cenfri

Public sector data classification frameworks

Module 2

of a report submitted in relation to a public sector data cataloguing, classification and sharing project undertaken as part of the Rwanda Economy Digitalisation Programme

March 2023

In partnership with

mastercard
foundation

GOVERNMENT
OF
RWANDA

Republic of Rwanda
Ministry of ICT & Innovation

Author(s)

Prof. Patrick McSharry

Claude K. Migisha

Cenfri South Africa

Tel. +27 21 913 9510
Email: info@cenfri.org

The Vineyards Office Estate
Farm 1, Block A
99 Jip de Jager Drive
Bellville, 7530
South Africa

PO Box 5966
Tygervalley, 7535
South Africa

Cenfri Rwanda

Tel. +250 788 312 132
Email: info@cenfri.org

www.cenfri.org

mastercard foundation

GOVERNMENT OF RWANDA

Republic of Rwanda
Ministry of ICT & Innovation

Table of contents

List of tables

List of figures

Introduction

Public sector data can be leveraged to increase access to services, make service delivery more efficient and inform policymaking. For instance, early childhood education data can help the government better plan for education development (Adeleke & McSharry, 2022). Real-time data on traffic and urbanisation can inform urban planning efforts including mobility, energy and water consumption, etc. (Uwizera et al., 2022). Combining machine learning techniques with access to data can enable solutions for predicting the incidence of malaria (Habamwabo & McSharry, 2016), forecasting maize productivity (Uwizera & McSharry, 2017) or predicting microgrid electricity consumption (Otieno et al, 2018) among many other things.

The Rwanda Economy Digitalisation Programme, implemented by Cenfri in partnership with the Rwandan Ministry of ICT and Innovation and the Mastercard Foundation, is an initiative that involves collaborating with stakeholders to leverage insights from data analysis to improve policymaking, catalyse innovation and, ultimately, improve livelihoods.

As part of this programme, Cenfri hired a team of consultants to work with the Government of Rwanda (GoR) – specifically the Rwanda Information Society Authority (RISA) and the Ministry of Education (MINEDUC), the Ministry of Agriculture (MINAGRI) and the Ministry of Finance and Economic Planning (MINECOFIN) in three domains:

- Data cataloguing
- Data classification
- Data sharing

This module (Module 2) covers some of the principles of data classification. Data cataloguing and data sharing are covered in modules 1 and 3 respectively.

While this consultancy assignment responded to the context in Rwanda, the needs of the government institutions with whom we worked, and specific programme objectives, this overview is likely to be of use to any public sector entity that is starting out on its data-for-decision-making journey. The datasets, examples and use cases referenced in this document are for indicative purposes and should be substituted with whatever is appropriate in the context in which these data cataloguing principles are being applied.

We recommend that you refer to modules 1 and 3 for a more holistic understanding of the nature of this assignment, the methodology that was adopted, and the aspects that might be applicable in your context.

Data classification

Data cataloguing is the first step to documenting what kind of data exists. Data classification is the second step, which helps set up procedures for managing the data and ensuring that it can be shared without generating unnecessary risks to data subjects or organizations. The objective is to develop a system of classification of the various government data assets to support standardized data management practices for the data already collected or available to policymakers and other public agencies. This requires establishing data sensitivity levels to differentiate how to treat various types of classified data and create a taxonomy to identify each category of data depending on their sensitivity levels (i.e., High, Medium, Low Sensitivity) which will determine the level of access and sharing of each category; as well as clearly define the classification criteria.

The proposed system of classification must be broad enough to accommodate most types of data being generated by the ministries and agencies across the Government of Rwanda.

The proposed "Data Classification Matrix" includes in-depth requirements for each category of information identified, including (but not limited to):

- Provisions on how each category's data is created, accessed, processed, shared, edited, archived, and deleted
- Ownership of data
- Personal data pseudonymization requirements – that's (according to Rwanda's new personal data protection law) the processing of personal data in such a manner that the data can no longer be attributed to a specific data subject without the use of additional information kept separately. Anonymization requirements,
- Obligations of each party with regard to data sharing
- Data usage and distribution
- Data protection and security (including the development of a Data Protection Impact Assessment (DPIA) process to help the institution identify and address the data protection risks for data sharing of each category)
- Any other relevant data governance provisions that can be adopted across the Government

Data Category	Storing	Accessing	Transferring
Public	Backed up	Public website	No encryption
Internal	Email; OneDrive	Username + password	Email to authorised staff
Confidential	Stored in encrypted format; Backed up on secure local drive in locked fireproof room	Authorised staff using username + password	Encryption
Restricted			

Table 1: Requirements for storing, accessing, and transferring each data category

Furthermore, Rwanda's public sector data classification may follow a three-level model namely

1) public
2) intra-institutional (within the Government of Rwanda)
3) internal

This classification is often presented using a traffic-light colour-coded graphic (Figure 1). From the perspective of storing and securing data, it is extremely simple in that once data is classified, it is clear where it can be shared. (Note GoR in this figure refers to the Government of Rwanda but would be substituted with the applicable public sector entity.)

