THE UNITED REPUBLIC OF TANZANIA

PRESIDENT'S OFFICE,
PUBLIC SERVICE MANAGEMENT

e-GOVERNMENT AGENCY

e-GOVERNMENT INFORMATION ARCHITECTURE – STANDARDS AND TECHNICAL GUIDELINES

Document Number eGA/EXT/IFA/001

Issued by eGovernment Agency - November 2017

THE UNITED REPUBLIC OF TANZANIA

PRESIDENT'S OFFICE,
PUBLIC SERVICE MANAGEMENT

e-GOVERNMENT AGENCY

Document Title

eGovernment Information Architecture – Standards and Technical Guidelines

Document Number eGA/EXT/IFA/001

APPROVAL	Name	Job Title/ Role	Signature	Date
Approved by	Dr. Jabiri Bakari	Chief Executive Officer		

Table of Contents

1.0 OVERVIEW

1.1. Introduction

The e-Government Agency (eGA) is established under the Executive Agencies Act No.30, 1997, Cap. 245 as a semi-autonomous Institution under President's Office Public Service Management. eGA is charged with the mandate of providing coordination, oversight and provision of e-Government initiatives and enforcement of e-Government standards to Public Institutions. In executing its duties, eGA shall implement and maintain coordinated government operations for Information and Communication Technology (ICT) that include the formulation of standards and guidelines to effectuate the purposes of the Agency.

To realize the vision of e-Government in Tanzania and successfully implement eGovernment Strategy, it is of paramount importance that **"e-Government Standards and Guidelines"** are formulated. The e-Government Standards and Guidelines' aim is to assist in the delivery of more consistent and cohesive services to citizen and support the more cost effective delivery of ICT services by Government. A worldwide agreeable practice for conducting Government wide eGovernment analysis, design, planning and implementation, using a holistic approach at all times, for the successful development and execution of eGovernment Strategy is known as **"eGovernment Enterprise Architecture".** The e-Government Standards and Guidelines Structure is hereby designed to cover most requirements of eGovernment Enterprise Architecture. This means that eGovernment Enterprise Architecture is incorporated in "eGovernment Standards & Guidelines".

Management of e-Government Standards and Guidelines requires categorisation. There are **nine** categories/areas covering all aspects of eGovernment. The **fifth** area is **eGovernment Information** Architecture. Information Architecture defines the structure of the Government's information assets and address data management considerations. It reflects the domain entities, their relationships and establishes accountability for data integrity. The eGovernment Information Architecture Standards and Technical Guidelines document has been derived from the *e-Government Enterprise Architecture as referred in e-Government Architecture Vision - Standards and Technical Guidelines (eGA/EXT/AVS/001).*

1.2. Rationale

The Information Architecture defines a common set of practices to access data which leads to efficiency and effectiveness in decision making. This enables timely response to information requests and service delivery by Public Institutions.

1.3. Purpose

In line with the above rationale, eGovernment Information Architecture enables easier, more efficient exchanging and processing of information. It also removes ambiguities and inconsistencies in the use of data across Public Institutions. The standards and guidelines will apply to all systems within the Government. This will ensure that the defined eGovernment Information Architecture Standards and Technical Guidelines are adopted across the Public Institutions.

1.4. Scope

This document applies to all Public Institutions. The Public Institution Accounting Officers (Heads of Institutions), Head of ICT Departments, Business Process Owners, Database Architects, Report Designers, Application Architects, Application Developers, Business Analysts and System Engineers shall be responsible for ensuring the effective implementation of these specific standards and technical guidelines associated with Information Architecture within their respective Institutions.

2.0 e-GOVERNMENT INTEGRATION ARCHITECTURE

2.1. e-Government Integration Architecture Reference Framework

The eGovernment Information Architecture is designed adhering to the recommended overarching Information Architecture Principles (e.g., data creation and accessibility, data availability, data security, confidentiality, integrity, data ownership, standard common data definitions etc.) and Data Reference Model (DRM).

The Data Reference Model (DRM) provides a structure that facilitates the development of Government data that can be effectively shared across Public Institutions for better and more effective service delivery, improved decision making and improved mission performance. The DRM is a service-oriented model that provides the pathway for "Services to Citizens" to become operational. At the same time, the DRM provides an impetus for Public Institutions to better understand their data, how it fits in the total realm of Government information.

Deriving from the DRM is the Information Architecture Reference Framework as illustrated in Figure I. The Information Architecture Reference Framework has been defined to depict the decomposition of information/data architecture framework components that is essential to manage the data assets of the Government.

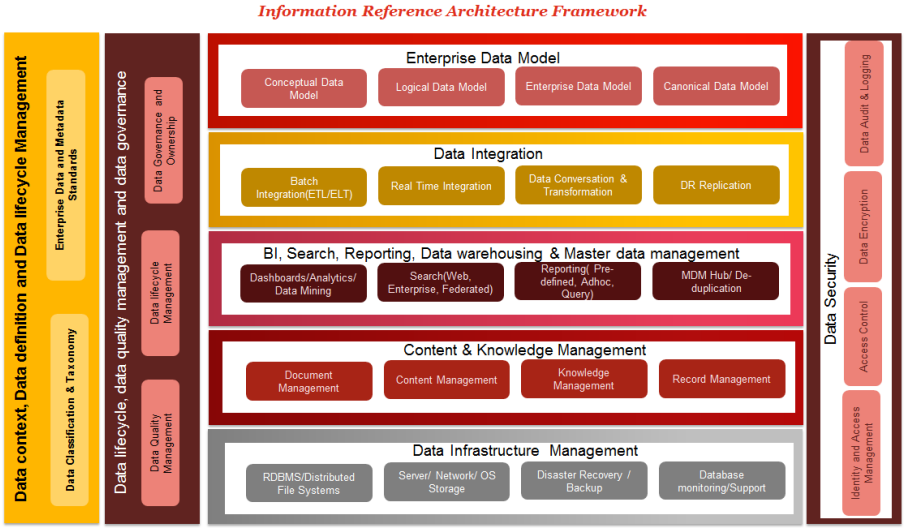

Figure I: Information Reference Architecture Framework

The diagram in Figure I illustrates the Information reference architecture framework with the components in Table I:

Table I: Components of Information reference architecture framework

Components	Description
Data context and data definition	This component deals with defining the context of the data by classifying data as per subject area and defining the enterprise data and metadata standards to ensure seamless interoperability by removing ambiguities and inconsistencies in the use of data across Public Institutions.

Enterprise data model Data life cycle, quality management and data governance	This component deals with data analysis and design of the underlying data structure. Data life cycle management deals with the management of structured data assets across the data life cycle, from creation and acquisition through archival and purge. Data quality deals with defining, monitoring and improving data quality. Data governance deals with planning, oversight, and control over data management and use of data
Data sharing and integration	This component deals with managing data transformation and data exchange across applications and data store, defining capabilities from batch-based to real-time integration (including extract, transform and load (ETL) and extract, load and transform (ELT)), event driven, message-driven, and real-time integration.
Data security	This component deals with managing data privacy, confidentiality and to prevent unauthorized data access, creation and change.

In summary the Information Architecture together with the reference framework provide a structured and comprehensive approach to information management which enables the effective use of data amongst the Public Institutions. Information Architecture can be defined based on DRM standards as follows;

i. Categorising information and establishing a classification of data with respect to how it supports the business. This will allow Public Institutions to share data in common formats with common definitions, resulting in consistent application of data across government; and

ii. Defining requirements for establishing standard information exchange formats that can be shared across Public Institutions. This will enable the collaborative development of information exchange, defining policies for development, storage and use throughout government.

