

# Data Governance Practices for Generative AI Powered Organizational Knowledge Management Systems Using Retrieval Augmented Generation

Tilman Friedrich  
University of Lausanne  
[tilman.friedrich@unil.ch](mailto:tilman.friedrich@unil.ch)

Karl Akbari  
Hochschule Bremen  
[karl.akbari@hs-bremen.de](mailto:karl.akbari@hs-bremen.de)

Daniel Fu"rstenau  
Freie Universita"t Berlin  
[daniel.fuerstenau@fu-berlin.de](mailto:daniel.fuerstenau@fu-berlin.de)

## Abstract

*This study examines how data governance supports the success of generative AI-based Knowledge Management Systems (KMS) using Retrieval-Augmented Generation (RAG) in large enterprises. Drawing on a multi-case study methodology, the research identifies 17 distinct data governance practices and synthesises them into a conceptual framework that theorises their contribution to KMS success. The adoption of these practices is shaped by the dynamically evolving technological affordances of generative AI and RAG, as well as the contextual challenges posed by the predominantly ingested semi-structured and unstructured textual data. While the identified practices enable value-add, they also introduce strategic trade-offs, particularly in balancing data protection and expected benefits. This study contributes to the evolving discourse on data governance by extending its scope beyond structured data and highlighting its dynamic, context-sensitive role in AI-enabled KMS.*

**Keywords:** Data Governance, Retrieval-Augmented Generation (RAG), Knowledge Management Systems, Unstructured Data

## 1. Introduction

Retrieval-Augmented Generation (RAG) enables dynamic integration of external knowledge into content generation processes of generative AI, enhancing the relevance and accuracy of outputs (Lewis et al., 2020). This capability has prompted large enterprises to explore RAG-powered Knowledge Management Systems (KMS) for more effective handling of unstructured organizational knowledge.

While RAG-based KMS offer semantic search and contextual reasoning capabilities, they also introduce data-related risks—including issues of quality, trust, and compliance. These challenges fall within the scope of data governance, broadly defined as the specification of rights, responsibilities, and controls over data assets (Abraham et al., 2019; Otto, 2011).

Data governance aims to maximise data value while mitigating risk—a tension increasingly recognised in the literature (e.g., Tallon et al., 2013; Vial, 2023; von Grafenstein, 2022). In the context of RAG-enabled KMS, both objectives are essential to the system's success. This raises the research question:

*Which data governance practices contribute to the success of generative AI-powered knowledge management systems using RAG in large enterprises?*

To answer this question, we pursue three complementary research objectives: (1) *identify relevant practices*, (2) *explore their contextual drivers*, and (3) *assess their implications*. This paper addresses these research goals by developing a conceptual framework that identifies key data governance practices, explains their contextual drivers, and links them to KMS success. This is grounded in a multi-case study that was conducted across six DAX40 (Germany's stock market index of 40 major companies) firms that implemented a RAG-enhanced KMS.

## 2. Background

### 2.1. Generative AI and RAG

Generative AI can employ retrieval-augmented generation (RAG) to enable the integration of proprietary data, real-time information access, source

citation, and the implementation of fine-grained access controls (Gao et al., 2024; Lewis et al., 2020). Initially introduced dominantly through *embedding-based retrieval* methods, RAG architectures are now increasingly incorporating *graph-based* approaches, which promise improved performance and contextual reasoning (Peng et al., 2023). Structurally, RAG systems are composed of two core components: the *data pipeline*, responsible for processing and preparing data from external sources for retrieval, and the *inference chain*, which spans the end-to-end workflow from user input to response generation.

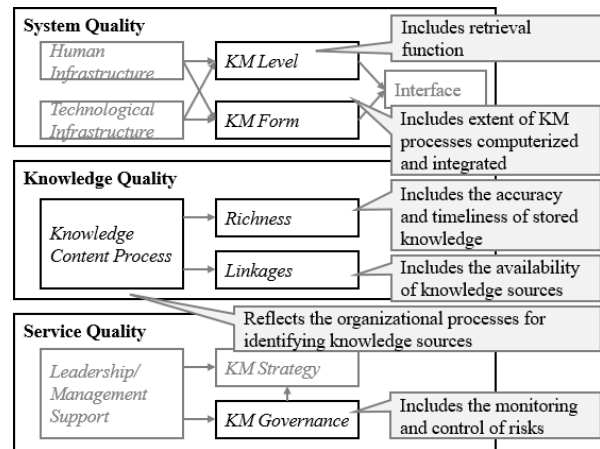
## 2.2. Organizational Knowledge Management Systems

Based on the foundational work of Alavi and Leidner (2001) and Meso and Smith (2000), Knowledge Management Systems (KMS) are defined as *IT-based systems designed to support the organizational processes of knowledge creation, storage/retrieval, transfer, and application*. To conceptualise the success of KMS, the Jennex–Olfman model (Jennex and Olfman, 2006) in its latest iteration (Jennex, 2020) is used. It is a knowledge management explication of the DeLone and McLean IS success model (DeLone and McLean, 2003), redefining information quality as knowledge quality and specifying constructs within the three quality dimensions. This study applies this model, referring to it as the KMS Success Model, to evaluate how data governance practices impact generative AI-powered KMS using RAG. The core assumption is that a data governance practice contributes to KMS success when it strengthens a construct within the quality dimensions of the KMS success model proposed by Jennex (2020). Enhancing system, knowledge, or service quality is expected to improve user satisfaction and usage intent, ultimately increasing net benefits. Consistent with the third research objective, we assess secondary impacts (e.g., multidimensional effects, implementation costs) to delineate additional implications, including potential trade-offs.