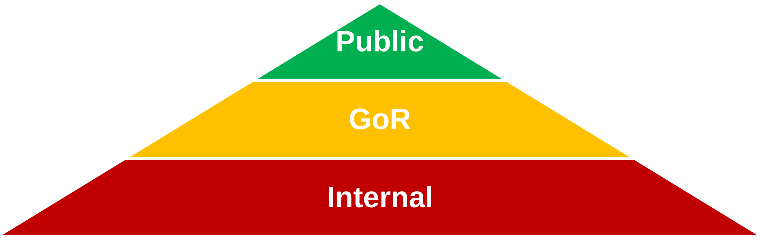
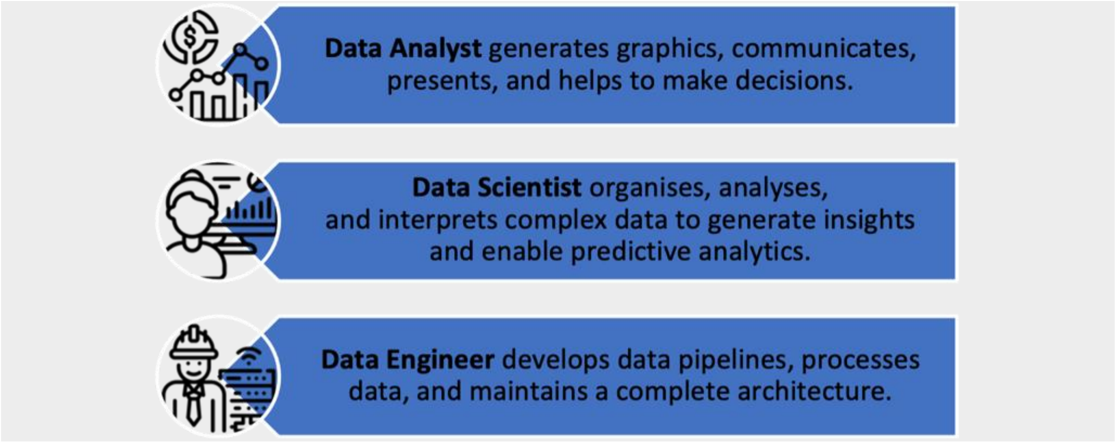


Figure 1: Three-level data classification system

The limitation of the simple three-level classification is that it ignores the humans that are involved in the processing and managing of the data. In all cases, it makes sense to limit access to sensitive data to those that absolutely need to have access in order to carry out specific duties that have been assigned to them. In this way, the more sensitive the data, the fewer people will have access. Furthermore, data usage should be considered when designing a classification system that is fit for a purpose and which involves not only security but also the ability to drive decision-making at multiple levels. By considering the specific roles of those that interact with the data, it is possible to both increase security and to also ensure that the data can support improved decision-making. This approach is best understood by first considering the three kinds of experts that will be working with data in institutions. These are data analysts, data scientists and data engineers as described below (Figure 2).



Data Analyst generates graphics, communicates, presents, and helps to make decisions.

Data Scientist organises, analyses, and interprets complex data to generate insights and enable predictive analytics.

Data Engineer develops data pipelines, processes data, and maintains a complete architecture.

Figure 2: Distinguishing between key staff roles that work with data assets

Data engineers are situated at the start of the pipeline, processing raw data, and therefore having access to personal data such as national IDs, contact details, addresses and individual transactions. They hash out sensitive information, aggregate over temporal and spatial dimensions and pass the pseudo-anonymized data up the chain. Ideally, this hashing process is carried out on the source system at the start of the data pipeline, thereby limiting the opportunity for data leakages. Data engineers play a critical role in protecting sensitive data and are now especially relevant in the light of compliance regulatory frameworks such as Rwanda's Data Protection Law (Number 058/2021 of 13-Oct-2021). In addition, data engineers can mask sensitive data using differential privacy (DP) systems, which make small substitutions in the database, in order that the query result cannot be used to infer much about any single individual, and therefore provides privacy.