2.1.1. Data, Metadata and Data Modelling

2.1.1.1. Data Standards

In the Government, data is standardized under three broad areas which are;

i. **Data Description** - The data description standardization area provides a means to uniformly capture the semantic and syntactic structure of data. It focuses on understanding data at two levels of abstraction:
   a) Metadata artefacts required to understand the data
   b) Aggregation of metadata artefacts to define a managed data asset.

ii. **Data Context** - The Data Context standardization area establishes an approach to the categorization/ classification of data assets using subject areas/ taxonomies and other descriptive information. It focuses on 2 management mechanisms to capture the context of a data:
   a) Taxonomies
   b) Data asset descriptions

iii. **Data Sharing** - The Data Sharing standardization area describes the access and exchange of common data sets. It covers two primary aspects of data sharing:
   a) Data Exchange: Fixed, recurring transactions between parties, such as the regular exchange of data among Public Institutions. These exchanges are implemented with data exchange services;
   b) Data Access; Requests for data services, such as the query a Data Asset. These requests are supported by Data Access Services. (Refer to Appendix – Illustration No. 1 for DRM)

Enterprise data standard catalogue needs to be defined and adopted across Public Institutions to ensure seamless interoperability, removal of ambiguities and inconsistencies in the use of data across Public Institutions and to enable easier and more efficient exchanging and processing of data.

Common and generic data entities used across Public Institutions (e.g. Name, Organization, Address, Email, Identifier etc.) needs to be identified and consistent data standards for these entities defined at the enterprise level and adopted across Public Institutions.

While defining common data standards, the Government will leverage on international universal formats and standards (such ISO, W3C, OASIS etc.) that may already exist to achieve interoperability between different systems, processes and platforms and eliminate data transformation required to convert common data from one proprietary format to another. Some examples of common international data standards are:

i. ISO 3166 Country Codes

ii. ISO Currency Codes

iii. ISO 8601 data elements and interchange format for representation of dates and times etc.

iv. OASIS Customer Information Quality (CIQ) Specifications v3.0 that defines a universal common format to describe the customer/citizen name, address, unique identifier and other customer attributes.

Refer to the *e-Government Interoperability Framework - Standards and Technical Guidelines (eGA/EXT/GIF/001)* for additional details.

Standard naming convention - A full business name will be assigned to each data standard using the following format: **OBJECT - QUALIFIER(S) - DESIGNATOR** where:

i. **OBJECT** is a keyword that describes the main object/entity/concept to which the standard relates, e.g. Person, Department, Public Institution, etc. This subject word is omitted where the name of the standard is otherwise sufficiently clear as to render a prefix superfluous;

ii. **QUALIFIER(S)** is a qualifying word(s) used to describe the standard uniquely. Each word used is itself meaningful to the standard being described. The order of the words is in decreasing order of importance from left to right

iii. **DESIGNATOR** is a keyword that designates the class or category of data to which the standard belongs, e.g. name, number, code etc. This designator word is at the end of names that are made up of several words.

iv. Standard designator used:
    a. Name -Alphanumeric data by which a data element is known. e.g. Person First Name, Person Last Name, Organization Name,
    b. Number - Alphanumeric data that is used for identification purposes, e.g. a Business Registration Number, National ID, Tax Identification Number (TIN)
    c. Description / Information - Alphanumeric data that is used to describe a specific data element e.g. Business Description
    d. Type - Data that categorizes a data element e.g. Organization Type
    e. Code - Data, maintained by the user, which readily identifies an occurrence of a data element e.g. Country Code, Currency Code
    f. Any of the standard Data Types e.g. Date, Time may be used as standard designators e.g. Person Birth Date, Business Registration Date

These standards needs to be be applied by all Public Institutions ICT systems and are for use in all other public sector interfaces. Compliance with these standards follows the e-GIF compliance policies.

2.1.2. Metadata Standards

Metadata means "data about data". Metadata provides a context for data assets in the form of core standardized and structured resource description that explains the origin, purpose, time reference, geographic location, creator, access conditions and terms of use of a resource. Metadata are typically used for resource discovery, providing searchable information that helps users to easily find existing data.

Public institution metadata standards will be primarily based on the international Dublin Core model (ISO 15836). The model provides standard for metadata and metadata element description and covers:

a. Dublin Core (DCMI Metadata Terms) standard based on ISO 15836 to be used for metadata description of websites, digital documents and objects.
b. Dublin Core Metadata Element Set - A simple and extensible metadata element set intended to facilitate discovery of electronic resources.

XML based metadata schema will be defined and standardized across Public Institutions. Every data item important for data exchange across the enterprise should have metadata. The use of standardized records in XML format brings key resource description together into a single document, creating rich and structured content about the data. Preference will be given to a centralized server-based source of metadata.

Metadata will be viewed with web browsers, used for extract and analysis engines and can enable field-specific searching. Metadata shall be harvested by Public Institutions for data sharing through the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) that supports access to web accessible material through interoperable repositories for metadata sharing, publishing and archiving.

There are two basic types of metadata recommended:

i. Logical Data Models to describe the Structured Data Resources

ii. Digital Data Resource metadata (such as Dublin Metadata Standard Elements) to describe Semi-Structured and Unstructured Data Resources.

The implementation of the Data Schema concept group would take the form of Entity-Relationship Diagrams, Class Diagrams, amongst others. Implementation of the Digital Data Resource could be records in a content management system or metadata catalogue. The context of the data will be captured by the following concepts in Table IV:

Table IV: Concepts for capturing context of the data

Components	Description
Taxonomy	Taxonomy provides a means for categorizing or classifying information within a reasonably well-defined associative hierarchical structure that can be used to describe the data entities. E.g. categorizing data based on subject area like "Transport Management Service", "Tax Administration Services", "Land Reforms and Management Services", "Civil Registry Services" etc. Taxonomy can be further categorized to Topic.
Topic	Topic is a category within Taxonomy. A Topic is the central concept for applying context to data. A Topic categorises a data asset, i.e. a sub-categorisation of data based on logical data components within a subject area. For example, "VAT Registration" is a topic within "Tax Administration Service" subject area.
Data Asset	A Managed container for data. In many cases, this will be a relational database; however, a Data Asset may also be a website, a document repository, directory or data service, etc.

Implementation of taxonomies will take the form of eXtensible Markup Language (XML) Topic Maps, Web Ontology Language (OWL) hierarchies or ISO 11179 Classification Schemes. Implementation of a Data Asset inventory could be records in metadata registry. The Data Sharing abstract model will be represented by concepts in Table V:

Table V: Concepts of representing Data Sharing abstract model

Components	Description
Exchange Package	A description of a specific recurring data exchange between a Service Provider and a Service Consumer. An Exchange Package contains information, (metadata) relating to the exchange (such as Service Producer ID, Service Consumer ID, Validity period for data, etc.) as well as the reference to the message content (Payload) for the exchange. For example description of the data exchange between Service Producer and Service Consumer department to get the information.

Payload	An electronic definition that defines the requirements for the Payload (data) that is exchanged between a Service Provider and Service Consumer, e.g. a specific message set expressed as an XML schema containing information about an entity.
Service Producer	An entity (person or organization) that supplies data to a consumer
Service Consumer	An entity (person or organization) that consumes data is supplied by a Supplier
Query Point	An endpoint that provides an interface for accessing and querying a Data Asset. A concrete representation of a Query Point may be specific URL at which a query Web Service may be invoked.

The Data Sharing standardization area is supported by the Data Description and Data Context standardization areas in the following ways:

i. Data Description: Uniform definition of Exchange Packages and Query Points supports the capability to effectively share them within and between government organization and Public Institutions.

ii. Data Context: Categorisation of Exchange Packages and Query Points supports their discovery, and their subsequent use in the data access and data exchange. Refer to Illustration No. 3 DRM for examples.

Implementation of Exchange Packages will be standard XML messages or EDI transaction sets. Implementation of Query Points could be descriptions in a Universal Description, Discovery and Integration (UDDI) or XML registry of a data access web service.