Figure 1 highlights the key concepts of the model relevant to this research. Although not explicitly modeled, security is addressed via governance practices within the *KM Governance* construct, in line with prior recommendations (Jennex and Durcikova, 2014).

## 2.3. Data Governance Practices

Although definitions of data governance differ across the literature, a widely accepted understanding centres on the specification and formalisation of decision rights and accountabilities over organizational



**Figure 1. Quality dimensions and their constructs**

data assets (e.g., Abraham et al., 2019; DAMA International, 2009; Otto, 2011). Contemporary conceptualisations emphasise its role in reconciling competing interests (von Grafenstein, 2022), balancing the maximisation of data value with associated risks and costs in the context of digital innovation (Vial, 2023). The conceptual data governance framework proposed by Abraham et al. (2019) uses the differentiation between *structural* (e.g., define roles and responsibilities), *procedural* (e.g., decision-making related to data processes), and *relational* mechanisms (e.g., alignment and collaboration between stakeholders) along three dimensions of organizational, data, and domain scope. Drawing on this framework, this study defines data governance practices as *initiatives situated within data governance domains, enforced through data governance mechanisms, including their technical implementations*.

To understand current insights about data governance practices for generative AI systems using RAG, literature was screened continuously between December 2024 and April 2025. Given the field’s novelty and rapid development, peer-reviewed publications are rare; therefore, preprints and grey literature were included. For this, academic databases (e.g., ACM, AIS, Web of Science), preprint libraries (ArXiv) and practitioner sources (e.g., vendor documentation, blogs) were searched based on search strings encompassing terms around AI (e.g., GenAI, LLM, foundation models), RAG (e.g., retrieval-augmented, embedding-based, vector), and KMS (e.g., knowledge systems, wiki, knowledge base). Because of limited results on KMS, relevant literature from adjacent RAG-enabled domains was added. The aggregated results with one guiding reference because of space constraints are displayed in Table 1. The complete literature search is documented in the appendix.

Data Governance Practice	Governance Mechanism	KMS Success Dimensions	Guiding Reference
Measures to assess, monitor, ensure, and enhance data quality, including technical validation, manual review, quality metrics, data handling guidelines, and KPIs	Procedural	Knowledge and Service Quality	Microsoft, 2024
Guidelines to optimise content for RAG (explain graphics in text, add summaries, simplify tables, ensure consistency)	Procedural	System Quality	Packowski et al., 2024
Metadata governance via confidentiality and licence metadata; use of data catalogues to extend RAG metadata	Procedural	Knowledge and Service Quality	Ramachandran, 2024
Access controls for embeddings using ACLs, complemented by GenAI-based contextual access decisions alongside RBAC/PBAC	Procedural	Service Quality	Akkiraju et al., 2024
Privacy and sensitivity management through data classification, PII handling, and data masking	Procedural	Service Quality	Haridasan, 2024
Monitoring data use via audit logs, dashboards, and response-toxicity scoring	Procedural	Service Quality	Panda and Mukherjee, 2025
Improving accessibility and oversight through API governance policies, cataloguing, labelling, and central repositories	Procedural	Knowledge, System, and Service Quality	Pahune et al., 2025
Training initiatives to raise awareness on data privacy, confidentiality, and effective system usage	Relational	Service Quality	Papagiannidis et al., 2023

**Table 1. Data governance practices for RAG-enhanced LLMs in literature**

### 3. Methodical Approach

Given the novelty of enterprise RAG applications and the limited empirical research available, this study adopts a qualitative multiple-case study design, following the roadmap of Eisenhardt (1989).

**Case selection** based on three criteria: the use of generative AI in the KMS, the application of RAG to access internal data, and implementation within a large enterprise, namely DAX40 corporations. Although many systems were in early or proof-of-concept (PoC) stages due to the recency of RAG adoption, they were included to capture emergent governance mechanisms.

Six cases were selected through targeted web and LinkedIn searches, complemented by referrals from personal networks. While the cases differed in application domain, design, and maturity, they shared a common reliance on semi-structured textual data (Table 2), ensuring comparability of governance practices while allowing for contextual variation in their adoption and impact.

**Data collection** was primarily gathered through semi-structured interviews, guided by the five-step framework of Kallio et al. (2016) and refined iteratively. Interviews began with open-ended questions on which practices were perceived to contribute to each KMS success dimension, followed by questions on the practices' adoption rationales, enforcement mechanisms, and scope. To ensure coverage, additional items drew on the relevant constructs of the KMS success model and literature insights. Participants were selected for their roles in system development, with an emphasis on data-related expertise (Table 3). Secondary data sources included internal project documents and public technical documentation.

