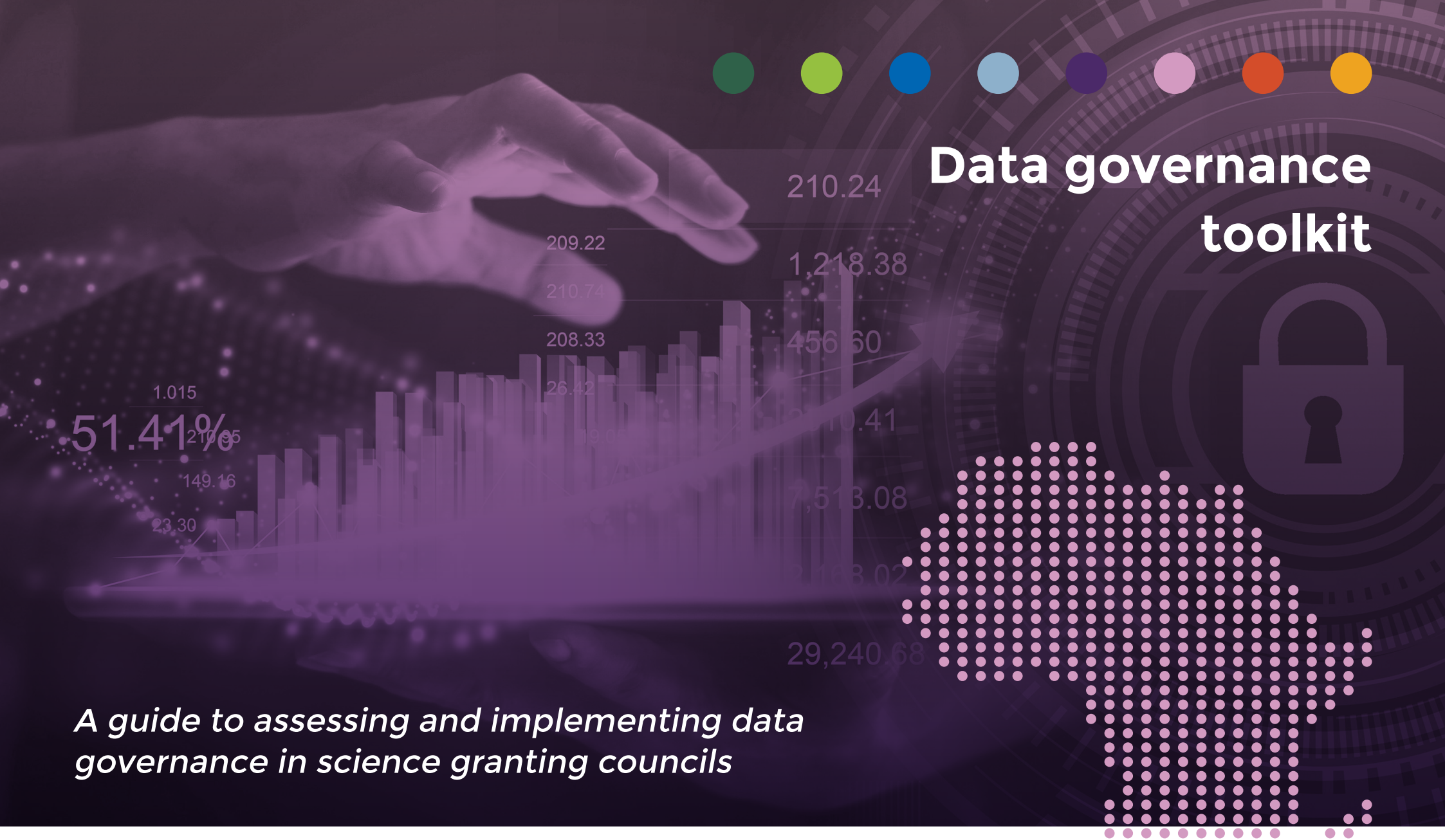


Data governance toolkit

A guide to assessing and implementing data governance in science granting councils

SCIENCE GRANTING COUNCILS INITIATIVE SGCI IOSRS INITIATIVE DES ORGANISMES SUBVENTIONNAIRES DE LA RECHERCHE SCIENTIFIQUE

Compiled as part of the Evi-Pol project, funded under the Science Granting Councils Initiative (SGCI) Phase II (grant number: 109573).

Evi-Pol aims to enable more effective use of evidence in policy and decision-making by African science granting councils. The SGCI is funded by the Canadian International Development Research Centre (IDRC), the South African National Research Foundation (NRF), the Swedish International Development Cooperation Agency (SIDA), the British Foreign Commonwealth and Development Office (FCDO), the German Research Foundation (DFG), and the Norwegian Agency for Development and Cooperation (Norad). Evi-Pol is led by the African Centre for Technology Studies (ACTS) in Kenya in partnership with the Centre for Science, Technology and Innovation Indicators (CESTII) of the Human Sciences Research Council (HSRC) in South Africa and Université Cheikh Anta Diop de Dakar (UCAD) in Senegal.

First published: May 2023

Authors
Yasser Buchana, Mbongeni Maziya, Marco Davids (Evi-Pol IT Architecture Specialist Consultant), and Il-haam Petersen.

Illustrations
Tebogo Matshana
Antonio Erasmus

Copy editing
Katharine McKenzie

Design and layout
Tracey Watson

Please cite this publication as:
Buchana, Y, Maziya, M, Davids, M, and Petersen, I. 2023. *Data governance toolkit: a guide to assessing and implementing data governance in science granting councils*. South Africa: CeSTII-HSRC

Project team
Glenda Kruss, CeSTII Executive Head and Evi-Pol co-Principal Investigator
Il-haam Petersen, Chief Research Specialist and Evi-Pol Project Lead at CeSTII
Nazeem Mustapha, Chief Research Specialist, and Evi-Pol Data Management Systems Work Package Lead at CeSTII
Darryn Whisgary, Research Manager, and Evi-Pol Project Manager at CeSTII
Moses Sithole, Research Director, CeSTII
Yasser Buchana, Senior Research Specialist, CeSTII
Atoko Kasongo, Research Specialist, CeSTII
Gerard Ralphs, Programme Manager and Senior Policy Analyst, CeSTII
Mbongeni Maziya, PhD Research Intern, CeSTII

Technical reviews
Moses Sithole, Research Director, CeSTII
Gerard Ralphs, Programme Manager and Senior Policy Analyst, CeSTII

Disclaimer

The views expressed herein are strictly those of the authors and do not necessarily represent those of the IDRC or its Board of Governors. This data governance toolkit is provided as a guide only and is not intended to be a substitute for professional advice. While every effort has been made to ensure the accuracy and reliability of the information contained herein, the creators of this toolkit make no representations or warranties, express or implied, as to the accuracy, completeness, reliability, suitability, or availability of the information or related graphics contained in the toolkit for any purpose. Any reliance you place on such information is therefore strictly at your own risk.

The creators of this toolkit disclaim any liability for any loss or damage including without limitation, indirect or consequential loss or damage, or any loss or damage whatsoever arising from loss of data or profits arising out of, or in connection with, the use of this toolkit.

This toolkit may contain links to third-party websites that are not under the control of the creators of this toolkit. The creators of this toolkit have no control over the nature, content, and availability of those sites. The inclusion of any links does not necessarily imply a recommendation or endorsement of the views expressed within them.

Every effort is made to keep the toolkit webpages up and running smoothly. However, the creators of this toolkit take no responsibility for, and will not be liable for, the toolkit being temporarily unavailable due to technical issues beyond our control.

This disclaimer statement is subject to change without notice. By using this toolkit, you agree to be bound by the most current version of the disclaimer statement.

Evi-Pol
Enabling more effective use of evidence in policy and decision making by science granting councils

African Centre for Technology Studies
Knowledge for Better Livelihoods

HSRC Human Sciences Research Council
CeSTII Centre for Science, Technology & Innovation Indicators

Contents

Preface

In sub-Saharan Africa, national systems of innovation take distinctive forms. Science granting councils play a central role and balance multiple mandates to set national research agendas, manage funds for research and innovation activities, gather evidence on science, technology and innovation (STI) and advise on STI policy. They typically do this with limited funding, human resources, and organisational capacity.

To strengthen their capacities to better perform these intermediary functions, the Science Granting Councils Initiative (SGCI) was launched by a consortium of international funding agencies, led by the International Development Research Centre (IDRC).

Currently, 16 sub-Saharan African countries participate in the SGCI from East, West and Southern Africa. The Evi-Pol project, which ran from November 2020 to February 2023, responded to one theme under the SGCI Phase Two, through a consortium led by the African Centre for Technology Studies (ACTS) in Kenya. It focused on strengthening the role that science granting councils play in identifying, managing and using evidence in policy and decision making.

