TOGAF® Series Guide

Information Architecture: Metadata Management

Prepared by Guillaume Hervouin and Pastel Gbetoho

THE *Open* GROUP

TOGAF® Series Guide

Information Architecture: Metadata Management

ISBN: 1-957866-16-1

Document Number: G234

Contents

Preface

The Open Group

The Open Group is a global consortium that enables the achievement of business objectives through technology standards. With more than 900 member organizations, we have a diverse membership that spans all sectors of the technology community – customers, systems and solutions suppliers, tool vendors, integrators and consultants, as well as academics and researchers.

The mission of The Open Group is to drive the creation of Boundaryless Information Flow™ achieved by:

- Working with customers to capture, understand, and address current and emerging requirements, establish policies, and share best practices

- Working with suppliers, consortia, and standards bodies to develop consensus and facilitate interoperability, to evolve and integrate specifications and open source technologies

- Offering a comprehensive set of services to enhance the operational efficiency of consortia

- Developing and operating the industry's premier certification service and encouraging procurement of certified products

Further information on The Open Group is available at www.opengroup.org.

The Open Group publishes a wide range of technical documentation, most of which is focused on development of Standards and Guides, but which also includes white papers, technical studies, certification and testing documentation, and business titles. Full details and a catalog are available at www.opengroup.org/library.

The TOGAF® Standard, a Standard of The Open Group

The TOGAF Standard is a proven enterprise methodology and framework used by the world's leading organizations to improve business efficiency.

This Document

This document is the TOGAF® Series Guide to Information Architecture: Metadata Management. It has been developed and approved by The Open Group.

It addresses the Metadata Management capability, as introduced and defined in the TOGAF Series Guide: Information Architecture – Introduction (in development). It is recommended that the reader is familiar with the latter document before proceeding.

The information shared in this document has been gathered from the lessons learnt on the ground (best practices and pitfalls) from several projects.

TOGAF® Series Guide (2023)

More information is available, along with a number of tools, guides, and other resources, at www.opengroup.org/architecture.

About the TOGAF® Series Guides

The TOGAF® Series Guides contain guidance on how to use the TOGAF Standard and how to adapt it to fulfill specific needs.

The TOGAF® Series Guides are expected to be the most rapidly developing part of the TOGAF Standard and are positioned as the guidance part of the standard. While the TOGAF Fundamental Content is expected to be long-lived and stable, guidance on the use of the TOGAF Standard can be industry, architectural style, purpose, and problem-specific. For example, the stakeholders, concerns, views, and supporting models required to support the transformation of an extended enterprise may be significantly different than those used to support the transition of an in-house IT environment to the cloud; both will use the Architecture Development Method (ADM), start with an Architecture Vision, and develop a Target Architecture on the way to an Implementation and Migration Plan. The TOGAF Fundamental Content remains the essential scaffolding across industry, domain, and style.

Trademarks

ArchiMate, DirecNet, Making Standards Work, Open O logo, Open O and Check Certification logo, Platform 3.0, The Open Group, TOGAF, UNIX, UNIXWARE, and the Open Brand X logo are registered trademarks and Boundaryless Information Flow, Build with Integrity Buy with Confidence, Commercial Aviation Reference Architecture, Dependability Through Assuredness, Digital Practitioner Body of Knowledge, DPBoK, EMMM, FACE, the FACE logo, FHIM Profile Builder, the FHIM logo, FPB, Future Airborne Capability Environment, IT4IT, the IT4IT logo, O-AA, O-DEF, O-HERA, O-PAS, Open Agile Architecture, Open FAIR, Open Footprint, Open Process Automation, Open Subsurface Data Universe, Open Trusted Technology Provider, OSDU, Sensor Integration Simplified, SOSA, and the SOSA logo are trademarks of The Open Group.

Dublin Core is a trademark of the Dublin Core Metadata Initiative.

Google is a trademark of Google, Inc.

All other brands, company, and product names are used for identification purposes only and may be trademarks that are the sole property of their respective owners.

TOGAF® Series Guide (2023)

About the Authors

AXA Data Architecture Community

This document is inspired from Data Architecture deliverables donated to The Open Group by AXA. They are the results of the Data Architect community's work led by the group and with the contribution of affiliates mainly in France, Germany, Belgium, United Kingdom, Japan, US, Hong Kong, Spain, Switzerland, and Italy from 2015 to 2021. The aim of this community is to share Data Architecture best practices across the AXA organization to improve and accelerate architecture delivery.

AXA is a global insurer (www.axa.com).

Guillaume Hervouin

Guillaume Hervouin has a Bachelor's degree in Statistics and Business Intelligence (awarded in 2002). He was a Solution Architect at Capgemini from 2002 to 2015, and since 2016 he has been a Data Architect at AXA.

Pastel Gbetoho

Pastel Gbetoho holds an Engineer's degree in Information Systems and Telecommunication (www.utt.fr), awarded in 2007. He was a Business Intelligence Consultant for Capgemini Technology Services and Capgemini Consulting until 2012, leading functional and technical design endeavors on large data analytics programs. Since 2013, he has provided consulting services on Data Architecture and Data Management. He has been supporting the AXA Group Data Architecture team since 2017 as an external consultant.

Acknowledgements

Referenced Documents

The following documents are referenced in this TOGAF® Series Guide:

- ArchiMate® 3.2 Specification, a standard of The Open Group (C226), published by The Open Group, October 2022; refer to: www.opengroup.org/library/c226

- Data Management Body of Knowledge (DMBOK2), DAMA International, published by Technics Publications, July 2017; refer to: https://dama.org/content/body-knowledge

- Dublin Core™ Metadata Element Set (DCMES), Dublin Core Metadata Initiative (DCMI); refer to: www.dublincore.org

- ISO 639-1:2002: Codes for the Representation of Names of Languages – Part 1: Alpha-2 Code; refer to: https://www.iso.org/standard/22109.html

- ISO 9000:2015: Quality Management Systems – Fundamentals and Vocabulary; refer to: https://www.iso.org/standard/45481.html

- ISO/IEC 19773:2011: Information Technology – Metadata Registries (MDR) Modules; refer to: https://www.iso.org/standard/41769.html

- The TOGAF® Standard, 10th Edition, a standard of The Open Group (C220), published by The Open Group, April 2022; refer to: www.opengroup.org/library/c220

- TOGAF® Series Guide: Information Architecture – Introduction (in development)

TOGAF® Series Guide (2023)

1 Introduction

In this document, information architecture best practices are shared to support architecture work on the specific topic of Metadata Management among other guides that deal with IT-based data management capabilities described in the TOGAF Series Guide: Information Architecture – Introduction (in development). It aims to provide a framework to:

- Accelerate the delivery of architecture study

- Inform Data Architects and other data professionals about good practices

- Provide a common language to describe the Metadata Management capability to data professionals: Chief Data Officers, Data Managers, Architects, Data Engineers, and Data Scientists

Metadata Management is instrumental in delivering value from data. This document intends also to emphasize the benefits of data documentation, and the necessary efforts required to set up an effective Metadata Management capability within an organization.

2 Relationship to Existing Standards from The Open Group

2.1 The ArchiMate® Specification

This document uses the formal notation from the ArchiMate Specification to describe the Metadata Management architecture reference models.

The ArchiMate elements used in this document are shown in Figure 1.
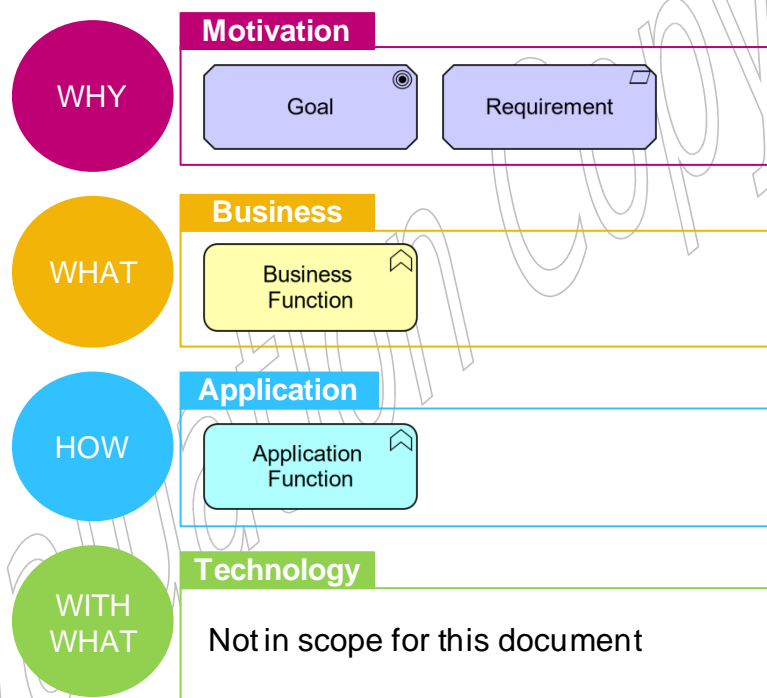
Figure 1: The ArchiMate Modeling Notation Elements Used in This Document

Definitions taken from the ArchiMate Specification are as follows:

- Goal: represents a high-level statement of intent, direction, or desired end state for an organization and its stakeholders

- Requirement: represents a statement of need defining a property that applies to a specific system as described by the architecture

- Business function: represents a collection of business behavior based on a chosen set of criteria (typically required business resources and/or competencies), closely aligned to an organization, but not necessarily explicitly governed by the organization

- Application function: represents automated behavior that can be performed by an application component

2.2 The TOGAF® Standard and Data Architecture

This document proposes:

- A focus on Metadata Management to guide the design of the capability including reference models with business function and application function descriptions (see Section 5.1 and Section 5.2)

- An adaptation of the TOGAF ADM phases to support the execution of a Metadata Management transformation program (see Section 4.1)

3   Metadata Management Capability

Simplistically, metadata can be called "data about data". For example, metadata indicates who is the author/creator of the data, when it was created, its security classification, its data quality attributes, and so on. According to the Data Management Body of Knowledge (DMBOK2), pp.417 – see Referenced Documents:

"Metadata helps an organization understand its data, its systems, and its workflows. It enables quality assessment and is integral to the management of databases and other applications. It contributes to the ability to process, maintain, integrate, secure, audit, and govern other data."