**Data analysis** was conducted iteratively, starting in parallel with ongoing data collection, as suggested by Eisenhardt (1989). Drawing on the two-stage approach by Bhattacharjee (2012), first the within-case and secondly cross-case analyses were conducted. An inductive coding strategy was applied using MAXQDA comprising open, axial, and selective coding (Strauss and Corbin, 1998) iteratively. Within-case analysis focused on identifying data governance practices and contextual conditions, such as adoption drivers and their

Case Nr.	KMS Description	KMS Maturity	Industry of Firm
1	Chatbot for customer support employees to assist in research tasks by using RAG to access documents such as guidelines, process descriptions, and training materials.	Deployed over 1 year	Financial services
2	Chatbot for software developers to assist in translating code from legacy languages into modern ones (e.g., Python) using RAG to access translation examples and internal documentation.	Proof-of-Concept	Technology
3	LLM extension of a decision-tree-based HR chatbot, using RAG to access documents, email trails, and relational databases.	Testing phase	Automotive
4	Chatbot to assist employees in applying internal and regulatory requirements, using RAG to access service level agreements, guidelines, and regulatory publications.	Prototype	Financial services
5	Implementation of MS Copilot 365 to assist in general tasks like question answering using RAG over MS 365 documents.	Pilot Phase	Industrial goods
6	Chatbot to assist employees in researching internal documentation using RAG over Confluence, Notion, and intranet content.	Deployed over 1 year	Automotive

**Table 2. Overview of the six cases selected for study**

Case Nr.	Interview Participant(s)	Involvement in KMS Development	Length (min)
1	Data Engineer Lead, Senior	Main development team of KMS	51:44
2	AI Lead	Involvement in strategic aspects	27:48
3	Data Engineer	Project lead of the system	61:02
4	Consultant; Data Engineer	Co-project lead of project; Involved in technical implementation	44:15; 59:51
5	Data Governance Employee	Involved in CoPilot data curation process, and implementation	60:26
6	Lead Developer (External)	Involved in initial technical design and development of chatbot	39:51

**Table 3. Interview participants**

implications, through systematic coding of the gathered case data. Cross-case analysis abstracted these practices into higher-order categories and core themes, which were then mapped by the researchers to constructs and dimensions of KMS success based on their anticipated effects. Only practices showing a consistent relationship with a construct within a success dimension were grouped together, ensuring conceptual coherence. The coding strategy is presented in Figure 2.

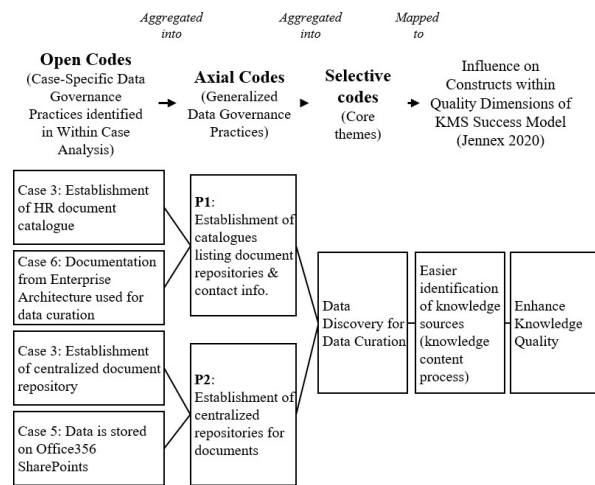
## 4. Findings

The findings are structured to support the development of the conceptual framework, beginning with data governance practices and their contribution to KMS success. These are followed by an exploration of their contextual drivers and associated trade-offs.

### 4.1. Data Governance Practices and Contribution to KMS Success

Across all six cases, 43 case-specific data governance practices were identified. By comparing and aggregating these, 17 generalized practices were derived (Table 4), reflecting case-specific measures that differ only slightly in design and scope, while using predominantly procedural mechanisms. These can be categorized into six identified core themes. These themes reflect recurring organizational responses to managing semi-structured data and ensuring the reliability, accessibility, and compliance of knowledge inputs in generative AI-enabled KMS using RAG.

**Theme 1: Amplifying Data Discoverability During Data Curation:** Organizations implemented measures to make semi-structured knowledge easier to locate and curate. This included cataloguing documents (P1) and centralising document storage to streamline access (P2). These practices strengthen the *Knowledge Content Process* (construct in the knowledge quality dimension) by enabling faster and more consistent



**Figure 2. Coding strategy**

identification of relevant sources. In one case, the use of standardised repositories significantly reduced manual search effort during curation.

**Theme 2: Formalisation of Quality Assurance for Knowledge Content:** To ensure the integrity of ingested documents, firms introduced responsibilities for content quality (P3), manual checks (P4), structured templates (P5), and AI-supported validation methods (P6). These efforts improve the *Richness* (construct in the knowledge quality dimension) of knowledge by increasing trust in its accuracy and completeness—an essential prerequisite for effective reuse by generative models.

**Theme 3: Definition of Metadata for Context Enrichment:** Where AI systems struggled with ambiguous or context-poor documents, firms enriched their data with additional context. This was done by creating document glossaries (P7) or introducing new metadata fields (P8). Such measures enhance *Linkages* (construct in the knowledge quality dimension) by improving the contextualisation of knowledge content. By defining and populating metadata, they support the integration of additional knowledge sources and enable more accurate AI-driven retrieval.

