Journal of Theoretical and Applied Information Technology
15th March 2025. Vol.103. No.5

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

HOW CAN MODEL-DRIVEN ARCHITECTURE AUTOMATE DATA CATALOGS FOR ENHANCED DATA MANAGEMENT?

ASMAE BOUFASSIL[1] , FADWA BOUHAFER[2] , AMINE EL HADDADI[3] , MOHAMED CHERRADI[4] , ANASS EL HADDADI[5]

[1,2,3,4,5] Data Science and Competetive Intelligence Team (DSCI), ENSAH, Abdelmalek Essaadi University (UAE) Tetouan, Morocco.

[1]asmae.boufassil@etu.uae.ac.ma, [2]f.bouhafer@uae.ac.ma, [3]m.cherradi@uae.ac.ma,

[4]amine.elhaddad@gmail.com, [5]a.elhaddadi@uae.ac.ma

ABSTRACT

In recent years, the scalability and maintenance of data catalogs posed significant challenges that hindered organizational efficiency and data accessibility. This paper examined these issues, highlighting the need for an automated approach. It advocated for the use of Model-Driven Architecture (MDA) to streamline the creation and maintenance of data catalogs. Through this approach, key data catalog components were automatically generated from higher-level models, minimizing manual work and improving both data integrity and system functionality. The findings indicated a considerable reduction in errors and operational demands, along with notable improvements in manageability and scalability. This integration of MDA into data catalog frameworks thus presented a compelling solution to the persistent challenges of data management, setting a new standard for efficiency and effectiveness in managing organizational data.

Keywords: *Data Catalog, Model-Driven Architecture (MDA), Data Management, Automated Data Catalog, Metadata Management.*

1. INTRODUCTION

In our increasingly data-driven environment, effective data management has become a cornerstone of organizational success. Data is now considered one of the most valuable assets of any organization, driving decision-making processes, strategic planning, and operational efficiencies. Central to this data-centric paradigm is the role of data catalogs, which serve as comprehensive repositories that facilitate data discovery, governance, and quality assurance. Data catalogs enable organizations to manage their metadata effectively, ensuring that data assets are easily accessible, understandable, and usable. However, the maintenance and scalability of data catalogs present significant challenges, often resulting in labor-intensive processes prone to human error. This article aims to establish a territory within the domain of data management by exploring the critical need for automated data catalog systems. By leveraging Model-Driven Architecture (MDA), we propose a novel approach to automate and enhance data catalog functionalities, thereby establishing a niche in the field of data catalog automation.

Traditional methods of maintaining data catalogs involve manual processes that are not only time-consuming but also susceptible to inaccuracies and inconsistencies. These challenges can lead to outdated or incomplete metadata, which in turn hampers data discovery and governance efforts. As organizations grow and their data environments become more complex, the limitations of manual data catalog management become increasingly apparent. This situation underscores the necessity for innovative solutions that can streamline and automate the processes involved in data catalog management. Our research seeks to occupy this niche by offering a structured methodology that integrates MDA to automate the data catalog lifecycle. Model-Driven Architecture allows for the generation of data catalog components from high-level models, thus reducing the need for extensive manual intervention and enhancing both maintainability and scalability. This approach not only ensures the accuracy and reliability of the data catalog but also makes it a more effective tool for

data governance and decision-making. Key functionalities addressed in our proposed methodology include metadata management, data discovery, governance implementation, data lineage tracking, fostering collaboration, documentation, data quality assurance, maintaining robust security, and automating maintenance and updates.

To substantiate our approach, we conducted a comprehensive literature review to identify the essential functionalities of data catalogs and explore existing automation techniques. Based on this review, we developed detailed guidelines for automating these functionalities using MDA. A hypothetical use case is presented to illustrate the practical application and benefits of our proposed methodology. This scenario exemplifies how the structured approach effectively addresses current challenges in automating data catalogs, paving the way for future advancements. The insights and methodologies presented in this article are poised to significantly advance the field of data management, providing organizations with the tools needed to fully harness their data assets. By occupying this niche, we aim to set a new standard in data catalog automation, driving future research and innovation in this critical area. Ultimately, our work highlights the potential for improved data management practices through the automation of data catalogs using MDA, offering a scalable, maintainable, and reliable solution that can adapt to the evolving needs of organizations in a data-centric world.

2. RESEARCH METHOD

Our research methodology comprises two primary components: a systematic literature review to gather and analyze existing knowledge on data catalogs, and a structured methodology for developing guidelines on automating data catalogs.

2.1 Research Question

The first component of our research involves a systematic literature review. This review aims to consolidate existing knowledge on data catalogs and provide a foundation for developing automation guidelines. The literature review focuses on two main research questions:
(**RQ1**) What are the essential functionalities of a data catalog?
(**RQ2**) How can a data catalog be automated?

2.2 Search Strategy

The main objectif of the literature review (SLR) is to consolidate existing knowledge on data catalogs and a detailed methodology for developing guidelines on automating data catalogs using Model-Driven Architecture (MDA), presented in the figure 1.

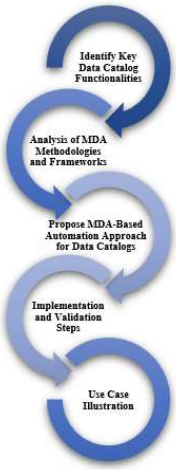


Figure 1:Illustrating Our Research Methodology.

A literature review aims to synthesize the existing state of knowledge on a specific phenomenon, making it a suitable research methodology for extracting functional requirements for data catalogs as a form of codified design knowledge. Following established guidelines for a systematic, concept-centric literature review, we conducted a comprehensive search. We applied refined search terms, "data catalog*" OR "metadata catalog*," across six major databases: Web of Science[1], SCOPUS[2], SpringerLink[3], ACM Digital Library[4], IEEEXplore[5], and AISeL[6]. The publication period was set from 2000 to 2024. Our focus was on accessible, peer-reviewed journal articles and conference papers, excluding incomplete texts and non-peer-reviewed articles. Relevant articles were selected based on their alignment with our research questions and their contributions to the field. The selected literature was then analyzed and synthesized to extract key insights on the functionalities of enterprise data catalogs. This

[1] https://access.clarivate.com/
[2] https://www.scopus.com/home.uri?zone=header&origin=sourceinfo
[3] https://link.springer.com/
[4] https://dl.acm.org/
[5] https://ieeexplore.ieee.org/Xplore/home.jsp
[6] https://aisel.aisnet.org/

synthesis provided a comprehensive understanding of current knowledge and identified gaps for further research. Many studies have focused on metadata management and data catalogs, but unfortunately, they often take manual approaches resulting in inefficient and error-prone processes. With respect to previous studies, they have highlighted the roles of data governance and metadata quality without explaining much how this can benefit from an automatic solution, for example, through Model-Driven Architecture (MDA). Evaluating these limitations underlines the strong need for novel research that interjoins cutting-edge technologies to tackle the scalability and maintenance problems of data catalog systems. This work addresses these issues by presenting a structured approach to automate data catalogs a requirement for addressing modern data management issues.

The second phase focuses on developing a structured methodology to automate data catalogs using Model-Driven Architecture (MDA), with the goal of enhancing their maintainability, scalability, and flexibility. This methodology aims to systematically facilitate the implementation of automation processes within data catalog frameworks. We start with a thorough analysis of existing methodologies and frameworks pertinent to MDA and data catalog automation, identifying optimal practices and theoretical foundations. Drawing on insights gleaned from our literature review, we propose a step-by-step approach for implementing MDA in automating data catalogs. This involves selecting suitable modeling techniques, defining meta-models, and establishing transformation processes to streamline catalog management and improve data accessibility and usability. Our methodology underscores iterative development and validation to ensure alignment with stakeholder requirements and operational efficiency. By documenting guidelines and recommendations derived from empirical findings and industry practices, our aim is to provide practical insights into the realms of data management and automation. To illustrate the practical application and benefits of our proposed methodology, we will present a hypothetical use case scenario. This demonstration will exemplify how our structured approach effectively addresses current challenges in automating data catalogs, paving the way for future advancements and applications in data management systems.

3. PROPOSED MDA APPROACH AND ESSENTIAL FEATURES IN DATA CATALOGS

Following a detailed examination of existing methods for data catalog creation and maintenance, this section progresses to outline the critical functionalities necessary for an effective data catalog. It then introduces a comprehensive framework for automating data catalogs using Model-Driven Architecture (MDA), highlighting how MDA can streamline and enhance data management processes. Finally, it provides a step-by-step guide for deploying an automated data catalog with MDA, offering practical insights and methodologies to ensure successful implementation. Collectively, these subsections present a structured approach to leveraging MDA for optimizing data catalog systems, addressing common challenges, and improving overall data governance and usability.

