other interested stakeholders in the PDE to be done through the system or using a simple ASCII text file.

4) Propose changes, if feasible, in which the current legal and/or regulatory framework might need to be amended to allow for anonymous access to individual citizens' data by authorized entities for statistical analysis.

   Findings: There is currently no legal mechanism for data sharing or protection. However, there are two bills which will affect data protection and sharing in Kenya once passed into law. The two bills are: the Data Protection bill, 2012 and Freedom of Information bill, 2012 which are currently being debated; an exercise being co-ordinated by the CIC.

Therefore it is clear that this research met the objectives of the study.

References and bibliography

Algorithms and Privacy Research Unit, Democritus University of Thrace, 2012. *Polis Project.* [Online]
Available at: https://euclid.ee.duth.gr
[Accessed 25 January 2012].

Bain & Company, 2010. *PDE Model.* [Art] (Bain & Company).

Beck, K. et al., 2001. *Manifesto for Agile Software Development.* [Online]
Available at: http://agilemanifesto.org/
[Accessed 26 May 2012].

British Columbia Public Service, n.d. *Citizens @ The Centre: B.C Government 2.0,* s.l.: British Columbia Public Service.

Cavoukian, A., 2005. *Creation of a Global Privacy Standard.* Montreux, Information and Privacy Commissioner, Ontario, Canada.

Cavoukian, A., 2011. *The 7 Foundational Principles.* Ontario: Privacy by Design.

Commission for the Implementation of the Constitution , 2012. *The Freedom Of Information Bill.* [Online]
Available at:
http://cickenya.org/sites/default/files/bills/Freedom%20of%20Information%20Bill%20Revised%2010%20th%20Jan%2C2012%20%281%29.pdf
[Accessed 20 May 2012].

Commission for the Implementation of the Constitution, 2012. *Data Protection Bill.* [Online]
Available at:
http://cickenya.org/sites/default/files/bills/Data%20Protection%20Bill%20Revised%2010th%20Jan%2C2012_0.pdf
[Accessed 20 May 2012].

Creative Research Systems, 2012. *Sample Size Formulas for our Sample Size Calculator.* [Online]
Available at: http://www.surveysystem.com/sample-size-formula.htm
[Accessed 5 June 2012].

Creative Research Systems, 2012. *The Survey Systems website.* [Online]
Available at: http://surveysystem.com/sscalc.htm
[Accessed 20 April 2012].

Davis, M., Martinez, R. & Kalaboukis, C., 2010. *Rethinking Personal Information - Workshop Pre-read,* Geneva: The World Economic Forum.

Eclipse, 2011. *The Higgins Open Source Framework,* s.l.: Eclipse.

Efraimidis, P. S., Drosatos, G., Nalbadis, F. & Tasidou, A., 2009. Towards Privacy in Personal Data Management. *Information Management & Computer Security (IMCS),* 17(4), pp. 311-329.

eGovernment Directorate, 2012. *E-government Implementation.* [Online]
Available at: http://www.e-government.go.ke/index.php?option=com_content&view=article&id=89&Itemid=93
[Accessed 15 February 2012].

Electronic Frontier Foundation, 2012. *International Privacy Standards.* [Online]
Available at: https://www.eff.org/issues/international-privacy-standards
[Accessed 21 January 2012].

Facebook Limited, 2011. *Announcing Facebook Connect - Facebook developers.* [Online]
Available at: https://developers.facebook.com/blog/post/108/
[Accessed 10 December 2011].

Government of Kenya, 2010. *The Constitution of Kenya.* Nairobi: National Council for Law Reporting.

Government of Kenya, 2012. *Open Data Initiative.* [Online]
Available at: www.opendata.go.ke
[Accessed 2 February 2012].

IBM, 2012. *Data Growing Rapdily.* [Online]
Available at: www-01.ibm.com/software/data/bigdata/
[Accessed 15 February 2012].

Information & Privacy Commissioner, Ontario, Canada, 2011. *Privacy by Design.* [Online]
Available at: www.privacybydesign.ca
[Accessed 21 December 2011].

International Conference of Data Protection and Privacy Commissioners, 2011. *International Conference of Data Protection and Privacy Commissioners - 2011.* [Online]
Available at: http://www.privacyconference2011.org/
[Accessed 21 January 2012].