Rather than follow a traditional model in which experts *parachute* in to *transfer* skills and knowledge, the Evi-Pol project took a different approach to providing technical assistance. The project design was based on a participatory approach, that emphasised consultation from the start, the co-creation of solutions, bringing in local consultants and building local networks. Flexibility in the design and process was encouraged. Using this model, much of the first year of the project was spent developing work plans, frameworks, and instruments through (virtual) consultative meetings and workshops. The technical assistance provided was thus demand driven and customised to the needs and capabilities of each science granting council, and included interactive workshops, peer-to-peer learning opportunities and one-on-one coaching.

In collaboration with partners in the Université Cheikh Anta Diop de Dakar (UCAD) in Senegal, CeSTII led activities aimed at supporting science granting councils to strengthen their data management systems. The work was led by Nazeem Mustapha, with a team of CeSTII researchers and statisticians, and a consultant IT Architecture Specialist. Glenda Kruss and Il-haam Petersen were responsible for overall project conceptualisation, oversight and co-ordination at CeSTII. During the first year, Gerard Ralphs, Amy Kahn and Moses Sithole drafted a project process document that set out the framework for the work on data management systems. This was informed by a needs assessment survey designed by Yasser Buchana and completed by nine science granting councils. Darryn Whisgary, as project manager, played a key role in team co-ordination, liaising with the science granting councils and keeping the project activities on track.

Working with data managers and staff at the participating science granting councils, the team produced a set of toolkits to help build sound data management systems that align with the mandates and capabilities of science granting councils:

- Digital Transformation Roadmap: a guide for African science granting councils to support the development and maintenance of data management systems, with a particular focus on digitalisation
- Data Governance Toolkit: a guide to assessing and implementing data governance in science granting councils
- Guide to Data Curation

The toolkits were designed for use as interrelated guides for *enabling more effective use of evidence in policy and decision making by African science granting councils*.

This Data Governance Toolkit guides science granting councils through a process to assess their current data governance systems and develop a strategy to strengthen data governance. Data governance is a key part of data management and is therefore integral to a science granting council's roadmap to digital transformation. Science granting councils may use the Digital Transformation Roadmap as a guide to create their own roadmaps to develop digital data management systems. The Guide to Data Curation was created in response to a specific need by science granting councils for a guide on best practice to curate grants and research-related data for decision making and public use.

A big thank you to the data managers, SGCI co-ordinators and leadership at the science granting councils who contributed to the creation of these toolkits as resources accessible beyond the Evi-Pol project.

ABOUT THIS **TOOLKIT**

Why this toolkit

A common need expressed by African science granting councils (SGCs) is to develop their own consolidated digital data management systems. Effective data management is key and must be underpinned by sound policies and structures governing how data is accessed, controlled, shared, and used.

In the modern technological era of big data and digital technologies, effective data governance is crucial to ensure that data is secure, accurate, accessible, and usable. As key national institutions responsible for funding research and managing large volumes of research-related data, SGCs need to ensure the protection of sensitive data from unauthorised access. This data includes surveys, administrative profiles and the grants awarded to researchers or research institutions. Data protection requires the establishment of effective security protocols and procedures to ensure data confidentiality, integrity, and accessibility. SGCs may damage their reputations, incur financial loss, legal damages, and lose the trust of stakeholders if they fail to adequately protect data.

Developing an effective and comprehensive data governance system requires a well-defined framework and tools that SGCs can use to implement data governance best practice. This toolkit provides a step-by-step guide for SGC data managers to implement effective data governance systems and protocols tailored to their specific needs.

WHY DATA GOVERNANCE IS IMPORTANT:

- Ensures the security and confidentiality of sensitive information, while meeting the SGC's legal and regulatory obligations.
- Helps to build trust with data users.
- Helps to protect the reputation and brand of the science granting council.

What this toolkit is about

The toolkit is based on the data governance framework developed by the Data Governance Institute, a widely recognised industry standard. The framework provides a comprehensive set of guidelines and best practices that can be used to develop and implement effective data governance systems. It covers all aspects of a data governance framework, including data quality, data management, data security, data privacy, and compliance.

Who this toolkit is for

This toolkit is designed for data managers in SGCs who are responsible for managing research-related data. It is a practical guide that can be used to develop and implement a data governance system tailored to the specific needs and challenges that face SGCs. The toolkit is designed to be accessible and user-friendly to SGCs with different levels of data governance experience and expertise.

How to use this toolkit

The toolkit is designed to guide SGC data managers through the process of implementing a data governance system. Step by step, it covers the key aspects of the data governance framework, including establishing a data governance structure, developing policies and procedures, implementing data quality controls, managing data security and privacy, and ensuring compliance with relevant regulations.

The toolkit also includes a **data governance maturity assessment tool** to evaluate the effectiveness of the data governance system and identify areas for improvement. This flexible tool can be adapted to the specific needs and priorities of individual SGCs. A template is provided, and an editable version is included in an accompanying Excel file. The sections of this toolkit are outlined in Figure 1.

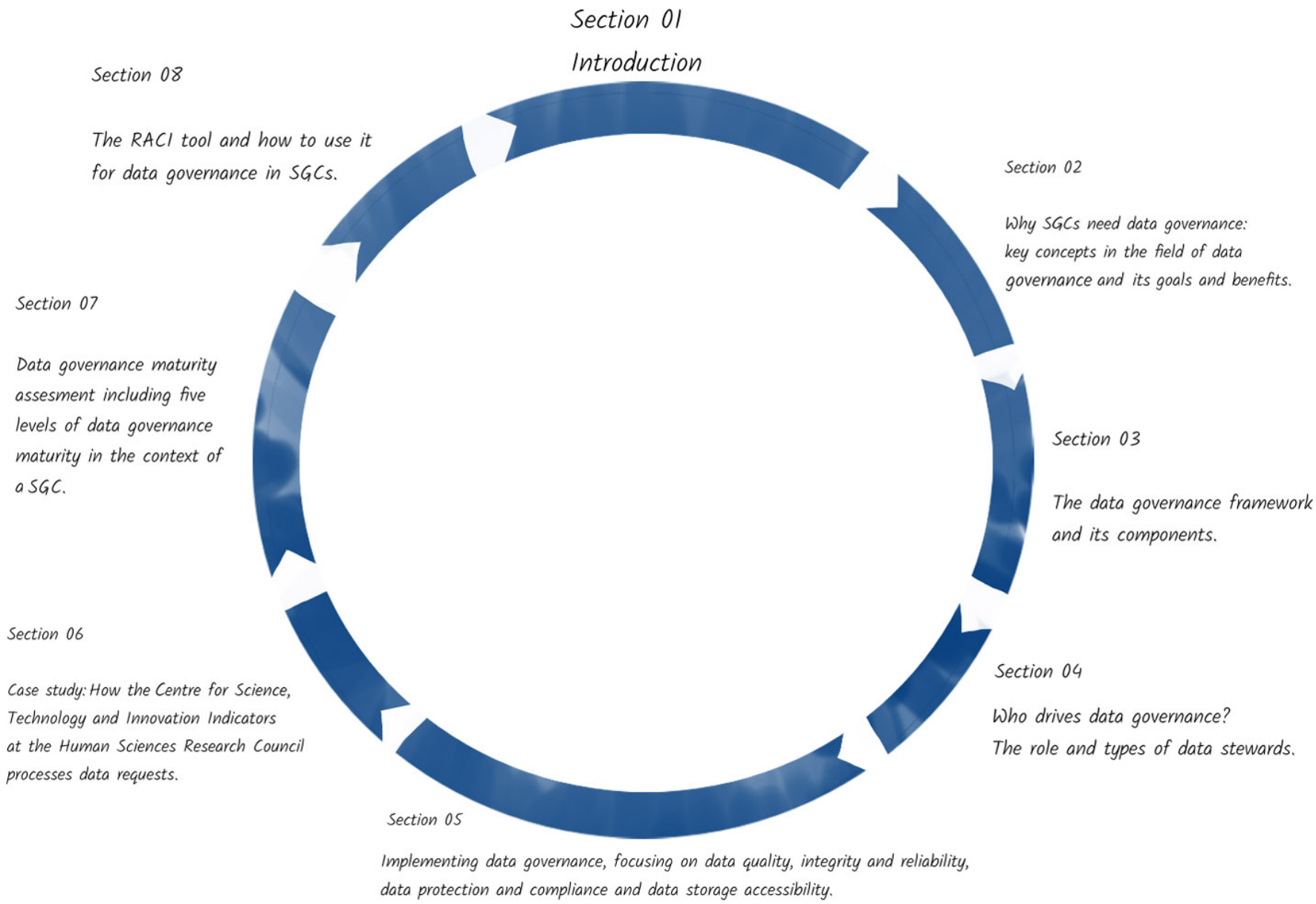


Figure 1
Overview of this toolkit

DATA GOVERNANCE FRAMEWORK **AND PRINCIPLES**

Why SGCs need data governance

Understanding the difference between data management and data governance

The simplest definition of data management is ensuring that each employee in an organisation knows how to and can access the specific data they manage, and that the quality of this data is of the highest standard. For instance, a funding manager in an SGC may require detailed information on financing for a particular research project, the type of research project, and the researcher's rating. This information helps with decision-making about the funding of that particular project. Other data stakeholders in SGCs have their own unique demands and specifications.