**Theme 4: Metadata for Data Enrichment Optimising Retrieval:** Beyond the general context, some organizations developed metadata specifically aimed at enhancing system-level retrieval effectiveness. These included the creation of abbreviation glossaries (P9) and trustworthiness scores (P10) or the restructuring of documents (P11) to reflect hierarchical logic. These practices contribute to *KM Level* (construct in the system quality dimension) by optimising retrieval and contextual interpretation. By embedding structured metadata into the data pipeline or inference process, they improve the system's ability to

Theme	Data Governance Practice	Annotations	Governance Mechanism	In Cases
Theme 1: Amplifying Data Discoverability During Data Curation	P1: Establishment of catalogues listing document repositories & contact information	Effects mainly data curation processes	Procedural	3, 6
	P2: Establishment of centralized repositories for documents	Effects mainly data curation processes	Procedural	(3), 5
Theme 2: Formalization of Quality Assurance for Knowledge Content	P3: Assignment of responsibilities for data quality	Only refers to integrated un-/semi-structured textual data	Structural	1, 3
	P4: Establishment of quality checks for documents	Refers to manual reviews before document publication or during data curation	Procedural	1, 2, 4, 5, 6
	P5: Establishment of procedures to assess accuracy of document content	Includes metrics and using LLMs to assess plausibility	Procedural	(1), 3, 6
	P6: Definition of document templates to ensure entry of all required information	Integrated during data creation, aligned with requirements	Procedural	(1), 2, 4
Theme 3: Definition of Meta- data for Context Enrichment	P7: Definition of data glossaries and dictionaries for context enrichment	Integration of their definitions as additional metadata to retrieve	Procedural	(3)
	P8: Definition and population of new helpful tags for context enrichment	Integration of metadata during data pipeline and inference chain	Procedural	4
Theme 4: Definition of Metadata for Data Enrichment Optimizing Retrieval	P9: Development of firm-specific abbreviations glossaries	Replacing abbreviations in data pipeline or as context during inference chain	Procedural	1, 6
	P10: Definition and population of trustworthiness score for included documents	Integration as retrieved context with prompt augmentation	Procedural	(1), 3
	P11: Definition of hierarchical document structure	Enables hybrid retrieval approaches (integrating graph-based retrieval)	Procedural	(4)
Theme 5: Guidance for Integration of Data Containing Sensitive Information	P12: Establishment of data inclusion review and approval procedures	Conducted after data curation, before data integration	Procedural	3, 4, 6
	P13: Establishment of data sensitivity classifications	Used in data curation processes, implemented on technical layer	Procedural	2, (5)
	P14: Definition of data eligible for inclusion in AI and RAG systems	Used in data curation processes, implemented via policies	Procedural	2, 5
Theme 6: Data Protection Enforcement and Monitoring	P15: Implementation of algorithms to remove/mask PII or sensitive information	Integrated directly in data pipeline (automatic process)	Procedural	(1), (3), 4, 6
	P16: Implementation of access controls on embedding-level	Technical integration in RAG system, via for example access control lists	Procedural	3, 4, 5, 6
	P17: Logging of prompts and responses for monitoring/auditing purposes	Integrated in inference chain	Procedural	2, 4, 5, 6

**Annotations:** (Case Nr.) = Practice is discussed/planned/in development but not implemented

**Table 4. Data governance practices across cases**

apply and retrieve relevant knowledge.

**Theme 5: Guidance for Integration of Data Containing Sensitive Information:** In response to the legal and reputational risks associated with generative AI, organizations introduced approval procedures (P12), data classification schemes (P13), and formal definitions of eligible content (P14). These practices are central to *KM Governance* (construct in the service quality dimension) as they help ensure that only compliant and risk-adequate data enters the system.

**Theme 6: Enforcement and Monitoring of Data Protection Measures:** To operationalise data protection, technical mechanisms such as data masking (P15), access controls (P16), and prompt logging (P17) were deployed. These measures support ongoing oversight and traceability, further reinforcing *KM Governance* (construct in the service quality dimension) by enabling granular risk control and ensuring oversight of how sensitive data is accessed and used within KMS.

## 4.2. Drivers for Adopting the Practices

Although these themes provide a structured overview of the governance practices observed across cases, they raise a further analytical question: What factors prompted and shaped their adoption in the first place? The analysis reveals two overarching factors that shape the adoption and configuration of data governance practices in RAG-based KMS implementations: First, evolving data governance requirements triggered by generative AI and RAG technologies, and second, data-centric contextual challenges specific to each implementation setting. While the former creates a general demand for new governance solutions, the latter influences how these solutions take shape in practice.

### Key Driver 1: Evolving Data Governance Requirements Induced by Generative AI and RAG Technical Affordances:

Generative AI and RAG technologies introduce new functional requirements that reframe what effective data governance entails. This can be further separated into

two groups:

Firstly, they increase the importance of data discoverability, contextualisation, and structure for semi-structured (e.g., documents with titled subsections) and unstructured (e.g., email content) textual data, which traditional governance mechanisms often overlook. This is evident in Themes 1 to 4. Secondly, the integration of generative AI and RAG into enterprise systems raises new data protection and compliance risks, prompting organizations to rethink the integration of sensitive information and enforce data protection (Themes 5 and 6).

This can be illustrated with the case data. For example, in Case 3, the need to access and curate large volumes of potentially relevant documents was hindered as these were distributed along unmanaged data silos because the RAG-enhanced KMS is the first system to utilize such data, prompting the adoption of document catalogues (P1) and discussing centralized document repositories (P2). Similarly, the inability of AI systems to disambiguate retrieved documents was attributed to missing metadata and unclear contextual cues. This prompted the introduction of document glossaries (P9) and efforts to enrich documents with descriptive metadata (P7).