2.1.3. Data Modelling Standards

Common data definitions (e.g., citizen's demographic profile) across Public Institutions as per the data standards are essential to eliminate redundancies. Data owners need to be identified to be responsible for common data definition, ensuring data integrity and protecting data from misuse and destruction.

Conceptual Data Model - The conceptual data model (CDM) identifies the semantics of data assets from the business context point of view and depicts the highest level significant business data entities, along with their relationship, to emphasize the business rules.

i. UML Class Diagram, Entity Relationship Diagram or Object Role Model (ORM) could be used as the modelling technique for depicting the conceptual data model.

ii. The CDM data entities typically do not contain any attributes and is technology and application dependent.

iii. Segregate the conceptual data model into multiple views based on the subject area.

iv. Abstract out the common data entities used across Public Institutions e.g., person (citizen, employee) profile, company profile, name, address, contact, email, unique identifier etc.

v. Define a generic enterprise core common conceptual data model based on the common data entities and industry standard best practices that describe the core generic data entities at the enterprise level to be exchanged and shared across Public Institutions.

vi. Define Public Institution specific segment wise conceptual data model based on specific business entities required to support the business process and government services. Leverage the common data entities defined in the earlier step to support these business services in each Public Institution.

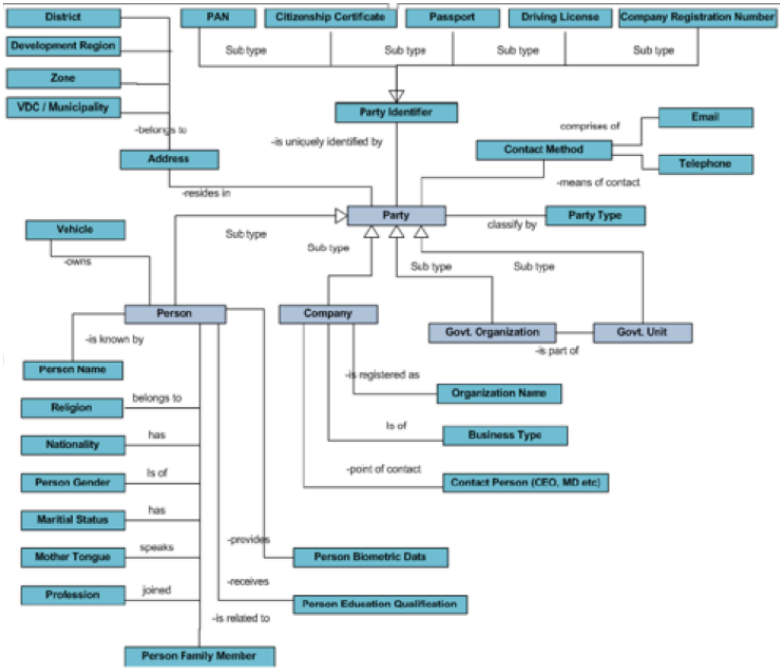

Figure III: Conceptual Model

Logical Data Model - A logical data model depicts the logical view of the conceptual data model by representing the data in as much detail as possible, without regard to how they will be physical implemented in the database. Enterprise core and segment wise logical data model will be derived from the enterprise core common and segment wise conceptual data model respectively (developed in the earlier stage) providing the logical view with respect to:

i. Identifying the data elements/attributes for each data entity,

ii. Defining the data type for each data elements / attributes

iii. Applying normalization, applying generalization / inheritance where applicable defining super type and specialized sub type data entities, and

iv. Absorbing relationships as attributes applying cardinality or multiplicity.

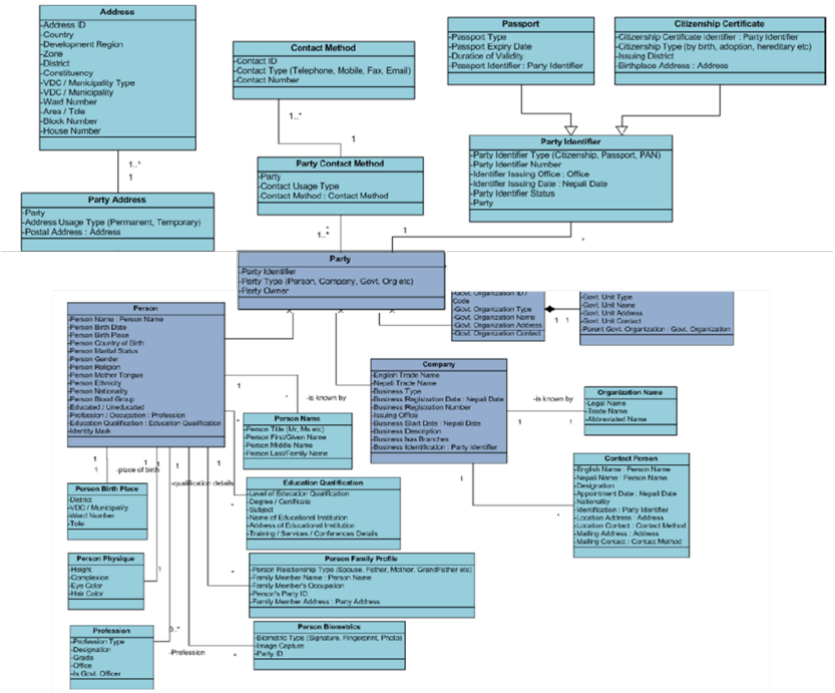

Figure IV: Logical Data Model

Canonical Data Model - Canonical data model is an enterprise application integration design pattern that defines a standardized and agreed common data definition and model to enable Public Institutions shall maintain different data formats to communicate and share in a standard data format. This model is essential in defining the data integration architecture for service delivery gateway.

Citizen Centric Service Information Flow

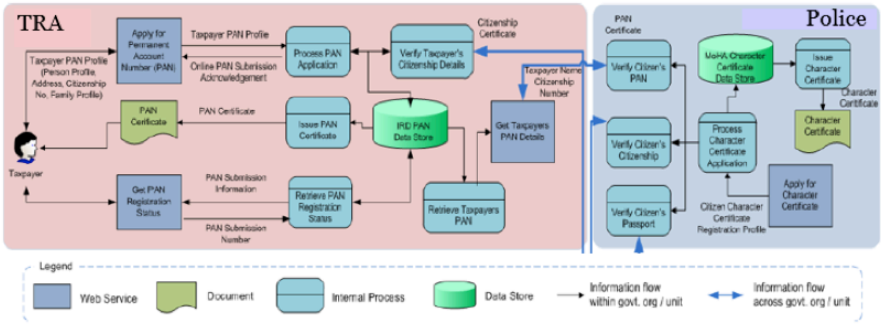

Figure V: Citizen Centric Service Information Flow – TRA Example

Business Centric Service Information Flow

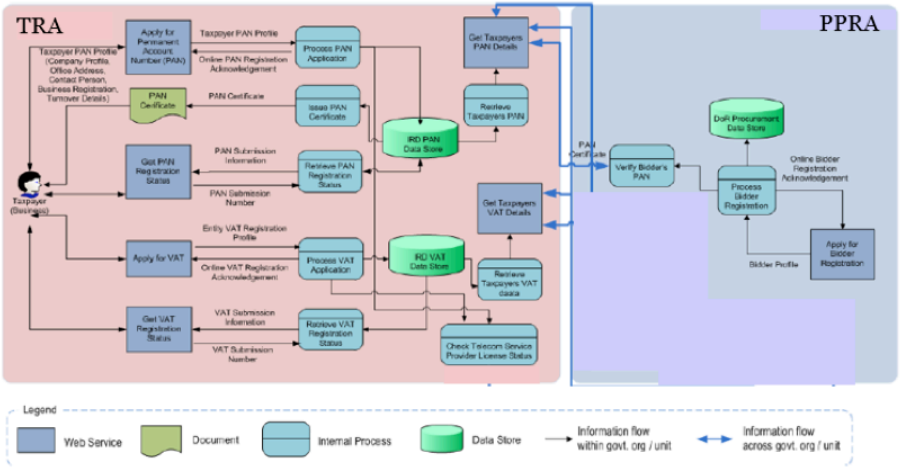

Figure VI: Business Centric Service Information Flow – TRA and PPRA Example

2.1.4. Data Classification and Taxonomy

The following are the data and taxonomy standards;

i. Data is a critical asset for Government and must be protected. The Government will adopt a data classification scheme to classify its information as per the sensitivity to business continuity and security concerns are critical.