Data governance is one part of the overall discipline of data management, though an important one. It is about the roles, responsibilities, and processes that ensure accountability for and ownership of data assets. While data management is a commonly used term within the discipline, it is sometimes referred to as data resource management or enterprise information management.

Successful data governance has clear objectives, processes, metrics, and standards. Overall, it contributes to the success of an organisation by embedding a culture of compliance and adherence to good governance practices. Hence it is mandatory for every organisation with responsibility for the management of data. Data governance is the starting point for managing data. A credible data governance programme addresses issues such as the availability and accessibility of data, as well as its provenance, meaning and trustworthiness. In essence, data governance is a responsibility that should be shared by all constituents of an organisation and facilitate best practice.[i]

Why data governance is important for SGCs

Data that is inconsistent, out of date, incorrect, and poorly safeguarded is a concern for organisations because it can result in bad decision making and be easily misused. Digital transformation has encouraged organisations to use data analytics to inform decision-making across organisations.

> "A well-designed data governance programme provides the right ownership and accountability model to get to the root cause and resolution of data issues."
>
> **Allison Sagraves, Chief Data Officer, M&T Bank**

Most organisations have some form of data governance for individual applications, business units, and functions, even if the processes and responsibilities are informal. As a practice, data governance is about establishing systematic, formal control over these processes and responsibilities. Doing so helps organisations remain responsive, especially as they grow to a size that means it is no longer efficient for individuals to perform cross-functional tasks. Several of the overall benefits of data management can only be realised after the organisation has established systematic data governance. Some of these benefits include:

1. Better and more comprehensive decision support as a result of consistent, uniform data management across the organisation.
2. Clear rules for changing processes and data that help the business and IT section become more agile and scalable.
3. Reduced costs of data management through the provision of central control mechanisms.
4. Increased efficiency through the ability to reuse processes and data.
5. Improved confidence in data quality and data processes.
6. Improved compliance with data regulations.

Overview of the data governance framework

Data governance can be thought of as a function that supports an organisation's overarching data management strategy. For SGCs, a data governance framework provides a holistic approach to collecting, managing, securing, and storing data.

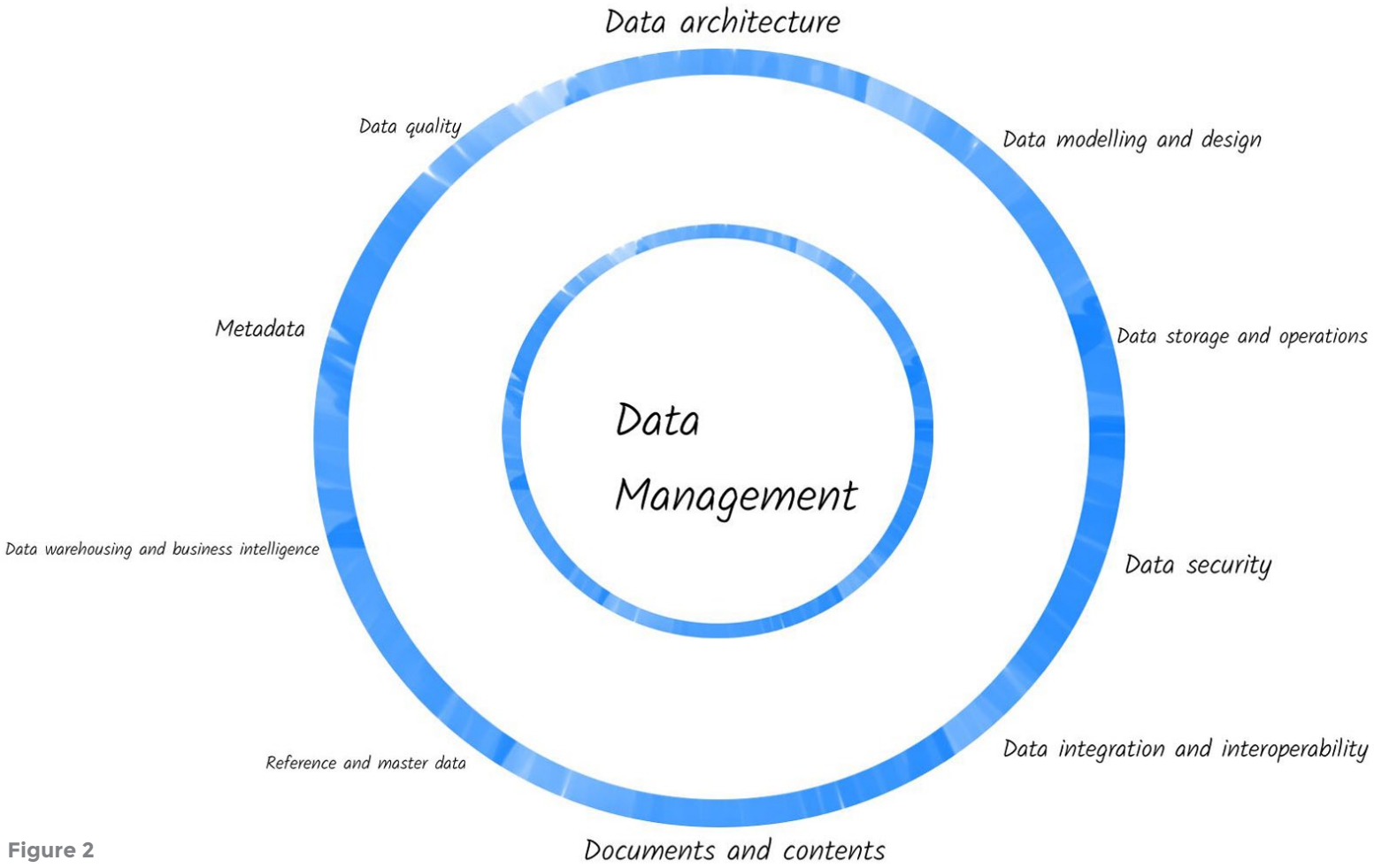


Figure 2
Data management knowledge areas

Source: Adapted from Earley, Henderson, and Data Management Association (DAMA) (2017)

To help understand what a data governance framework should cover, data management can be seen as a wheel with data governance as the hub from which ten data management knowledge areas radiate including data architecture, quality, modelling and design, storage and operations, security, integration and interoperability, documents and content, as well as reference and master data, data warehousing and business intelligence, and metadata (see Figure 2).[ii]

A strong data governance framework addresses problems that may compromise the data assets of an SGC. This part of the toolkit defines the key structure and methods by which different SGCs may apply ownership over their own data, wherever the data resides. Figure 3 presents a data governance framework that has been adopted for this toolkit, based on a framework developed by the Data Governance Institute. The data governance framework has three important layers:

1. **Processes, policies and procedures** are the mechanisms that drive governance.
2. **Rules and rules of engagement** include in-house control mechanisms, such as assigning who makes decisions on the different levels of data governance and accountability (who does what).
3. **People and organisational** bodies include the different groups of individuals within the organisation that perform different functions and have a stake in the data.