#### **Key Driver 2: Data-Centric Contextual Challenges Specific to KMS Implementation:**

In parallel, organizations face case-specific data challenges shaped by their internal structures, governance maturity, and application contexts. These challenges—six of which were identified across the cases—interact with the evolving technical affordances to influence which practices are adopted. These include dispersed repositories (Theme 1), lacking governance maturity (Theme 2), additional context required for knowledge application (Theme 3), specific data characteristics that aggravate the retrieval function (Theme 4), uncertainty if sensitive information is compliant to include (Theme 5), and requirements for compliant integration of data (Theme 6).

This becomes evident in the case data. For instance, organizations were primarily motivated by compliance uncertainty on what data is compliant to be integrated, leading to the adoption of data sensitivity classifications. In Case 5, the respondent stated “*We also cannot say which project files may not contain any protected data.*”, why they plan to deploy tools like Microsoft Purview for classifying content based on sensitivity (P13).

### **4.3. Trade-offs**

Although these two key drivers explain the emergence and variation of governance practices, their

implementation was not without tension. These describe inherent trade-offs implied by the implementation of the data governance practices. Across the cases, two recurring tensions emerged:

**Trade-off 1: Corresponding Implementation and Ongoing Operational Costs:** All practices demanded organizational efforts and technical resources, resulting in both implementation and operational costs. Case 4 illustrated this clearly: although hierarchical structuring (P11) was seen as technically beneficial, the respondent stated “*We quickly decided against implementing it because it would be too complex and would be disproportionate to the costs.*” This reflects a broader tension between anticipated contribution to KMS success and resource feasibility, particularly in settings with limited technical capacity. Even in cases where technical benefits were clear, organizations had to prioritise within the limits of existing governance capacity and infrastructure.

**Trade-off 2: Diminishing Extent of Integrated Knowledge:** Practices in Themes 5 and 6 were designed to safeguard against regulatory, reputational, and ethical risks. However, they also created friction in the knowledge inclusion process, sometimes excluding content that could have been valuable to the system. Case 5 exemplified this tension. Here, eligibility rules (P14) led to the exclusion of documents containing potentially sensitive client information, stating “*management does not want any customer data in AI systems*” because of a strong risk aversion. This approach, however, significantly constrained the KMS’s utility. As one respondent noted, CoPilot could be “*great to search through old project files and use them to create drafts or give recommendations,*” but this potential remained unrealised due to the restrictive data policy. More inclusive approaches, meanwhile, increase coverage but require robust technical safeguards or increase risk.

### **4.4. Consolidating Findings**

The findings are synthesized into a conceptual framework (Figure 3) that illustrates how technical affordances and contextual conditions interact to shape data governance practices, which in turn influence KMS success while introducing specific trade-offs. At the core of this framework is the observation that the implementation of RAG-enhanced LLMs in knowledge management systems triggers evolving governance needs. These stem not only from the unique characteristics of generative AI and RAG technologies but also from contextual challenges tied to the nature and use of organizational data. Together,