ii. There will be a Government-wide consistent approach to capture the context of the data assets and define a data classification scheme as outlined in the Data Reference Model (DRM) and as per regulatory requirements. This includes Data Context standardization area. The Data Context standardization area establishes an approach to the categorization / classification of data assets focusing on two management mechanisms to capture the context of the data:
   a. Taxonomies and Topics
   b. Data asset / entity description

iii. There will be a data entity catalogue to be maintained in the Government based on the data classification scheme as outlined below by categorizing data assets based on subject areas context (taxonomy) at the highest level and logical data components (topics) within each subject area as illustrated below:

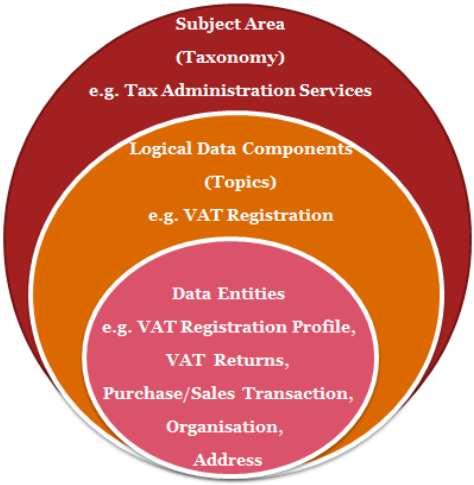

Figure II: Data Classification Scheme

a) Taxonomies provides a high-level set of categories that groups the logical data components and data entities based on the business/domain area they most closely align with, the stakeholders they impact, the extend or degree to which they are dependent on each other and the need to be managed as a unit.

b) The logical grouping of data components within each subject area provides a boundary zone that encapsulates related data entities to form a logical grouping.

2.1.5. Data Life Cycle, Quality Management and Data Governance

2.1.5.1. Data Life Cycle

As illustrated on figure VII below, data life cycle management is the process of managing the Public Institution data throughout its life cycle: collection, creation, storage, transmission, data usage, data sharing, data retention and disposal as depicted below:

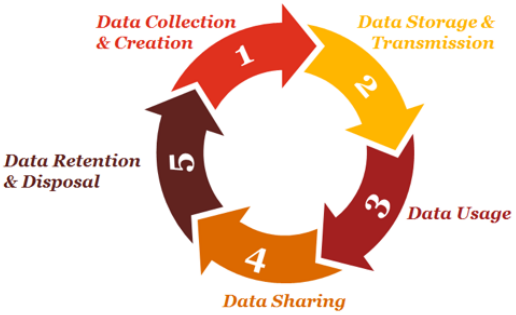

Figure VII: Data Management Cycle

Table III: Description of Data lifecycle

Data Components	Description
1. Data collection and creation	i. This is the first step of the data life cycle process when the structured data will be captured electronically in the source system. Some of the means of data collection and capture includes data collected from online web forms and eForms, data acquired from external sources, initial data migration from legacy systems to new systems etc. ii. Rigorous data quality checks, such as, data validation, cleansing, standardization and de-duplication check at the data source during the process of data collection is essential to improve the quality of the data being captured. This will reduce the propagation of erroneous, inconsistent, incomplete and potential duplicate data to other downstream systems.

	i. Online forms should be pre-filled with available data to minimize data entry error.
2. Data Storage and Transmission	i. To ensure secure transmission of sensitive data secure for transmitting it must be encrypted to an appropriate standard. Only data confirmed as un-classified or public should be transmitted in unencrypted form. Encryption ensures data security during transmission. ii. Pretty Good Privacy (PGP), an industry-standard encryption technology could be leveraged. Using this method, encrypted data can be transferred via portable media or electronically via file upload or email. iii. After data is collected and transferred to its storage location, it must be protected from unauthorized access by both internal and external sources to prevent identify theft. iv. A data storage procedure is essential because digital storage media are inherently unreliable, unless they are stored appropriately, and all file formats and physical storage media will ultimately become obsolete. v. Some important physical dataset storage and archiving considerations for electronic/ digital data include: • Server Hardware and Software – a. Type of database needed for the data b. Physical system infrastructure to be set up or use of existing infrastructure c. Whether system be utilized for other projects and data d. Responsibilities to oversee the administration of this system e. Protection of database from unauthorized access by both internal and external sources to prevent identity theft. • Network Infrastructure – a. Connection of database to a network or to the Internet b. Bandwidth required to serve the target audience c. Data Accessibility timeframes

	• Size and Format of Datasets – a. The size of a dataset will be estimated so that storage space can properly be accounted for. b. The types and formats will be identified to avoid any surprises related to database capabilities and compatibilities. • Database Maintenance and Updating – a. A database or dataset will have carefully defined procedures for updating. b. If a dataset is live or on-going, this will include such things as additions, modifications, and deletions, as well as frequency of updates. c. Versioning will be extremely important when working in a multi-user environment. • Database Backup and Recovery Requirements – i. To ensure the longevity of a dataset, the requirements for the backing up or recovery of a database in case of user error, software / media failure, or disaster, should be clearly defined and agreed upon. ii. Mechanisms, schedules, frequency and types of backups, and appropriate recovery plans should be specified and planned. This can include types of storage media for onsite backups and whether off-site backing up is necessary.
3. Data Usage	Good data management requires on-going data audit to monitor the use and continued effectiveness of existing data. Data audit trails should be maintained with the ability to generate audit reports.
4. Data Sharing	Data and information will be readily accessible to those who need them or those who are given permission to access them. Some issues to address with access to data and a database system include: i. Relevant data rules, and data ownership issues regarding access and use of data. ii. The data sharing needs along with various types and differentiated levels of access needed and as deemed

	appropriate. The need for single-access or multi-user access, and subsequent versioning issues associated with multi-user access systems. iii. The cost of providing data (such as enhancement of existing applications) versus the cost of providing access to data (such as charging service fees to access data). iv. Standard format for data exchanges along with the transformation service required as appropriate for end-users. v. System design considerations, including any data (if any) that requires restricted access to a subset of users. vi. Liability issues will be included in the metadata in terms of accuracy, recommended use, use restrictions, etc. A carefully worded disclaimer statement can be included in the metadata so as to free the provider, data collector, or anyone associated with the data set of any legal responsibility for misuse or inaccuracies in the data.
5. Data Retention and Disposal	i. The Government will plan and define data retention and disposal policies in accordance with regulatory requirements. ii. Access control mechanisms must also be utilized to ensure that only authorized users can access data to which they have been granted explicit access rights during the disposal process. iii. All computer systems, electronic devices and electronic media must be properly cleaned of sensitive data and software before being transferred outside the government. Having a strategy for reliably erasing data files is a critical component of managing data securely and is relevant at various stages in the data cycle. iv. For hard drives, which are magnetic storage devices, simply deleting does not erase a file on most systems. Files need to be overwritten to ensure they are effectively erased.

	v. Software is available for the secure erasing of files from hard discs, meeting recognized standards of overwriting to adequately erase sensitive files. vi. The most reliable way to dispose of data is physical destruction. Risk-adverse approaches for all drives are to encrypt devices when installing the operating software and before first use; and physically destroy the drive using a secure destruction facility approved by the government when data need to be destroyed. vii. Shredders certified to an appropriate security level should be used for destroying paper and CD/DVD discs. Computer or external hard drives at the end of their life can be removed from their casings and disposed of securely through physical destruction. viii. Adoption of a consistent backup policy across all Public Institutions will be considered. The institution responsible for ICT can identify the Public Institutions which are currently dependent on a single server and arrange for backing up of data on one common backup platform.