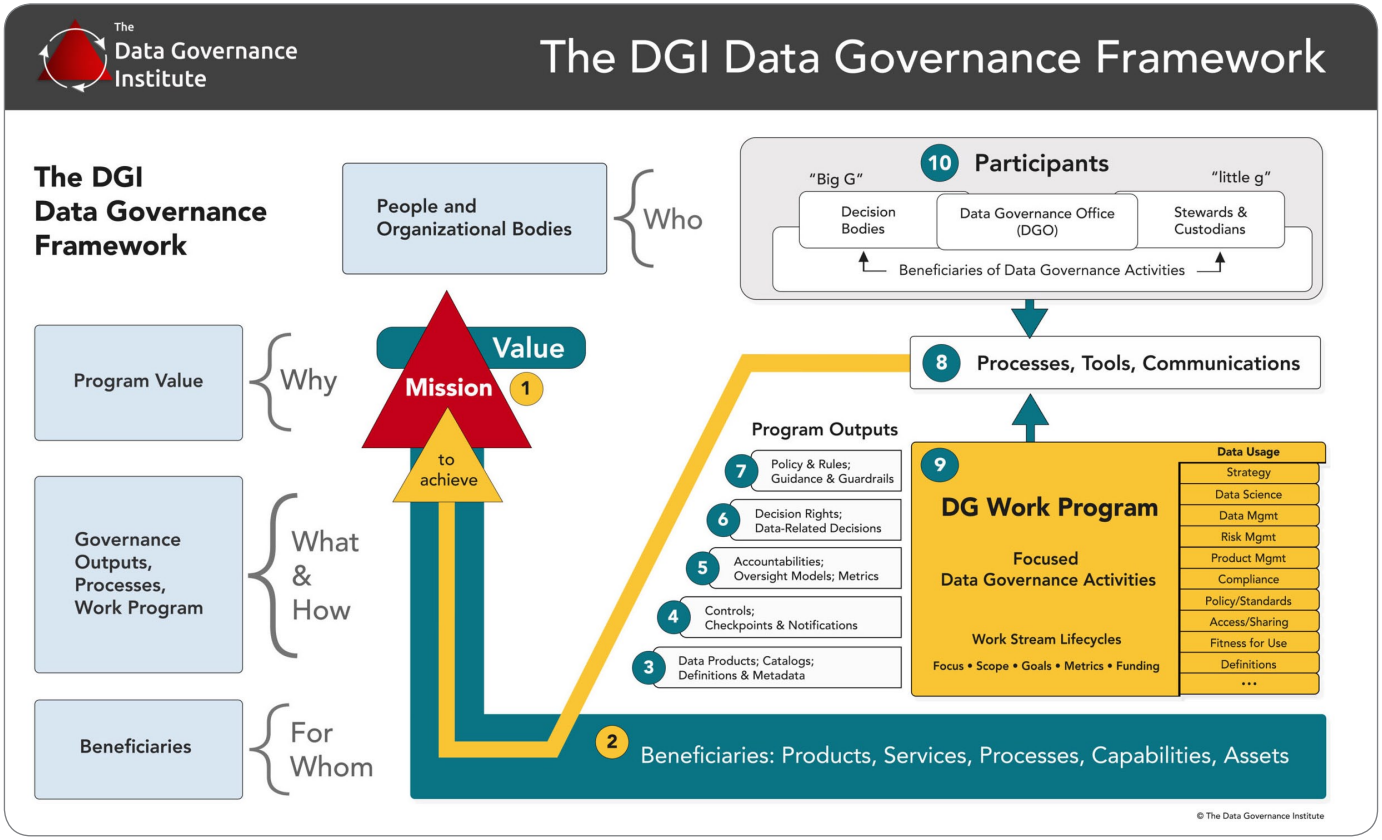


Figure 3
Data governance framework. *Source: The Data Governance Institute (2023)*

Goals of data governance

Data governance goals are value-based objectives that organisations put in place to manage data.[iii] These goals are the meanings that data represent to the SGC. As public bodies, a straightforward goal for SGCs is to fill knowledge gaps by funding research activities. Other goals include: inform policy making, address societal challenges, innovate and adopt a responsible approach to data. The value created through these goals is economic growth and improved service delivery to citizens.

DATA GOVERNANCE FOCUS AREAS:

1. Data governance goals support an organised system to manage data effectively and ensure that data is clean and consistent.
2. Data governance metrics are measures of success.
3. Data governance processes require funding from the SGC. A data governance line item should be included in the budget.

Data governance metrics

Metrics are indicators used to track performance in a system and help to ensure that the system yields the desired outcomes. Table 1 presents the goals of data governance and the metrics that can be adopted by SGCs. Individual SGCs can have their own data governance metrics depending on the nature and scope of their operations.

Table 1 Data governance metrics

Goal	Metrics
Ensure data quality	Percentage of data attributes that meet data quality dimensions such as completeness, correct format, valid values, non-duplicates or currency Increased customer satisfaction in areas such as: • the trustworthiness of data • the availability of data • conformance of data with industry standards • the extent to which data is easy to understand, manipulate and apply to different tasks
Build data infrastructure	Number of consolidated and shared data sources Increased speed in delivering data to stakeholders
Deliver data services to stakeholders	Adherence to terms specified by organisation in terms of the lead time between data request and delivery Time taken to complete data tickets such as response to data-related inquiries from stakeholders, improvement requests for common data entities, access requests or other requested data updates
Improve decision-making	Reduction in time for making data-driven decisions Reduction in number of bad decisions due to bad data

Source: Pierce (2022)

Data governance processes[iv]

Prior to the implementation of a data governance programme, it is necessary to construct a comprehensive data governance mechanism. This begins with policy development to define governance goals and strategies, followed by the establishment of data governance structures for the organisation. The final step is the establishment of data governance procedures, to enhance data governance through process optimisation. This section on data governance processes was largely drawn from Khatri and Brown (2010). A description of these steps is provided here:

1. **Develop data governance policies**: Since the data collected by the SGCs primarily consists of data from the private citizens/organisations, it is necessary to develop regulations for the governance of the data (e.g., how the data can and should be used), the categories that are available for big data analytics (e.g., personal information), and data management mechanisms.
2. **Establish organisational structures for data governance**: Data may be sourced from each organisation's departments; therefore, it is necessary to establish data management mechanisms and leadership structures responsible for defining data governance strategies, goals and budgets, which encompass the entirety of the organisation. It is also necessary to establish data governance working groups, which are responsible for the implementation of data governance policies.
3. **Establish data management procedures**: The aim of data governance procedures is to formulate a systematic and standardised set of processes and usage rules. Although the data governance procedures of each industry may differ, they must strictly regulate every aspect of data, at every level. This includes the quality, standards, safety, framework, model, and lifespan of the data, as well as master data and metadata management. After the establishment of a comprehensive set of data governance procedures, it is necessary to confirm whether data governance has improved and the goals achieved. Therefore, a result appraisal must be included as the final step in a data governance system.

In summary, data governance procedures may be further divided into four steps: setting objectives, defining assessment criteria, appraising results, and auditing and improving.

Who drives data governance? The role of data stewards

Data stewards are the people and organisational bodies in charge of how data is collected, stored, and utilised. Their actions frequently determine whether data governance succeeds or fails. They work hand-in-hand with data stakeholders and the data governance office.

Data stewardship formalises responsibilities for managing data so that those responsible for this important task are held accountable for the loss or misuse of the data. The role of data stewards includes making sure that data quality metrics are maintained throughout the lifecycle of the data. Since the data is monitored on an ongoing basis, where loopholes emerge, data stewards provide feedback to the IT department so that defects are removed. Data stewards who are accountable for the data and protocols help ensure that decisions pertaining to the data are made in the best interest of all data users. Figure 4 identifies four types of data stewards:[v]

1. **Lead data stewards**: The lead data stewards report to the data governance office and their role is to establish the data domain ownership of business elements and make important group decisions.

2. **Business data stewards**: Business data stewards in a specific department of the SGC are responsible for data that belongs to their department.

3. **Operational data stewards**: These are project leaders in the SGC interacting with the data on a day-to-day basis. They are the first to notice loopholes or deficiencies in the data. These stewards can report back to business stewards if they identify a problem.

4. **Project data stewards**: Project data stewards work on specific projects and report back to the operational data stewards when data issues arise or when new data must be governed. These stewards handle the data and provide feedback when there's a need to strengthen data governance processes.

5. **Technical data stewards**: Technical data stewards are information technology specialists who store the data in computer applications so that it can be easily handled by other data stewards. They work with the other data stewards to improve data governance systems in the SGC.

DATA GOVERNANCE OFFICE

A *data governance office* refers to a group of individuals responsible for measuring successes and gathering metrics.

