cenfri

Public sector data cataloguing frameworks

Module 1

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 covers some of the principles of data cataloguing. Data classification and data sharing are covered in subsequent modules.

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.

Methodology

This methodology section provides an initial proposal for use-cases, a process to prioritize use-cases based on stakeholder engagement, and the steps required to arrive at the data catalogue. An agile approach was used to divide the project into several small manageable iterations with constant stakeholder collaboration and engagement throughout. While working with institutions to scope the role of data in facilitating decision-making and development processes, it was important to consider all the dimensions given that data represents both risks and rewards. The former is easier to identify, and data classification plays an important role in making this more tangible. It is, however, critical to also think about the potential rewards, which are often underestimated and overlooked.

The consultants worked with institution representatives to itemise potential rewards that, ultimately, outweigh the risks. These rewards were deemed crucial for unlocking resources, leveraging goodwill, and building successful partnerships. Three interlinked and important components were considered to unlock the value of data. These are
(1) use-cases: demonstrating explicitly the purpose of leveraging digital assets,
(2) institutions: gate-keepers responsible for the security of the digital assets; and
(3) sectors: economic sectors where the use-cases offer tangible rewards.

The participating institutions, with the help of the consultants, identified existing datasets and then created a searchable and explorable "Data Catalogue" of the datasets they are currently collecting and processing. These data catalogues include, amongst others, details on technical, analysis and operational metadata for each of the included datasets. This exercise served to equip teams within these institutions with a holistic understanding of the data currently collected, benchmark their current status and to facilitate both internal and external data sharing.

With reference to the Rwanda Economy Digitalisation Programme's four priority sectors of focus namely, agriculture, retail, education and tourism, several use-cases were proposed (see Table 1). These were compiled based on the consultants' knowledge of the local data ecosystem and extensive research over the last decade.

Use-case	Institution(s)/ Data holders	Sector	Description
Precision agriculture	MINAGRI, RSA	Agriculture	Use of remote sensing (satellite, drone) and surveys to enhance crop yield.
GDP estimation	MINECOFIN, NBR		Historical crop production and market prices at the district level to estimate sub-national GDP from agriculture.
Climate Smart Investment	MINAGRI, RSA, RRA		Historical crop production at the district level and climate scenarios to quantify the climate risk of particular agricultural crops.

Use-case	Institution(s)/ Data holders	Sector	Description
Retail classification	RRA, RDB, NBR	Retail	Point of sales EBM billing machine data to classify business operations, monitor sales and generate revenue for RRA.
Economic Impact analysis	RRA, NBR		Google mobility data to quantify the economic impact of lockdowns to balance public health and the economy for future pandemics.
Route planning	RDB, RURA	Tourism	Mobile phone call detail records (CDR) and transport infrastructure to advise on activities and recommend travel routes to tourist destinations.
Education Dashboard	MINEDUC, REB	Education	Monitoring educational performance using grades and disaggregated data by subjects, teachers, and schools.

Table 1. Proposed use-case examples demonstrating institution interlinkages and sectors

Importantly, to obtain a full grasp of existing digital assets and data systems within target institutions, a set of questions were posed to key managers and teams.

1. What are your most important digital assets and why?
2. Which datasets are viewed as being a high risk if there was a data breach?
3. Which datasets require the most resources and why?
4. Which datasets take up the largest amount of storage space in your organisation and how do you currently store your data?
5. Which datasets are used the most by your organisation?
6. Which datasets are used the most by other organisations and require data sharing?
7. Which datasets could be placed in the public domain?
8. Which datasets should be used for internal purposes only?
9. Which datasets are confidential (contain private data and pose a privacy risk)?
10. Which datasets are restricted (if leaked could result in charges, fines, or other damages)?
11. Describe any existing data security measures in your organisation?
12. Describe any existing controls to reduce privacy risks in your organisation?
13. With regard to data security, do you feel that the current approach is adequate?
14. What data protection or security would be most useful right now?
15. How is data in your organisation organised and catalogued at present?
16. Does your organisation currently have a data classification system?
17. Would a data classification system help your organisation? If yes, how?