Internet Society, 2008. *Preserving the User Centric Internet,* Geneva: Internet Society.

Internet Society, n.d. *The Internet Society.* [Online]
Available at: www.internetsociety.org
[Accessed 2 October 2011].

Kantara Initiative, 2010. *Kantara Initiative Identity Assurance Framework,* s.l.: Kantara Initiative.

Keele, B. J., 2009. Privacy by Deletion: The need for Global Data Deletion Principle. *Indiana Journal of Global Legal Studies,* 16(1), pp. 363-384.

Kenya Government, 2009. *Census 2009 Data,* Nairobi: Kenya Government.

Kenya ICT Board, 2011. *OpenData.go.ke Initiative Launched by H.E. the President Mwai Kibaki.* [Online]
Available at: http://www.ict.go.ke/index.php?option=com_content&view=article&id=355:open-data&catid=74:books&Itemid=414
[Accessed 5 December 201].

Kenya ICT Board, 2012. *Kenya ICT Board.* [Online]
Available at: www.ict.go.ke
[Accessed 1 February 2012].

Kothari, C. R., 2004. Research Process. In: *Research Methodology, Methods and Techniques.* 2nd Revised Edition ed. New Delhi: New Age International Publishers, pp. 14,176.

Mary, R. et al., 2010. *The Open Identity Trust Framework,* s.l.: Open Identity.

Mydex , n.d. *The Case for Personal Information Empowerment: The rise of the personal data store,* s.l.: Mydex CIC.

National Information Infrastructure, 1995. *Privacy and the National Information Infrastructure: Principles for Providing and Using Personal Information.* [Online]

Available at: http://aspe.hhs.gov/datacncl/niiprivp.htm
[Accessed 5 January 2012].

Nissenbaum, H., 1998. Protecting Privacy in an Information Age: The Problem of Privacy in Public. *Law and Philisophy,* Issue 17, pp. 559-596.

Open Identity Exchange, 2010. *The Open Identity Trust Framework 2010.* [Online]
Available at: http://openidentityexchange.org/sites/default/files/the-open-identity-trust-framework-model-2010-03.pdf
[Accessed 12 January 2012].

Organization for Economic Cooperation and Development, n.d. *OECD Privacy Guildelines.* [Online]
Available at: www.oedprivacy.org
[Accessed 28 December 2011].

Osterwalder, A., 2006. *Business Model Design.* [Online]
Available at: http://business-model-design.blogspot.com
[Accessed 20 December 2011].

P.S, E., Drosatos, G., Nalbadis, F. & Tasidou, A., 2012. *Polis Project.* [Online]
Available at: https://euclid.ee.duth.gr/wordpress/?page_id=250
[Accessed 20 February 2012].

Personal Data Ecosystem Consortium, 2012. *Personal Data Ecosystem Consortium.* [Online]
Available at: www.http://personaldataecosystem.org/
[Accessed 13 January 2012].

Privacy By Design, 2011. *Privacy by Design.* [Online]
Available at: www.privacybydesign.ca
[Accessed 10 December 2011].

Regan, P., 1995. *Legislating Privacy: Technology, social values and public policy.* 1st ed. North Carolina: University of North Carolina Press.

Republic of Moldova, 2011. *Open Government Data.* [Online]
Available at: http://data.gov.md/en
[Accessed 13 December 2011].

The Economist, 2012. *India's Unique Identity Project.* [Online]
Available at: http://www.economist.com/node/21542763
[Accessed 13 January 2012].

The Kenya ICT Board, 2012. *The Kenya ICT Board.* [Online]
Available at: www.ict.go.ke
[Accessed 1 January 2012].

The Standard Group, 2012. *LION 2.* [Online]
Available at: http://www.standardmedia.co.ke/politics/InsidePage.php?id=2000052714&cid=14
[Accessed 21 February 2012].

The White House, 2011. *'Blue Button' provides Access to Downloadable Personal Health Data.* [Online]
Available at: http://www.whitehouse.gov/blog/2010/10/07/blue-button-provides-access-downloadable-personal-health-data
[Accessed 21 November 2011].