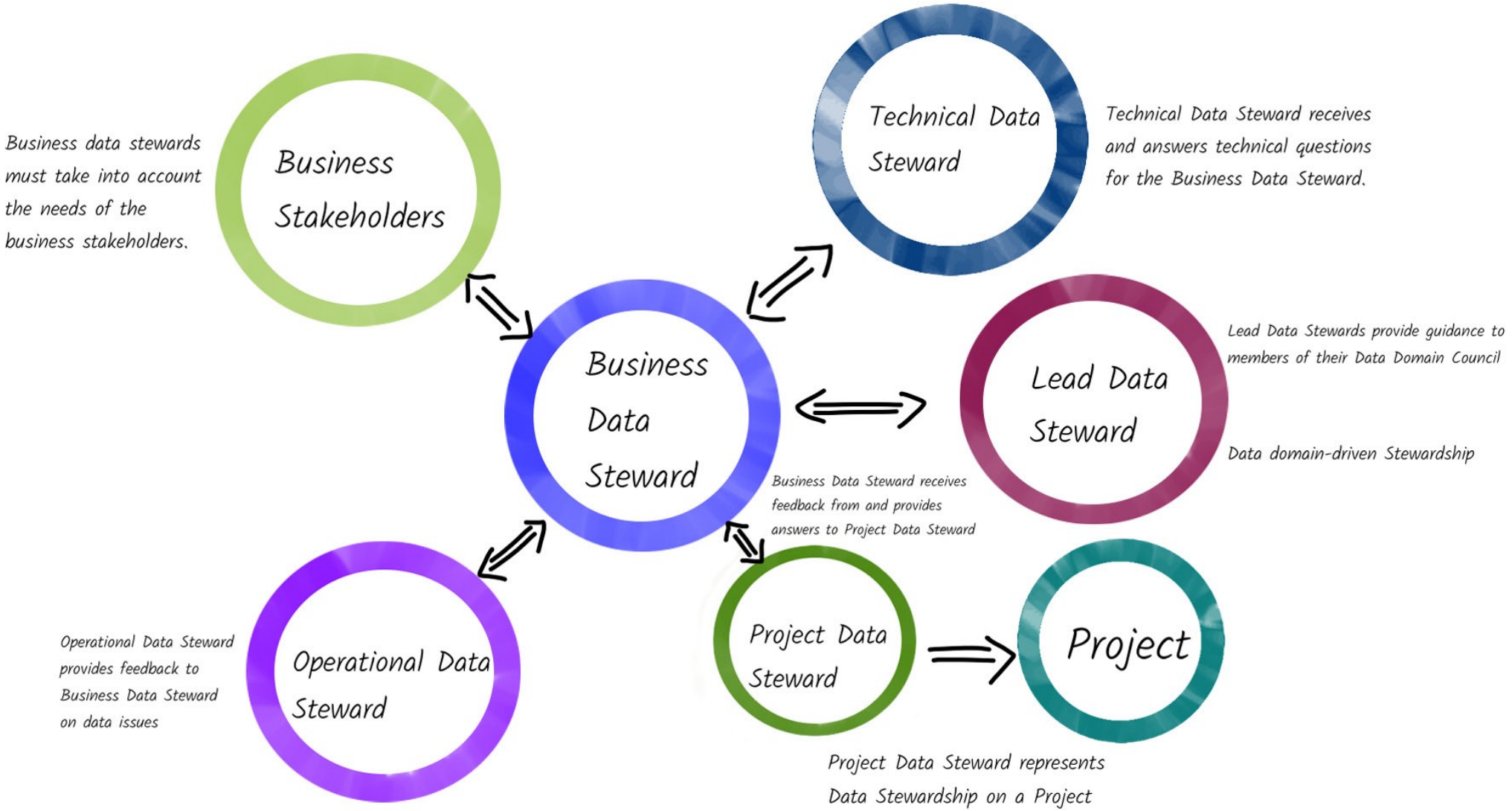


Figure 4
Types of data stewards
Source: Adapted from Plotkin (2021)

Implementing data governance principles

The data governance principles covered here include data security, accessibility, quality, integrity and reliability. SGC's should establish data governance principles that are relevant to their organisations. For example, an SGC may decide to focus on data quality and data reliability.

Data security

SGCs need to implement data protection strategies for data availability and data management. Data must be protected and secured so that it is available when users request it. One way to protect data is by using a backup system. Data can be stored in an offline storage facility (i.e., disk or tape). The advantage of on-site storage is that data is safe from cyber-attacks. However, there has been a shift from on-site storage systems as a result of digital transformation and improved IT systems. As a result, many organisations have moved to cloud backup systems as these offer additional advantages as compared to on-site storage facilities.

CLOUD STORAGE

Organisations are increasingly shifting to cloud storage facilities as they offer additional advantages. Examples of cloud storage include DropBox, iCloud, Google Drive, Microsoft One Drive, IDrive, Mega, Box, and pCloud.

Organisations need to implement a comprehensive data recovery and backup strategy. Snapshots and replications are the main methods used to recover data and make it possible to recover data much faster. In a case where a server fails, data from the backup array is used in place of the primary storage.

The misuse of information of private citizens and organisations has been a major concern for governments around the world. Regions and countries globally have adopted legislation to protect the misuse of personal data. For example, recently, South Africa introduced the Protection of Personal Information Act (POPIA). This legislation is based on these five core principles:

1. **Security**: Data protection measures must be taken by organisations to protect data from loss, theft, as well as unauthorized access.
2. **Disclosure**: Personal data may only be given to third parties if the owner of the data has given consent during the time when the data was collected.
3. **Notice and choice**: Organisations are mandated to inform individuals about the personal data that is being processed, the reason for processing, the third parties to whom the data user may disclose the personal data, and whether submitting the personal data is required or voluntary.
4. **Data integrity**: Organisations must take steps to ensure that personal data is accurate, comprehensive and not misleading.
5. **Access**: An individual should be allowed access to and be able to correct personal information at any given time.

Data storage and accessibility

International standards that pertain to data governance, such as ISO 2001, requires organisations that store data to create systems that define data storage rules, standards, protocols, and procedures for the storage and dissemination of data. Besides adhering to best practices, data backup is necessary since computers have limited storage. Backup also helps when a computer experiences a problem and data loss becomes a reality. There are two major types of data storage:[vi]

1. **Direct-attached storage (DAS)**: direct attached storage includes storage devices that connect directly to a computer including CD/DVD, flash drives, hard drives, magnetic tapes, and solid-state drives
2. **Network-based storage (NAS)**: network-based systems allow the storage of data from various devices and can be based on a network attached storage and storage-area network. Network-based storage systems are recommended for SGCs because data is centralised, easy to share, there is a single cost to data storage and better control mechanisms.

HIGH-LEVEL DATA STORAGE GUIDELINES

In the magazine, CIO Africa, Jeniffer Lonoff gives 14 high-level guidelines for data storage. This requires organisations to understand the value, amount, and type of data they intend to store before they decide where and how they will store it.[vii] These guidelines are:

(1) Know your data.
(2) Do not neglect unstructured data – combine multi-structured data from your transactional systems with semi-structured or unstructured data from your email servers, network file systems, etc.
(3) Understand your compliance needs.
(4) Establish a data retention policy.
(5) Look for a solution that fits your data, not the other way around.
(6) Do not let upfront costs dictate your decision.
(7) Use a tiered storage approach – save money by only using your fastest storage facility for data you actively use.
(8) Know your clouds – use clouds optimised to handle archiving and those for primary data storage systems, accordingly.
(9) Carefully vet storage providers.
(10) Do not store redundant data.
(11) Make sure your data is secure.
(12) Leverage technologies that use deduplication, snapshotting and cloning – these can save a fair amount of space.
(13) Make sure you can find the data you need once it has been stored.
(14) Have a data recovery plan – and constantly test it.

Data quality, integrity, and reliability: an organisational strategic asset

Data governance formalises not just behaviour associated with the definition, production, and use of data, but also its quality.[viii] Data governance ensures that data can be trusted with specific people responsible for data quality. These data stewards can be held accountable for any adverse events emanating from poor data quality. Governance also includes establishing who in the organisation holds the decision rights for determining standards for data quality.

Data can be an organisation's most strategic asset if it is trustworthy while untrustworthiness emanates from poor data quality. Data that is of poor quality can result in skewed analysis, incorrect insights, and reckless recommendations. Data should be assessed on a continuous basis to ensure that its quality is consistent throughout its lifecycle. The identification of data quality dimensions can be a starting point to build the basis for continuous improvement. The five common dimensions of data quality[ix] are illustrated in Figure 5.

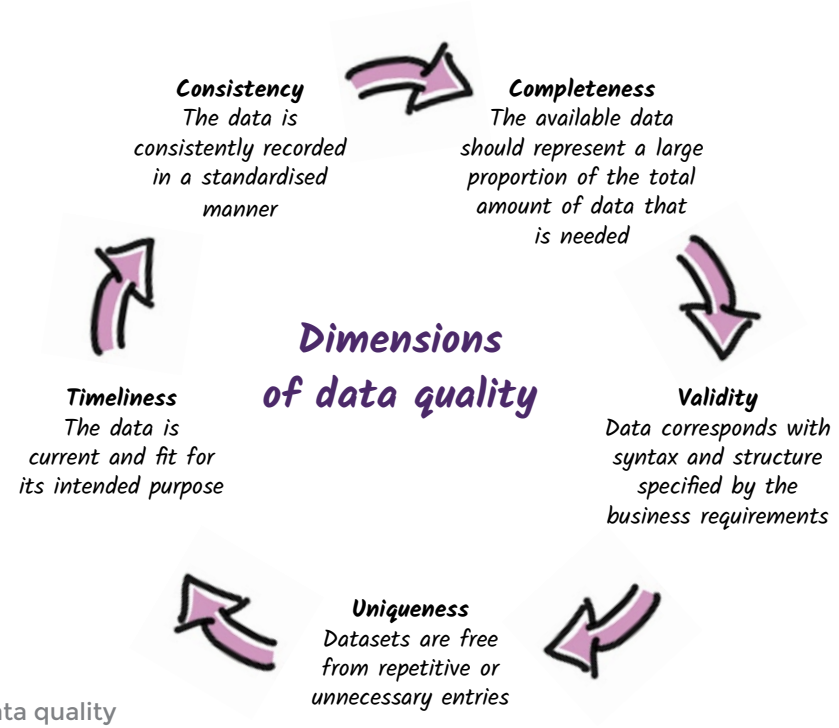


Figure 5
Dimensions of data quality

Source: Adapted from Cichy and Rass (2019)

Data integrity is another important component of data governance. It pertains to the consistency and accuracy of data stored in databases. Data integrity guarantees and secures the searchability and traceability of data to its original source. Organisations should apply data integrity constraints to maintain the quality of data. Organisations can also preserve the integrity of their data through the following processes:[x]

1. **Data cleaning and maintenance**: Maintenance and cleaning of data should be done on a regular basis. Data cleaning ensures that data is free of errors and is suitable for use. Data that is not cleaned may result in skewed findings and improper policy recommendations.