3.1 Overview of Existing Approaches to Data Catalog Creation and Maintenance

A data catalog is a centralized repository that comprehensively lists and organizes the data assets available within an organization. It serves as a reference point for data users to discover, understand, and access relevant data sources, datasets, databases, and other data-related items. A data catalog typically includes metadata, such as descriptions, information about data provenance, data quality, and other relevant attributes, to facilitate effective data discovery and use [1].

Managing a data catalog involves navigating a multitude of complex challenges. At the forefront lies the imperative task of ensuring comprehensive data discovery and availability. Organizations grapple with the daunting prospect of identifying and maintaining an up-to-date repository of available data assets, a task exacerbated by the dispersed nature of data across various systems and platforms. Concurrently, maintaining data quality and consistency within the catalog emerges as a formidable challenge. The dynamic nature of data necessitates a continuous cycle of updates to metadata, ensuring alignment with shifting data structures, semantics, and availability. Yet, erroneous or incomplete metadata severely undermines the reliability and comprehensibility of data assets, inevitably impacting the accuracy of data-driven analyses and decision-making processes [2]. Furthermore, effective metadata management remains a cornerstone of a functional data catalog. However, establishing consistent standards,

definitions, and classification mechanisms for metadata across diverse data assets often presents hurdles. These inconsistencies significantly impede efforts in data integration, elevating the risk of misinterpretation and considerably limiting the exploitation of the data catalog's full potential. Moreover, the critical issue of data security and privacy looms large over catalog management endeavors. The inclusion of sensitive and confidential information within the catalog necessitates stringent measures encompassing robust access controls, encryption methodologies, and data anonymization techniques. The arduous task of ensuring compliance with stringent data protection regulations further complicates data catalog management, demanding meticulous handling of personally identifiable information and other sensitive data elements. Beyond the technical aspects, the success of a data catalog hinges significantly on fostering collaboration among diverse stakeholders, including data custodians, stewards, analysts, and business users. Establishing streamlined governance processes and delineating clear responsibilities is paramount in maintaining the catalog's accuracy, integrity, and continual maintenance. Conversely, a lack of robust collaboration, communication, and adherence to established data governance practices inevitably leads to inconsistencies, outdated information, and severely curtailed usability of the data catalog. As data volumes soar and the spectrum of data assets diversifies, scalability and performance pose substantial concerns. Managing a vast and intricate data catalog, brimming with myriad data sources and intricate relationships, invariably strains the catalog's performance and responsiveness. Addressing these challenges demands leveraging scalable infrastructure, implementing efficient indexing techniques, and optimizing query mechanisms to cater to the burgeoning demands placed on the data catalog. Lastly, the pivotal aspect of user adoption and usability cannot be understated. The efficacy of a data catalog is inexorably tied to user acceptance. An unintuitive interface or lack of intuitive search functionalities could significantly deter user engagement and utilization. Thus, organizations must invest significantly in comprehensive user training programs, intuitive interfaces, and robust search functionalities to augment the catalog's usability and encourage widespread adoption.

Developing and maintaining a comprehensive data catalog is a critical component of effective data management within an organization. This task can be approached through a variety of methods

- **Manual Approach**: With this strategy, users manually create and maintain the data catalog. The information regarding data sources, metadata, schemas, etc. is gathered and manually entered into the catalog. Spreadsheets, special cataloging software, and database management tools can all be used to do this. Although this method offers total control, it is prone to human mistake and necessitates continual work to maintain the catalog current.

- **Semi-Automated Approach**: This method combines manual and automated components. In order to automatically extract metadata from existing data sources like databases, flat files, APIs, etc., it requires tools and scripts. The data catalog is then updated using the extracted metadata. Although manual labor is reduced, the information still has to be verified and completed by humans.

- **Tool-Based Data Management Approach**: Many data management products come with built-in data cataloging functions. These technologies make it possible to create and maintain data catalogs by offering features like automated data source discovery, metadata extraction, schema management, documentation, etc. Additionally, they provide governance and collaboration tools to streamline data management inside the company.

- **AI-based approach**: New avenues for data catalog generation and upkeep have been made possible by advances in AI, notably in natural language processing and machine learning. From unstructured documents like data specs and other unstructured documents, AI approaches may be utilized to automatically extract metadata. By offering semantic search and recommendation functionality, AI may also improve the catalog's search and discovery capabilities.

The choice of technique depends on the particular demands of the business in terms of the construction and upkeep of data catalogs. Each approach has benefits and drawbacks. When choosing a strategy, it is important to take into

account the resources at hand, the complexity of the data environment, the security and governance needs, as well as the long-term goals for data management.

3.2 Critical Functionalities of Data Catalog

In today's data-driven landscape, a well-managed data catalog is essential for organizations wishing to optimize their data management practices. A data catalog is a comprehensive repository that centralizes and organizes metadata, facilitating data discovery, governance and quality assurance. By leveraging a data catalog, organizations can improve collaboration, ensure data integrity and support informed decision-making processes. The following sections examine the essential functions of a data catalog, highlighting its role in improving data management and accessibility.

- **F1 - Gathering and Managing Metadata**

A data catalog's primary function is to collect and arrange information from different data sources in a way that is useful for managing information. This involves assembling comprehensive information regarding datasets, including detailed descriptions, data schemas, classifications, and annotations. By centralizing this information, organizations can maintain a single, comprehensive repository of their data that provides a comprehensive understanding of their data assets. This promotes increased data management and accessibility, and ensures that all users have a comprehensive understanding of the available data within the organization [3] [4] [5].

- **F2 - Streamlining Data Discovery**

One of the key features of a data catalog is its ability to facilitate data search and discovery. Advanced search functionalities, such as keyword searches and filters, allow users to locate datasets that are relevant quickly. The catalog's intuitive and user-friendly design allows users to easily navigate and explore the data, which promotes informed decision making [6] [7].

- **F3- Implementing Data Governance**

Data governance is a crucial element of a data catalog. It involves establishing policies to guarantee data quality, safeguard privacy, and ensure regulatory compliance. This entails defining roles and permissions to manage data access and changes. Through these governance policies, organizations can shield sensitive data, preserve data integrity, and align with industry regulations [3].

- **F4- Unraveling Data's Journey: A Catalog's Comprehensive Tracking**

Comprehending data's lifecycle is paramount, and a data catalog's detailed lineage feature offers invaluable insights. This function meticulously traces the origin, transformations, and utilization of data, empowering users to understand its evolution. By offering transparency and accountability, data lineage simplifies the monitoring of data flow and changes, a crucial aspect for auditing and troubleshooting purposes [8].

- **F5- Fostering Collaboration and Documentation**

A data catalog's robust documentation and collaborative features create a synergistic environment. Users can document datasets by adding notes, descriptions, and links, while engaging in comments, tags, and ratings. This collective effort enhances the overall understanding and usability of data, as insights and feedback are shared directly within the catalog [9].

- **F6 - Ensuring Data Quality and Profiling**

Maintaining data quality is a vital component of a data catalog. Automated data profiling analyzes datasets, assessing their completeness, validity, and consistency. The catalog generates reports on data statistics, anomalies, and trends, providing valuable insights into data health. This proactive approach to data quality management enables organizations to identify and address issues promptly, ensuring the reliability and accuracy of their data [10] [6] [8].

- **F7- Robust Security and Controlled Access**

A trustworthy data directory upholds rigorous security procedures to protect delicate details. It oversees access privileges based on user duties and company guidelines, ensuring that just approved individuals can see or alter information. By controlling access and applying security precautions, the data directory helps shield data from unapproved utilization and potential vulnerabilities, thus preserving data security and reliability [11].