2.1.5.2. Data Quality

Data Quality is the process of measuring the quality or accuracy of the data within the Government.

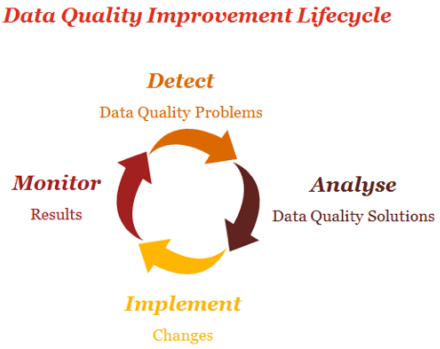

Figure VIII: Data Quality Improvement Lifecycle

The data quality improvement lifecycle involves the following steps:

i. **Detect** – Profile of the data and identify significant problems contributing to data quality degradation, such as, inconsistent and incomplete data, invalid data, duplicate or near duplicate records, violation of business rules etc.

ii. **Analyse** - Investigate the cause(s) of a problem and analyse the data improvement solutions that will cost effectively eliminate the problem without introducing new quality issues. Examples of data improvement solutions could be defining data catalogue, better communication of business rules, defining data validation, cleansing and standardization policies, strong data validation checks during insert and edit of transactional data in the source system during the data collection process etc.

iii. **Implement** - Implement the changes to reduce the data quality problem. This could include publication of data standards, enforcement of business rules, enforcement of more rigorous data validation, cleansing and standardization policies, tight edit checks in the data collection processes, centralization of master data to create single version of the truth etc. Examples of some of the data scrubbing activities include:

    a) Format fields - Ensure consistent terms and formats across a given field.
    b) Parse components - Break down strings of data into multiple fields to more effectively standardize data elements with greater accuracy.
    c) Check content - Some records include accurate information that is embedded in the wrong fields. Other fields may appear populated but are not accurate (for example, a phone number field that looks like: "111-111-1111"). The data cleansing process and rules should identify and correct these anomalies so the data is fit for use. Consider valid values and data range checks at the field levels.
    d) Eliminate duplicates - Identify matches and eliminate duplicate records. Once the data is standardized, this can be done with a high degree of confidence.

iv. Measure and Monitor – Define the data quality metrics and service levels to monitor the data quality conditions over time. Examples of possible data quality metrics to be monitored over time include reduction in business rule error rate, improvement in the consistency and validity of data over time, reduction in duplicate records etc. Using statistical methods monitor the results to ensure that the implemented changes are having the desired effect and to begin the cycle again with the next most significant quality reducer. Having an automated or repeatable process in place will enable the government to measure accuracy on a regular basis, show an improvement and eventually reaching and maintain its

target accuracy level.

Data governance refers to the administration and processes around data protection, sharing of data on different levels across the organization and data quality based on the agreed defined data principles. Having a data governance model is important as it ensures that:

i. Data principles are defined, updated, maintained and implemented.

ii. A logical view of the data is created and the entities and attributes all have definitions.

iii. The physical implementation of the data does not break any legal requirements that the Public Institution must comply to.

iv. The day to day management of the data includes activities such as storage, replication, access, interfaces, audit, reconciliation, archiving, backup and recovery all of which are done using common standards and processes.

Data governance standards include:

i. Defining the scope and goals of their data set.

ii. Defining a governance structure - A data governance structure is the central decision-making body for data governance. Members are accountable for creating and enforcing policies and procedures, establishing governance processes, settling disputes over data, refining workflows and other data governance decisions. Without a formally assigned and approved council, there will be no coordination of the work of data governance.

iii. Establishing the data stewardship - Before data can be formally managed, stewardship of the data must be formally established. Data stewardship must be formally recognized and assigned in order to sanction the role and define its scope. Its definition will vary according to what data is most important to the Public Institution. Data stewards will be accountable for all aspects of the data under their control. They will establish data quality thresholds; monitor the cost of poor data quality, deciding who should have access to the data, for what purpose and under what conditions; collect and verify all applicable metadata; and act as a representative of the data in resolving all usage conflicts.

iv. Segregation of duties - Data steward should not have responsibilities for any aspect of managing business processes. Doing so puts the data steward in the untenable position of managing both data and the process in which it was created. This is equally true of data quality verification. Data stewards should be involved in the

tactics of data correction and cleansing, but the business should define rules for its respective data sets.

v. Establishing controls and measurable key performance indicators (KPIs) to monitor the progress.

vi. Ensuring transparency specifically while in the management of data quality. The results of the process of measuring the progress of governance should be made available to all stakeholders.

vii. Addressing data quality problems at source. It is always recommended to get to the root cause of data quality issues.

viii. Implementing change management. Implicit in any well-designed change management system is an issue escalation and resolution process, and a feedback mechanism that keeps the originator of the issue informed about progress.

2.1.6. Data Structure and Semantics

The data structure and semantics concepts are represented as shown in Table III;

Table III: Concepts to represent data structure and semantics

2.1.7. Information Sharing and Data Capabilities

Setting up information sharing and data integration capabilities will be as follows:

Concepts	Description
Data Schema	Describes a structured data set and the representation of its meta data. The data artefacts for data schema could be the conceptual and logical data models.
Data Entity	An abstraction for a person, place, object, event or concept described (or characterized) by common elements/ attributes. For example "Person" and "Company" are data entities. An instance of an entity represents one particular occurrence of the data entity, such as Citizen, Taxpayer, etc.
Data Elements/ Attributes	Properties of a data entity that contain information about its state.
Data Type	A constraint on the type of physical representation that an instance of a data element/ attribute may hold (e.g. "string" or "integer").

Data Relationships	Describes the relationship between two data entities e.g. the "Person" entity is related to the "Person Name" entity by relationship "is known by".

i. Defining an effective data integration strategy that will go beyond ETL that supports both analytical and operational processes and data.

ii. Considering leveraging a universal data integration solution that will support the following architectural considerations:
   a) Extract, Transform and Load (ETL) - ETL is leveraged across all of the data integration programs to access data from one system (preferably the source), transform it and load it in the target system. The ETL technology brings together underlying services, normally through a design interface, to build re-usable services and support various data integration initiatives.
   b) Data Quality - Data quality checks are essential when data is loaded from the source to the target. Ensure the data integration solution provides a graphical environment for data stewards that allows them to bring together the underlying services to profile, parse, enrich, cleanse and match data to create the business rules to be applied either in real-time/batch as a part of data integration/migration/ synchronization process.
   c) Data Synchronization - Ensure data synchronization to enable data to be moved from source to the target by enabling the use of message queues, triggers, Change Data Capture, and more.
   d) Data Federation - Ensure data federation capability that allows data to remain in place and be integrated and accessed as needed. Due to its dynamic nature, this technique lends itself to solving potential problems where there is need to access large amounts of data or data from many underlying systems.
   e) Metadata Mapping - Ensure metadata mapping capabilities that will allow the import of metadata from various systems and exchange metadata with other systems.
   f) Enterprise Connectivity - As part of data integration initiative ensuring connectivity with the interfacing systems is essential to allow data transfer across systems. Ensure the data integration solution supports a wide variety of connectivity techniques such as native access using standard utilities and open standard access (such as ODBC) to all major structured data sources, including relational databases, flat files, ERP systems, and mark-up languages such as XML for reading and writing. Support for connecting to and reading and writing data from message queues and the ability to receive and send data to and from Web services should also be supported by the solution to provide complete connectivity.