The WhiteHouse, 2011. *National Strategy For Trusted Identities in Cyberspace,* Washington: The WhiteHouse.

The World Economic Forum, 2010. *Rethinking Personal Data,* Geneva: The World Economic Forum.

The World Economic Forum, 2011. *Personal Data: The Emergence of a New Asset Class,* Geneva: The World Economic Forum.

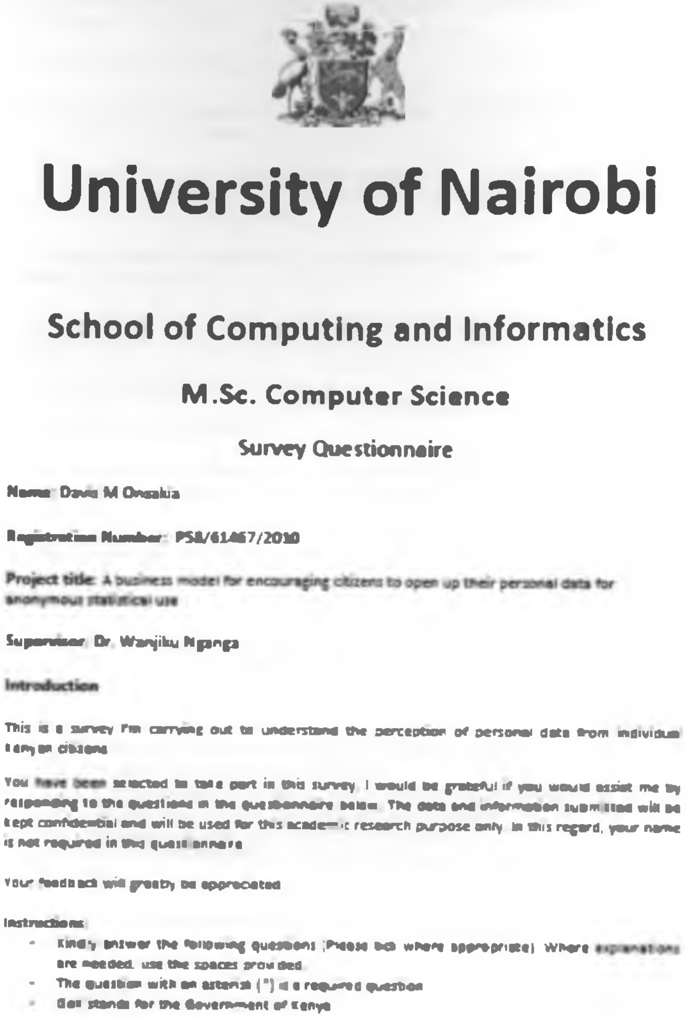
The World Economic Forum, 2012. *Big Data, Big Impact:New Possibilities for International Development,* Geneva: The World Economic Forum.

Wikipedia, 2011. *Ann Cavoukian.* [Online]
Available at: http://en.wikipedia.org/wiki/Ann_Cavoukian
[Accessed 11 November 2011].

World Wide Web Consortium, 2002. *Platform for Privacy Preferences Project.* [Online]
Available at: www.w3.org/P3P/
[Accessed 20 December 2011].

Appendix

Paper-based survey questionnaire



University of Nairobi

School of Computing and Informatics

M.Sc. Computer Science

Survey Questionnaire

Name: Davis M Onsakia

Registration Number: P58/61467/2010

Project title: A business model for encouraging citizens to open up their personal data for anonymous statistical use