- **F8- Automated Maintenance and Updates**

Automation [8] [5] is essential for keeping a data catalog up-to-date and accurate. The data catalog serves as a centralized repository that provides comprehensive information about an organization's data assets, including datasets, their metadata, and the underlying source systems. Automating the collection and updating of this metadata is crucial, as it ensures that the catalog reflects the latest changes and modifications made to the datasets. Through continuous updates and

synchronization, the automated processes maintain the catalog's currency, keeping the information accurate and timely. This is particularly important in today's dynamic data landscape, where datasets are constantly evolving, new sources are being added, and existing ones are being retired or modified. Without automation, the manual effort required to track and update all these changes would be immense, leading to outdated and unreliable catalog information. The automated maintenance of the data catalog not only reduces the manual workload but also enhances the trustworthiness of the catalog as a resource for data management and discovery. Users can rely on the catalog to find accurate and up-to-date information about the available data, its lineage, quality, and other relevant details. This trust in the catalog's integrity is essential for empowering data-driven decision-making and fostering a data-centric culture within the organization. Moreover, the automated processes can also include advanced features, such as automated data profiling, data quality monitoring, and data lineage tracking. These capabilities further enhance the catalog's value by providing deeper insights into the organization's data assets, enabling more informed data governance and data stewardship practices. The automation of data catalog maintenance is a critical component in ensuring the catalog's relevance, accuracy, and trustworthiness. By continuously and efficiently updating the catalog, the automated processes enable the data catalog to serve as a reliable and comprehensive source of information, supporting the organization's data management initiatives and data-driven decision-making.

In the dynamic realm of data management, the significance of a meticulously maintained data catalog cannot be overstated. A data catalog serves as the backbone for any organization striving to harness the full potential of its data assets. However, the real challenge lies in ensuring that this catalog remains current and reflective of the latest data updates and changes. This is where automation steps in as a pivotal component, transforming the maintenance process from a daunting task into an efficient and reliable system. Automation not only streamlines the collection and updating of metadata but also significantly reduces the manual effort involved, ensuring that the catalog remains an accurate and trustworthy resource. The integration of automated processes into the data catalog's maintenance framework offers numerous benefits. By continuously syncing with underlying data sources, automation ensures that the catalog reflects the most recent modifications and additions. This ongoing synchronization is vital for maintaining the catalog's relevance and accuracy, particularly in environments where data is constantly evolving. Moreover, automated maintenance enhances the overall trustworthiness of the catalog, providing users with confidence that the information they access is up-to-date and reliable. This trust is essential for fostering a data-driven culture within the organization, where informed decision-making is supported by accurate and timely data insights.

3.3 Framework for Automating Data Catalog

The importance of MDA in managing data catalogs cannot be overstated. It creates a metadata model that details the stored data, facilitating understanding through depiction of various types plus relationships and automates related tasks (code generation) from the model. This specifies properties and business rules within the metadata model so that the quality of data can be known by the high standards and user requirements— thus promoting quality information. MDA also helps promote better collaboration between development and business teams by providing a common metadata model, which leads to an understanding of what users need. With these enhancements achieved at implementation levels while still maintaining their effectiveness after deployment, MDA supports scalability plus maintenance for future modification or new addition easily and effectively to ensure success in managing data catalog.

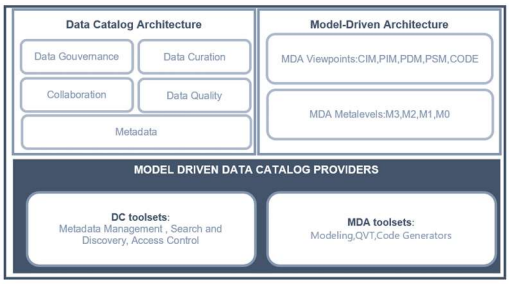


Figure 2:Automated Data Catalog design framework.

The proposition outlines an integrated approach that merges Data Catalog Architecture and Model-Driven Architecture (MDA) concepts to facilitate effective data management and software development. Let's delve into a detailed explanation of this approach [12]:

- **Data Catalog Architecture:**

The crux of the **Data Catalog Architecture** delves into multifaceted components vital for robust data management. It encapsulates various pivotal aspects, notably data governance, data curation,

collaboration, data quality, and metadata. These facets interplay harmoniously within the architecture, forming the backbone of its functionality [13].

Data governance within the Data Catalog Architecture orchestrates the establishment of policies, procedures, and guidelines. These frameworks meticulously define data ownership, access rights, and overall management strategies, ensuring compliance with regulatory requirements and organizational standards. It navigates the intricate landscape of data privacy, security, and ethics, reinforcing the integrity and confidentiality of sensitive data.

Data curation involves the active and continuous refinement of data assets, encompassing processes for data cleaning, transformation, and enrichment. It involves the application of domain-specific expertise to enhance the accuracy, completeness, and relevance of data, rendering it more valuable and actionable [14].

Collaboration within this architecture facilitates seamless interaction among diverse stakeholders. It cultivates an environment where data custodians, analysts, and business users converge, fostering a collaborative ecosystem. Effective collaboration mechanisms ensure shared understanding, streamlined workflows, and coherent data usage, amplifying the catalog's utility and efficiency.

Data quality is a cornerstone element, emphasizing the meticulous validation and assurance of data accuracy, consistency, and reliability. Data quality mechanisms embedded within the architecture continually monitor and assess data, identifying anomalies or inconsistencies, thereby ensuring high-quality, trustworthy data assets [15].

Metadata forms the bedrock of the Data Catalog Architecture. It encapsulates detailed information about data assets, encompassing descriptions, lineage, usage, and other contextual information. Comprehensive metadata management practices streamline data discovery, understanding, and utilization.

The Data Catalog Architecture is a sophisticated system designed to handle the complexities of diverse data sources, including structured, semi-structured, and unstructured data. Structured data, with its predefined schemas, is neatly organized within relational databases. Semi-structured data, such as JSON and XML, possesses a more flexible structure, while unstructured data, encompassing raw text and images, lacks a defined structure. These varied data sources form the foundational layer of the architecture, residing within databases, data warehouses, and expansive data lakes, providing flexible storage solutions for unorganized data.

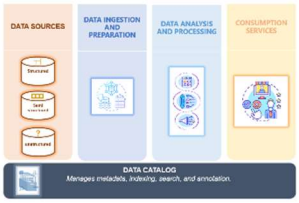


Figure 3:Architecture of Data Catalog.

At the heart of this architecture lies the data catalog, which orchestrates the ingestion, preparation, analysis, and consumption of data. The ingestion and preparation phase involves extracting, transforming, and loading (ETL) data from various sources into the catalog, ensuring it is ready for analysis. Analytical tools then extract meaningful insights from the data, transforming it into actionable information through visualizations like graphs and dashboards. The services layer empowers end-users with the ability to search, explore, and access data via APIs and SQL queries. The Data Catalog Architecture serves as a linchpin in contemporary data management paradigms, fostering an environment where disparate data sources coalesce seamlessly, facilitating informed decision-making, and leveraging the full potential of organizational data assets.

- **Model-Driven Architecture (MDA):**

MDA, or model-driven architecture, embodies a crucial paradigm in software engineering, emphasizing diverse **viewpoints** of a system across multiple abstraction levels. These viewpoints, represented by various models, elucidate distinct layers of the system's architecture, enabling comprehensive understanding and effective development.

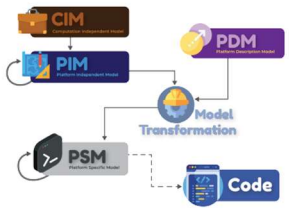
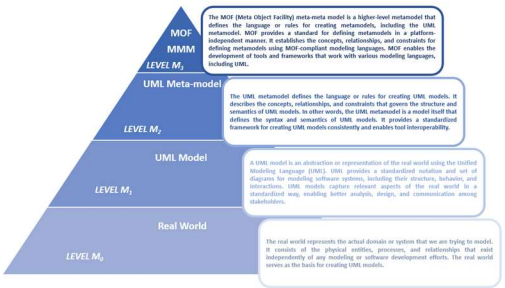


Figure 4:Overview of the MDA approach.

At the highest level stands the **computation-independent model** (**CIM**). This model operates in a domain-specific realm, detached from specific technological details. It holistically captures the system's conceptual aspects, delineating its functionalities, business processes, and overarching requirements. The CIM caters to stakeholders, offering a panoramic yet abstract view of the system and fostering shared comprehension across diverse

teams. Transitioning towards more concrete representations, the **platform-Independent Model** (**PIM**) refines the CIM. It focuses on the system's design and logical architecture, abstracting the structure, interactions, and functionalities at a higher level. Yet, it remains agnostic to technological specifics, concentrating on system behavior and data flow in a technology-neutral manner. This model aids developers in crafting system designs aligned with the CIM's conceptualization. Further down the abstraction hierarchy, the **platform-dependent model** (**PDM**) emerges. This model bridges the gap between the abstract PIM and the system's concrete implementations on specific platforms or technologies. It encapsulates technology-specific details, software components, and infrastructure specifications, providing a blueprint for system realization aligned with the PIM's specifications. Descending into the realm of implementation specifics, the **platform-Specific Model** (**PSM**) takes center stage. Unlike the preceding models, the PSM dives into technicalities, specifying system implementations tailored to particular platforms or technologies. It encompasses specific coding details, configuration specifics, and technology-dependent nuances essential for actual implementation. The code level, or **platform-specific details** (**PSD**), culminates the hierarchy. This viewpoint offers precise coding instructions, configuration details, and technology-specific nuances necessary for direct system implementation. It serves as a direct guide for developers, translating the high-level designs of preceding models into concrete, executable code.