2.1.8. Business Intelligence

Business Intelligence refers to the set of capabilities supporting the extraction, aggregation and presentation of Government data to facilitate decision analysis. It provides information that pertains to the history, current status or future projections of the government to help analyse data for the purpose of supporting risk assessment and ad hoc queries etc. BI broadly covers the following capabilities:

i. Demand and Forecasting - Facilitates the prediction of sufficient production to meet a Public Institution's demand for a service.

ii. Balanced Scorecard - Supports the strategy performance framework pulling together financial and non-financial data.

iii. Decision Support and Planning - Supports the analysis of information and predict the impact of decisions before they are made.

iv. Data Mining - Provides for the efficient discovery of non-obvious, valuable patterns and relationships within a large collection of data.

2.1.9. Data Search

Search provides the ability to locate sources of specific data (i.e., structured, usually in operational systems) or information (i.e., unstructured, usually in content repositories or internet/intranet stores). The following are Search Capabilities:

i. Web Search – The ability to identify and retrieve content across the intranet/ internet.

ii. Enterprise Search – The ability to produce a consolidated list ranked by relevance of multiple types of content across a variety of sources.

iii. Federated Search – Search capabilities across multiple applications or using multiple search applications.

iv. Application Specific Search – Search capabilities within a specific application.

2.1.10. Reporting and Reporting Tools

Reporting provides the ability to report and query data held within the Business Intelligence. It can also be used to conduct operational reporting on source systems where the use of the service is deemed appropriate. Data reporting involves;

i. Pre-Defined Reporting- Pre-defined reports created for users to meet regular requirements.

ii. Ad hoc Reporting - User created reports for infrequent requirements and support the use of dynamic reports on an as-needed basis.

iii. Query and Analysis - Allows users to query and analyse data, e.g. drill down, 'slice and dice'.

iv. Application Specific Reporting - Limited to and within a specific application.

During selection of reporting tools, the following should be considered;

i. The tool have comprehensive presentation mechanism such as charts, graphs, query, text etc., which allows users to call up pre-defined reports or create ad hoc reports. The tool should also have the customization capability to allow users to customize and pre-set the presentation of report to ensure adherence with Public Institution standards.

ii. The tool have transformation techniques, support business logic to convert raw data into useful information.

iii. The tool support generating reports in various file formats like rtf, CSV, XML, etc.

iv. The reporting tools have data source connection capabilities and support standard access mechanism. The tool should have the flexibility to provide support for data integration from various databases, web services, flat files, objects etc.

v. The tool have the requisite security features to prevent unauthorized access. The tool should also have the capability to seamlessly integrate with external user authentication and single sign-on frameworks.

vi. The tool have flexible export capabilities supporting excel, flat file and PDF export format.

vii. The tool is platform independent.

viii. Version control features and change control features is available.

ix. Scheduling and distribution capabilities of a reporting tool are value added features.

x. Scheduling of reports to run daily, weekly and distribution of the reports generated to target audience either by emails or web publishing means are encouraged.

2.1.11. Document, Content and Knowledge Management

The use of electronic documents facilitate the exchange of information and improving collaboration across the government. Converting paper-based document repositories into electronic format will allow easy access and retrieval of information in electronic form

for government staffs and citizens improves process efficiency and reduces operational cost. The following document, content and knowledge management capabilities will be considered as follows:

i. Moving from paper-based forms to electronic online forms or e-Forms that will allow citizens to fill in the form data online thereby less likely to be error prone and reduce manual data entry effort.

ii. For existing paper forms and documents, scanning and OCR/ICR of the content is a fast way to capture the data with greater efficiency than manual entry and lower error rates.

iii. Use of electronic Document Management Solution (DMS) will enable the governments to share information within and across Public Institutions.

iv. The use of a Government wide central document repository to consolidate documents across Public Institutions that will facilitate faster content based searching of documents. Appropriate security controls must be enforced to prevent unauthorized access to confidential documents. In case of multiple DMS rollout out in different Public Institutions, a unified single document gateway could be conceptualized, integrating all the Public Institutions DMS.

v. Leverage industry standard based out of the box document management solutions (preferably open source avoiding proprietary solutions and technologies wherever possible) that will facilitate faster deployment of the solution with minimum turnaround time and less development effort

vi. Along with the technology consideration, there is a need for cultural change with a paradigm shift in the mind-set of the people using it (i.e., citizens, government staff etc.). Proper training should be imparted to all level of staff within government.

vii. Some of the emerging trends in knowledge management:
    a) Deploy knowledge portals for internal government staff for information sharing and national portals for citizen access for information dissemination as well as access to services.
    b) Leveraging enterprise 2.0 technology e.g., surveys, polls, discussion forums.
    c) Use of social media and networks to deepen interaction between government and the citizenry by moving beyond a traditional broadcasting model to active engagement on issues, programs and decision making.
    d) Integration of mobile devices into communication planning (e.g. text messaging for notifications and transactions).

2.2. eGovernment Information Architecture Standards

2.2.1. Table II provides principles under which the eGovernment Information Architecture is designed. Institutional Enterprise Architectures should also be designed basing on these principles:

Table II: Information Architecture Design Principles

Principle #1	***Data is an asset***
Rationale	i. Information / Data are a national asset that has high value to the Government as data is the foundation of all decision making. ii. Effective and careful data management is of high importance and priority to the Government to ensure where it resides, rely on its accuracy, and obtain it as and when required.
Implications	i. Public Institutions will establish effective data management to ensure effective decision making and improved performance. ii. Public Institutions shall organise and manage their key data assets to drive operations.
Principle #2	***Data Creation, Accessibility & Availability***
Rationale	i. Users – public servants and citizens alike – must have access to accurate, relevant and timely data to render or consume an effective government service. A well informed citizenry is necessary to the constitutional democracy; and accurate information to authorized users is critical to effective decision making, improved performance, and accurate reporting. ii. Creation: All enterprise information / data will be captured once at the point of its creation / source. iii. Update: New / updated data entry should be restricted to the designated source system who is the owner of the data.

	iv.	Accessibility: Data will be accessible to users to perform their respective business functions. Effective use of information must be considered from an enterprise perspective to allow access by a wide variety of users.
	v.	Availability: Government wide enterprise data will be made readily available (real-time), so as not to delay the business processes, and will enable appropriate timely sharing across the organization.
Implications	i.	Government will have a Dictionary of ICT Terms and Definitions on ICT that is freely shared and collectively owned by ICT practitioners.
	ii.	Government will leverage on the Data Reference Model to define their data catalog (a schema that contains the data entities and their definitions), a meta-data model (a schema that defines relationship between the data entities) and a meta-data store (an electronic repository to store it).
	iii.	Data definition scheme must comply with a common data definition standard as prescribed by the Interoperability Standard (e-GIF). Refer to e-Government Interoperability Framework and Standards *(e-GIF) Ref No: eGA/EXT/ ARC/002 for further details.*
	iv.	Data definitions and ICT Dictionary of terms must be available to the whole of Government to enable use, integration and common understanding.
	v.	Government meta-data management discipline must be established, and data standardization initiatives are coordinated throughout government.
	vi.	Government wide data catalogue (inventory) has to be developed and used to identify authoritative sources of high quality information that can be made available for access to empower public servant and citizens alike.
	vii.	Default Access control to data will be set to "open for all" and made available to all through any means, unless security policies requires access restrictions (i.e. application software

	should not unnecessarily restrict users to access data). viii. "Search" or "Find" functionality shall exist for all end-user applications/web portals to improve access to data sources. ix. Access to data sources will be available via various interfaces, channels such as mobile, portal etc to improve the convenience for the user. x. Public Institutions will key in information once and re-use it across the organisation. This will reduce costs, promote the efficiency, accuracy, consistency of data and assures quality. Readily available enterprise data will facilitate timely data access at every level of the organization and provide timely response to information request and effective service delivery. xi. Public Institutions will provide wide access to data. This will lead to efficiency and effectiveness in decision-making, and enables timely response to information requests and service delivery.
Principle #3	***Data is Shareable***
Rationale	Government wide enterprise data must be shared across Public Institutions on need to know basis.
Implications	i. Public Institutions will share data electronically. This will result in increased efficiency as existing data entities shall be used, without re-keying, or re-creating create new entities. ii. Public Institutions will rely on authorized sources of more accurate and timely managed data. This will help in improved decision making. iii. Public Institutions will have access to the necessary shared data required for their respective business functions. Shared data should be centrally controlled and managed at the appropriate Public Institution level.