18. Can you provide feedback on the proposed four-level classification system (based on the risk quantification of datasets)?
19. How is data currently accessed and shared within your organisation?
20. Are there any Application Programming Interfaces (APIs) available in your organisation for granting access to datasets?

Data catalogues

With the increasing availability and adoption of technology such as mobile phones, computers, point of sales devices, card readers, transactions, the Internet of Things, satellite imagery and unmanned aerial vehicles (UAV), there are several data categories to consider which are not mutually exclusive (see Table 2). These categories capture information about the data format and source and serve to provide a sense of the difficulty in managing, using, and updating each one.

Data Category	Description	Examples
Survey	Data from a predefined group of respondents to gain information and insights into various topics of interest.	Census, Demographic & Health (household) surveys (DHS)
Time series	Data where time ordering defines the sequence that each data point was either captured (event time) or collected (processing time).	GDP, inflation
Machine data	Digital exhaust created by systems, technologies, and devices.	Mobile calls, log records, digital payments
Spatiotemporal	Data that describe the location and time for an event or observation.	GIS, plots, roads, buildings
Unstructured	Social media, news, blogs, audio, videos.	Tweets, Instagram
Real-time	Data made available as soon as it is observed (via an API) which supports live decision-making.	Stock prices, weather

Table 2: Data formats, description, and examples

Furthermore, data are typically distinguished as being either quantitative, implying it is numerical and may be represented as discrete or continuous values, or qualitative which is categorical and can be nominal or ordinal (see Figure 2).

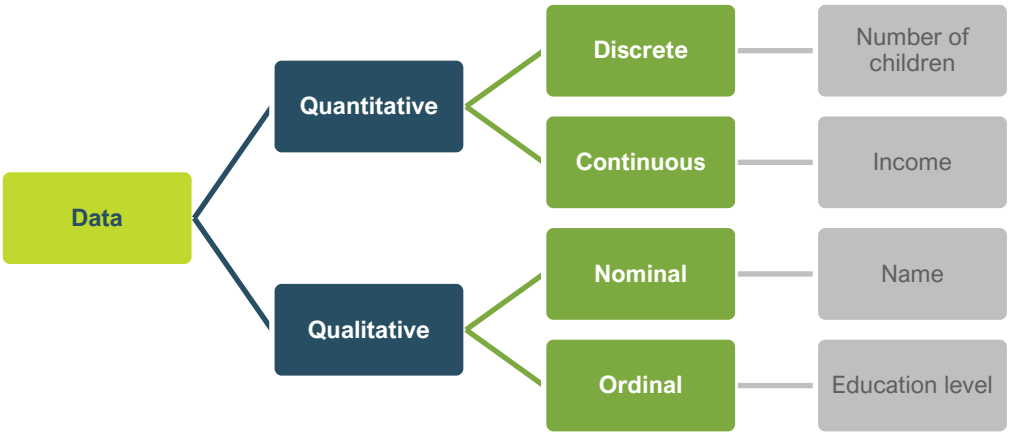

Figure 1: Data types and examples

The general characteristics of the data catalogue include but are not limited to storage size, spatial and temporal coverage, machine readability, source, ownership, quality, used terminology, data classification, usage, sharing obligations, availability of anonymization, completeness, and tagging nomenclature. This includes provisions for how each category's data is created, accessed, processed, shared, edited, archived, and deleted. Additionally, the data catalogues provide information about the names and numbers of different datasets. Therefore, the data catalogue itself can be considered as "metadata" and should be published as structured data, so that third parties can extract information about the datasets (see Table 3).

Metadata structure of the data catalogue	Metadata consistency	Metadata availability
Structure, consistency, and availability of the metadata that makes up the catalogues. **General**: title, description, publisher, frequency, release date, update date, temporal coverage, geographic coverage, licence, data dictionary, granularity, metadata update. **Categorisation**: theme, tags/keywords **Other**: references and citations, quality characteristics, data collection characteristics	Metadata consistency can be ranked on a qualitative scale. Date fields are considered consistent if they follow consistent syntactical formats within a catalogue. Other properties are considered consistent if their values are drawn from a fixed set of options (controlled vocabulary).	Metadata availability is critical for data cataloguing. It is important to provide an overview of available values for given common properties across catalogues which highlights the usefulness of the metadata.