2. **Data validation rules**: Validation rules are systems that limit input errors by restricting the values that users can enter into a system.

3. **Data entry training**: Data users should be trained to enter and maintain data so that they take responsibility for data quality. For example, survey or data on grants if captured incorrectly can result in skewed reports.

Case study: data governance at the Centre for Science, Technology and Innovation Indicators (CeSTII)

This case study illustrates how CeSTII governs the use of survey data on behalf of South Africa's Department of Science and Innovation. Although CeSTII is not responsible for grants management, it collects and manages STI survey data that is made publicly available. CeSTII's data governance process is managed by internal stewards.

CeSTII conducts statistical surveys including the Business Innovation Survey (BIS) and the South African Research and Experimental Development (R&D) Survey. Statistical data collected by CeSTII is public data and is therefore available to the public on request. The sharing of innovation data is essential to avoid survey duplication by different organisations in South Africa. Within CeSTII, two main teams are responsible for coordinating the approval of data requests:

- The role of the **data committee** is to oversee data requests and engage with the data team and approve data requests from internal and external stakeholders. The committee consists of the chairperson, data team members, senior CeSTII staff members and survey project leaders (depending on the nature of the data request). A data request is received by the chairperson who then liaises with the internal or external stakeholder interested in the data. The chairperson communicates with senior CeSTII staff members responsible for managing the specific data requested.
- The **data team** consists of statisticians and its role is to carry out data extraction, data cleaning and data analysis.

How to access data at CeSTII

Stakeholders who wish to make use of CeSTII's data are requested to complete a data request form outlining the details of the request for the chairperson of the data committee.

If access to the requested data is granted, the stakeholder making the request must complete and sign a confidentiality agreement. Data access is granted if the request does not violate the confidentiality of firms in the survey data. This is because surveys conducted by CeSTII are undertaken in accordance with the prescripts of the South African Statistics Quality Assessment Framework (SASQAF) which requires that data sharing be governed by data access protocols, including confidentiality.

Figure 6 shows the procedure that is followed when internal and external stakeholders request innovation data at CeSTII. Stakeholders are requested to enter into a data sharing agreement with CeSTII if the shared data will be curated in online platforms or data visualisation portals. When data is shared through a data sharing agreement, the cost of preparing the data are governed by the conditions of the agreement. This ensures that the cost of preparing the data request does not exceed the budget.

TIP:
If the aggregation level of data requested by a stakeholder compromises the confidentiality of the respondent, it may be possible to provide data at a higher aggregation level. If data cannot be shared with a stakeholder, the data committee should provide a reason for the refusal. The process may take two to six weeks from the receipt of the request to granting access to the data.

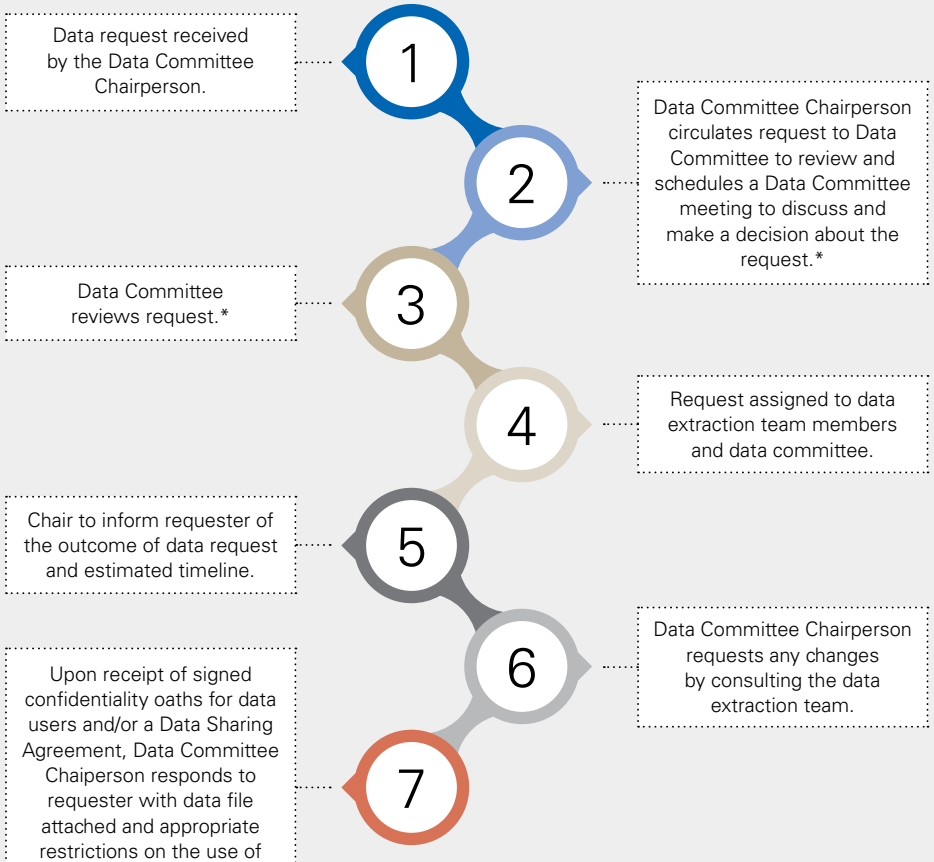


Figure 6
Data request process at CeSTII

TOOLS FOR ASSESSING AND **IMPLEMENTING DATA GOVERNANCE**

Tool 1: Data Governance Maturity Assessment Framework

The data governance maturity assessment framework is a tool to evaluate an organisation's data governance practices and can be used to identify gaps in data governance.

The data governance maturity assessment framework identifies five levels of data governance maturity in a science granting council, as shown in Table 2.[xi]

Table 2 Data Governance Maturity Assessment levels

Maturity level	Description
Level 1	Initial (localised and ad hoc)
Level 2	Reactive (defined but not complete)
Level 3	Managed (no formal process in place)
Level 4	Proactive (defined across enterprise, replicable, metrics)
Level 5	Transformational (implemented, monitored, used proactively across council)

Initial (localised and ad hoc), level one, describes a situation in which there is no organisation-wide oversight or awareness of the need for data governance. Business units govern data separately or in silos, with no clear ownership or accountability structure for data governance. At this level, there are no formal policies, procedures, or guidance around data governance, management, sharing, or other procedures, and trust in data quality is low.

Reactive (defined but not complete), level two, refers to an organisation in which some oversight and governance policies are in place in individual units and divisions, but they are not well-known or consistent across the organisation. Data governance is not yet fully defined, but there is some documentation defining the structure of data governance. Some ad hoc standards are in place for common data, and there is ad hoc sharing of data through CSV exports. No effort is made to maintain or improve data quality beyond the immediate business need.

Managed (no formal process in place), level three, indicates that the organisation has defined data governance at the enterprise level, with some awareness of the benefits of data governance. Some procedures are in place to evaluate gaps and requirements across divisions, and some defining documentation exists. The data stewards and other roles are identified. The team meets regularly, and a data governance body with membership across the council (business and IT) meets and receives reports from the lead data steward and others as required. At this stage, data classification has been defined and articulated in council policy, procedures, and standards.

Proactive (defined across the organisation, replicable, metrics are in place), level four, the organisation has established organisation-wide awareness and a central definition of data governance, with an established data governance body or authority. The organisation has defined responsibilities and ownership of data governance, and the authority structure is documented and clearly understood across the council. Data management policies are comprehensive, and there is a defined and documented process for data requests. Training is in place for new and current employees in how to handle confidential data, council criteria for sharing data, and the roles and responsibilities of all actors in the process.

Transformational (data governance is implemented, monitored, and used proactively across council), level five, refers to a council with consistent rules across all divisions and buy-in and support from council leadership. Communication originates from the top of the organisation, with a clear and documented authority structure in place with escalation points and regular engagement from executive leadership. The stewardship structure is active, engaged and functions smoothly, and stewards have duties incorporated into performance review and mechanisms for feedback, with backups trained and knowledge management processes in place. The stewardship model is widely understood and used throughout the council, individuals know where to go to ask questions and support the operational elements of data governance.