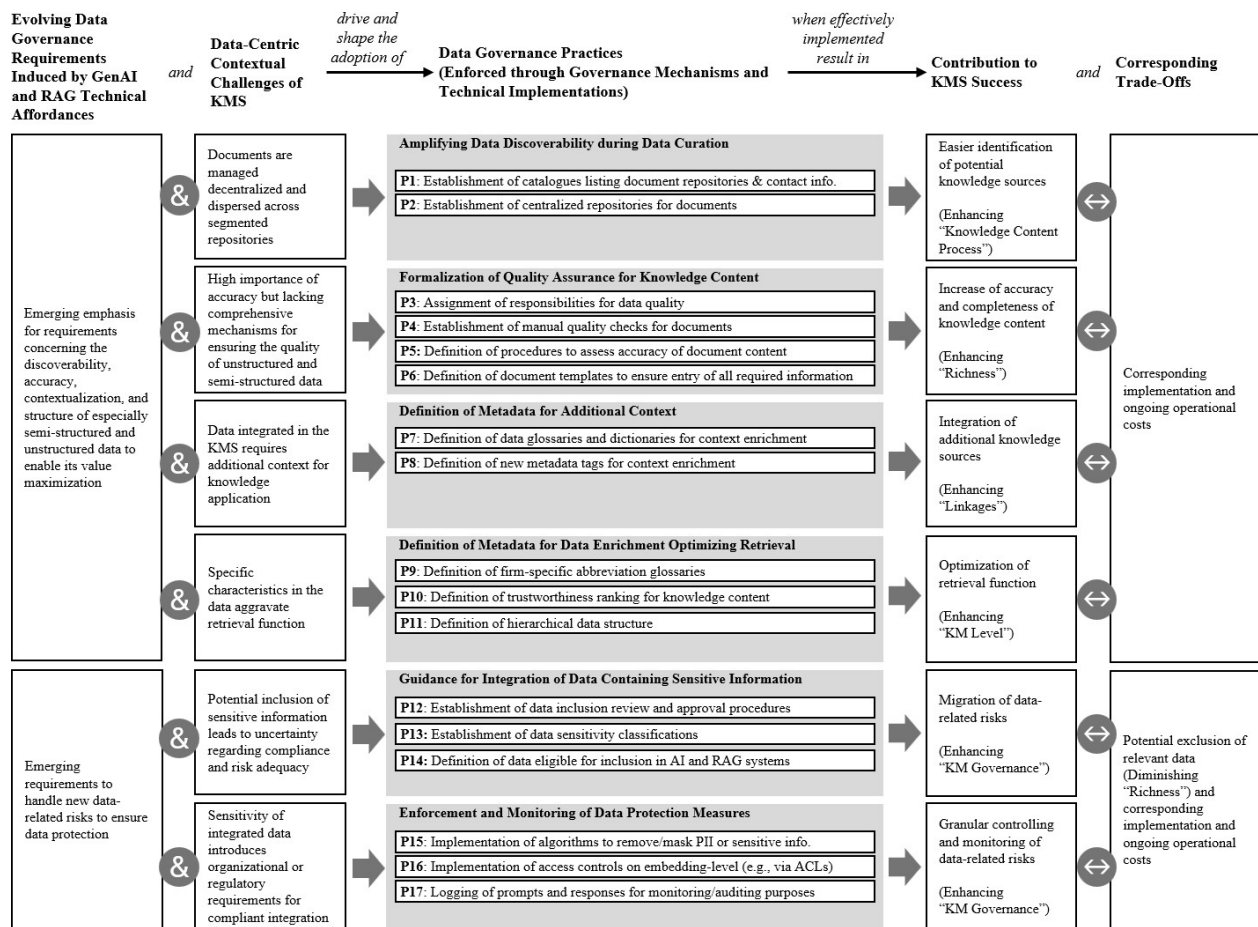


Figure 3. Overview of data governance practices, their drivers, contribution to KMS success and trade-offs

these drivers create pressures and opportunities that organizations respond to by adopting targeted data governance practices. The study identifies 17 such practices, clustered into six thematic groups, each addressing a specific pattern of governance need. These practices rely primarily on procedural governance mechanisms, often reinforced by technical tools, rather than structural arrangements. When enacted effectively, they support KMS success by enhancing distinct quality dimensions—namely system, knowledge, and service quality—based on the framework by Jennex (2020). Yet, these benefits do not come without costs: each practice introduces trade-offs that organizations must navigate, highlighting the balancing act inherent in governing AI-augmented knowledge systems.

## 5. Discussion

**Interpretation of Findings:** This study developed a conceptual framework explained in the previous section. It shows that data governance in RAG-based

KMS is not governed by a single, prescriptive model. Instead, organizations implement a diverse set of practices tailored to the technical affordances of generative AI and the specific characteristics of their data landscapes. The prevalence of practices grounded in procedural governance mechanisms suggest that organizations predominantly rely on formalized processes and routines rather than organizational structures or roles to govern data in RAG-based KMS implementations.

The findings suggest a strong link between data governance practices and KMS success across all three quality dimensions. Given the centrality of high-quality proprietary unstructured data for generative AI value creation, governance becomes a strategic enabler—enhancing content richness, retrieval performance, and compliance. This aligns with the dual role of governance in enabling innovation and managing risk, as emphasised by Vial (2023).

The results also suggest the emergence of unstructured data governance as a distinct concern,

catalysed by the demands of RAG-based KMS. However, current practices remain narrow in scope and largely reactive. Only one practice (P6) directly targets content creation, indicating a limited deployment of systematic, long-term approaches to governing unstructured knowledge assets. This underscores a broader need for dedicated unstructured data governance frameworks in enterprise AI contexts.

Two findings diverge from expectations:

*Firstly*, no practices were observed to align or structure data at source for improved pipeline integration, indicating that organizations prefer post-hoc transformation through technical architectures rather than proactive data design.

*Secondly*, while procedural mechanisms (e.g., checks, classifications) dominate, relational mechanisms—such as training or awareness-building—are almost entirely absent, despite being an obvious guess and their occurrence in prior work identified in the grey literature (e.g., Papagiannidis et al., 2023). This under-representation suggests that relational mechanisms remain overlooked or under-reported in enterprise generative AI initiatives. However, two factors may account for the absence of relational mechanisms in the findings. *Firstly*, the questionnaire was designed to elicit mechanisms perceived as directly contributing to KMS success, whereas relational mechanisms typically exert their influence more indirectly. *Secondly*, four of the six cases examined were either prototypes or in pilot and testing phases, where relational mechanisms such as awareness campaigns usually occur only after full deployment.

**Comparison with Literature:** A comparison between the identified practices in the case studies and the findings from the literature review (Table 1) reveals that current academic literature does not yet fully reflect the methods employed in practice. While several practices observed in the cases align with insights from RAG-specific studies or broader data governance literature, others—such as trustworthiness scoring (P10) and the use of abbreviation glossaries (P9)—are largely absent from existing scholarly discourse. This discrepancy likely reflects the rapidly evolving nature of RAG applications in enterprise settings, which has yet to be comprehensively explored in academic research. Conversely, certain practices highlighted in the literature—such as API governance and advanced technical approaches like toxicity scoring—were not observed in the studied cases. This suggests that more sophisticated techniques have yet to gain traction in practical implementations.