The concept of metadata is not new and is widely used in the IT industry to define data about data. In ancient times long before the digital age, metadata such as author, title, number of tablets, etc. was already used to catalog collections of clay tablets. As such, metadata provides additional information to better know and use the data.

Metadata Management is the planning, implementation, and control activity that enables access to high-quality, integrated metadata (Source: DMBOK2, pp.419). It enhances the data with additional information so that it can be used to:

- Know your data:

    — Reduce uncertainty and possible misinterpretation of the data

    — Know the quality of the data available and associated business rules (data quality business rules, transformation business rules, etc.)

    — Data origin and usage of the data

    — Improve auditing of Business Intelligence (BI) reports

    — Monitor data volume across the information system to reduce environmental impact

- Manage the data risk and compliance with regulatory requirements:

    — Leverage data classification (confidentiality, privacy) to support data risk management (e.g., to support data protection impact assessment), base data access model on defined metadata (e.g., roles, responsibility of the user, purpose of data processing, etc.)

    — Leverage security and privacy metadata for the enterprise's most valuable asset; for example:

        — Security metadata enables the transition from network-centric to attribute-centric security embodied in new standards such as zero trust that treat data as a corporate asset

        — Enterprise identity management relies on a federated identity (common attributes) to control human or machine access to data

        — Privacy legislation in many countries and/or groups of countries (e.g., the European Union) makes the dissemination of personally attributable data illegal

TOGAF® Series Guide (2023)

- — Improve auditing of data lineage and usage to meet regulatory requirements

- Manage data lifecycle:

  - — Need metadata to indicate what is required to adhere to enterprise policies and directives

  - — Need to know what must be archived and retained as a record (often a function of national legislation)

  - — Need to identify retention lifecycle to trigger disposition (deletion)

- Improve time-to-market:

  - — For any new business initiative, reduce data scoping effort with the support of good data documentation (glossary, dictionary, classification)

  - — Identify relevant data elements for use-cases

- Manage data quality and sharing :

  - — Enable information interoperability/sharing to facilitate enterprise and extraprise collaboration, shared (situational awareness), and decision support

  - — Identify redundant and obsolete data

  - — Support the business to leverage external data

Metadata is all too often delegated to IT, whereas it is essentially a business concern. Subsequent to numerous investigations in the mid-2000s and new government legislation, metadata is a key corporate asset that has to be maintained to ensure shareholder transparency, corporate security, and client privacy.

The rise of NoSQL databases (relaxed data schemas) and the recent big data trends with the associated promise to unlock value from any kind of data, have led to focus solely on building the capabilities to collect, store, and analyze the data. Very little effort was put on metadata which is often depicted as irrelevant or available automatically using machine learning tools. Beyond market trends, Metadata Management is still a challenge in many organizations for multiple reasons:

- IT complexity:

  - — Legacy applications are poorly documented: reference data is lacking, and data changes its meaning during the lifecycle of the application

  - — Mergers and acquisitions are increasing the IT landscape and therefore the scope of metadata to be documented is changing

  - — The replication factor: this concept defines how many times a data element (e.g., customer first name) is replicated in the information system; the higher the replication factor, the more challenging the metadata documentation effort

- Data complexity:

  - — Variety and velocity of data collected (e.g., social media feeds or logs collected near real-time, text and image collected for machine learning purpose, etc.) require new techniques to collect the relevant metadata

- — Dynamic schemas make it challenging to maintain up-to-date metadata documentation

- IT marketing trends:

  - — Microservices/micro applications often challenge data lineage and increase the replication factor

  - — "Agile" delivery is incorrectly interpreted as no requirement for documentation

- Manual work:

  - — How to capture the information with business and IT Small and Medium Enterprises (SMEs)?

  - — What is the right granularity and priority criteria?

  - — Metadata validation is not automatic and needs a human review

- Lack of data standards:

  - — Naming conventions

  - — Data types and unit conventions

However, metadata is key from a business perspective to having good knowledge and understanding of the data already available within the organization. Therefore, metadata is an important foundational data management capacity for any organization looking to efficiently leverage its data.

In the effort to design this capability, we have focused on the minimum set of functions required to effectively deliver Metadata Management. As an organization becomes more mature in the practice, it is possible to go beyond the baseline defined in this framework. We also believe that ethical use of data is becoming more and more important for every organization. It is reflected in the regulations taken in several parts of the world following the General Data Protection Regulation (GDPR),[1] which came into force in the EU on May 25, 2018. Even more than compliance to regulatory requirements, it can be a competitive advantage for the organizations that document data usage by design to enable data governance over ethical use of data. Therefore, we have included in this baseline framework "data usage documentation" in the Metadata Management business functions, with an emphasize on personal data.

This section covers the high-level reference model functions and their definitions. More detailed functions are available in the dedicated reference model part of this document in Section 5.1 and Section 5.2.

---

[1] Refer to: https://en.wikipedia.org/wiki/General_Data_Protection_Regulation.

TOGAF® Series Guide (2023)

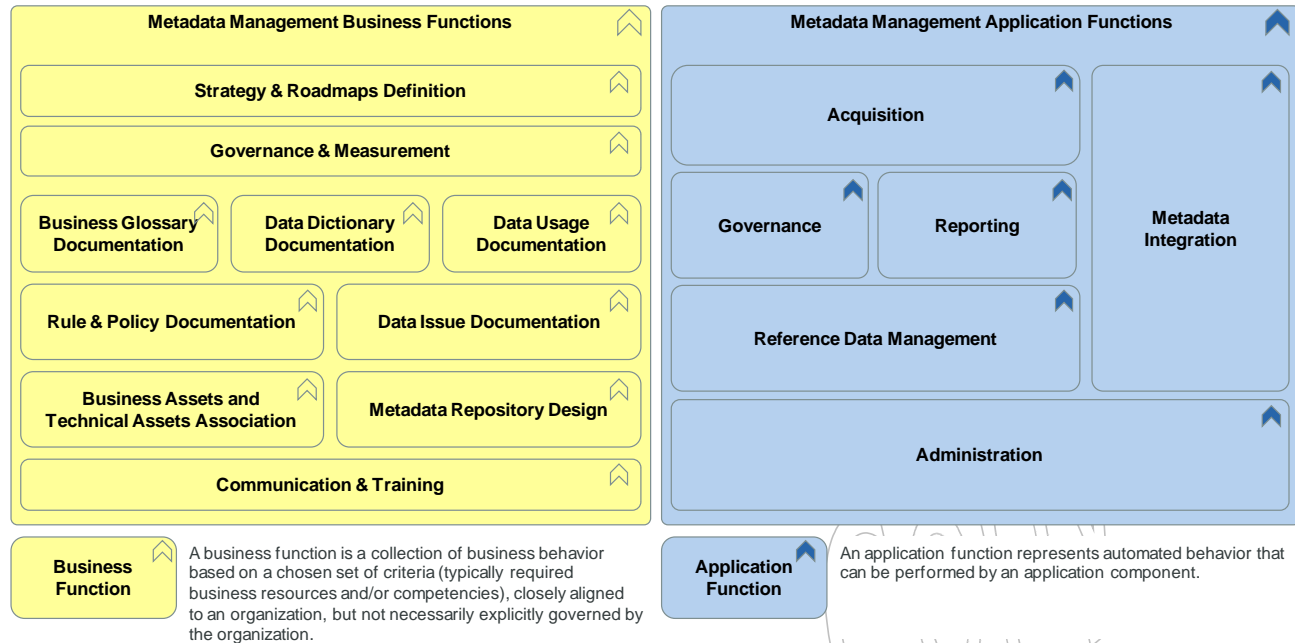## 3.1 Metadata Management High-Level Reference Model

Figure 2: Metadata Management Capability – High-Level Functions

3.2 Metadata Management High-Level Business Functions

Business Function	Description
Strategy & Roadmaps Definition	Document the vision and how it contributes to the organization's objectives. Formalize this vision through a well-defined Metadata Management strategy and roadmap.
Governance & Measurement	Govern the implementation and execution of the Metadata Management capability: • Define data domain • Define governance committee(s) (frequency, scope, participants) to enforce metadata strategy • Identify and onboard key actors • Measure the quality of the metadata documentation
Business Glossary Documentation	Capture relevant business concepts and classify them to serve data risk management (privacy, security, other regulatory requirements).
Data Dictionary Documentation	Capture physical representation of the data across the information system landscape to increase knowledge about the data (lineage, usage, lifecycle, etc.). It is advantageous to combine the Business Glossary and the Data Dictionary to facilitate collaboration and precision in the elaboration of business rules.