These distinct viewpoints within the MDA framework encapsulate varied layers of abstraction, catering to stakeholders and developers across



different stages of system development. They facilitate a comprehensive understanding and effective translation of high-level concepts into tangible, well-aligned software systems.

The MDA Meta-levels delineate distinct tiers of abstraction, each crucial in comprehending and structuring the software development process. At the zenith lies the M3 level, recognized as the Meta-meta model.

The stratified layers offer a profound insight into the MDA's hierarchical structure. **M3** (**Meta-meta model**) anchors the topmost tier, embodying the epitome of abstraction, governing the principles and methodologies of modeling techniques. Descending to **M2** (**Meta model**), this layer focuses on defining the language used to craft specific models, establishing the rules and semantics guiding their creation. Further down, **M1** (**Model**) embodies the concrete representation of the system, detailing

Figure 5:Interpreting the OMG multi-level modeling stack

its functionalities, design, and behavior. Ultimately, at **M0** (**Object-level**), tangible system components, such as databases or software modules, manifest, reflecting the realized implementation of the higher-level models. These meta-levels orchestrate a structured approach, facilitating a seamless transition from abstract conceptualization to tangible system realization.

- **Model-Driven Data Catalog Providers:**

Underpinning the **Data Catalog** Architecture are sophisticated **toolsets** designed to augment its functionalities. These encompass metadata management, search and discovery, and access control.

Metadata management tools facilitate the creation, storage, and organization of metadata. They enable the enrichment of data descriptions, lineage tracking, and attribute classification, ensuring comprehensive and accurate metadata records.

Search and discovery toolsets empower users to navigate the vast data landscape efficiently. Through intuitive search functionalities and advanced discovery mechanisms, these tools enable users to pinpoint and comprehend relevant data assets swiftly.

Access control mechanisms within the toolsets govern data accessibility. They enforce stringent access policies, authenticate user permissions, and ensure adherence to privacy and security protocols, safeguarding sensitive data assets.

This sophisticated interplay of architectural components and accompanying toolsets accentuates the comprehensive nature of the data catalog architecture, underscoring its pivotal role in contemporary data management endeavors.

MDA toolsets represent a specialized array of software applications meticulously crafted to uphold the tenets of model-driven architecture (MDA). These tools are purpose-built to facilitate and enhance the MDA methodology, empowering

developers and architects with a suite of capabilities essential for seamless model development and transformation into executable systems. These toolsets typically encompass an arsenal of functionalities, including sophisticated modeling capabilities that aid in conceptualizing and refining system designs across various abstraction levels. Additionally, they offer robust code generation features, automating the translation of models into executable code [34]. At the core of MDA toolsets lies their adherence to industry standards, notably the Query, View, and Transformation (QVT) standard. This standard ensures consistency and compatibility, enabling smooth interoperability between different tools and platforms within the MDA ecosystem. QVT serves as the linchpin for transformation operations, facilitating the conversion of models from one form to another across diverse abstraction layers. This standardization fosters a cohesive environment, streamlining the process of model manipulation and refinement while ensuring adherence to MDA principles throughout the software development lifecycle.

The integrated approach brings together specific vendors and tools in the field of data cataloging, exploiting the fundamental principles of model-driven architecture (MDA) to create complex, resilient data management systems. These solutions draw on advanced modeling techniques and different levels of abstraction to meticulously design and administer data catalogs. The focus is on global governance, impeccable data quality, transparent collaboration, and optimized use of metadata [16]. In adopting this approach, the intention is to merge the best methodologies from the fields of data management and software development. This convergence enables the adoption of streamlined, standardized, and model-centric methodologies for navigating the intricacies of data cataloging within organizational frameworks. To orchestrate a harmonious synergy between the rigors of data management and the structured, methodical paradigms inherent in software development, thus optimizing the management and use of data catalogs in diverse organizational landscapes.

3.4 Steps for deploying an automated Data Catalog with MDA:

Deploying an automated data catalog using Model-Driven Architecture (MDA) is a methodical and phased approach aimed at achieving seamless integration and optimal performance. This section offers a comprehensive guide to the essential deployment steps. It begins with assessing needs and available resources, proceeds through selecting metadata management tools, conducting data modeling, implementing automation, and concludes with user training and adoption. Each step is meticulously designed to ensure the data catalog meets both functional and non-functional requirements, scales effectively, and adapts to future organizational demands. Following these steps enables organizations to mitigate risks, maximize benefits, and fully capitalize on the advantages offered by automated data catalogs employing MDA.

3.4.1. Evaluation of needs and resources:

Automating the data catalog starts by carefully assessing the essential needs and resources for a successful implementation. This includes pinpointing the data management requirements, such as identifying data assets, defining metadata specifications, and determining data governance necessities. Concurrently, an inventory of existing tools, technologies, and team capabilities is conducted to ensure alignment with project objectives and available resources. Additionally, engaging stakeholders from both business and technical domains is crucial. This collaborative effort enables a comprehensive understanding of project requirements and lays the groundwork for effective planning and execution of the data catalog automation initiative.

3.4.2. Choosing the right metadata management tools:

As the data catalog automation process advances, the focus shifts to identifying the most appropriate metadata management tools. The primary aim is to select tools and technologies that align with the project's objectives and requirements. This process begins with a comprehensive evaluation of various Metadata Management and Data Catalog solutions available on the market. These tools are scrutinized based on a range of factors, including features, scalability, and compatibility with the existing systems. Following this initial assessment, a Proof of Concept is conducted to validate the functionalities of the shortlisted tools in real-world scenarios. This empirical testing helps gauge the effectiveness and suitability of the tools for the organization's specific needs. Additionally, a thorough vendor comparison is undertaken to evaluate various aspects, such as cost, support services, integration capabilities, and the robustness of the user community. This comparison enables informed decision-making, ensuring that the tool selected not only meets

technical requirements, but also aligns with budgetary constraints and long-term strategic objectives. By adopting a systematic approach to metadata management tool selection, organizations can establish a solid foundation for the successful implementation of data catalog automation.

3.4.3. Data modeling:

Building data models that align with Model-Driven Architecture (MDA) principles is crucial in data modeling. This involves employing customized modeling methods suited to various data sources, guaranteeing that data models precisely depict data attributes and connections for effective data cataloging and administration. Organizing and managing data effectively hinges on the art of data modeling. This technique is especially crucial when building comprehensive data catalogs that encompass diverse data formats. By embracing the principles of the MDA approach, you can establish a well-structured framework for designing and constructing data catalogs that can seamlessly integrate a wide range of data inputs. The realm of data cataloging encompasses a wide spectrum of information sourced from diverse origins and formats. Data can originate internally within companies, systematically collected and stored in databases as structured data. Conversely, unstructured data arises from external or disparate sources, lacking organizational structure. This includes data available on the web or information gathered without direct user input, spanning text, audio, video, images, and more. The diversity of data formats underscores the complexity within the data catalog landscape. Within the domain of Data catalog, three primary data source classifications exist:

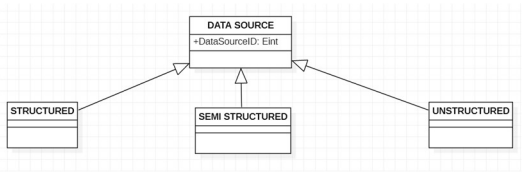


Figure 6: Model for the Three Types of Data Sources

Unstructured data comprises information not confined to a specific structure type. This encompasses textual data like emails, presentations, and documents, along with non-textual data in the form of media files like images, audio, and videos. Structured data, on the other hand, encompasses information residing. Developers structure this data, enabling its manipulation and utilization through specific formats. Semi-structured data stands as an intermediary form, lacking intricate organization for sophisticated access and analysis. However, they often contain accompanying metadata tags facilitating elements' identification within the data. In a data catalog, these diverse data types from various sources are integrated and managed. This involves processes such as ingesting, processing, and storing data to ensure accessibility and usability for users.