**Theoretical Contributions:** This study contributes

to data governance literature in three ways. *First*, it addresses the limited empirical insight into the governance of semi-structured and unstructured data in RAG-based systems, offering 17 practices across six themes grounded in real-world implementations. *Second*, it challenges normative assumptions in existing frameworks by highlighting underexplored yet widely used practices (e.g., trustworthiness scoring, abbreviation glossaries) and the absence of others commonly discussed in the literature (e.g., formal user training). *Third*, prior research tends to emphasize centralized, formal mechanisms, whereas our findings highlight how organizations increasingly rely on informal, embedded, and technical safeguards. This includes substituting traditional governance artifacts with integrated system features and collaborative routines. The proposed framework captures these emergent patterns and helps explain how governance adapts under conditions of rapid technical change and limited maturity models. Together, these insights advance a more practice-oriented and context-sensitive understanding of governance in generative AI settings.

**Limitations:** The findings should be interpreted in light of several limitations. *Firstly*, the analysis is based on only six cases, most of which involve low-sensitivity data. This may constrain the transferability of the results to high-stakes or highly regulated settings. While some practices were explicitly employed to respond to regulatory risks (e.g., PII reduction algorithms), further practices, particularly those driven by regulatory requirements such as extended risk assessments, are likely to emerge in such contexts. *Secondly*, as the framework illustrates, the technical affordances of GenAI and RAG shape the adoption of data governance practices. Given the rapid evolution of generative AI, including potential moves beyond the embedding-based retrieval approaches used in all examined cases, certain practices risk becoming obsolete as the underlying technologies advance. *Thirdly*, the predominance of prototypes or early-stage deployments among the examined cases limits the generalizability of the findings. This is evident in the absence of proactive approaches and the lack of relational mechanisms, which are typically observed at later stages of system maturity. *Fourthly*, the qualitative design does not permit the quantification of the effectiveness of individual mechanisms, and relational practices may therefore be under-represented.

These limitations highlight opportunities for future research, including longitudinal studies that examine how advances in GenAI and RAG reshape governance practices over time, and how organizations transition from reactive toward more proactive measures.

Nonetheless, the proposed framework represents an important step toward understanding the evolution of data governance in the context of generative AI. It offers a foundation for future work to investigate more advanced system contexts, longitudinal dynamics, and the effectiveness of specific governance mechanisms.

## 6. Conclusion

This study investigated which data governance practices contribute to the success of generative AI-powered KMS that leverage RAG in large enterprises. The main contribution is the development of the conceptual framework, grounded in a multi-case study. It gives an answer to the research question, identifying 17 distinct data governance practices that enhance quality dimensions of the KMS success. It shows how these practices strengthen the three quality dimensions in the Jennex (2020) KMS success model via six mechanisms: (1) amplifying data discoverability during data curation, (2) formalising quality assurance for knowledge content, defining metadata (3) to enable additional source integration and (4) to enable data enrichment for optimised retrieval, (5) providing guidance for handling sensitive data, and (6) enforcing and monitoring data protection at the system level. The contextual drivers shaping the adoption of these practices are the combination and interaction of dynamically evolving technology-induced requirements (particularly stemming from the reliance on semi-structured data) and contextual data-related challenges. While they enhance knowledge, system, and service quality, they also introduce strategic and operational trade-offs, notably due to their resource intensity and constraints on data integration for the sake of data protection.

In sum, the findings show that for data governance to enable the success of RAG-enabled KMS, it must rely on context-sensitive mechanisms tailored to knowledge content typically in the form of semi-structured textual data. These mechanisms play a critical and evolving role in maximising content value while balancing emerging risks and operational costs. Furthermore, the findings highlight the need for future research on topics such as informal governance mechanisms and the temporal evolution of governance practices in response to emerging technologies within the dynamic landscape of enterprise generative AI systems.

## 7. References

Abraham, R., Schneider, J., & Vom Brocke, J. (2019). Data governance: A conceptual framework, structured review, and research

- agenda. *International Journal of Information Management*, 49, 424–438. <https://doi.org/10.1016/j.ijinfomgt.2019.07.008>
- Akkiraju, R., Xu, A., Bora, D., Yu, T., An, L., Seth, V., Shukla, A., Gundecha, P., Mehta, H., Jha, A., Raj, P., Balasubramanian, A., Maram, M., Muthusamy, G., Annepally, S. R., Knowles, S., Du, M., Burnett, N., Javiya, S., Boitano, J. (2024, July). FACTS About Building Retrieval Augmented Generation-based Chatbots [arXiv:2407.07858 [cs]]. <https://doi.org/10.48550/arXiv.2407.07858>
- Alavi, M., & Leidner, D. E. (2001). Review: Knowledge Management and Knowledge Management Systems: Conceptual Foundations and Research Issues. *MIS Quarterly*, 25(1), 107. <https://doi.org/10.2307/3250961>
- Bhattacharjee, A. (2012). *Social Science Research: Principles, Methods, and Practices*. Global Text Project.
- DAMA International. (2009). *The DAMA guide to the data management body of knowledge: DAMA-DMBOK guide* (M. Mosley, M. Brackett, S. Earley, & D. Henderson, Eds.; First edition). Technics Publications, LLC.
- Delone, W., & McLean, E. (2003). The DeLone and McLean Model of Information Systems Success: A Ten-Year Update. *Journal of Management Information Systems*, 19(4), 9–30. <https://doi.org/10.1080/07421222.2003.11045748>
- Eisenhardt, K. M. (1989). Building Theories from Case Study Research. *The Academy of Management Review*, 14(4), 532. <https://doi.org/10.2307/258557>
- Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2024, March). Retrieval-Augmented Generation for Large Language Models: A Survey [arXiv:2312.10997 [cs]]. <https://doi.org/10.48550/arXiv.2312.10997>
- Haridasan, P. K. (2024). The Salesforce Einstein Trust Layer for Retrieval-Augmented Generation (RAG) for Enterprise Applications. *International Journal of Scientific Research in Engineering and Management*, 08(10), 1–3. <https://doi.org/10.55041/IJSREM28465>
- Jennex, M. E. (2020). A Re-Examination and Re-Specification of the Jennex Olman Knowledge Management Success Model: In M. E. Jennex (Ed.), *Advances in Knowledge Acquisition, Transfer, and Management* (pp. 1–29). IGI Global. <https://doi.org/10.4018/978-1-7998-2189-2.ch001>