TIP:

Use the **Data Governance Maturity Assessment template** to assess an SGC's level of data governance maturity. On page 38, scan the QR code or follow the link to access the Data Governance Maturity Assessment tool in Excel. The tool asks five questions based on an SGC's data governance context to provide an assessment of maturity.

Categories	Assessment questions	Tick the most appropriate response that applies to the SGC [scores in brackets]				
		Level 1: Initial	Level 2: Reactive	Level 3: Managed	Level 4: Proactive	Level 5: Transformational
Data Leadership/ Executive Support	To what extent is data governance leadership and executive integrated within your SGC?	We have a single leader who is responsible for data governance initiatives in our SGC **[1]**	Data Governance policies have been defined and communicated to relevant stakeholders **[2]**	We have formal data governance processes or procedures in place **[3]**	We have data governance metrics being tracked and reported to executive leadership **[4]**	Data governance is integrated into all relevant business activities and decision-making processes **[5]**
Data Stewardship	To what extent are data stewards defined and integrated into data management and governance in your SGC?	We have data stewards assigned on an ad- hoc basis for specific projects or initiatives in our SGC **[1]**	We have data stewards responsible for defining data quality standards and rules in our SGC **[2]**	We have a formal process for identifying and managing data stewards across our SGC **[3]**	Data stewards are actively involved in identifying and resolving data quality issues in our SGC **[4]**	Data stewards are involved in all aspects of data management and governance, including policy development and decision-making **[5]**
Data Processes, Policies and Procedures	What is the status of data processes, policies, and procedures within your SGC?	We have documented data processes, policies, and procedures in place in our SGC **[1]**	Our data processes, policies, and procedures are defined and documented but are currently incomplete **[2]**	Our data processes, policies, and procedures are managed in an ad hoc manner with no formal process in place **[3]**	Our data processes, policies, and procedures are currently defined across the enterprise and repeatable with metrics in place to monitor their effectiveness **[4]**	Our data processes, policies, and procedures are implemented, monitored, and used proactively across our SGC **[5]**
Data Management	What is the current state of data management processes within your SGC?	Our data management is handled on an ad hoc, localized basis with no formal process in place **[1]**	Our data management processes have been defined but are not complete **[2]**	Our data management processes are managed in an ad hoc manner with no formal process in place **[3]**	Our data management processes defined across the enterprise and repeatable with metrics in place to monitor their effectiveness **[4]**	Our data management processes are well implemented, monitored, and used proactively across the council **[5]**
Value Creation	What is the level of formality and standardization in the value creation processes across your SGC?	Our value creation processes are currently in an ad hoc and localized with no formal process in place **[1]**	Our value creation processes have been defined but are not complete **[2]**	Our value creation processes are managed in an ad hoc manner with no formal process in place **[3]**	We currently have metrics in place to monitor their effectiveness **[4]**	Our value creation processes are well implemented, monitored, and used proactively across the council **[5]**
Privacy, Security, Regulatory Control and Risk	Does your SGC have formal processes to handle privacy, security, regulatory control, and risk management with defined metrics for monitoring effectiveness?	Our privacy, security, regulatory control, and risk management handled in an ad hoc, localized basis with no formal process in place **[1]**	Our processes for privacy, security, regulatory control, and risk management have been defined but are not complete **[2]**	Our privacy, security, regulatory control, and risk management processes are currently managed in an ad hoc manner with no formal process in place **[3]**	Our privacy, security, regulatory control, and risk management processes are well defined across the enterprise and repeatable with metrics in place to monitor their effectiveness **[4]**	Our privacy, security, regulatory control, and risk management processes are well implemented, monitored, and used proactively across the council **[5]**

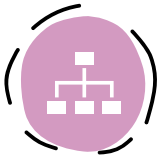
Overall Maturity Level
[Average score: add category scores and divide by 5]

2

Tool 2: The RACI Matrix

*RACI stands for **R**esponsible, **A**ccountable, **C**onsulted, and **I**nformed. It is a method for identifying, assigning and tracking roles and responsibilities for specific tasks, activities, and projects.*

For SGCs, the responsibility assignment matrix, RACI, can be used to identify the roles of the different data stewards and their level of engagement in data governance activities.[xii]



Responsible:
the project data steward performs the work and is responsible for fulfilling the activity until it is finished and approved by the operational or business data steward

Accountable:
the operational data steward has the authority to decide on a problem or approve the answer to a decision taken and accounts for all decisions in the project

Consulted:
this can either be the operational or business data steward who provides advice and opinions that inform the decisions taken

Informed:
these are data stewards who are notified after a decision has been made

The RACI matrix can be used as a tool to include data managers in decision-making about the assignment of roles and responsibilities. It can be shared with stakeholders and the data office. It is useful for assessing work packages and identifying missing roles, providing an opportunity for early detection and correction of mistakes.

TIP:
Use the **RACI matrix for data governance implementation** to assign roles and responsibilities for data governance stewards. This template can be adapted to the structure of the SGC.

Template 2 RACI matrix for data governance implementation

Data governance role	Business data steward	Technical data steward	Data quality analyst	Technical data stewards	Data modeller	Executive steering committee
Metadata	R					
Data lineage	R					
Data retention and purge	R					
Data access, security, and privacy	R					
Data classification	R					
Data sharing	R					
Data integration	C					
Data model	C					
Reporting and business intelligence	I					
Business glossary	R					
Data policies, standards, and guidelines	R/I					
Data governance adherence	I/C					

Source: Adapted from Khan and Quraishi (2014)

Conclusion

The aim of this toolkit is to offer a set of guidelines and tools for assessing and implementing data governance in science granting councils. Sound data governance helps to ensure that data is trustworthy, consistent, and properly used by stakeholders within the council. It sets responsibilities so that data governance stakeholders know their roles. Science granting councils can use the maturity assessment tool to measure progress in their data governance journey while the RACI tool can be used to assign and manage roles and responsibilities for data stewards. Data governance should be a shared responsibility of all constituents of the council to facilitate best practice across the organisation.

ADDITIONAL **RESOURCES**

References and further reading

Ahmad, S., Kumar, A. and Hafeez, A. 2019. The importance of data integrity and its regulation in pharmaceutical industry. The Pharma Innovation Journal 8(1): 306-313.

Al-Badi, A., Tarhini, A. and Khan, A.I. 2018. Exploring Big Data Governance Frameworks. Procedia Computer Science 141: 271-277.

Cichy, C. and Rass, S. 2019. An overview of data quality frameworks. IEEE Access 7:24634–24648.

Earley, S., Henderson, D., and Data Management Association. 2017. Dama-Dmbok: Data Management Body of Knowledge (Second). Bradley Beach, NJ: Technics Publications.

GCIS. 2021. Draft National Policy on Data and the Cloud. Pretoria: Department of Communications and Digital Technologies (Available at https://www.google.com/search?q=GCIS.+2021.+Draft+National+Policy+on+Data+and+the+Cloud.+Pretoria%3A+Department+of+Communications+and+Digital+Technologies.&rlz=1C1GCEU_enZA987ZA987&oq=GCIS.+2021.+Draft+National+Policy+on+Data+and+the+Cloud.+Pretoria%3A+Department+of+Communications+and+Digital+Technologies.&aqs=chrome..69i57.1760j0j9&sourceid=chrome&ie=UTF-8)

Geng, L. 2015. The research of digital library mass information storage system architecture. Advances in computer science research. Proceedings of the 2015 international symposium on computers and informatics.

Gupta, U. and Cannon, S. 2020. Data Governance Frameworks: A Practitioner's Guide to Data Governance, Emerald Publishing Limited, Bingley, pp. 101-122. https://www.cio.com/article/2382585/14-things-you-need-to-know-about-data-storage-management.html (Accessed on 20 June 2022).

Khan, P.M. and Quraishi, K.A. 2014. Impact of RACI on delivery and outcome of software development projects. Fourth international conference on advanced computing and communication technologies held from 8-9 February 2014 in Rohtak, India.

Khatri, V. and Brown, C.V. 2010. Designing data governance. Communications of the ACM 53(1):148–152.

Koltay, T. 2016. Data governance, data literacy and the management of data quality. International Federation of Library Associations and Institutions 42(4): 303-312.

Micheli, M., Ponti, M., Craglia, M. and Suman, A. B. 2020. Emerging models of data governance in the age of datafication. Big Data and Society, July–December: 1–15.

Pierce, E. 2022. A balanced scorecard for maximizing data performance. Front Big Data 5: 821103.

Schiff, J. L. 2013. 14 Things You Need to Know About Data Storage Management. Retrieved June 4, 2021, from CIO Africa.

Seiner, R.S. 2014. Non-Invasive Data Governance: The Path of Least Resistance and Greatest Success. Basking Ridge, NJ: Technics Publications.