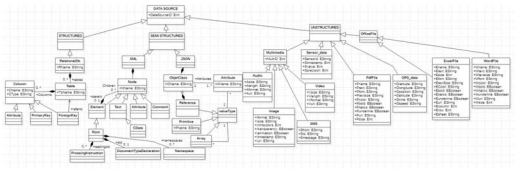


Figure 7: Meta-model of Data Sources layer with different data stereotypes

This diagram (Fig. 7.) presents a structured depiction of the different data types or categories within the Data Sources layer of the data catalog. It illustrates how data is classified, organized, and administered, offering a standardized framework for managing heterogeneous data sources. These visual representations are crucial tools for comprehending the intricate processes involved in data catalog management. They underscore the significance of categorizing, organizing, and integrating diverse data types to ensure efficient data management and maximize data utilization.

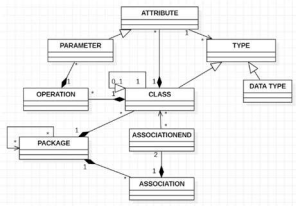


Figure 8: Meta-Meta model of Data Sources layer (MOF).

MOF metamodel offers a versatile framework for building metamodels, such as the specific UML metamodel tailored to data catalogs from various sources. It establishes guidelines based on MOF specifications for defining concepts, relationships, and constraints within these metamodels, which in

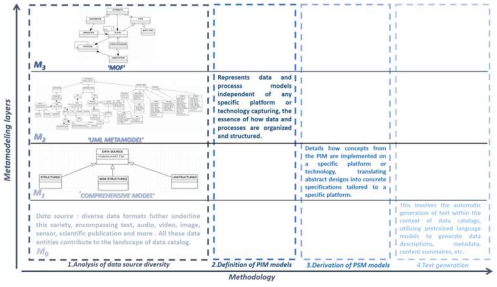


Figure 9: MDA approach for developing Data Catalogs across diverse data types.

turn define the UML models of the data catalog. Fig. 8. illustrates how data is classified, organized, and managed within this layer. It underscores the importance of categorizing, organizing, and integrating diverse data types to ensure efficient data management and maximize their utility. On On On this basis, the global model stage consists of translating abstract designs into concrete specifications adapted to specific platforms. This stage includes the automatic generation of text in the context of data catalogs, taking into account a wide range of data formats. By leveraging the power of pre-trained linguistic models, the derivation stage of PIM models takes a major step forward. By generating data descriptions, metadata and content summaries, this step enriches data catalogs with valuable, insightful information, making data more accessible and meaningful to users. PSM model derivation further automates the process, generating descriptions, metadata and other relevant content essential for data cataloging, based on the concepts of the UML metamodel. This streamlined approach ensures that data catalogs are complete and well-structured, and meet the diverse needs of data consumers. Finally, the text generation stage describes a methodical approach to the generation of textual content in data catalogs. Acknowledging the diversity of data sources and types, this step gives priority to flexibility, adaptability and the ability to adapt to the needs of different users.

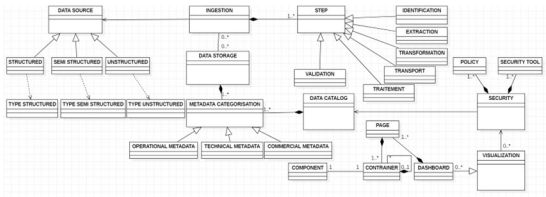


Figure 10: A generic meta-model for hetegeneous sources in a Data Catalog with UML.

Navigating the complexities of heterogeneous data environments can be a daunting task, but this meta-model offers a robust and standardized approach to managing metadata. Rooted in the Unified Modeling Language (UML), this foundational structure serves as a guiding light, illuminating the key components that work in harmony to create a cohesive data ecosystem.

At the heart of this model lies the Data Source, representing the diverse range of structured, semi-structured, and unstructured data available. The Ingestion process seamlessly brings this data into the fold, where it can be further refined and enriched. The Step component outlines the intermediate processes, such as extraction, transformation, and loading (ETL), which shape the data into a more usable form. Identification assigns metadata to data attributes, providing crucial context about type, ownership, and creation date. Validation ensures data quality and compliance, while the Data Lake offers a centralized repository for raw, unstructured data, allowing for flexible and dynamic processing down the line. The Transport component facilitates the movement of data between different stages, ensuring a smooth and efficient flow. Processing, in turn, executes the necessary operations for data transformation and enrichment, unlocking the true potential of the information at hand. Metadata Categorization organizes this wealth of metadata into operational, technical, and business-centric classifications, making it readily accessible and actionable. The Data Catalog serves as the heart of this ecosystem, storing metadata and enabling effortless data search and discovery. Visualization tools then bring the data to life, presenting it through intuitive dashboards and graphical representations. Underpinning this entire framework is the crucial element of Security, ensuring that data access rights are meticulously managed, safeguarding the integrity and confidentiality of the information.

This meta-model offers a comprehensive and standardized approach to metadata management, empowering organizations to navigate the complexities of their data environments with confidence and efficiency. By leveraging this robust framework, businesses can unlock the full potential of their data, driving informed decision-making and fueling innovation.

3.4.4. Implementation of Automation

Deploy and configure the selected tools to automate data cataloging processes with Model-Driven Architecture (MDA). The implementation begins by integrating the data catalog tool with existing heterogeneous data sources, ETL processes, and data pipelines. This integration ensures seamless connectivity and data flow between different components of the data ecosystem, creating a unified platform for metadata collection. Using MDA, the next step is to develop and configure automation workflows that continuously collect, update, and manage metadata. MDA facilitates the creation of a robust and adaptable data model that can effectively

handle the dynamic nature of data, ensuring that the catalog remains up-to-date with minimal manual intervention. These workflows should be meticulously designed to support data lineage tracking, governance, and consistency across various data domains.

After establishing these workflows, perform extensive testing to validate the functionality, accuracy, and performance of the automated data catalog. This testing phase is critical to identify and rectify any issues, ensuring that the automation processes operate smoothly and deliver the expected outcomes. The validated data catalog, structured with MDA principles, then becomes a reliable source for analytical tools. The data, transformed back from the MDA structure, is ready for comprehensive analysis and visualization. This process completes the data lifecycle from heterogeneous sources, through model-driven automation, to insightful visual representations that support informed decision-making.

Figure 11: Automated Data Catalog architecture

3.4.5. Training and Adoption

Ensure successful adoption of the data catalog through user training and support. To ensure the successful adoption of the data catalog, it is essential to develop and deliver training programs tailored to different user roles, such as data stewards, analysts, and business users. These training programs should be designed to equip each user group with the knowledge and skills needed to effectively utilize the data catalog. Comprehensive documentation must be created, including user guides, best practices, and troubleshooting tips, to serve as a reference for users and support their learning process. Additionally, establishing a robust support mechanism is crucial for providing ongoing assistance and addressing any issues that users may encounter. Gathering feedback from users will also be instrumental in continuously improving the data catalog, ensuring it meets the evolving needs of the organization and enhances overall data management practices.

4. USE CASE: IMPLEMENTING AUTOMATED DATA CATALOGS IN 12 MOROCCAN UNIVERSITIES

In the context of deploying an automated data catalog for the 12 principal universities in Morocco, we will explore two distinct scenarios: one using Amundsen alone and the other combining Amundsen with Model-Driven Architecture (MDA). This approach aims to efficiently manage metadata, enhance data discoverability, and support data governance within these institutions. Morocco boasts 12 major universities, each with specific needs and resources for managing their academic and administrative data. The implementation of automated data catalogs in these universities involves a detailed extraction and management of metadata, which is essential for improving data accessibility and governance across various academic and administrative activities.

Implementing automated data catalogs in Moroccan universities requires a meticulous extraction and management of metadata, which is crucial for enhancing data accessibility and governance across diverse academic and administrative activities. Below is a table outlining the types of metadata extracted from each of the 12 universities, each having its own specialized needs and resources.