Business Function	Description
Data Usage Documentation	Keep an inventory of how the data is used to meet regulatory requirements.
Rule & Policy Documentation	Document the rules (data quality rules, data deletion rules, data archiving and retention rules, data transformations rules, etc.) and policies that should be enforced by the data governance.
Data Issue Documentation	Document any relevant business data issues (data quality, master data, reference data, etc.).
Business Assets and Technical Assets Association	Provide business context to the technical metadata. Metadata association provides a first level of understanding of the data lifecycle from its origin as well as the movements and transformations it has been through in the information systems.
Metadata Repository Design	Translate the business requirements into metadata requirements by defining: • The set of metadata that is relevant to be captured and stored within the metadata repository • The relationships and constraints that should be implemented between the metadata objects • The processes and workflows to create and maintain the metadata within the repository
Communication & Training	Communicate around the metadata repository value; organize training plan for end users (business and technical).

3.3 Metadata Management High-Level Application Functions

Application Function	Description
Acquisition	Provide functions to collect (create, infer, maintain), classify, and associate metadata.
Governance	Functions required to support data documentation governance: workflow validation, task assignment, metadata activities monitoring, change impact analysis.
Reporting	Support reporting and dash-boarding activities using stored metadata.
Reference Data Management	Create, maintain, and administer a normalized list of values.
Administration	Administrative tasks required to support the Metadata Management capability, such as access management, environment management, internationalization and localization, event logging management, database versioning.

Application Function	Description
Metadata Integration	Provide functions that allow the ingestion of the metadata from various sources across the organization or the distribution of the metadata to other applications.

4 Process and Methodology

4.1 TOGAF Adaptation for a Metadata Management Capability

This chapter describes a synthesis of key actions related to the TOGAF ADM phases (Preliminary, A, B, C, D, E, F) that should be performed in the context of a Metadata Management capability delivery.

| TOGAF ADM Phase	Metadata Management Recommendation
Preliminary Phase	Make sure that the data owners or any business stakeholders with clear metadata business requirements are identified to support the architecture work.
	Make sure that there is an architecture governance board well integrated with other data governance and operational data boards. This is to ensure that the architecture decisions are consistent with the organization data strategy and data management activities.
	Ensure that the outputs of the architecture work can be operationalized during the implementation phase via a design authority.
Phase A: Architecture Vision	Clarify the business drivers to document metadata and the business goals that will be achieved.
	Capture the business use-cases that will be supported by the Metadata Management capability, the priorities, and the business actor requirements.
	Assess the maturity of the organization in relation to Metadata Management practice, existing pain points if any, and expected improvements.
	Develop an overview of: • The metadata sources available within the organization • The existing communities (business and technical actors) around each metadata source • The metadata lifecycle (create, modify, delete) within each source and related Metadata Management processes of each community
	Define the scope of metadata required to deliver the business requirements and to support data governance in accordance with the organization's data strategy. Clarify the level of metadata, community, and metadata process integration required.
	Create an enterprise metadata policy including a conceptual metadata model.
	Develop the value proposition of the target Metadata Management capability taking into consideration the business use-cases, organization maturity, existing pain points, metadata sources available, and the level of integration required.

TOGAF ADM Phase	Metadata Management Recommendation
Phase B: Business Architecture	Leverage known regulatory requirements, other internal and external compliance requirements, business use-case requirements, business data issues, and available data sets to formalize metadata requirements. Define key metadata assets to be documented to support the business requirements. Capture the level of metadata quality required per scope of metadata. Identify the as-is (baseline) and to-be (target) business users of the Metadata Management capability, the related roles, and access requirements. Formalize the organization and detailed Metadata Management processes that should be implemented in the target with the support of the Data Management Office. Clarify the set of business functions that are required to deliver the business requirements and perform the baseline *versus* target gap analysis. The business reference model provided in Section 5.1 can be used as a guide for this step. Refer to Section 4.2 for further details about the metadata requirements, the metadata model, and Metadata Management processes.
Phase C: Information Systems Architectures	Extend the conceptual metadata model into the logical level (attributes and domains specified) and then map the conceptual metadata entities or attributes to the logical data model entities, records (tuples/instances), or attributes. Define the metadata integration model, clarifying: • The scope of metadata that is managed within the target metadata repository that will support data governance • Primary/secondary roles between the target metadata repository and the existing metadata sources within the organization • The collaboration contract for metadata exchange between the target repository and other sources of metadata Define the metadata collection approach that is relevant in the target: manual ingestion, pushing metadata from sources, pulling metadata from sources. Clarify the metadata inference requirements (if any) and the level of complexity pertaining to this inference. Data variety (structured, semi-structured, unstructured) and IT landscape are two important factors of metadata inference complexity. One of the key challenges about data documentation is to document the data lifecycle from its origin as well as the movements and transformations it has been through in the information systems. This is done through metadata associations and typically requires manual effort. Therefore, clarify metadata association requirements based on business requirements to concentrate the effort on the relevant scope. Clarify the set of application functions that are required to deliver the business requirements and perform the baseline *versus* target gap analysis. The application reference model provided in Section 5.2 can be used as a guide for this step. Refer to Section 4.3 for more details about this step.

TOGAF ADM Phase	Metadata Management Recommendation
Phase D: Technology Architecture	Full coverage of the metadata IT capability may require multiple technologies to be combined. Data integration vendors in general provide optimized connectors to extract the technical and operational metadata from data integration repositories, database dictionaries, and data patterns inside semi-structured and unstructured data. Data governance solutions are less focused on technical metadata extraction, but provide functions for governance such as metadata change validation workflow, compliance reporting, etc. Define whether a technology or a combination of technologies is required to deliver the Metadata Management capability. Ensure that the information exchange platform services include the relevant metadata for every exchange internal and external to the enterprise. This is particularly important to incoming data from external vendors and/or partners.
Phase E: Opportunities & Solutions	Use the application reference model provided in Section 5.2 to challenge solution providers.

TOGAF® Series Guide (2023)

TOGAF ADM Phase	Metadata Management Recommendation
Phase F: Migration Planning	First, start documenting the metadata assets in a spreadsheet focusing on a small scope to become familiar with the concepts and foster a documentation culture within the team. Metadata documentation requires manual effort: business metadata is managed by people, technical metadata exists in the data structures of the IT systems, operational metadata is generated at a specific point in time. This first step should also aim to prove the value of documenting metadata. Next, move to a centralized repository to enforce metadata governance (stewardship, ownership, validation workflow, etc.), scale metadata documentation, and ease metadata access and administration management. A centralized repository has several benefits, such as the ease of knowledge sharing and re-use by bringing the different stakeholders (business data owner, data steward, internal controllers, data protection officer, data manager, etc.) into the same repository, the ease of metadata association as all the assets are already collected into a single repository, and the centralized metadata stewardship to support metadata governance. However, a strong governance model enforced by the Data Management Office and the onboarding of the different stakeholders (business data owner, data steward, internal controllers, data protection officer, data manager, etc.) can be challenging as the repository may not be dedicated to their role and specific needs. The more mature the different stakeholders become, the more advanced requirements might emerge such as automated data discovery (e.g., record linkage) and catalog, data flow visualization, automated reverse engineering of data schemas, etc. At this stage, it can make sense to consider dedicated repositories for different families of contributors to support their specific requirements. This means that a dedicated system can be defined per logical group of assets. For example, a repository for a business glossary, a repository for data dictionaries, a repository for data quality rules documentation and implementation, etc. It is therefore possible to provide advanced features dedicated to each metadata documentation function and to support specific user interfaces customized for each role or family of users. However, a dedicated effort should be put into metadata integration to link the metadata assets together, ensure consistency between the repositories, and avoid silos. Also, collaboration between the different stakeholders can be limited.
Phase G: Implementation Governance	Organize a compliance review with Metadata Management projects (metadata repository implementation, metadata integration) to check that they respect the principles and target architecture.