Principle #4	*Data Ownership & Primary Data Source*
Rationale	i. Each data entity / item will be owned by a Public Institution. The Public Institution will be responsible for data definitions, domain, values, integrity and security. Owner will be identified for each data entities and its related data services. ii. Primary Data Source: Government wide enterprise data entity will have an authoritative, official, primary data source that is the location for all create, update and delete actions. All copies of enterprise data will be considered secondary and will not be updated as part of business transactions.
Implications	i. In order for Public Institutions data to be managed effectively, there can be only one primary source for each data entity so that data entity could be traceable back to the source system. Otherwise, inconsistent, erroneous and out-of-date data may result. Public Institutions will, therefore, identify ownership of their respective data entities to avoid ambiguity and create clear responsibility and accountability for their data. ii. Public Institutions will identify data owners and point of contact that will be responsible and accountable for all changes in the data entities & data services and the approval of the same. Data integrity is at its highest level when the central management of changes to data is done by an authoritative source of record.
Principle #5	*Data Security & Permission*
Rationale	Data will only be available to users who require the information as part of their role. The duty to protect and secure sensitive information must be balanced against the duty to share and release public information.
Implications	i. Public Institutions will adhere to open information sharing and release of information via relevant legislation must be balanced against the need to restrict the availability of classified, proprietary, and sensitive information.

	ii.	Public Institutions will consider security architecture is an integral part of business, data, application and technology architectures.
	iii.	Public Institutions will ensure compliance with legislation and government data and security policies.
	iv.	Public Institutions will adopt role based authorization such that the right level of information is shared to users.
Principle #6		***Standard, common vocabulary and data / metadata definitions***
Rationale	i.	Public Institution data and metadata standards will be defined to ensure seamless interoperability while interchanging data e.g. definition of eGIF & Metadata Standards.
	ii.	Every data item important for data exchange across the government will have metadata.
	iii.	A centralized server-based source of metadata is required.
Implications	i.	Public Institutions will have a common definition of data to be exchanged across the government, agreed format and meaning of the data items. A common vocabulary will facilitate effective communications and enable sharing of data. In addition, it is required to interface systems and exchange data.
	ii.	Public Institutions will maintain a centralized metadata provides single point for maintaining the metadata. This provides a consistent and quality metadata modelling.

2.2.2. Public Institutions will define their respective Information Architecture based on DRM standards

2.2.3. Public Institutions will standardize their data under 3 broad areas:

i. Data Description
ii. Data Context
iii. Data Sharing

2.2.4. Public Institutions will create logical data components to group data entities into encapsulated modules for governance, security, and deployment purposes.

2.2.5. Public Institutions will adhere to data classification and taxonomy standards Public Institutions will adopt a Government-wide consistent approach to capture the context of the data assets and define a data classification scheme as outlined in the Data Reference Model (DRM) and as per regulatory requirements.

2.2.6. Public Institution will maintain data entity catalogue based on the data classification scheme by categorizing data assets based on subject areas context (taxonomy) at the highest level and logical data components (topics) within each subject area

2.2.7. Public Institutions will adhere to enterprise data and metadata standards as well as data modeling standards. Public Institutions will adhere to Data Life Cycle, Quality Management and Data Governance Public Institutions will adhere to Data Quality Management Standards.

2.2.8. Public Institutions will adhere to Data Governance Standards.

2.2.9. Public Institutions will identify Data owner for ownership of datasets and its related data services and must be accountable for the effective and efficient management of such data. This includes:

i. Establish policies and agreements for data security and data sharing.

ii. Ensure important datasets are developed, maintained and accessed within the defined specifications adhering to data guidelines.

iii. Maintain appropriate levels of data security and ensure data accessibility to authorized users.

iv. Ensure adequate and agreed-upon data quality and metadata metrics are maintained on a continuous basis.

v. Ensure periodic data audits to assure on-going data integrity.

vi. Ensure fundamental data maintenance across the data life cycle including data storage and data archival.

2.3. eGovernment Information Architecture Technical Guidelines

2.3.1. Public Institutions will represent data structure and semantics concepts:

2.3.2. Public Institutions will define their information architecture components

2.3.3. Public Institutions will set up information sharing and data integration capabilities

2.3.4. Public Institutions shall adhere to specified data reporting mechanism in Data reporting.

2.3.5. Public Institutions will adopt specified consideration during selection of a reporting tool.

2.3.6. Public Institutions will make use of electronic documents to facilitate the exchange of information and improving collaboration across the government.

2.3.7. Public Institutions will consider document, content and knowledge management capabilities.

2.3.8. Public Institutions will refer to the eGovernment Security Architecture - Standards and Technical Guidelines (eGA/EXT/ISA/001) for additional information.

3. IMPLEMENTATION, REVIEW AND ENFORCEMENT

3.1. This document takes effect once approved in its first page.

3.2. This document is subject to review at least once every three years.

3.3. Any exceptions to compliance with this document should be approved in writing by Chief Executive Officer (CEO) of e-Government Agency.

4. GLOSSARY AND ACRONYMS

4.1 Glossary

None

4.2 Acronyms

Abbreviation	Explanation
BCP	Business Continuity Planning
CDM	Conceptual Data Model
CIQ	Customer Information Quality
DCMI	Data center infrastructure management
DMS	Document Management Solution
DRM	Data Reference Model
ELT	Extract, Load and Transform
ERP	Enterprise Resource Planning
ETL	Extract, Transform and Load
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting
ODBC	Open DataBase Connectivity
OWL	Web Ontology Language
PGP	Pretty Good Privacy

5. RELATED DOCUMENTS

5.6. **eGovernment Integration Architecture - Standards and Technical Guidelines (eGA/EXT/ITA/001)**

5.7. **eGovernment Infrastructure Architecture - Standards and Technical Guidelines (eGA/EXT/IRA/001)**

5.8. **eGovernment Security Architecture - Standards and Technical Guidelines (eGA/EXT/ISA/001)**

5.9. **eGovernment Architecture Processes and Governance - Standards and Technical Guidelines (eGA/EXT/PAG/001)**

6. DOCUMENT CONTROL

Version	Name	Comment	Date
Ver. 1.0	eGA	Creation of Document	February 2016
Ver. 1.1	eGA	Alignment with eGovernment Guideline	November 2017

APPENDIX

Illustration No.1 ICT Organisation Capability

Table AI is a generic list of ICT organizational capabilities required for any Public Institution:

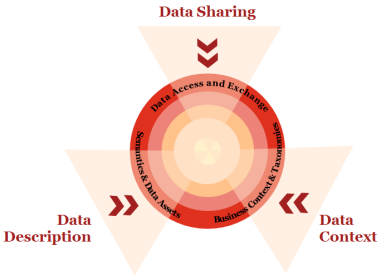

Figure AI - Example of Data Reference Model

Data Description

Purpose: Data description standardization addresses key questions like:

i. How do you understand what data is available and what it means?