State of Oregon - Enterprise Information Services. 2021. Oregon's Data Strategy Unlocking Oregon's Potential 20212023. Retrieved from: https://www.oregon.gov/das/OSCIO/Documents/68230_DAS_EIS_DataStrategy_2021_v2.pdf (accessed 12 June 2023)

Thomas, G. 2006. The DGI data governance framework. The Data Governance Institute. Available at https://datagovernance.com/wp-content/uploads/2020/07/dgi_data_governance_framework.pdf (Accessed 7 June 2023).

Some of the content in this toolkit was adapted from the many resources available online on the website of the Data Governance Institute at https://datagovernance.com/

Table 3 Data Governance Maturity Model

Categories	1. Initial (Localised and Ad hoc)	2. Reactive (Defined but not complete)	3. Managed (No formal process in place)	4. Proactive (Defined across council, repeatable, metrics)	5. Transformational (Implemented, monitored, used proactively across council)
Data leadership/ executive support	No council-wide oversight or awareness of the need for data governance. No clear ownership or accountability structure for data governance within the council. Business units govern data separately or in siloes.	Some oversight and governance policies in place in individual council units and divisions. Not well known or consistent across division lines. Data governance is known but not fully defined within the council.	Council has defined data governance at the enterprise level. Awareness of benefits of data governance at enterprise level, some efforts in place to evaluate gaps and requirements across divisions. Some defining documentation.	Council has established enterprise-wide awareness and central definition of data governance with established data governance body or authority. Positions in the council with defined responsibilities/ ownership of data governance. Authority structure is documented and clearly understood across the council.	Consistent rules across all divisions, with buy in and support from council leadership.
Data stewardship	No defined stewardship model/ little business involvement so there is a lack of understanding of its value.	Lead data steward appointed. Stewardship concepts and value are defined and understood within units, but do not connect to a larger stewardship structure within the council.	Council stewards and other roles identified and meeting regularly.	Stewardship model in place, and stewards/ other roles meet regularly to set standards and policies. Data governance body has set vision and priorities for stewardship and is building metrics to measure success of data governance programme. Stewards' duties (and other roles) are well documented, with onboarding and training provided to new stewards.	Communications originate from top of organisation, with processes in place to facilitate council wide sharing, agreements, and communications.
Data processes, policies and procedures	No formal policies, procedures, or guidance has been established around data governance, management, sharing, or other procedures. No centralised guidance for data sharing or council data sharing procedure is not well understood.	Ad hoc or business unit level policies and procedures. Some individuals within council business units understand data sharing protocols or privacy considerations. Data sharing templates are similar, but not broadly agreed on.	Chartered Data Governance Body with membership across the council (business and IT) is meeting and receiving reports from lead data steward and other roles as required.	Some data management guidelines and requirements are in place and have been adopted, with multiple business units already implementing or starting to implement. Consistent training, outreach, or other communication methods are in place to educate employees on data management guidelines and procedures.	Clear and documented authority structure is in place with escalation points and regular engagement from executive leadership.

Continues overleaf...

Categories	1. Initial (Localised and Ad hoc)	2. Reactive (Defined but not complete)	3. Managed (No formal process in place)	4. Proactive (Defined across council, repeatable, metrics)	5. Transformational (Implemented, monitored, used proactively across council)
Data management	Documentation missing for data models; data is in silos with duplicate and near duplicate data; no create, read, update and delete (CRUD) controls; no retention policy; data cannot be extracted in a manipulatable form (e.g., printed reports); trust in data quality is low.	Ad hoc standards in place for common data; ad hoc sharing of data through csv exports; some awareness of duplicate data in multiple datasets; minimum documentation of critical datasets (ER diagram); quality audits may be performed only on datasets in current use; effort not made to maintain or improve data quality beyond the immediate business need; datasets backed up in some cases.	Initial council-wide data governance scope has been established with an existing Data Governance Body. Data classification has been defined and articulated in council policy, procedures and standards. Limited data management policies have been authored for use or are authored but not yet implemented. Effective data sharing occurs in some locations across council. Legal requirements for data sharing or management may be interpreted differently in different units or subject to different interpretations amongst business areas.	Data capture standards are adopted and in use across departments; data capture standards are integrated in the data creation/ intake process; master data in use; APIs are specified where possible for accessing data in source systems; quality audits are automated; backups occur on all datasets; recovery plans in place for all datasets; audits in place to evaluate CRUD access; and data retention policy governs most datasets.	Stewardship structure is active and engages and functions smoothly.
Value creation	No central inventory of data exists, metadata capture is left to individual discretion (if at all). Council has not sought out internal or external datasets for analytic value and has not identified “high value” data assets for the council. Classification and identification of systems of record/ single source of truth is minimal or inconsistent.	Inconsistent inventories/ awareness of data, metadata procedures, identification of business value of data beyond specific regulatory purposes. Some individuals may be combining datasets for use, but these uses are ungoverned and do not have feedback from individuals with authority and responsibility for managing data or engagement from data stewards.	Data capture standards are documented and available for use; most data models are documented at the team or department level; efforts are underway to combine and standardise duplicate data in critical datasets (master data); ad hoc shared data repositories exist; quality standards in place for high value datasets; CRUD controls in place for high value datasets; datasets backed up at a regular cadence; recovery plan drafts exist; data retention policy exists.	Metadata repositories maintained at unit/ division levels. Data governance committee structure is responsible for measuring value of data. Individuals know where to find the data they need at the right time and do not need to “reinvent the wheel”. Metadata, data dictionaries and process for publishing data to repository are defined across council. Data catalogues for operational, reference, and analytic data are complete. Periodic extract, transform and load processes are in place from vendor systems to shared repositories.	Stewards have duties incorporated into performance review and mechanisms for feedback, with back-ups trained and knowledge management/ sharing.

Continues overleaf...

Categories	1. Initial (Localised and Ad hoc)	2. Reactive (Defined but not complete)	3. Managed (No formal process in place)	4. Proactive (Defined across council, repeatable, metrics)	5. Transformational (Implemented, monitored, used proactively across council)
Privacy, security, regulatory control and risk	Individuals are aware of regulatory requirements or overarching policy, but implementation is inconsistent. No risk management documentation, policies, procedures in place, and lack of understanding of how best to achieve them. If a risk event occurs, there are no procedures in place. Little to no leadership communication.	Regulatory requirements are documented within specific business units but not shared with the larger data governance programme or data stewardship body. Stewards implement and assess regulatory requirements individually with minimal centralisation. Risk management, if executed, is executed reactively. Senior management lacks appreciation for what risk management would achieve for them and does not require these activities be performed.	Collection of and repositories built of associated metadata, to enhance understanding of business value.	Clear, auditable and measurable procedures are in place for all confidential or sensitive data. Council has identified privacy considerations and concerns when engaging in data sharing, integrated data, or analytics projects, and documents data limitations or quality issues that impact privacy and decision-making. Proactive approach to reviewing risks periodically and eliminating risk via management decisions to accept, avoid, or mitigate risk. Risk management practiced with a plan of action at departmental level. Documentation, policies, procedures, training, and communication on risk management is well defined. Subject matter experts are utilised to train operational personnel.	Stewardship model is widely understood and used throughout the council; individuals know where to go to ask questions and support the operational elements of data governance.

Source: State of Oregon - Enterprise Information Services (2021)

Link to editable templates and tools that may be customised to purpose

To access the digital version of this toolkit, Excel templates and other Evi-Pol outputs and resources, scan the QR code below.

Alternatively, you can access these tools by visiting the CeSTII webpages at **https://hsrc.ac.za/divisions/centre-for-science-technology-and-innovation-indicators/** , or the HSRC's Research Output Repository at **https://repository.hsrc.ac.za**.

i Koltay (2016)

ii Earley, S., Henderson, D., and Data Management Association (2017)

iii Micheli et al. (2020)

iv This section on data governance processes was largely drawn from Khatri and Brown (2010).

v Plotkin (2021)

vi Geng (2015)

vii Schiff (2013)

viii Seiner (2014)

ix Cichy and Rass (2019)

x Ahmad et al. (2019)

xi Based on the State of Oregon Data Governance Maturity Assessment model, retrieved from https://www.oregon.gov/das/OSCIO/Pages/Index.aspx (accessed 12 June 2023).

xii Based on the RACI responsibility matrix developed by Khan and Quraishi (2014)

SCIENCE GRANTING COUNCILS INITIATIVE SGCI

INITIATIVE DES ORGANISMES SUBVENTIONNAIRES DE LA RECHERCHE SCIENTIFIQUE IOSRS

Evi-Pol

Enabling more effective use of evidence in policy and decision making by science granting councils

African Centre for Technology Studies
Knowledge for Better Livelihoods

HSRC Human Sciences Research Council

CeSTII Centre for Science, Technology & Innovation Indicators