Table 3: Data Catalogue Metadata attributes[1]

1 Enabling Interoperability of Government Data Catalogues: https://hal.inria.fr/hal-01056576/document

With a long-term goal to develop an inventory of the digital assets of the Government of Rwanda and to ensure that the existing data are secured, protected, organised, and can be shared in a way that facilitates decision-making, a series of meetings with crucial GoR stakeholders were conducted to define the required metadata for the respective data catalogues. Data cataloguing serves to discover the nature of these digital assets and paves the way for successful data classification. Simply put, a data catalogue is a repository containing information on data holdings within an organisation which serves to make data discoverable. Catalogues are designed to help people find data and understand how such data can be used, typically including capabilities to register, classify, and detail the metadata describing the data assets.

A data catalogue adds a context layer with a focus on discovery, search, metadata management, lineage, collaboration, and governance. In practice, a data catalogue accelerates the generation of insights by ensuring that the data is readily accessible, with sufficient context, and clearly specified access permissions. By working with Chief Digital Officers (CDOs) and teams from the institutions, respective data catalogues were developed to help find answers to the following questions:

- What data do we have?
- Where does it come from?
- Who is the owner?
- How clean is the data are there any gaps?
- How is it classified?
- Is the data good enough for running analysis?

The developed data catalogue hinges on three components namely, the Glossary, Metadata and Dictionary. A glossary provides definitions of the key terms, acronyms, terminology, calculation rules, and any other related information that might be helpful for understanding the data assets. The metadata then describes the dataset, its purpose, and its source. A data dictionary contains the metadata about the database and can be viewed as its documentation. It provides analysts with the context behind tables, rows, columns, and data fields. Without data dictionaries, one would have to rely on colleagues or read through manuals and queries. The data dictionary is an integral part of a data catalogue. The following serves as a template while featuring examples from MINAGRI and MINEDUC for what is expected in these three components.

Data glossary

Term	Definition	Calculation Rule
Promotion Rate	The number of pupils entering a given level of education as a percentage of the pupils who were enrolled in the previous year at the previous level. It shows the percentage of pupils promoted to the next grade in the following school year.	Number of pupils promoted to grade G+1 in year t+1/ Number of total pupils who were enrolled in grade G in year t x 100

Term	Definition	Calculation Rule
Net Enrolment Rate	Enrolment of the official age-group for a given cycle of education expressed as a percentage of the corresponding population	Number of pupils of specific age at a given level in year t/ Population of school age in year t x 100
Repetition Rate	Proportion of pupils enrolled in a given school year who study in the same grade the following school year.	Number of pupils repeating the grade G in year t+1/ Number of total pupils who were enrolled in grade G in year t x 100
Dropout Rate	Proportion of pupils from a cohort enrolled in a given grade at a given school year who are no longer enrolled in the following school year. Dropout rate can also be obtained by subtracting the sum of promotion rate and repetition rate from 100 in a given school year.	Number of pupils who are no longer enrolled / Number of total pupils who were enrolled in grade G in year t x 100

Data metadata

Item	Description	Examples
Asset ID	Unique identifier of dataset	ESOKO1, CAMIS1
Title	Title of the dataset	Transactions from Jan 2000 to Sep 2022
Description	Around 100 words describing the sector, key fields, owner, and purpose	The platform was established to link sellers, buyers, and exporters of tea, coffee horticulture, and emerging value chain products to support effective trading, storage, and sharing of information. The system aims to digitize the link between agricultural and animal product sellers with buyers in the international and local markets
Keywords	Keywords to identify when searching	Commodity; crops; agriculture; prices
Owner	List the owner(s)	MINAGRI
Curator	List the entity that produced the resource	MINAGRI
Contributor	Entity that produced the dataset	E-soko management
Consumers	List the consumer(s)	Public, farmer, dealers
Application	What is the application name?	E-soko