4.2 Metadata Management Guidelines for a Business Architecture (ADM Phase B)

Metadata is available in multiple sources within an organization (see Figure 3). The main challenge is to bring together different kinds of metadata, from different communities, with different processes and lifecycles, in order to achieve specific business goals such as data compliance, data quality monitoring, and data accessibility.
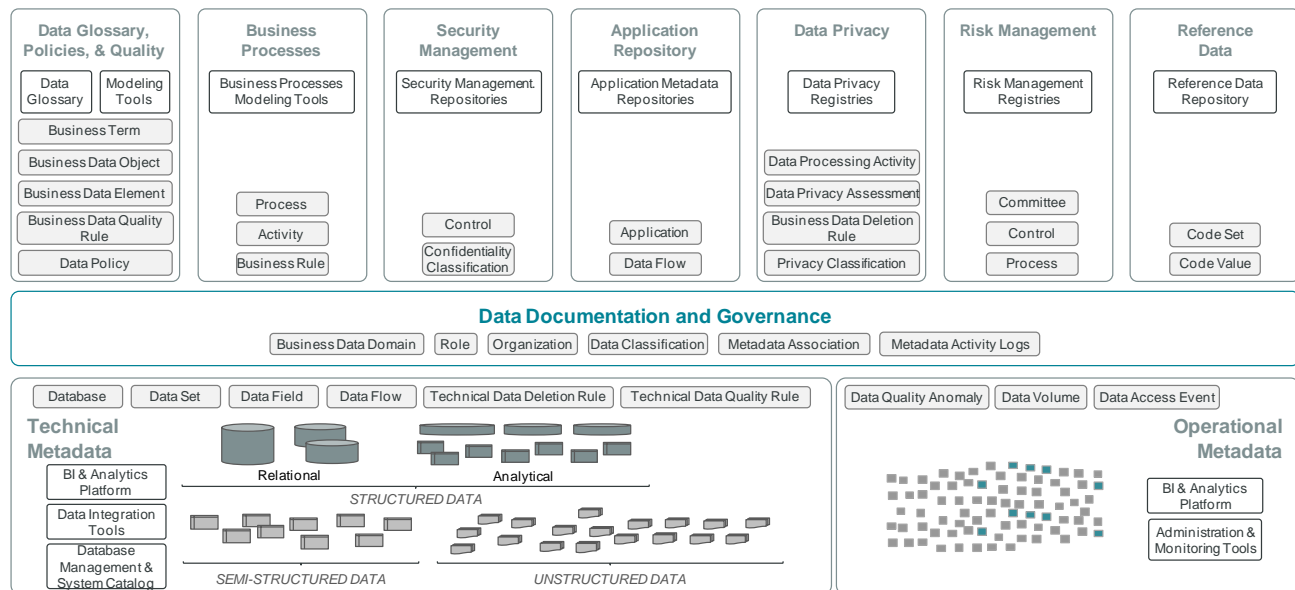
(Metadata domains have been harvested from use-cases and applied to the context of this document.)

Figure 3: Metadata Sources Illustration

Our recommendation is to engage in a step-by-step progressive approach to build the metadata documentation rather than attempt to document all the assets identified in the existing metadata sources at the same time.

4.2.1 Collect Metadata Requirements

There are different approaches to define metadata requirements (see Figure 4):

- A top-down approach starting with compliance or other business use-cases to formalize metadata requirements; for example:

  — Compliance: Regulation (GDPR, Solvency II,[2] IFRS 17),[3] internal (digital sustainability, security rules)

  — Business use-cases: get customer interactions data to improve up-sell

- A bottom-up approach based on business data issues or data set requirements; for example:

  — Data set: document existing Customer Information File (CIF) to support BI & Analytics use-cases

  — Business data issue: customers are invoiced at the wrong frequency generating frustrations and complaints; investigate root causes to tackle the issue

---

[2] Refer to: https://en.wikipedia.org/wiki/Solvency_II.
[3] Refer to: https://en.wikipedia.org/wiki/IFRS_17.
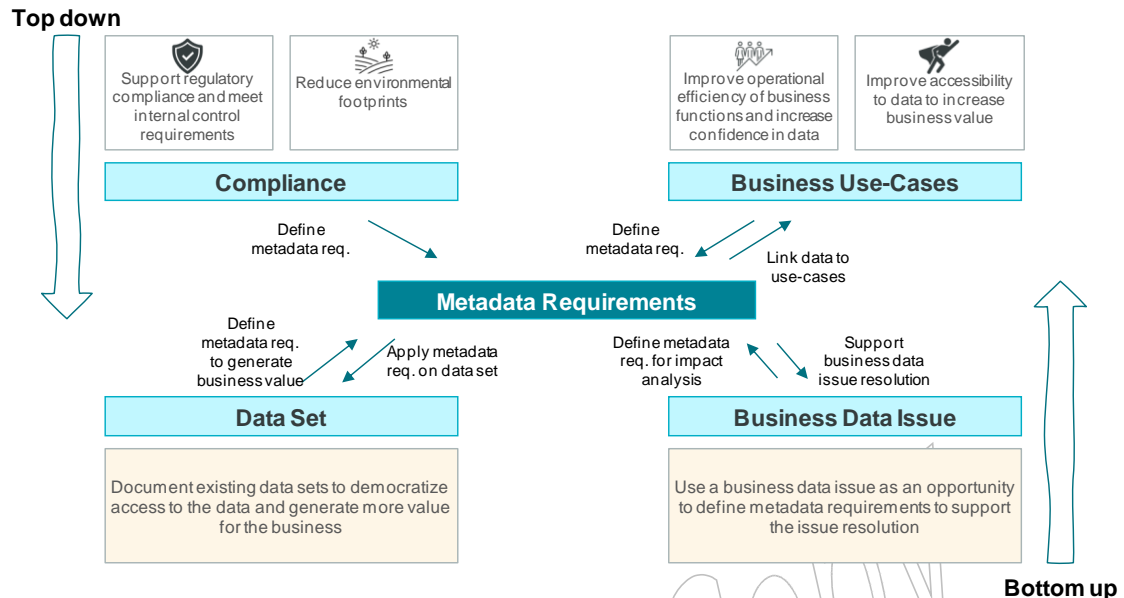
TOGAF® Series Guide (2023)

Figure 4: Metadata Requirements Collection Illustration

The top-down approach (compliance, business use-cases) often has strong business support; metadata consistency is required and guaranteed by validation processes and workflows. The bottom-up approach is more opportunistic and does not always have strong business support. For this approach it is important to show the value with concrete examples before further automation and governance. In both cases it is important to clarify the metadata requirements including the expected level of metadata quality required.

4.2.2   Formalize Metadata Management Processes

Metadata quality is enforced by processes and validation workflows that can vary per metadata scope. There are mainly two approaches (see Figure 5):

- The normative approach, which aims to guarantee precision, accuracy, integrity, and consistency of the metadata

  In this approach the metadata is created by a few, published as read-only, static (few changes overtime), and its lifecycle is managed via validation workflows. It fits well with metadata subject to regulatory constraints and central organization.

- The collaborative approach is more appropriate in the data democratization approach where there is a need to have multiple contributors maintaining the technical metadata and, when required, to associate the technical metadata with business metadata to provide a business context: created by many, edited by many, dynamic, eventual consistency
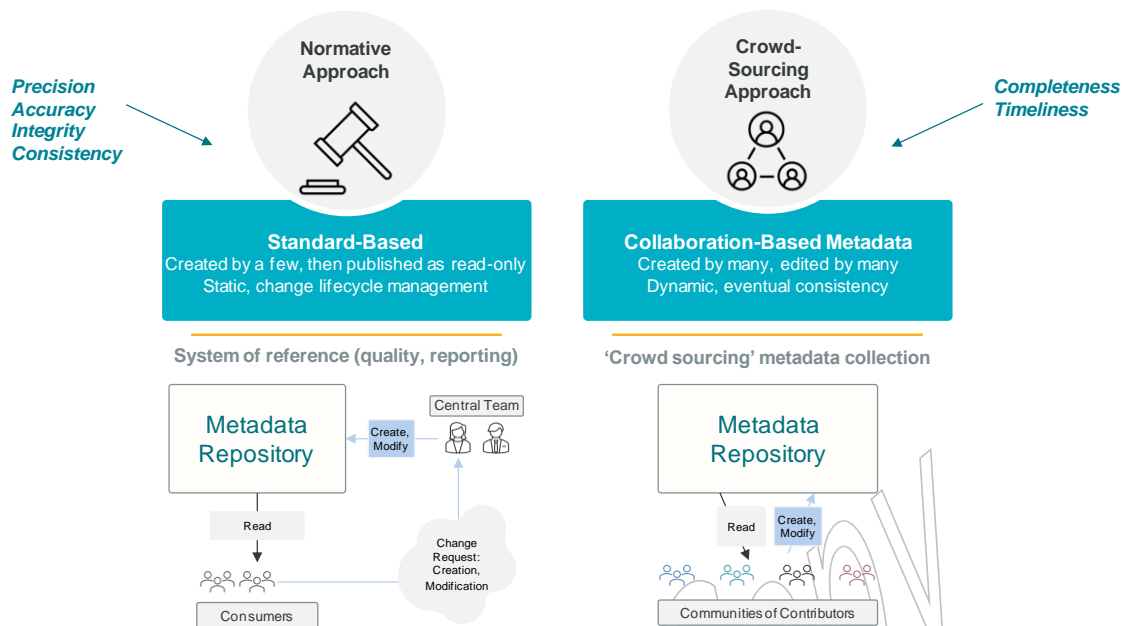
Figure 5: Metadata Quality Enforcement Approaches

4.3 Metadata Management Guidelines for Information Systems Architectures (ADM Phase C)

The metadata repository is supported by a Metadata Management IT solution, which is a combination of technologies (set of functions), a data model, and processes (metadata update and validation workflows) in order to support Metadata Management activities.
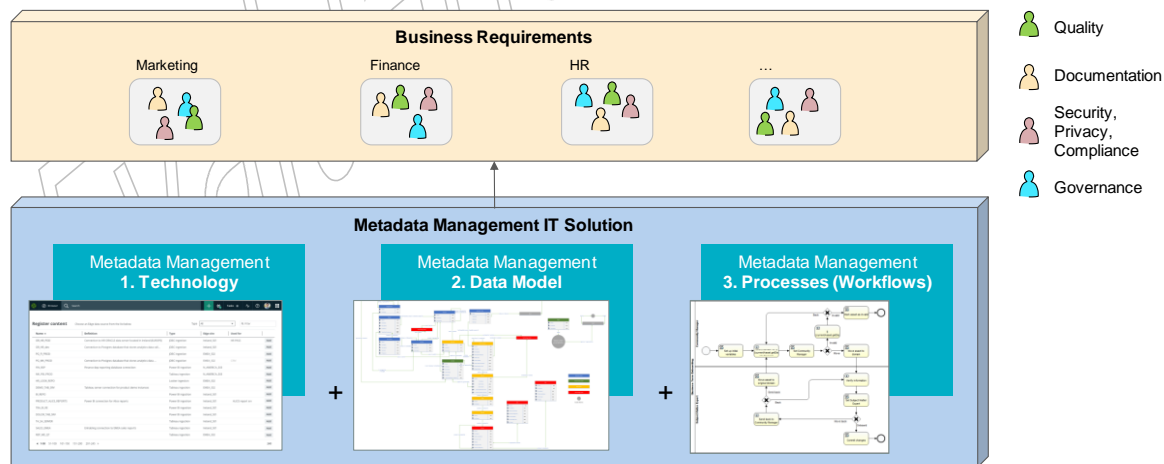

Figure 6: Metadata Management IT Solution

The metadata data model formalizes the metadata requirements by clarifying the set of metadata assets that will be documented in the metadata repository and their relationships. As described in Section 4.2, the metadata validation workflows are designed to support metadata quality requirements.