Supervisor: Dr. Wanjiku Nganga

Introduction

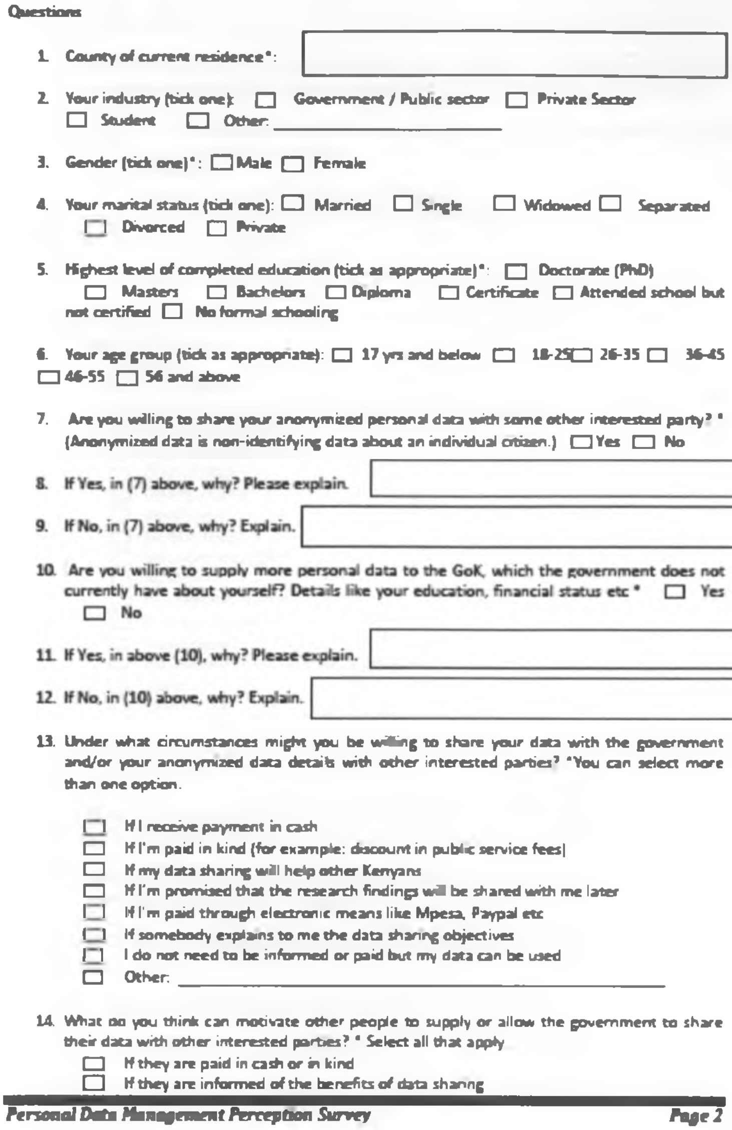
This is a survey I'm carrying out to understand the perception of personal data from individual Kenyan citizens

You have been selected to take part in this survey. I would be grateful if you would assist me by responding to the questions in the questionnaire below. The data and information submitted will be kept confidential and will be used for this academic research purpose only. In this regard, your name is not required in this questionnaire

Your feedback will greatly be appreciated

Instructions

- Kindly answer the following questions (Please tick where appropriate). Where explanations are needed, use the spaces provided
- The question with an asterisk (*) is a required question
- GoK stands for the Government of Kenya



Questions

1. County of current residence*:

2. Your industry (tick one): □ Government / Public sector □ Private Sector □ Student □ Other: ______________________

3. Gender (tick one)*: □ Male □ Female

4. Your marital status (tick one): □ Married □ Single □ Widowed □ Separated □ Divorced □ Private

5. Highest level of completed education (tick as appropriate)*: □ Doctorate (PhD) □ Masters □ Bachelors □ Diploma □ Certificate □ Attended school but not certified □ No formal schooling

6. Your age group (tick as appropriate): □ 17 yrs and below □ 18-25□ 26-35 □ 36-45 □ 46-55 □ 56 and above

7. Are you willing to share your anonymized personal data with some other interested party? * (Anonymized data is non-identifying data about an individual citizen.) □ Yes □ No

8. If Yes, in (7) above, why? Please explain.

9. If No, in (7) above, why? Explain.

10. Are you willing to supply more personal data to the GoK, which the government does not currently have about yourself? Details like your education, financial status etc * □ Yes □ No