Data scientists do not need to work directly on the raw data and by keeping a security wall in place, the risk of personal data leaking out is therefore minimized. Data analysts work with derived data and focus on delivering insights across the institution and enabling others to monitor key performance indicators (KPIs). At the highest level within government, managers will not need to have direct access to any data, but they will use a dashboard to check progress and track KPIs. Finally, at the public level, individuals, researchers, NGOs, and those working in business will have open access to high-level data and summary statistics.

From the perspective of organisational roles, it now becomes clear that there are five distinct classification levels. These five levels are related to the simpler three-level classification in that the public sit at the top, but the Government of Rwanda (GoR) and internal levels are each split into two more levels to acknowledge the specific data experts that will be working with the data. The intra-institutional data sharing among the GoR is divided into two levels, which are data analysts and managers with classifications "Restricted Access" and "Controlled Access" respectively. Internal is also divided into two levels which are data scientists and data engineers with classifications "Restricted Access" and Confidential & Restricted Access". The link between the three-level and five-level systems is maintained visually by the colour-coded stratification (Table 2).

Role	Objectives	Data Usage	Classification	Access
Individuals	Understand, learn, build applications for public good	Summary statistics with anonymised and aggregated data	Open Access	Public
Managers	Monitor KPIs and evaluate progress	Dashboard	Controlled Access	GoR
Data Analysts	Communicate, present and support decision-making	Create graphics	Restricted Access	GoR
Data Scientists	Generate insights and predictions	Data-driven analytics and Machine learning	Restricted Access	Internal
Data Engineers	Maintain data architecture	Develop pipelines with raw data	Confidential & Restricted Access	Internal

Table 2: Roles, objectives, usage, classification, and access

This classification approach allows us to not only assign a classification level (numbers from 1 through 5) to data but also states the role that can have access privileges and provides example use-cases (Table 3).

Classification	Level	Role	Example
Open Access	1	Public	Quantify national educational attainment trends and monitor sustainable development goals (SDGs)
Controlled Access	2	Manager	Monitor KPIs on economic growth at a district or sector level and evaluate progress
Restricted Access	3	Data Analyst	Support decisions about which district to select for piloting an intervention
Restricted Access	4	Data Scientist	Generate insights identifying which variable is responsible for improving educational performance at primary level
Confidential & Restricted Access	5	Data Engineer	Develop data pipeline to access names, grades and locations of primary students and pseudo anonymise for use by the data scientist team

Table 3: Classification level, roles, and specific use-case examples

It is also important to acknowledge that highly sensitive data can be transformed to become less sensitive. Essentially, this means that a data field can be modified so that its classification passes through all five levels. This ability to change classification level in a systematic way is critical as it means that many functions required for use-cases can still be carried out once the correct transformation has been applied. This process of changing classification levels will now be explained with some clear examples that can help to underpin specific use cases. These examples are generic across government in that many databases pertain to people and privacy law aims to protect these individuals. The chain of responsibility is also clearly demonstrated by considering the five roles that are indicated at each level.

The first example focuses on the field date of birth, which will be recorded in a format giving the day of the month, the month, and the year. Having access to a date of birth and other information such as address could easily be used to identify a unique individual. For this reason, the date of birth expressed in the format YYYY-MM-DD, such as 2000-10-01 is classified as level 5 “Confidential & Restricted Access” and can only be accessed by a Data Engineer. To avoid situations where the DD and MM get mixed up, which can be misleading, a better way to ensure date formats are standardised is to apply an ISO code, namely ISO 8601 which is internationally recognised.

Having this highly sensitive classification, however, does not mean that nobody else can use information relating to ages to derive valuable actionable insights. There are a large number of insights that can be generated using ages and hence it makes sense to find a mechanism to transform this sensitive information to protect privacy and still facilitate decision-making. The hierarchy of classification levels and transformations required is illustrated in Table 4 and demonstrates how coarsening the temporal and spatial resolution in steps addresses both privacy risks and permits the sharing of relevant data with the assigned roles specified in Table 2.