Metadata is often divided into three categories that can be summarized as follows, based on the DMBOK2:

- Business metadata focuses largely on the content and condition of the data and includes details related to data governance

- Technical metadata provides information about the technical details of data, the systems that store data, and the processes that move it within and between systems

- Operational metadata describes details of the processing and accessing of data

Business metadata is managed by people, technical metadata exists in the data structures of the IT systems, and operational metadata is generated at a specific point in time. Figure 7 shows a baseline proposal of key metadata assets (see the definitions in Appendix A) to be captured to support data governance.
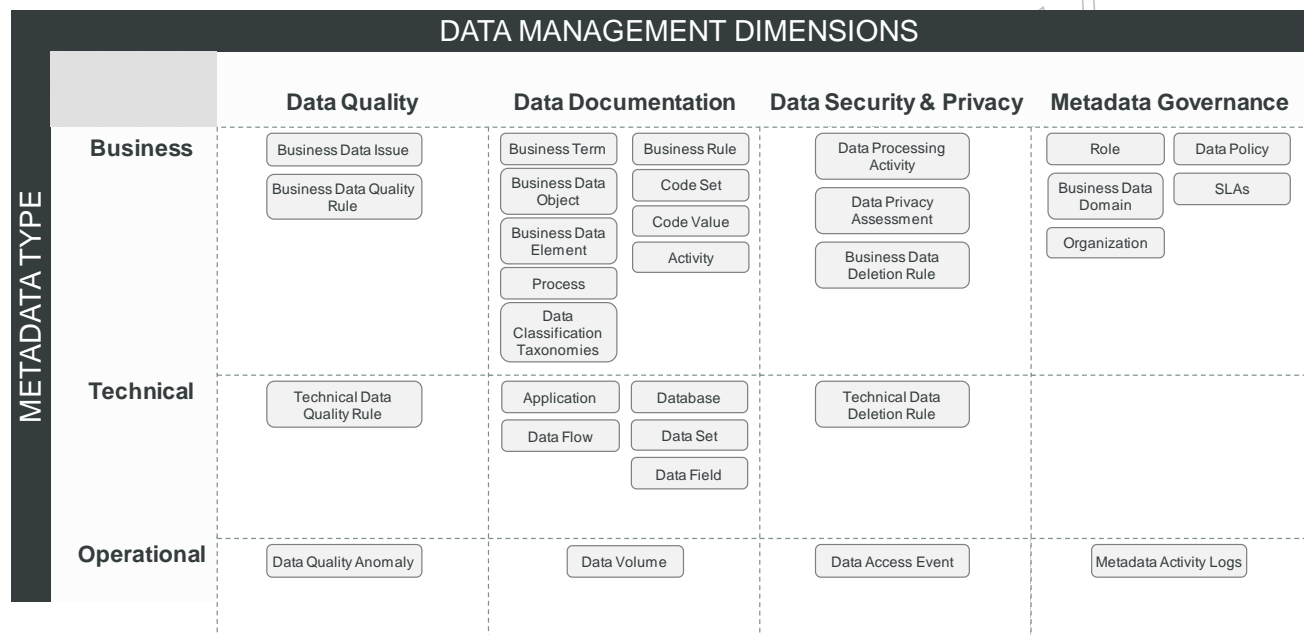

DATA MANAGEMENT DIMENSIONS

METADATA TYPE	Data Quality	Data Documentation	Data Security & Privacy	Metadata Governance
Business	Business Data Issue; Business Data Quality Rule	Business Term; Business Data Object; Business Data Element; Process; Data Classification Taxonomies; Business Rule; Code Set; Code Value; Activity	Data Processing Activity; Data Privacy Assessment; Business Data Deletion Rule	Role; Data Policy; Business Data Domain; SLAs; Organization
Technical	Technical Data Quality Rule	Application; Data Flow; Database; Data Set; Data Field	Technical Data Deletion Rule	
Operational	Data Quality Anomaly	Data Volume	Data Access Event	Metadata Activity Logs

Figure 7: Key Metadata Assets

After identifying the metadata data assets to be documented, it is important to leverage the Dublin Core™ Metadata Element Set (DCMES) at the attribute level to effectively describe each asset. For instance, the subject field is key to making sure that metadata is universally discoverable throughout the enterprise. It is useful to have the tags correspond (map) to the logical data model elements, especially the generalization-specialization constructs (e.g., Sedan is a type of Car that is a type of vehicle) so that searches (and business rules) can find all of the requisite data for whatever purpose.

Once the scope of metadata assets to be collected is defined, the metadata integration model needs to be clarified taking into consideration the broader scope of existing metadata sources and the business architecture outputs (business actors to be onboarded, metadata validation processes, data governance operating model).

As stated in ISO/IEC 19773, *"... it is a general description framework for data of any kind, in any organization, and for any purpose. ISO/IEC 19773 does not address other data management*
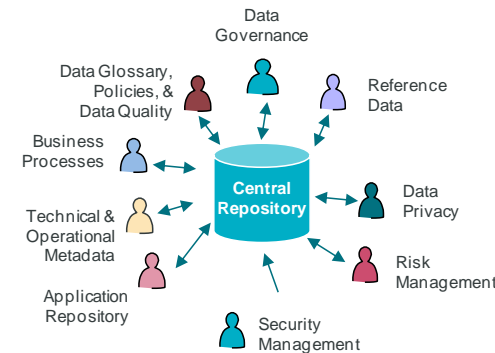
needs, such as data models, application specifications, programming code, program plans, business plans, and business policies."

The proposed framework here is to go beyond the ISO/IEC 19773 model by making explicit the key assets that are relevant to support data semantics documentation but also data management needs.

Table 1 shows three common models observed and their implications.

Table 1: Metadata Integration Model Examples

Metadata Integration Model	Illustration	Implications
Consolidation	 **Figure 8: Metadata Integration: Consolidation Model**	The metadata repository provides a unique point of access to consume the metadata available from different sources. It holds only a copy of the business metadata. The processes to create and maintain this metadata are external to the repository.
Hybrid	 **Figure 9: Metadata Integration: Hybrid Model**	For a subset of metadata, the repository holds the master copy. It is also designed to support the processes and validation workflows required to maintain this metadata. The existing tools are expected to be decommissioned after migration to the metadata repository. For the remaining metadata scope, the existing tools hold the master copy and synchronize with the metadata repository.

Metadata Integration Model	Illustration	Implications
Unified	 **Figure 10: Metadata Integration: Unified Model**	The repository unified for the full scope, both the metadata and the processes and validation workflows to maintain the metadata. An effort can also be made to provide a dedicated user experience to each community of users according to their requirements. It may involve a dedicated user interface.

Legend

 Existing Metadata Sources

The metadata exchange can be done by one of the following:

- Manual ingestion: each source exports the metadata in the agreed format and it is uploaded manually in the repository

- Pushing metadata: each source takes the responsibility to push to the metadata repository an agreed scope according to the collaboration contract (format, data structure, frequency, etc.)