11. If Yes, in above (10), why? Please explain.

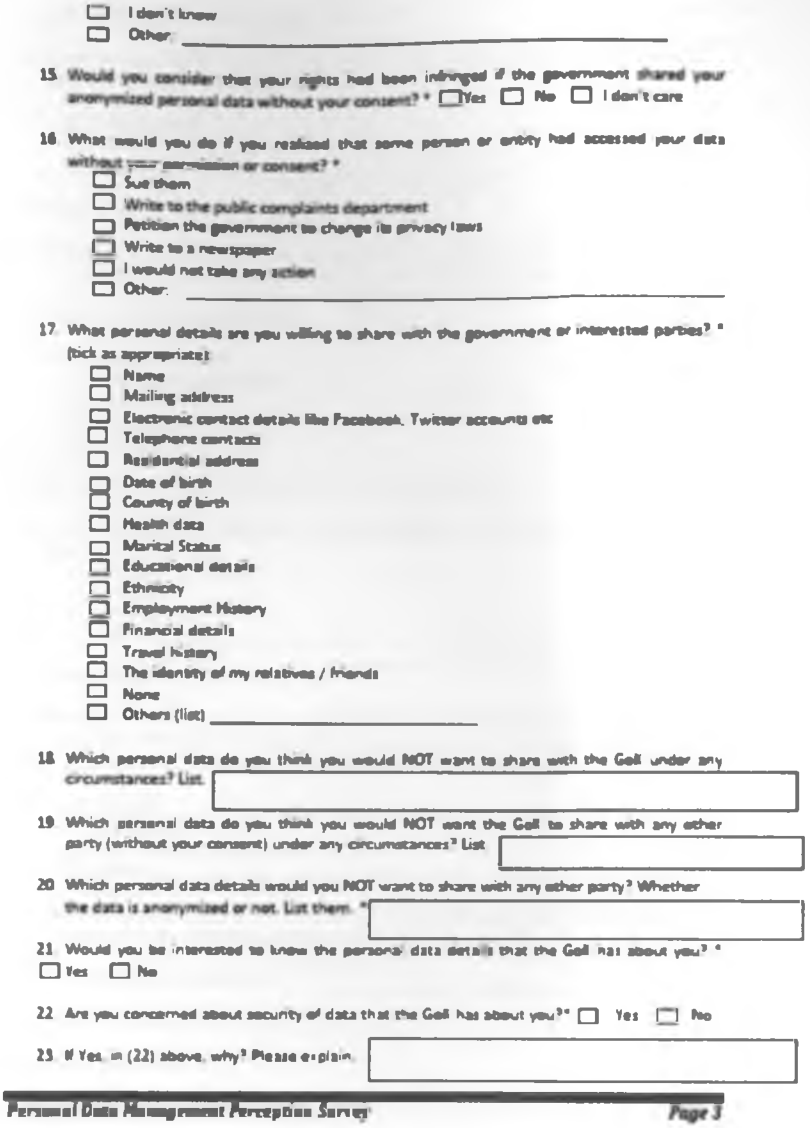
12. If No, in (10) above, why? Explain.

13. Under what circumstances might you be willing to share your data with the government and/or your anonymized data details with other interested parties? *You can select more than one option.

    □ If I receive payment in cash
    □ If I'm paid in kind (for example: discount in public service fees)
    □ If my data sharing will help other Kenyans
    □ If I'm promised that the research findings will be shared with me later
    □ If I'm paid through electronic means like Mpesa, Paypal etc
    □ If somebody explains to me the data sharing objectives
    □ I do not need to be informed or paid but my data can be used
    □ Other: ____

14. What do you think can motivate other people to supply or allow the government to share their data with other interested parties? * Select all that apply

    □ If they are paid in cash or in kind
    □ If they are informed of the benefits of data sharing



□ I don't know
□ Other: ______________________________

15 Would you consider that your rights had been infringed if the government shared your anonymized personal data without your consent? * □ Yes □ No □ I don't care

16 What would you do if you realised that some person or entity had accessed your data without your permission or consent? *
□ Sue them
□ Write to the public complaints department
□ Petition the government to change its privacy laws
□ Write to a newspaper
□ I would not take any action
□ Other: ______________________________

17 What personal details are you willing to share with the government or interested parties? * (tick as appropriate)
□ Name
□ Mailing address
□ Electronic contact details like Facebook, Twitter accounts etc
□ Telephone contacts
□ Residential address
□ Date of birth
□ County of birth
□ Health data
□ Marital Status
□ Educational details
□ Ethnicity
□ Employment History
□ Financial details
□ Travel history
□ The identity of my relatives / friends
□ None
□ Others (list) ______________________

18 Which personal data do you think you would NOT want to share with the GoK under any circumstances? List.

19 Which personal data do you think you would NOT want the GoK to share with any other party (without your consent) under any circumstances? List