Classification	Level	Transformation	Data
Open Access	1	Aggregate nationally using five-year bins	National histogram of ages in 5-year bins
Controlled Access	2	Aggregate across district	Histogram of ages in one year bins for each district
Restricted Access	3	Reduce temporal resolution to age	23 years
Restricted Access	4	Reduce temporal resolution: MMM-YYYY	1990-10
Confidential & Restricted Access	5	Raw data: Date of birth in format DD-MMM-YYYY	1990-10-01

Table 4: Classifying and protecting personal data – Age

A second example considers location, which may be measured using a digital device or a human enumerator. In the former case, it can be precisely recorded as a global positioning system (GPS) coordinate giving the latitude and longitude with an accuracy to within a few metres. In the latter, the location for a household survey is likely to be represented as a street name and number, sector, district, and province. Either of these raw data entries is sufficiently accurate to identify a unique individual or family and hence is classified as level 5 “Confidential & Restricted Access” and can only be accessed by a Data Engineer. The recommendation is to adopt ISO 6709 and specify coordinates in decimal degrees with latitude coming before longitude. It is of course extremely useful to be able to analyse spatial patterns and to assess how products and services are being accessed and consumed by geographical location. Rather than allowing the level 5 classification to prevent such analyses, a hierarchy of transformations is used to coarsen the spatial resolution and pass through all five levels as shown in Table 5 below. By modifying the location from precise GPS coordinates to address to the sector, district and province, an appropriate classification level is achieved and yet the data can be successfully shared with the assigned roles specified in Table 2.

Classification	Level	Transformation	Data
Open Access	1	Reduce spatial resolution to province	Kigali, Rwanda
Controlled Access	2	Reduce spatial resolution to district	Gasabo, Kigali, Rwanda
Restricted Access	3	Reduce spatial resolution to sector	Kacyiru, Gasabo, Kigali, Rwanda
Restricted Access	4	Reduce spatial resolution: Street, Sector, District, Province, Country	KG 7 Avenue, Kacyiru, Gasabo, Kigali, Rwanda
Confidential & Restricted Access	5	Raw data: GPS coordinates	Latitude: -1.9444501 Longitude: 30.0896764

Table 5: Classifying and protecting personal data – Location

As a third example, we explore how the GoR can share “Restricted Access” data between institutions to directly benefit a specific ministry by obtaining important information from other institutions. Suppose the Ministry of Education (MINEDUC) wants to know how household income, access to finance, and spending patterns affect education levels. Managers within MINEDUC may request a data scientist to generate insights and the data analyst to answer key questions such as:

- How does household income affect grades?
- How does electricity consumption affect grades?
- How does airtime consumption affect grades?
- Does having a bank account improve grades?
- Does using mobile money improve grades?

MINEDUC already collects education performance data, which is stored internally and only the data engineers need to have access to this raw data that would identify individual students. By masking all personal data using hashes, MINEDUC can match certain fields of multiple databases that sit in different institutions. These institutions, such as The Rwanda Utilities Regulatory Authority (RURA), The National Bank of Rwanda (BNR), and The National Institute of Statistics (NISR) have valuable predictive variables that can be matched and merged using a national ID. The national ID is hashed and will remain confidential through the data merge. It is recommended to follow ISO/IEC 10118-3:2004 for hashing IDs. Age in years and location by the district is shared at a lower resolution to protect the privacy of each student.

In practice, this would work in the following way. MINEDUC would prepare internally generated data (Table 6) and would request that other institutions within the GoR share relevant variables that can be matched using the hashed version of the National ID (Table 8). The merged data would then combine Table 6 and Table 7 for all the students across the country while protecting their privacy and respecting Rwanda's Data Protection Law. Furthermore, this approach would provide a holistic picture of how household income, access to finance and spending patterns affect education levels. It would allow a data scientist to calculate summary statistics, generate scatter plots, run regressions, and estimate the relationships between variables that are necessary to provide a quantitative answer to each of the questions above. In addition, the statistical significance of each result could be assessed so that action could be recommended when confidence is high.

Name	Description	Data
nationality_id	National Id	9caf77d971cf673e5a59a193be31d203
number_subjects	Number of subjects the student has	8
avg_grade	Average grade	89.7
Age	Age	12
District	District	Gasabo
year_repeats	Repeats of years	0
best_subject	Best subject	Maths

Table 6: MINEDUC internally generated data (example).