- Pulling metadata: the IT solution supporting the metadata repository provides metadata agents that can be deployed directly at source to collect or infer the metadata

    This is usually the case for technical and operational metadata. Business metadata requires manual effort.
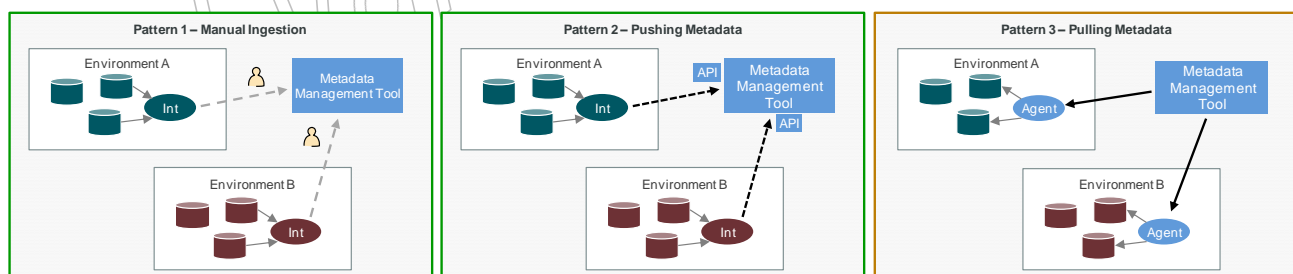

Figure 11: Metadata Integration Patterns

Each approach has some implications to be taken into consideration:

- Manual ingestion:

    — Manual work with an impact on operational efficiency

— Possible limited vendor support to metadata integration

— Initial effort to build the extractors for each metadata source

- Pushing metadata:

— Initial effort to set up the extractors for each metadata source

— Possible additional tool(s) required to support metadata integration

- Pulling metadata:

— The metadata solution may have access to all the data; indeed, sometimes the tool will have access to the data files to be able to infer the technical metadata

— The metadata agent can be an intrusive component for critical applications

— Security audit is required to ensure compliance with security policies as the metadata solution will have read access to multiple applications in the information system landscape

The complexity to collect technical and operational metadata depends on the data variety and IT landscape, with a main difference when the data schema and operational metadata are defined by-design *versus* when the metadata is extracted *a posteriori* by leveraging analytical techniques (metadata inference, machine learning, etc.).
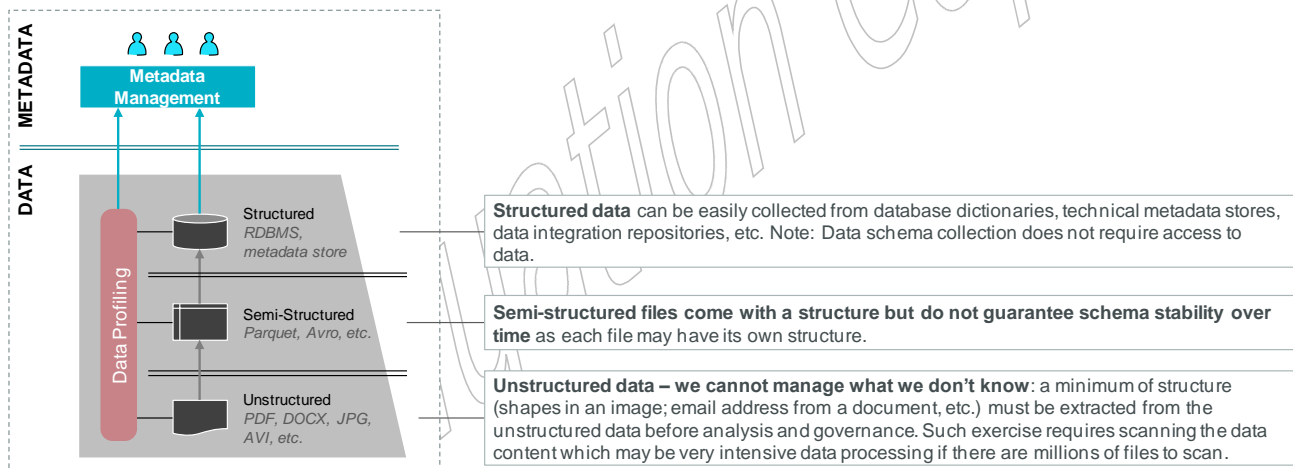

Figure 12: Technical Metadata Inference

- Structured data can be easily collected from database dictionaries, technical metadata stores, data integration repositories, etc.

    Note that data schema collection does not require access to data.

- Semi-structured files come with a structure but do not guarantee schema stability over time as each file may have its own structure

- Unstructured data: we cannot manage what we do not know

    A minimum structure (shapes in an image; email address from a document, etc.) must be extracted from the unstructured data before analysis and governance. Such an exercise

requires scanning the data content which may involve intensive data processing if there are millions of files to scan.

Multiple techniques/libraries/tools exist to extract technical and operational metadata from data; the tool selection depends on the kind of metadata that is needed, the complexity to extract the information, and the performance/volume of data to scan.

To a certain extent, managing metadata for unstructured data means first capturing the technical metadata; multiple structures can be extracted over time from the same unstructured data set. The main challenge is to maintain the associations with other metadata to put this unstructured data in context:

- Business meaning: association with the business terms, business data objects, business data elements

- Context of use: association with data processing activities

- Data quality requirement: association with business data quality rule, technical data quality rule, data quality anomaly

- Privacy requirement: association with data deletion rule, etc.

5 Reference Models

5.1 Metadata Management Reference Model – Detailed Business Functions

Figure 13 shows detailed business functions of the Metadata Management capability.
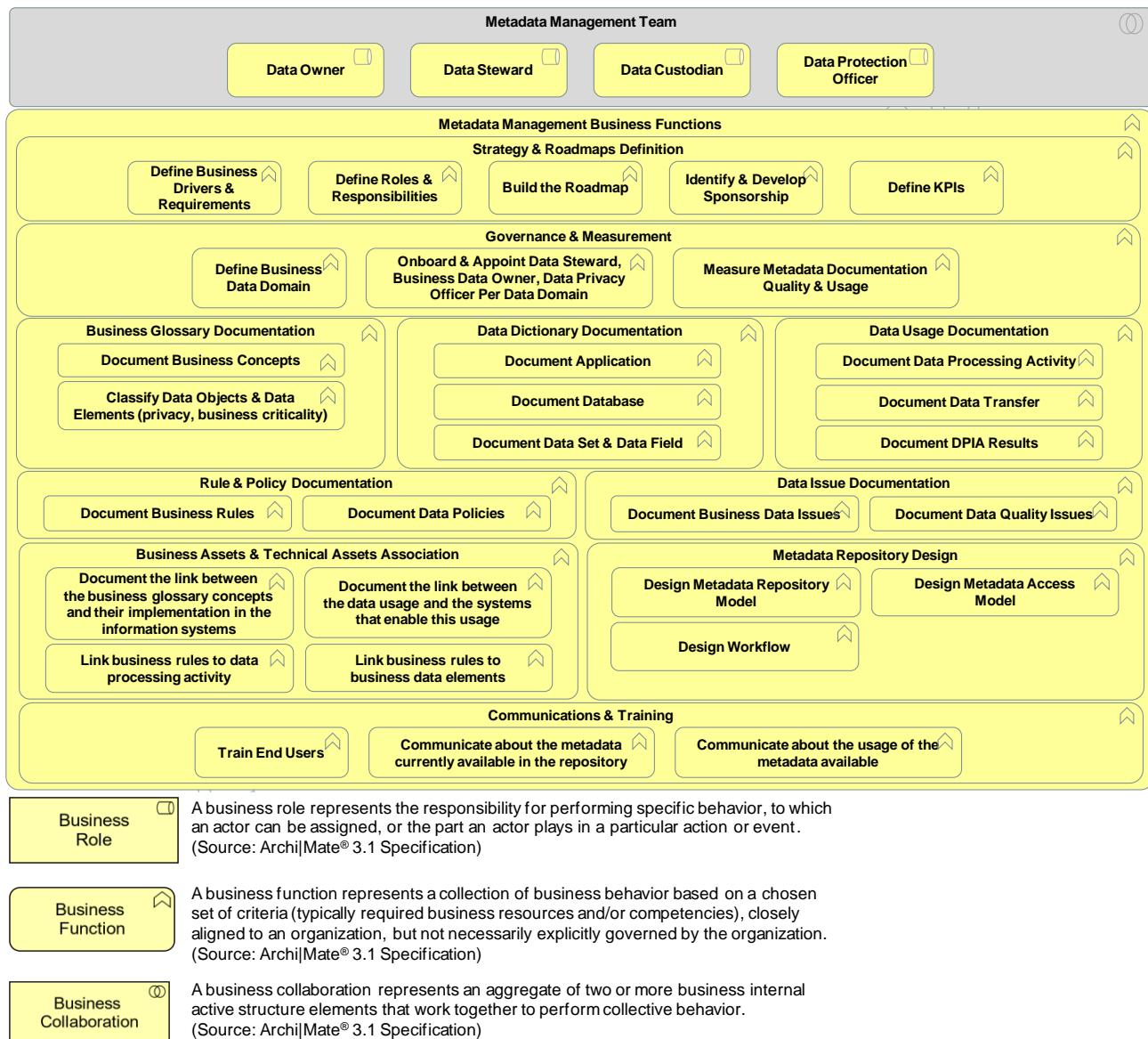

Figure 13: Metadata Management Detailed Business Functions

Table 2 provides detailed descriptions of the business functions shown in Figure 13.

Table 2: Detailed Business Functions

Business Function	Description
Strategy & Roadmaps Definition	
Define the business drivers and requirements	Describe the vision, context, and strategy that the Metadata Management capability has to support. Capture and document business use-cases and priorities. Maintain this list over time.
Define roles and responsibilities	Define the roles and responsibilities of each actor (business, IT) in the build and run of the Metadata Management capability.
Build the roadmap	Build and maintain the roadmap. Have a list of documented and prioritized business requirements to implement the Metadata Management capability. Define the target and overall journey.
Identify and develop sponsorship	Have the IT and business sponsors initiated in Metadata Management platform governance.
Define KPIs	Define Key Performance Indicators (KPIs) to measure and steer the effectiveness of the Metadata Management platform for delivering business goals.
Governance & Measurement	
Define business data domain	Define the data domain for governance purposes. A domain is a consistent grouping of data and metadata assets. It allows stewardship and ownership of the data and metadata assets.
Onboard and appoint data steward, business data owner, data privacy officer per data domain	Identify the key stakeholders that will maintain and govern the metadata repository.
Measure metadata documentation quality and usage	Define the KPIs to monitor the quality and usage of the metadata in the repository.
Business Glossary Documentation	
Document business concepts	Define the names and semantics of business concepts such as business term, business data object, business data elements, or any other meaningful business concept.
Classify data objects and data elements (privacy, business criticality)	Classify data for security and compliance purposes. Identify valuable data assets for the business (critical data, "crown jewels").