Item	Description	Examples
URL	Website name	www.esoko.gov.rw
Size	Storage capacity	10 GB
Number of fields	Total number of fields	30
Number of records	Total number of records	100000
Unstructured	Is the data unstructured?	Yes, No
Files	Type of data files	csv, excel, doc, PDF, txt
Units	Units of measurement	Commodity prices in RWF
Database type	Give the type of database	Flat file, Relational
Personal	Does it contain personal data?	Yes, No
Anonymization	Give the level of anonymization	Anonymized, Pseudo-anonymized, anonymized
StartDate	When was the data first collected?	01-Jan-2000
LastDate	Give the last date of data entry	01-Sep-2022
RefreshDate	The date of the next data collection	01-Oct-2022
Sampling Period	The sampling period of the data	Second, minute, hour, day, week, month, year
Refresh period	How often is the data refreshed?	Second, minute, hour, day, week, month, year
Spatial coverage	What is the spatial coverage?	Rwanda, Kigali
Spatial resolution	What is the spatial resolution?	National, district, sector, km2, hectare, m2
Aggregation	Does the dataset contain raw data or aggregated?	Raw, aggregated
Access	How is the data accessed?	API, download, email, negotiation

The data dictionary should reference the specific database and the structure can group fields by modules. For example, MINAGRI's E-soko database contains fields regarding agents, markets, commodities, and prices. Similarly, MINEDUC's CAMIS database contains fields regarding students, teachers, and assessments.

Data dictionary

Item	Description	Examples
Database	Name of the database	E-soko, CAMIS, MIS
Module	Name of the module	Student, Teacher, Agent, Commodity
Field	Name of the field	StudentID, AgentID, Marks, Price
Description	Description of the field	A unique identity number for each student
Type	Type of data entry	DATE, UUID, TEXT, BOOLEAN, INT, DOUBLE
Size	Number of characters or digits	1, 10, 20, 100
Min	Minimum value	0
Max	Maximum value	100
Completeness	Percentage of expected entries that are complete	

During the assignment period, the team of consultants worked closely with select public sector entities to develop suitable data catalogues encompassing existing datasets. The data cataloguing exercise identified observational level data that is either generated through the regular process of doing business, as a by-product of regular economic activity (but not currently seen as valuable) or as part of regulatory compliance. Moreover, for each data holding, the corresponding metadata was listed to make it easy for those who wish to be able to access the contents of the data, its purpose, scope, contact details, etc. The data catalogues were created in a searchable and explorable MS Excel sheet.

For information on data classification and data sharing please refer to modules 2 and 3 (forthcoming).

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

References

Adeleke, O. & McSharry, P.E. (2022). Female enrollment, child mortality and corruption are good predictors of a country's UN Education Index. International Journal of Educational Development 90, April 2022, 102561.

Habamwabo, D. & McSharry, P.E. (2016). Healthcare Monitoring based on Digital Transactions at Pharmacies: Malaria in Kigali. IEEE International Conference on Bioinformatics and Biomedicine, 15-18 Dec, Shenzhen, China. ISBN 978-1-5090-1612-9.

Otieno, F. & McSharry, P.E. (2018a). Forecasting energy demand for microgrids over multiple horizons. IEEE PES&IAS PowerAfrica Conference, Cape Town, South Africa.

Uwizera, D. & McSharry, P.E. (2017). Forecasting and monitoring maize production using satellite imagery in Rwanda. IEEE Technological Innovations in ICT for Agriculture and Rural Development, 7-8 Apr, Chennai, India. ISBN:978-1- 5090-4437-5.

Uwizera, D., Ruranga, C. & McSharry, P.E. (2022). Deep learning inter-city road conditions in East Africa for infrastructure prioritization using satellite imagery and mobile data. Research Journal of the South African Institute of Electrical Engineers (SAIEE) 114(1):14-24.