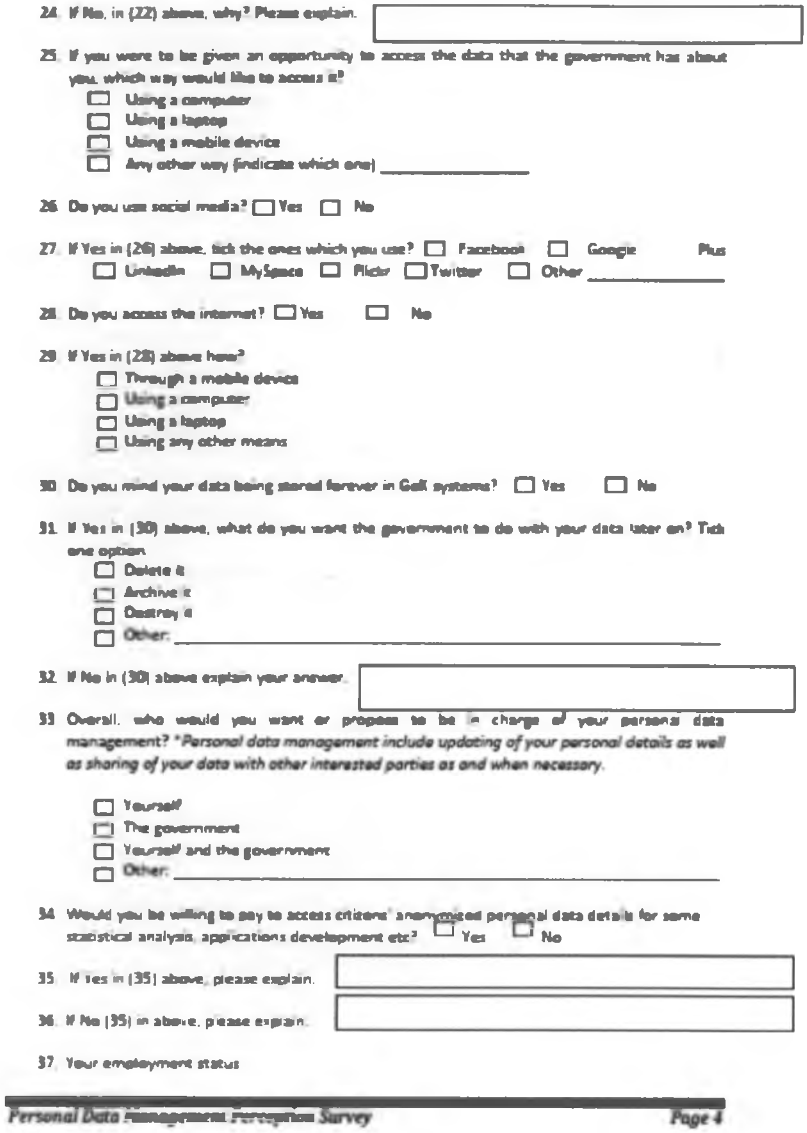
20 Which personal data details would you NOT want to share with any other party? Whether the data is anonymized or not. List them. *

21 Would you be interested to know the personal data details that the GoK has about you? *
□ Yes □ No

22 Are you concerned about security of data that the GoK has about you?* □ Yes □ No

23 If Yes, in (22) above, why? Please explain.

Personal Data Management Perception Survey *Page 3*



24. If No, in (22) above, why? Please explain. □

25. If you were to be given an opportunity to access the data that the government has about you, which way would like to access it?
    - □ Using a computer
    - □ Using a laptop
    - □ Using a mobile device
    - □ Any other way (indicate which one) ________________

26. Do you use social media? □ Yes □ No

27. If Yes in (26) above, tick the ones which you use? □ Facebook □ Google Plus □ LinkedIn □ MySpace □ Flickr □ Twitter □ Other ________________

28. Do you access the internet? □ Yes □ No

29. If Yes in (28) above how?
    - □ Through a mobile device
    - □ Using a computer
    - □ Using a laptop
    - □ Using any other means