Abstract Model:

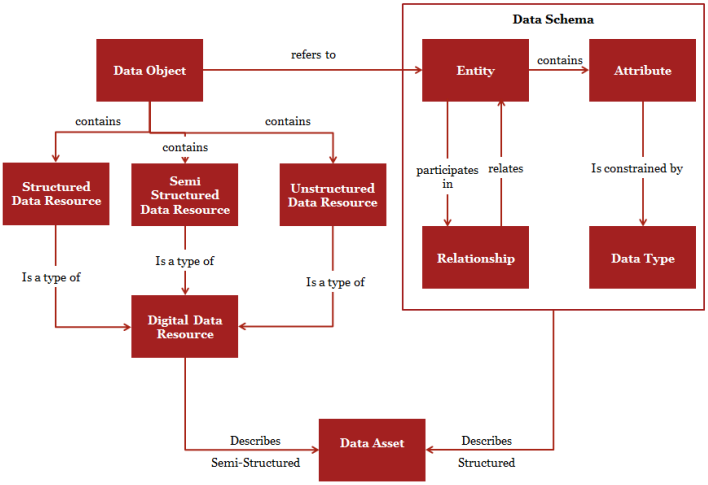

Figure AII – Abstract Model of Data Description

Data Context

Purpose: Data context standardization addresses key questions such as:

i. What are the data (subject areas/ topics and entities of interest) contained within the Data Asset that are needed to support the objective/ vision of e-government initiative?

ii. Which Public Institution is responsible for maintaining the Data Asset?

iii. What is the linkage to the business reference model?

iv. What core information or services are needed to make the data discoverable and establish governance (Refer to Data Sharing below)

Abstract Model:

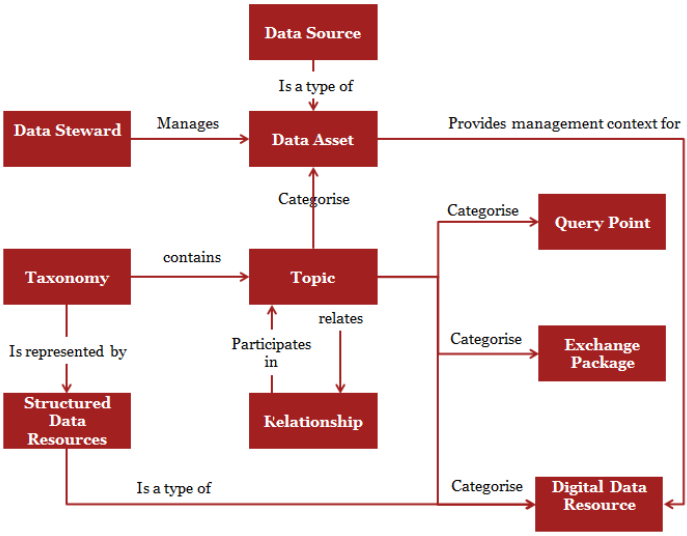

Figure AIII – Abstract model of Data Context

Example

i. Categorization of data based on subject area
   a. "Tax Administration Service" subject area that broadly covers the services provided by the Tanzania Revenue Authority

   b. "Civil Registry Service" subject area that broadly covers the services provided by the Registration, Insolvency and Trusteeship Agency (RITA).

   c. "Electoral Service" subject area that broadly covers the services provided by the Electoral Commission.

ii. Sub-categorisation of data based on logical data components within a subject area:
   a. For example "VAT Registration": Logical data components within the "Tax Administration Service" subject area that represents the logical grouping of data entities related to the registration for VAT.

   b. For example "Birth Registration" and "Marriage Registration": Logical components within the "Civil Registry Services" subject area that represent the logical grouping of data entities related to birth and marriage registration of citizens respectively.

   c. For example, "Voters Registration": Logical data components within the "Electoral Service" subject area that represent the logical grouping of data entities related to voter's registration and the capture of voter's information.

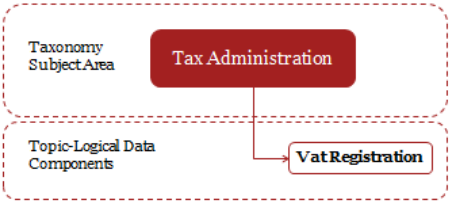

Figure AIV – Example of Data Context

Data Sharing

Purpose: The Data Sharing standardization area answers key questions like:

i. What is the data sharing architecture? (That is how will the data be made shareable?)
ii. What volume of data will be shared, frequency, etc.?

Abstract Model:

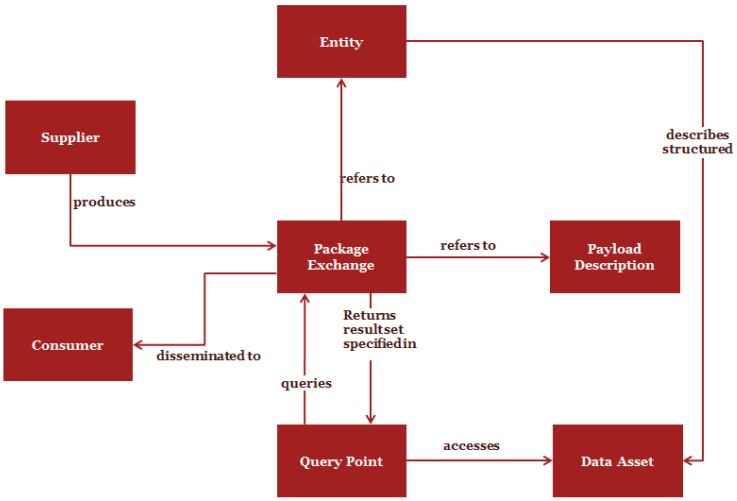

Figure AV – Example of Data Sharing Workflow

Example:

Figure AVI – Example of Data Sharing

An overall example of the above DRM structure from the Government Enterprise Architecture Framework perspective considering the Tanzania Revenue Authority (TRA) data entities, its categorisation, structure definition and exchange package service is illustrated below:

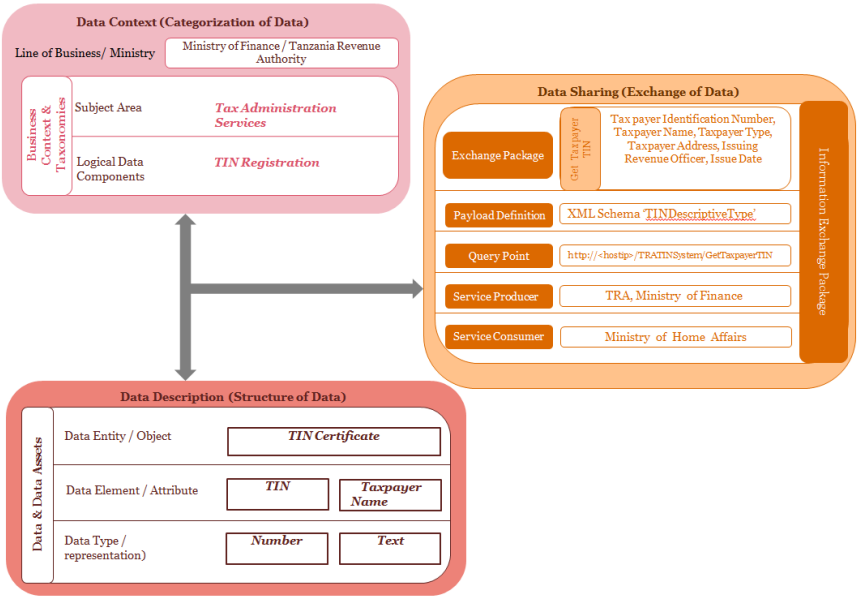

Figure AVII – Example of DRM at TRA

Issued by eGovernment Agency - November 2017