Table 1 :Metadata Selection

University Name	Website URL	Metadata Types
University Mohammed V (Rabat)	http://www.um5.ac.ma/	Institutional, Contact, Academic, Research, Event
University Cadi Ayyad (Marrakech)	http://www.uca.ma/	Administrative, Research, Resource, Infrastructure
University Hassan II (Casablanca)	http://www.univh2c.ac.ma/	Institutional, Alumni, Academic, Event
Al Akhawayn University (Ifrane)	http://www.aui.ma/	Contact, Academic, Administrative, Infrastructure
University Ibn Tofail (Kénitra)	http://www.uit.ac.ma/	Research, Resource, Event, Alumni
University Sidi Mohamed Ben Abdellah (Fès)	http://www.usmba.ac.ma/	Academic, Research, Administrative, Event
University Mohammed Premier (Oujda)	http://www.ump.ma/	Contact, Resource, Alumni, Infrastructure
University Chouaïb Doukkali (El Jadida)	http://www.ucd.ac.ma/	Institutional, Academic, Administrative, Research
University Hassan Premier (Settat)	http://www.uh1.ac.ma/	Event, Resource, Academic, Alumni
University Abdelmalek Essaâdi (Tétouan)	http://www.uae.ma/	Contact, Infrastructure, Academic, Research, Administrative, Resource
University Ibn Zohr (Agadir)	http://www.uiz.ac.ma/	Academic, Administrative, Research, Resource
University Moulay Ismail (Meknès)	http://www.umi.ac.ma/	Institutional, Event, Alumni, Infrastructure

This table showcases the variety of metadata types that are relevant to each university. The types include institutional details, which cover organizational structures and governance, contact information for facilitating communication, academic metadata for courses and programs, administrative data for operational details, research metadata to highlight projects and publications, resources for library and learning materials, event details for campus activities, infrastructure information regarding physical and IT assets, and alumni data to connect past students with the university.

The next step involves the implementation of appropriate data catalog tools, like Amundsen,

possibly integrated with Model-Driven Architecture (MDA), to handle these diverse types of metadata efficiently. This integration will facilitate improved searchability, governance, and utilization of data, fostering a more data-driven environment across these institutions.

4.1. Scenario 1: Implementing Amundsen Alone

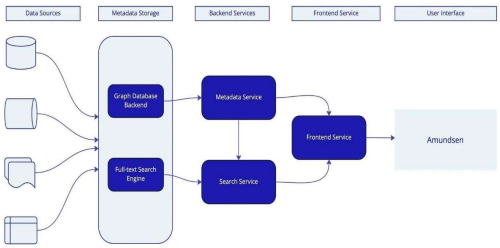


Figure 12: Amundsen architecture [17]

The primary objective of this initiative to implement Amundsen in twelve Moroccan universities is centered around creating a robust and centralized system for managing and cataloging the metadata associated with various academic and research resources. This endeavor seeks to significantly enhance the accessibility and usability of information, thus streamlining educational and research activities across these institutions. By leveraging Amundsen's DataBuilder utility, the project initiates a comprehensive integration process where it connects with multiple institutional databases, document management systems, and research directories.This crucial first step is aimed at extracting and consolidating diverse metadata from different university systems into a single, unified catalog, ensuring that all pertinent data is readily accessible in one centralized location.

Upon successfully extracting the metadata, it is then systematically stored in a graph database. This storage solution was specifically chosen for its superior capability to handle complex relationships between vast arrays of data points effectively. The inherent structure of a graph database not only supports a more interconnected data environment but also greatly enhances the system's search capabilities, making it an ideal choice for academic institutions where data is interlinked across various disciplines and departments. Following the storage of metadata, the implementation of a full-text search engine marks a significant enhancement in the system's functionality. This engine facilitates quick and efficient searches, allowing end users students, faculty, and researchers to perform rapid searches across the compiled metadata. This feature is particularly crucial as it underpins the quick retrieval of academic and research data, thereby boosting the productivity and efficiency of academic endeavors.

The culmination of this project is marked by the development of an intuitive and user-friendly interface, designed specifically to meet the needs of a diverse academic population. This interface is pivotal in ensuring that all users can navigate through the data catalog with ease, finding and utilizing the needed resources without technical hurdles. The design and functionality of the user interface focus on minimizing complexity and enhancing user engagement, which is essential for fostering an environment where data is not only accessible but also actionable. By simplifying access to information, this interface helps in maximizing the utility of the cataloged data, thereby contributing to a more dynamic and responsive academic and research landscape across the Moroccan universities. This approach not only showcases Amundsen's capabilities in managing large datasets but also highlights its adaptability to meet specific needs within an educational context, thus reinforcing the value of a well-implemented data catalog system.

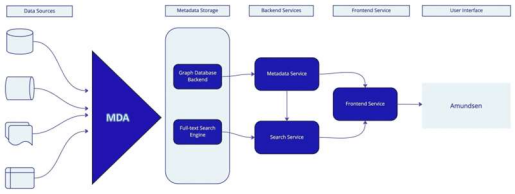


Figure 7:Amundsen _ MDA architecture

4.2. Scenario 2: Integrating Amundsen with MDA

In the second scenario, the deployment of Amundsen integrated with Model-Driven Architecture (MDA) across twelve Moroccan universities represents an advanced approach to establishing a highly sophisticated data management system. This integration seeks to harness the strengths of both Amundsen and MDA to create a versatile and scalable platform that enhances the cataloging, management, and utilization of metadata across academic and research resources. The objective is to build a system that not only meets the current data management needs of these universities but is also capable of adapting to future requirements and technological advancements, thereby ensuring long-term sustainability and efficiency.

The integration process starts with the utilization of Amundsen's DataBuilder utility, which is adept at connecting to various data sources to extract metadata. This capability is coupled with MDA's robust modeling frameworks, which

facilitate the structured organization and detailed representation of data relationships. By applying MDA's principles, the project effectively maps complex metadata into comprehensive models that outline clear relationships and dependencies among data points. This metadata, once extracted and modeled, is stored in a graph database designed to optimize the interconnectivity and accessibility of information. The graph database, with its inherent flexibility to manage complex data structures, is particularly well-suited for academic environments where data is interconnected across various disciplines and administrative sectors. This strategic storage solution ensures that updates, scalability, and modifications to the database can be handled with minimal disruption, reflecting changes in the academic landscape swiftly and accurately.

Following the modeling and storage of metadata, an advanced full-text search engine is implemented to navigate this structured data efficiently. This search engine is tailored to exploit the rich metadata models created through MDA, enabling precise and context-aware searches. This capability significantly enhances the discovery process, allowing students, faculty, and researchers to locate specific data with greater accuracy and relevance to their needs. Moreover, the system incorporates sophisticated algorithms that can interpret complex queries, thus providing more relevant and comprehensive search results.

The culmination of this integration is the development of a user-friendly interface, meticulously designed to cater to a diverse range of users from various academic backgrounds. The interface is not only intuitive but also leverages the structured data models to provide guided navigation through complex datasets, simplifying the user interaction with the system. It supports a variety of user activities, from straightforward data retrieval to intricate analytical tasks, making it an indispensable tool for academic research and administration. By enhancing the interface to accommodate the detailed and structured data models, the system ensures that all users regardless of their technical expertise can fully utilize the extensive capabilities of the catalog. This approach not only elevates the data management practices within the universities but also significantly boosts academic productivity and collaborative research opportunities, fostering a more data-driven culture across the institutions.

5. RESULTS & DISCUSSION

In this section, we explore the outcomes following the implementation of two automated data catalog scenarios across twelve major Moroccan universities. The first scenario solely utilizes Amundsen, while the second combines Amundsen with Model-Driven Architecture (MDA). The exclusive use of Amundsen facilitated the connection of various institutional data sources, enabling the extraction and integration of metadata into a centralized catalog, thus simplifying metadata management. Conversely, the scenario that integrates Amundsen with MDA allowed for more structured data modeling, providing significant advantages in terms of detailed data relationship representation, which enhanced data interoperability and governance through precise and maintainable models.

Both scenarios were assessed on several metrics, including accuracy, precision, and recall. The evaluations indicate that despite increased complexity, the integration of MDA significantly enhances metadata management, precision, and search capabilities. These enhancements are also evident in the user interface design, which has become more advanced and customizable, and in the robust management of metadata.

Table 2 :Evaluation Metrics

Metric	Amundsen	Amundsen with MDA
Accuracy(%)	82.35	89.72
Precision(%)	75.47	88.24
Recall(%)	70.38	85.62
Metadata Automation Timing	Fast	Faster

The improvements in accuracy, precision, and recall with the MDA-integrated scenario highlight the superiority of this approach in metadata processing quality. The reduction in metadata automation time per university also reflects significant process optimization, crucial in environments where speed and responsiveness are paramount.