30. Do you mind your data being stored forever in GoK systems? □ Yes □ No

31. If Yes in (30) above, what do you want the government to do with your data later on? Tick one option
    - □ Delete it
    - □ Archive it
    - □ Destroy it
    - □ Other: ____

32. If No in (30) above explain your answer. □

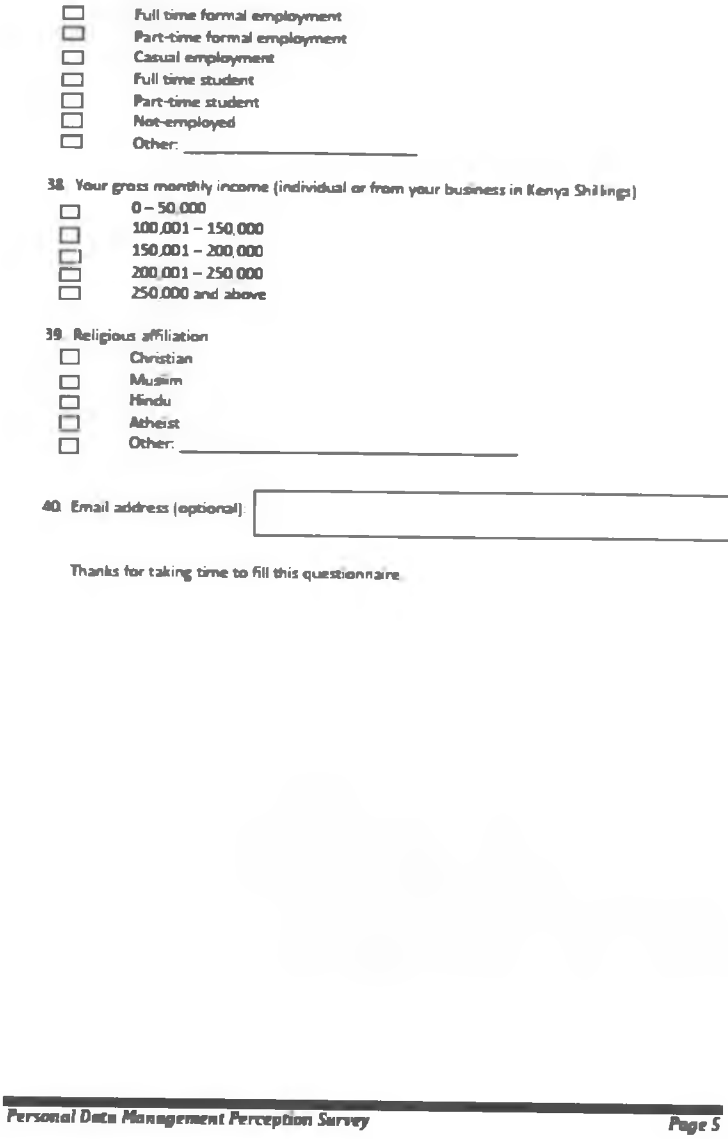
33. Overall, who would you want or propose to be in charge of your personal data management? **Personal data management include updating of your personal details as well as sharing of your data with other interested parties as and when necessary.*
    - □ Yourself
    - □ The government
    - □ Yourself and the government
    - □ Other: ____

34. Would you be willing to pay to access citizens' anonymized personal data details for some statistical analysis, applications development etc? □ Yes □ No

35. If Yes in (35) above, please explain. □

36. If No (35) in above, please explain. □

37. Your employment status



□ Full time formal employment
□ Part-time formal employment
□ Casual employment
□ Full time student
□ Part-time student
□ Not-employed
□ Other: ____________________

38. Your gross monthly income (individual or from your business in Kenya Shillings)
□ 0 – 50,000
□ 100,001 – 150,000
□ 150,001 – 200,000
□ 200,001 – 250,000
□ 250,000 and above

39. Religious affiliation
□ Christian
□ Muslim
□ Hindu
□ Atheist
□ Other: ____________________

40. Email address (optional): □

Thanks for taking time to fill this questionnaire

Personal Data Management Perception Survey *Page 5*

Online questionnaire

The online questionnaire was available at the link:

https://docs.google.com/spreadsheet/viewform?pli=1&formkey=dElISHFETFdhSHh1T1JsbWFTeXJFVUE6MQ#gid=0