Business Function	Description
Data Dictionary Documentation	
Document application	• Describe and maintain data flows between applications • Identify data location and data collection point • Document and maintain applications list and related IT managers • Contacts (including applications managed by third parties) • Collect and maintain SLAs for applications where critical data elements are stored or processed
Document database	Document the database system where the data is stored. Identify the related data custodians to be included in operational data governance. The notion of the database can be understood here as any means to store a collection of data.
Document data set and data field	Capture within each database system relevant data set and data field for the purpose of making data documentation available for everyone to foster re-use and enforce data governance and compliance.
Data Usage Documentation	
Document data processing activity	Document data usage and manipulation such as collection, recording, organization, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction.
Document data transfer	• Document the legal conditions binding the disclosure by transmission of any personal data to a third party in the context of data usage • Document the third parties receiving personal data in the context of data usage
Document DPIA results	Document the Data Privacy Impact Assessment (DPIA) results when personal data processing is involved in the data usage and manipulation.
Rule & Policy Documentation	
Document business rules	Document the data deletion, data quality, and other relevant business rules that apply to the data.
Document data policies	Document the data policies that should be enforced by the data governance board(s).
Data Issue Documentation	
Document business data issues	Document business problems and related business impacts (quantitative and qualitative).
Document data quality issues	Document the data quality problems identified during the information lifecycle analysis. Document the costs of poor data quality and the costs of improving data quality.

TOGAF® Series Guide (2023)

Business Function	Description
Business Assets & Technical Assets Association	
Document the link between the business glossary concepts and their implementation in the information systems	Document the link between business data object/data element/business term and the application/database/data set/data field.
Document the link between the data usage and the systems that enable this usage	Document the link between data processing activity and the application/database involved in the data usage.
Link business rules to data processing activity	Document the business rules involved in the data usage and/or manipulation.
Link business rules to business data elements	Document the data elements to be evaluated by each business rule.
Metadata Repository Design	
Design metadata repository model	Design the data model underlying the metadata repository.
Design metadata access model	Define, manage, and control the way in which metadata can be accessed.
Design workflow	Define the workflow and processes to review and validate the metadata to guarantee consistency and integrity of the metadata or to ease metadata collection (minimum of controls at metadata entry level).
Communication & Training	
Train end users	Train end users to ease adoption of the Metadata Management processes and tools.
Communicate about the metadata currently available in the repository	Share with a broad audience the scope of metadata documented and how it can be used to generate value for the organization.
Communicate about the usage of the metadata available	Share with a broad audience the actual usage of the metadata documented to generate awareness.

5.2 Metadata Management Reference Model – Detailed Application Functions

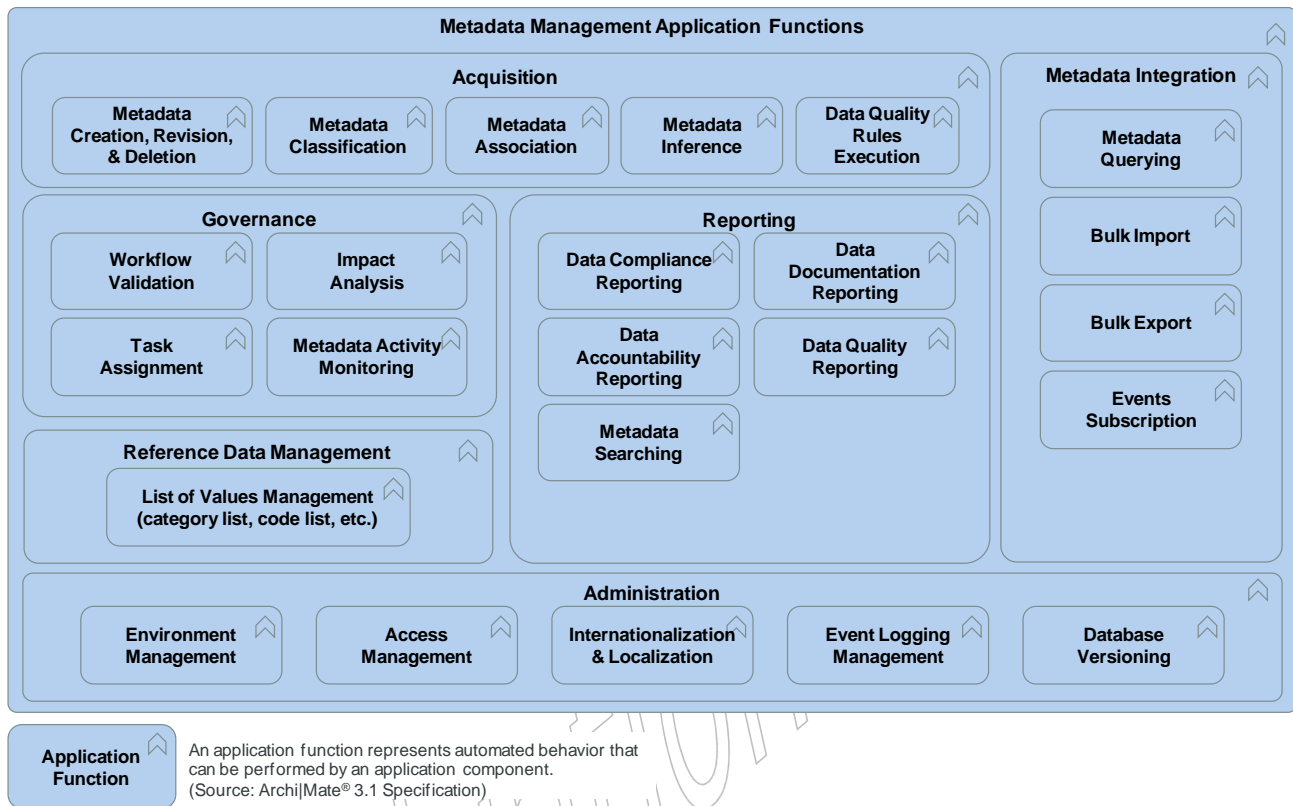Figure 14 shows detailed business applications of the Metadata Management capability.

Figure 14: Metadata Management Detailed Application Functions

Table 3 provides detailed descriptions of the application functions shown in Figure 14.

Table 3: Detailed Application Functions

Application Function	Description
Acquisition	
Metadata creation, revision, and deletion	Create and maintain metadata objects such as (but not limited to): • Data glossary: — Business data object — Business data element — Business term — Deletion/retention date • Data dictionary: — Environment — Application (and related technology) — Data set (database)

TOGAF® Series Guide (2023)

Application Function	Description
	— Table — Column • Data usage: — Data processing activity — Data recipient — Data transfer — Data protection impact assessment — Data provider — Data access and privileges — Add URL links to external documentation reference • Data classification: — Data confidentiality classification — Data privacy classification — Business value classification — Define custom metadata classification dimension • Rule and policy: — Data policy — Business data quality rule • Issue: — Business data issue including related business problem and business impact — Data quality issue reported following data quality rules execution • Governance assets: — Logical group of metadata, not limited to, but such as subject area (e.g., policy, claim, product), business area (e.g., home, motor, travel, health), geographical area (e.g., Italy, France, Spain, Paris) — Data domain as the combination of two or more groups of metadata — Data steward, data owner, data manager, data privacy officer, data custodian, application owner • Data quality rules execution: — Technical data quality rule
Metadata classification	Capability to classify metadata: • Data confidentiality classification • Data privacy classification • Define custom metadata classification dimension
Metadata association	Linking metadata assets to describe the full value chain of the data and data lineage.

Application Function	Description
Metadata inference	Documentation of metadata through an automated process such as machine learning or Artificial Intelligence (AI) algorithm; for example (but not limited to): • Automatically process available data to discover metadata relationships • Identify metadata from patterns (email, phone number, credit cards, etc.) • Identify relations between data (same values, repartition, etc.) • Identify metadata from dictionaries (company name, city, first name, last name, synonyms, etc.)
Data quality rules execution	Execute and capture the results of data quality rules defined in the metadata repository.
Governance	
Workflow validation	Create workflows, visualize status and tasks.
Task assignment	Assign tasks to people.
Impact analysis	View change impact. (For example, if I change the business definition, how many technical definitions are based on it ?.)
Metadata activity monitoring	Report on metadata usage.
Reporting	
Data compliance reporting	Visualize end-to-end data lifecycle (data origin, data transformation, data transfer, data classification) to support regulatory requirements.
Data documentation reporting	• Visualize metadata completion within the repository • Visualize metadata associations • Measure metadata quality and share it within the organization
Data accountability reporting	Reporting capability to visualize who is accountable for the data within the organization.
Data quality reporting	Measure data quality through time (history and present) and share it within the organization.
Metadata searching	• Multi-criteria search on metadata • Full-text search on metadata (Google™ like) • Search on metadata with auto-completion
Reference Data Management	
List of values management (category list, code list, etc.)	• Create and maintain normalized list of values (category list, code list, etc.) • Administrate reference data (list of values)

Application Function	Description
Administration	
Environment management	Manage multiple environments, each with its own governance (e.g., workflow validation) and administration (e.g., access management, database versioning) settings.
Access management	• Manage, create, delete, update metadata rights access • Create user, group, and give authorizations on metadata depending on roles (including batch or user role)
Internationalization and localization	Provide the means to adapt the Metadata Management tool to different languages and local requirements.
Event logging management	Log any event in the system. Keep history, source, and author of metadata modifications.
Database versioning	Support versioning of the metadata database.
Metadata Integration	
Metadata querying	Query metadata assets in the repository using native SQL or any other query language or interface (e.g., API).
Bulk import	Import multiple metadata in the metadata repository from a file (e.g., CSV, JSON, XML) or via an application interface (e.g., API).
Bulk export	Export multiple metadata from the repository in a non-proprietary file format (e.g., CSV), via an API or any other application interface.
Events subscription	Listen to one or multiple events from an event broker.