Table 3:. Impact of System Architecture on Performance

Architecture Component	Amundsen	Amundsen with MDA
User Interface	Simple, intuitive	Advanced, customizable
Metadata Management	Basic	Modeled, robust
Database	Traditional graph	Optimized graph
Search Engine	Standard full-text	Advanced contextual
System Integration	Limited	Extensive, with flexible APIs
Collaboration Support	Basic	Enhanced with advanced collaborative features
Adaptability and Scalability	Moderate	High, customizable

MDA provides substantial benefits in terms of user interface, metadata management, and collaboration support, demonstrating its capability to address the complex and evolving needs of academic environments. Each component has seen significant enhancements, illustrating a more robust and flexible data system architecture.

Deep integration of Amundsen with MDA has also surpassed the limitations of traditional graph databases by optimizing their structure for better management of complex data relationships. This optimization aids in more sophisticated queries and better aligns with the needs of end-users, maintaining high performance even with increasing data volumes. The search engine also benefits from this integration, transitioning from a standard full text to an advanced contextual search. This upgrade enhances the understanding of query contexts, delivering more accurate and relevant results that improve user experiences and academic research process efficiencies. System integration has broadened due to MDA's flexible APIs, enhancing connectivity with other systems and platforms within universities. This facilitates more effective collaboration and improved communication between various departments and services, thereby strengthening institutional synergy. Collaboration support has significantly improved, moving from basic functionalities to a suite of advanced collaborative features. These improvements enable richer user interactions, facilitating metadata sharing, annotation, and real-time communication, which are vital for collaborative research projects and academic development. The system's adaptability and scalability have also been markedly enhanced, allowing universities to tailor and expand their data catalog systems in line with their evolving needs without compromising performance. This flexibility is critical for adapting to rapid changes in the fields of higher education and research. These comprehensive enhancements, demonstrated across the two implementation scenarios, highlight the effectiveness of MDA integration in advanced data catalog management, providing a robust platform for academic data management and decision-making based on accurate and accessible information.

Existing studies on data catalogs have predominantly concentrated on metadata management, data discovery, and governance, often relying on manual or semi-automated processes. While research has explored aspects such as metadata enrichment, search optimization, and data lineage tracking, the potential of a fully automated approach using Model-Driven Architecture (MDA) remains largely underexplored. Unlike previous works, which typically employ predefined taxonomies and rule-based metadata extraction, our approach leverages MDA to automate schema evolution, metadata generation, and model transformations. This automation not only minimizes human intervention but also enhances metadata accuracy and ensures greater scalability. Moreover, our framework introduces a systematic methodology for integrating heterogeneous data sources across academic institutions, improving interoperability and governance beyond conventional cataloging techniques. By embedding automation at the core of data catalog management, this research establishes a novel paradigm that strengthens metadata quality, streamlines governance, and optimizes accessibility in large-scale institutional environments.

6. CONCLUSION AND FUTURE PERSPECTIVES

This research has thoroughly explored the crucial role of Model-Driven Architecture (MDA) in automating data catalogs, particularly within the context of academic institutions. By integrating MDA, data catalogs transform from static repositories into dynamic, scalable, and adaptive systems capable of addressing the evolving demands of institutional data governance. Through the implementation across twelve Moroccan universities, this study has demonstrated that the MDA-based approach significantly reduces human error, streamlines metadata management, and improves data accessibility. The structured automation ensures consistent metadata modeling, simplifies schema evolution, and facilitates managing complex data relationships critical elements in academic environments that deal with vast and heterogeneous data sources. By leveraging MDA alongside Amundsen, these institutions can now manage metadata more effectively, fostering a data-driven culture that enhances research and educational outcomes.

The research successfully achieved its primary objectives:

- **Automation & Accuracy**: MDA-driven automation substantially reduced inconsistencies in metadata management, ensuring a structured and reliable data catalog.
- **Scalability**: The proposed framework proved adaptable to evolving institutional needs, allowing seamless updates to metadata schemas without extensive manual intervention.
- **Interoperability**: The integration of MDA-based APIs enhanced system connectivity, promoting a unified data ecosystem across institutions.

- **Governance & Data Quality**: Automated metadata validation and lineage tracking strengthened data governance and ensured compliance with institutional policies.

While these achievements are significant, several limitations should be acknowledged:

- **Generalizability**: Although the implementation was successful across twelve universities, its applicability to large-scale enterprise environments requires further investigation.
- **Adoption Complexity**: Transitioning from traditional metadata management to an MDA-driven automated approach demands expertise and institutional commitment, which may hinder adoption.
- **Scalability in Big Data Environments**: While MDA improved scalability for structured metadata, its effectiveness in highly dynamic big data **ecosystems** remains an area for further exploration.
- **Integration with Emerging Technologies**: The study primarily focused on MDA and Amundsen, leaving opportunities to explore integration with AI-driven metadata enrichment and cloud-native data governance frameworks.

To address these limitations and further enhance the automation of data catalogs, several areas offer promising opportunities:

- **Integration of Artificial Intelligence and Machine Learning**: Advanced algorithms could automate complex tasks such as metadata categorization, tagging, and predicting data usage patterns, enhancing both the accuracy and utility of data catalogs.
- **Advanced Interoperability Standards**: Developing robust interoperability standards will be essential as data ecosystems grow more complex, ensuring seamless integration across diverse platforms, particularly within interconnected academic environments.
- **Adoption of Cloud Technologies**: Extending MDA-driven data catalogs to cloud environments will enable real-time data processing, scalability, and cost efficiency, supporting larger volumes of data.
- **Focus on Data Privacy & Security**: Strengthening data security mechanisms will be crucial in light of increasing regulatory requirements, ensuring the privacy and protection of sensitive academic data.
- **User-Centric Design**: Enhancing the user interface and experience of data catalogs will support non-technical users, democratizing data access and fostering a data-driven culture across academic institutions.

This study establishes a strong foundation for advancing data catalog automation through MDA, offering a structured, scalable, and interoperable approach to metadata governance. By addressing the identified limitations and embracing future technological advancements, academic institutions can further optimize their data management strategies, driving research and educational excellence in an increasingly data-driven world.

REFERENCES:

[1] What is Data Catalog and Why You Should Care?, KDnuggets (2019). https://www.kdnuggets.com/what-is-data-catalog-and-why-you-should-care.html (accessed June 11, 2023).

[2] A.-M. Järvenpää, J. Jussila, I. Kunttu, Barriers and Practical Challenges for Data-driven Decision-making in Circular Economy SMEs, in: A. Visvizi, O. Troisi, M. Grimaldi (Eds.), Big Data and Decision-Making: Applications and Uses in the Public and Private Sector, Emerald Publishing Limited, 2023: pp. 163–179. https://doi.org/10.1108/978-1-80382-551-920231011.

[3] N. Jahnke, B. Otto, Data Catalogs in the Enterprise: Applications and Integration, Datenbank Spektrum 23 (2023) 89–96. https://doi.org/10.1007/s13222-023-00445-2.

[4] L. Ehrlinger, J. Schrott, M. Melichar, N. Kirchmayr, W. Wöß, Data Catalogs: A Systematic Literature Review and Guidelines to Implementation, in: G. Kotsis, A.M. Tjoa, I. Khalil, B. Moser, A. Mashkoor, J. Sametinger, A. Fensel, J. Martinez-Gil, L. Fischer, G. Czech, F. Sobieczky, S. Khan (Eds.), Database and Expert Systems Applications - DEXA 2021 Workshops, Springer International Publishing, Cham, 2021: pp. 148–158. https://doi.org/10.1007/978-3-030-87101-7_15.