Name	Description	Data
nationality_id	National Id	9caf77d971cf673e5a59a193be31d203
family_income	family income per month	100,000
has_bank	Has a bank account	1
has_momo	Has a mobile money	1
average_electricity_spend	average_electricity_spend	500
average_airtime_spend	average_airtime_spend	1000

Table 7: GoR externally generated data (example)

If a data scientist wants to investigate the effect of household income on educational performance using average grades, then the data engineer would prepare specific merged datasets with the structure shown in Table 8. As the data scientist does not need to have a national ID, this hashed field can be replaced by a data count number. Important information such as the average grade, age, district, and family income is now available on a separate row for each student in the database. The data scientist can now create a scatter plot of avg_grade versus income or run regressions of avg_grade versus predictive variables such as income.

number	avg_grade	age	district
1	89.7	12	Gasabo
2	82.5	11	Kicukiro
:	:	:	:
100000	79.9	12	Nyarugenge

Table 8: Merged dataset for investigating the effect of family income on educational performance

This example has presented a particular use-case for MINEDUC but shows the generic applicability of being able to merge different databases across GoR and the tangible benefits that come from doing so. Of course, data classification relies on having access to a complete data catalogue in order for all the data assets to be considered and correctly classified which will, of course, involve a certain amount of data transformation as highlighted in Table 1.

Data cataloguing & classification

By listing all the steps required, it is possible to demonstrate how the data cataloguing involves many roles as indicated by having a designated person being held responsible (Table 9). In contrast, the data engineer (who has full visibility of the raw data) takes charge of protecting personal and sensitive data using the tools of hashing and aggregation and therefore is responsible for the series of five steps for the data classification process (Table 10). In both cases, it is recommended that the newly appointed Chief Digital Officers (CDO) approve and sign off these important processes and the annual updates that will be required to keep everything fresh and fit for purpose.

Checklist	Responsible	Tools	Outcome
Step1: Identify or map out existing data (across departments, etc.) at institution level	Data Engineer	List all datasets and fields in a spreadsheet. Compile a full list using internal communication tools (i.e., email, etc.)	Awareness of existing data assets
Step 2: Collect existing data (i.e., raw, aggregated) and the underlying metadata, if it exists, at institution level	Data Scientist	Email exchange, internal safe data sharing repositories, secured data sharing devices (i.e., USB stick) and consideration of secure API as a future option	Collection of existing internal data
Step 3: Adapt the data catalogue template and start populating it	Data Analyst	Complete data catalogue	Populated data catalogue
Step 4: Validate the data catalogue with CDO and department heads and agree on relevant data	Data Analyst, CDO	Internal meeting and CDO sign-off	Approved data catalogue
Step 5: Keep the data catalogue up to date	Data Engineer, Scientist, Analyst, CDO	Repeat steps 1-4 at annual intervals	Updated data catalogue

Table 9: Checklist for the data cataloguing process

Checklist	Responsible	Tools	Outcome
Step1: Identify personal data and any other sensitive fields.	Data Engineer	Searching for keywords such as national identify	Awareness of sensitive data assets
Step 2: Mask all personal and sensitive data by hashing	Data Engineer	Hashing process	Removal of sensitive data
Step 3: Coarsen resolution of sensitive data	Data Engineer	Aggregation	Transform classification level
Step 4: Validate the data classification with the CDO	Data Engineer, CDO	Internal meeting and CDO sign-off	Approved data classification
Step 5: Keep the data classification up to date	Data Engineer, CDO	Repeat steps 1-4 at annual intervals	Updated data classification

Table 10: Checklist for the data classification process

For information on data cataloguing and data sharing please refer to modules 1 and 3 (forthcoming).

For more information on this project, contact ***Marcellin Nyirishyaka.***

References

Brown, S. & McSharry, P.E. (2016). Improving accuracy and usability of growth charts: Case study of Rwanda. British Medical Journal Online 6: e009046.

Behar, J., C. Liu, K. Kotzen, K. Tsutsui, V.D.A. Corino, J. Singh, M.A. F. Pimentel, P.A. Warrick, S. Zaunseder, F. Andreotti, D. Sebag, G. Kopanitsa, P.E. McSharry, W. Karlen, C.K. Karmakar and G.D. Clifford (2020). Remote health diagnosis and monitoring in the time of COVID-19. Physiological Measurement.

Dushimimana, B., Wambui, Y., Lubega, T. & McSharry, P.E. (2020). Use of Machine Learning Techniques to Create a Credit Score Model for Airtime Loans. Journal of Risk and Financial Management 13 (8), 180.

Mutai, C.K., McSharry, P.E., I. Ngaruye, Musabanganji, E. (2021). Use of machine learning techniques to identify HIV predictors for screening in sub-Saharan Africa. BMC medical research methodology 21 (1), 1-11.

Njuguna, C. & McSharry P.E. (2016). Constructing spatiotemporal poverty indices from big data. Journal of Business Research 70: 318-327.