Information Architecture: Metadata Management   29

A   Metadata Asset Definitions

Category	Name	Description
Business	Activity	A business task or group of business tasks that are undertaken by the business to achieve a well-defined goal. (Source: Integrated Architecture Framework (IAF))
Business	Business Data Deletion Rule	A business rule related to the removal of a set of business data elements that can be executed to meet regulatory or business requirements. It is typically the responsibility of business data owners, as part of their data governance responsibilities, to define data deletion rules for the data domains they own. They should seek input from other data actors (business data owner, data steward, internal controllers, data protection officer, data manager, etc.) as they define these rules. These are an essential component for an ethical use of the data and build trust with the data providers and consumers.
Business	Business Data Domain	A combination of one subject area, one business area, and (when relevant) other dimensions (e.g., geographical area, organization) that constitutes a consistent grouping of data and/or metadata for data governance purposes. Fine-grained data domains usually focus on governance of data (e.g., the data domain in charge of the data quality of a claim) or highly specific metadata (e.g., one data element or data quality rule that makes sense in a restricted context). Data domains defined at macro level usually focus on governance of shared metadata (e.g., one data element or data quality rule that is widely used around the enterprise).

Category	Name	Description
Business	Business Data Element	A fundamental, conceptual unit of data. Often, the terms "variable" and "field" are used synonymously to mean a data element (e.g., Individual Contract Status, Claim Event Date, and Employee First Name). A data element is comprised of three components: • A data object – the idea, person, or thing that is of interest and is being described (e.g., the Customer) • A property – the property is the characteristic of the data object that is of interest (e.g., Customer Preferred Language) • A value domain – establishes the permissible values that can be used to represent the data element A value domain has a data type (e.g., boolean, decimal, and integer) and, optionally, a minimum length, a maximum length, and a decimal precision. A value domain can be enumerated (specified through a list of values of a least two individual permissible values) or non-enumerated. For example, the Customer Preferred Language is represented by a list of language codes from ISO 639-1:2002). Data elements can be thought of as belonging to one of the following groups: • Input: if a data element is of the input type, this indicates that the data is used or stored essentially as it is collected, without major modification; most data elements are of this type • Derived: one or more data elements which are modified in some way; for example, multiple data elements might be aggregated together for output or logical purposes See Table 2 for more information.
Business	Business Data Issue	Describes a business problem occurring as the result of a data issue and the impact of this problem, either qualitative or quantitative. Also describes the nature of the business impact such as financial, regulatory, legal, reputational, operational, etc.
Business	Business Data Object	A key idea, person, or thing that provides context for the data collected through the enterprise. Data objects aid understanding and are useful for defining objects against which multiple data elements are collected. For example, "contact" may be defined as a data object, which aids in the application of data elements, including contact date, contact delivery mode, and contact customer present status. The use of data objects reduces the repetition of large sections of text across multiple data elements.

Category	Name	Description
Business	Business Data Quality Rule	A business rule related to a business data element or a set of business data elements that can be tested to determine if the business data elements meet the data quality requirements. It is typically the responsibility of business data owners, as part of their data governance responsibilities, to define data quality rules for data domains they own. They should seek input from other data actors (business data owner, data steward, internal controllers, data protection officer, data manager, etc.) as they define the data quality rules. The definition, design, and implementation of data quality rules are at the heart of a solid data quality management program. Examples: Date of Birth must be a valid date. Surname must not be empty. Email address must match patternuser@domain.
Business	Business Rule	A rule that is practicable and that is under business jurisdiction. Derived from business policy. (Source: Object Management Group (OMG))
Business	Business Term	Words or phrases used to describe a thing to express a concept in language or branch of business.
Business	Code Set	Defines a list of values that enumerate the possible values for a data element. We can, for instance, define a code set for the business data element "Gender" of the business data object "Person" with the following possible values: male, female, unknown.
Business	Code Value	An instance or a single value of a code set.
Business	Data Classification Taxonomies	An important activity of the data documentation which requires collaboration between at least four professional groups: Information Security, Data Privacy, Information Risk Management, and Data Management. Consequently, it has impacts on security measures, on regulatory constraints on personal and sensitive data, and on data management and quality priorities. It has increased in importance following regulatory requirements such as GDPR. Data classification taxonomies are applicable labels that should be used to classify each data element. Even if they are essentially code set and code values, we have chosen to highlight data classifications taxonomies as they require alignment between multiple stakeholders.
Business	Data Policy	A set of rules, guidelines, standards, and patterns that are applied to the data.
Business	Data Privacy Assessment	A process designed to identify risks related to the use of personal data and define mitigation plans to be compliant with regulatory requirements.

TOGAF® Series Guide (2023)

Category	Name	Description
Business	Data Processing Activity	The collection and manipulation of data elements to produce meaningful information. It represents any operation or set of operations which is performed on data or on sets of data, whether by automated means, such as collection, recording, organization, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction. In the context of GDPR, the record of data processing activities is the subset of data processing activities which collect and manipulate personal data.
Business	Organization	Used to describe a group of stakeholders within the company that are bounded by a common purpose. It often describes organizational units such as branches, departments, teams, etc. Used to animate the operational governance within the metadata repository. It may be required in some context to specialize the business data domain per organization units.
Business	Process	A set of interrelated or interacting activities that use inputs to deliver an intended result. (Source: ISO 9000:2015)
Business	Role	A business role is the responsibility for performing specific behavior, to which an actor can be assigned, or the part an actor plays in a particular action or event. (Source: ArchiMate Specification)
Business	Service-Level Agreements (SLAs)	Essential components to operate data governance. They describe within each business context the applicable level of service required to support business use-cases. For example, a data quality rule on an individual email address may have different acceptable thresholds depending on the business operating model: B2C or B2B2C. In the second scenario (B2B2C), the quality of the email address may be irrelevant for the company if they do not have any direct relationship with the customer.
Operational	Data Access Event	Any event recorded in the information system to track access to a subset of data by a user account.
Operational	Data Quality Anomaly	An issue detected following the execution of a technical data quality rule within the information system.
Operational	Data Volume	Measure of the quantity of data within a data set or database.
Operational	Metadata Activity Logs	A set of events related to any activity undertaken within the metadata repository.
Technical	Application	A classification of computer programs designed to perform specific tasks, such as word processing, database management, or graphics.

Category	Name	Description
Technical	Data Flow	Describes the data flows between applications.
Technical	Database	An organized collection of data, generally stored and accessed electronically from a computer system. The Database Management System (DBMS) is the software that interacts with end users, applications, and the database itself to capture and analyze the data. (Source: Wikipedia)
Technical	Data Field	A place where data can be stored. Commonly used to refer to a column in a database or a field in a data entry form or web form. (Source: Wikipedia)
Technical	Data Set	A collection of related, discrete items of related data that may be accessed individually or in combination or managed as a whole entity. (Source: TechTarget)
Technical	Technical Data Deletion Rule	The physical executable component (in most cases an SQL SELECT statement) of the business data deletion rule description. It is the responsibility of IT to design and develop data deletion rules that can then be implemented, executed, and monitored within batch or near real-time applications across the enterprise. Output of a data deletion rule is the list of all business references (e.g., contract references/numbers, employee references/numbers, vendor references/numbers, etc.) that do not comply with the rule (e.g., the list of all customers contact details that should not be kept in the BI & Analytics platform).
Technical	Technical Data Quality Rule	The physical executable component of the business data quality rule. It is the responsibility of IT to design and develop data quality rules that can then be implemented, executed, and monitored within batch or near real-time applications across the enterprise. Output of a data quality rule is the list of all business references (e.g., contract references/numbers, employee references/numbers, vendor references/numbers, etc.) that do not comply with the rule (e.g., the list of all contract references for which the contract start date is not a valid date).

Acronyms & Abbreviations

ADM	Architecture Development Method
AI	Artificial Intelligence
API	Application Program Interface
B2C	Business-to-Consumer
B2B2C	Business-to-Business-to-Consumer
BI	Business Intelligence
CIF	Customer Information File
CSV	Comma-Separated Value
DBMS	Database Management System
DCMES	Dublin Core Metadata Element Set
DCMI	Dublin Core Metadata Initiative
DPIA	Data Privacy Impact Assessment
GDPR	General Data Protection Regulation
IAF	Integrated Architecture Framework
JSON	JavaScript Object Notation
KPI	Key Performance Indicator
MDR	Metadata Registries
OMG	Object Management Group
RDBMS	Relational Database Management System
SLA	Service-Level Agreement
SME	Small and Medium Enterprise
SQL	Structured Query Language
URL	Uniform Resource Locator
XML	Extensible Markup Language

Index