- Jennex, M. E., & Durcikova, A. (2014). Integrating IS Security with Knowledge Management: Are We Doing Enough? *International Journal of Knowledge Management*, 10(2), 1–12. <https://doi.org/10.4018/ijkm.2014040101>
- Jennex, M. E., & Olfman, L. (2006). A Model of Knowledge Management Success: *International Journal of Knowledge Management*, 2(3), 51–68. <https://doi.org/10.4018/ijkm.2006070104>
- Kallio, H., Pietilä, A.-M., Johnson, M., & Kangasniemi, M. (2016). Systematic methodological review: Developing a framework for a qualitative semi-structured interview guide. *Journal of Advanced Nursing*, 72(12), 2954–2965. <https://doi.org/10.1111/jan.13031>
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 9459–9474.
- Meso, P., & Smith, R. (2000). A resource-based view of organizational knowledge management systems. *Journal of Knowledge Management*, 4(3), 224–234. <https://doi.org/10.1108/13673270010350020>
- Microsoft. (2024, October). Improve RAG application quality. Retrieved January 3, 2025, from <https://learn.microsoft.com/en-us/azure/databricks/generative-ai/tutorials/ai-cookbook/quality-overview>
- Otto, B. (2011). Data Governance. *Business & Information Systems Engineering*, 3(4), 241–244. <https://doi.org/10.1007/s12599-011-0162-8>
- Packowski, S., Halilovic, I., Schlotfeldt, J., & Smith, T. (2024). Optimizing and Evaluating Enterprise Retrieval-Augmented Generation (RAG): A Content Design Perspective. *Proceedings of the 2024 8th International Conference on Advances in Artificial Intelligence*, 162–167. <https://doi.org/10.1145/3704137.3704181>
- Pahune, S., Akhtar, Z., Mandapati, V., & Siddique, K. (2025, April). The Importance of AI Data Governance in Large Language Models. <https://doi.org/10.20944/preprints202504.0219.v1>
- Panda, M., & Mukherjee, S. (2025). Enhancing Privacy and Security in Rag-Based Generative AI Applications. *AI, Machine Learning and Applications advances 2025*, 01–10. <https://doi.org/10.5121/csit.2025.150301>
- Papagiannidis, E., Enholm, I. M., Dremel, C., Mikalef, P., & Krogstie, J. (2023). Toward AI Governance: Identifying Best Practices and Potential Barriers and Outcomes. *Information Systems Frontiers*, 25(1), 123–141. <https://doi.org/10.1007/s10796-022-10251-y>
- Peng, C., Xia, F., Naseriparsa, M., & Osborne, F. (2023). Knowledge Graphs: Opportunities and Challenges. *Artificial Intelligence Review*, 56(11), 13071–13102. <https://doi.org/10.1007/s10462-023-10465-9>
- Ramachandran, A. (2024, August). Retrieval Augmented Generation (RAG) for Large Language Models Leveraging Enterprise Data (SAP, Salesforce, Workday). Retrieved December 29, 2024, from [https://www.researchgate.net/publication/383561133\\_Retrieval\\_Augmented\\_Generation\\_RAG\\_for\\_Large\\_Language\\_Models\\_Leveraging\\_Enterprise\\_Data\\_SAP\\_Salesforce\\_Workday](https://www.researchgate.net/publication/383561133_Retrieval_Augmented_Generation_RAG_for_Large_Language_Models_Leveraging_Enterprise_Data_SAP_Salesforce_Workday)
- Strauss, A. L., & Corbin, J. M. (1998). *Basics of qualitative research: Techniques and procedures for developing grounded theory* (2nd ed). Sage Publications.
- Tallon, P. P., Ramirez, R. V., & Short, J. E. (2013). The Information Artifact in IT Governance: Toward a Theory of Information Governance. *Journal of Management Information Systems*, 30(3), 141–178. <https://doi.org/10.2753/MIS0742-1222300306>
- Vial, G. (2023). Data governance and digital innovation: A translational account of practitioner issues for IS research. *Information and Organization*, 33(1), 100450. <https://doi.org/10.1016/j.infoandorg.2023.100450>
- von Grafenstein, M. (2022). Reconciling Conflicting Interests in Data through Data Governance. An Analytical Framework (and a Brief Discussion of the Data Governance Act Draft, the Data Act Draft, the AI Regulation Draft, as well as the GDPR) [Publisher: Zenodo]. <https://doi.org/10.5281/ZENODO.7390542>

## 8. Appendix

The appendix for this study can be accessed at: <https://doi.org/10.5281/zenodo.17057766>