[5] The Commonwealth Scientific and Industrial Research Organisation, Current Biology 7 (1997) R126. https://doi.org/10.1016/S0960-9822(97)70976-X.
[6] D. Welter, P. Rocca-Serra, V. Grouès, N. Sallam, F. Ancien, A. Shabani, S. Asariardakani, P. Alper, S. Ghosh, T. Burdett, S.-A. Sansone, W. Gu, V. Satagopam, The Translational Data Catalog - discoverable biomedical datasets, Sci Data 10 (2023) 470. https://doi.org/10.1038/s41597-023-02258-0.
[7] M. Yee, A. Surkis, I. Lamb, N. Contaxis, The NYU Data Catalog: a modular, flexible infrastructure for data discovery, Journal of the American Medical Informatics Association 30 (2023) 1693–1700. https://doi.org/10.1093/jamia/ocad125.
[8] D. Petrik, A. Untermann, H. Baars, Functional Requirements for Enterprise Data Catalogs: A Systematic Literature Review, in: S. Hyrynsalmi, J. Münch, K. Smolander, J. Melegati (Eds.), Software Business, Springer Nature Switzerland, Cham, 2024: pp. 3–18. https://doi.org/10.1007/978-3-031-53227-6_1.
[9] D.K. Shin, S.H. Lee, J. Kang, E.M. Park, Data Catalogue Standards Based on DCAT for Transportation Data: DCAT-Trans, jkst 37 (2019) 430–444. https://doi.org/10.7470/jkst.2019.37.5.430.
[10] P.P.F. Barcelos, T.P. Sales, M. Fumagalli, C.M. Fonseca, I.V. Sousa, E. Romanenko, J. Kritz, G. Guizzardi, A FAIR Model Catalog for Ontology-Driven Conceptual Modeling Research, in: J. Ralyté, S. Chakravarthy, M. Mohania, M.A. Jeusfeld, K. Karlapalem (Eds.), Conceptual Modeling, Springer International Publishing, Cham, 2022: pp. 3–17. https://doi.org/10.1007/978-3-031-17995-2_1.
[11] H. Dibowski, S. Schmid, Y. Svetashova, C. Henson, T. Tran, Using Semantic Technologies to Manage a Data Lake: Data Catalog, Provenance and Access Control, (2020).
[12] A.W. Brown, J. Conallen, D. Tropeano, Introduction: Models, Modeling, and Model-Driven Architecture (MDA), in: S. Beydeda, M. Book, V. Gruhn (Eds.), Model-Driven Software Development, Springer, Berlin, Heidelberg, 2005: pp. 1–16. https://doi.org/10.1007/3-540-28554-7_1.
[13] Data Catalog Architecture: Components, Integrations, & More, (2023). https://atlan.com/data-catalog-architecture/ (accessed December 25, 2023).
[14] Data Curation Definition, Explanation & Examples , Explanation & Examples | Secoda, (2022). https://www.secoda.co/glossary/data-curation (accessed November 30, 2023).
[15] E. Chazbani, Building a Winning Data Quality Strategy: Step by Step, IBM Blog (2023). https://www.ibm.com/blog/winning-data-quality-strategy/www.ibm.com/blog/winning-data-quality-strategy (accessed December 17, 2023).
[16] O. Rafique, M. Gesell, K. Schneider, Targeting different abstraction layers by model-based design methods for embedded systems: A case study, in: 2013 IEEE 19th International Conference on Embedded and Real-Time Computing Systems and Applications, 2013: pp. 334–337. https://doi.org/10.1109/RTCSA.2013.6732235.
[17] Amundsen Data Catalog: Features, Setup, Uses & Alternatives, (2023). https://atlan.com/amundsen-data-catalog/ (accessed July 10, 2024).
[18] J. Hilger, Z. Wahl, Data Catalogs and Governance Tools, in: J. Hilger, Z. Wahl (Eds.), Making Knowledge Management Clickable : Knowledge Management Systems Strategy, Design, and Implementation, Springer International Publishing, Cham, 2022: pp. 187–192. https://doi.org/10.1007/978-3-030-92385-3_11.
[19] B. Ait-Amir, P. Pougnet, A. El Hami, Meta-Model Development, in: Embedded Mechatronic Systems 2, Elsevier, 2015: pp. 151–179. https://doi.org/10.1016/B978-1-78548-014-0.50006-2.
[20] Data Integration vs Data Migration: A Comparative Study - Learn | Hevo, (2022). https://hevodata.com/learn/data-integration-vs-data-migration/ (accessed June 11, 2023).
[21] P. Subramaniam, Y. Ma, C. Li, I. Mohanty, R.C. Fernandez, Comprehensive and Comprehensible Data Catalogs: The What, Who, Where, When, Why, and How of Metadata Management, (2023). http://arxiv.org/abs/2103.07532 (accessed June 11, 2023).
[22] S.E. McCord, J.L. Welty, J. Courtwright, C. Dillon, A. Traynor, S.H. Burnett, E.M. Courtright, G. Fults, J.W. Karl, J.W. Van Zee, N.P. Webb, C. Tweedie, Ten practical questions to improve data quality, Rangelands 44 (2022) 17–28. https://doi.org/10.1016/j.rala.2021.07.006.
[23] R. Gupta, D. Saxena, A.K. Singh, Data Security and Privacy in Cloud Computing: Concepts and Emerging Trends, (2021).

http://arxiv.org/abs/2108.09508 (accessed June 11, 2023).
[24] E. Quimbert, K. Jeffery, C. Martens, P. Martin, Z. Zhao, Data Cataloguing, in: Z. Zhao, M. Hellström (Eds.), Towards Interoperable Research Infrastructures for Environmental and Earth Sciences, Springer International Publishing, Cham, 2020: pp. 140–161. https://doi.org/10.1007/978-3-030-52829-4_8.
[25] J.-N. Mazón, J. Trujillo, An MDA approach for the development of data warehouses, Decision Support Systems 45 (2008) 41–58. https://doi.org/10.1016/j.dss.2006.12.003.
[26] M. Soukaina, Model Driven Engineering (MDE) Tools: A Survey, AJSET 3 (2018) 29. https://doi.org/10.11648/j.ajset.20180302.11.
[27] A. Erraissi, A. Belangour, An Approach Based On Model Driven Engineering For Big Data Visualization In Different Visual Modes, 9 (2020).
[28] H. Dev, A. Seth, MDA based approach towards Design of Database for Banking System, IJCA 49 (2012) 32–37. https://doi.org/10.5120/7713-1129.
[29] A. Srai, F. Guerouate, N. Berbiche, H.D. Lahsini, MDA Approach for EJB Model, in: 2018 6th International Conference on Multimedia Computing and Systems (ICMCS), IEEE, Rabat, 2018: pp. 1–6. https://doi.org/10.1109/ICMCS.2018.8525924.
[30] M. Polo, I. García-Rodríguez, M. Piattini, An MDA-based approach for database re-engineering, J. Softw. Maint. Evol.: Res. Pract. 19 (2007) 383–417. https://doi.org/10.1002/smr.353.
[31] Z. Mohammadzadeh, H.R. Saeidnia, M. Kozak, A. Ghorbi, MDA Framework for FAIR Principles, in: J. Mantas, A. Hasman, M.S. Househ, P. Gallos, E. Zoulias, J. Liaskos (Eds.), Studies in Health Technology and Informatics, IOS Press, 2022. https://doi.org/10.3233/SHTI210888.
[32] J. Bézévin, Sur les principes de base de l'ingénierie des modèles, Objet 10 (2004) 145–157. https://doi.org/10.3166/objet.10.4.145-157.
[33] M.A. Resources Information, Software Design and Development: Concepts, Methodologies, Tools, and Applications: Concepts, Methodologies, Tools, and Applications, IGI Global, 2013.
[34] A. Brown, Model driven architecture: Principles and practice, Software and System Modeling 3 (2004) 314–327. https://doi.org/10.1007/s10270-004-0061-2.
[35] A.G. Kleppe, J. Warmer, W. Bast, MDA Explained: The Model Driven Architecture: Practice and Promise, Addison-Wesley Longman Publishing Co., Inc., USA, 2003.
[36] D. Wells, Introduction to Data Catalogs, (2019).
[37] C. Minotti, A. Ashton, S. Bliven, F. Bolmsten, S. Egli, M. Leorato, D. McReynolds, M. Novelli, T. Richter, L. Shemilt, Enhancing Data Management with SciCat: A Comprehensive Overview of a Metadata Catalogue for Research Infrastructures, (2023) 6 pages, 0.439 MB. https://doi.org/10.18429/JACOW-ICALEPCS2023-THMBCMO02.
[38] C. Labadie, C. Legner, M. Eurich, M. Fadler, FAIR Enough? Enhancing the Usage of Enterprise Data with Data Catalogs, in: 2020 IEEE 22nd Conference on Business Informatics (CBI), IEEE, Antwerp, Belgium, 2020: pp. 201–210. https://doi.org/10.1109/CBI49978.2020.00029.
[39] J. Martinez-Gil, Framework to Automatically Determine the Quality of Open Data Catalogs, (2024). http://arxiv.org/abs/2307.15464 (accessed June 7, 2024).
[40] V. Kopsachilis, M. Vaitis, GeoLOD: A Spatial Linked Data Catalog and Recommender, BDCC 5 (2021) 17. https://doi.org/10.3390/bdcc5020017.