Effectiveness of Google Cloud Vision, Amazon Rekognition, and Microsoft Azure's Computer Vision Capability for generating Metadata tags for Images

Table of Contents

Acknowledgments

This dissertation would not have been possible without the invaluable guidance and support of several esteemed academics and mentors. First and foremost, I would like to express my deepest gratitude to my supervisor, Dr. Hannah Ishmael, for her insightful feedback, unwavering encouragement, and steadfast commitment throughout the entire research process. Her expertise and dedication have been instrumental in shaping the direction and quality of this work.

I am also indebted to Dr. Laura Gibson, whose constant support and mentorship have been a source of inspiration and motivation. Her incisive perspectives and constructive criticism have significantly contributed to the development and refinement of this dissertation. I would also like to extend my sincere appreciation to Dr Daniel Chávez Heras and Dr. David Young for their valuable feedback and guidance in navigating the technical aspects of this research. Their expertise in coding and the nuances of digital methodologies have been instrumental in enhancing the rigor and implementation of this study. Furthermore, I am grateful to Dr. Erik Ketzan for his insightful feedback and thoughtful engagement throughout this endeavour. His erudite insights and recommendations have been invaluable in strengthening the overall quality and coherence of this work.

Moreover, I would like to express my gratitude to King’s College London. As an international student, I have been fortunate to experience the exceptional academic environment and resources provided by the university, which have greatly enriched my learning experience.

I am also thankful to the Department of Digital Humanities at King's College for the vibrant intellectual community and support, which have been instrumental in shaping my understanding and expanding the horizons of this research.

Lastly, I would like to express my heartfelt appreciation to my family, wife, and friends, whose unwavering belief and steadfast support have been the foundation for this accomplishment. Their encouragement and understanding have been a constant source of inspiration throughout this journey.

Abstract

This dissertation investigates the effectiveness of three leading cloud-based AI services— Google Cloud Vision, Amazon Rekognition, and Microsoft Azure Computer Vision—in generating accurate and comprehensive metadata tags for a diverse set of images. The study employs a robust mixed-methods approach, integrating quantitative metrics and qualitative analysis, to evaluate the performance of these AI services across various image categories, including people, art and creative craft, scenes of daily life, nature, objects, text and documents, and urban and rural settings.

The findings reveal significant variations in the performance of the AI services, with Microsoft Azure Computer Vision and Amazon Rekognition outperforming Google Cloud Vision in several categories. While the AI services demonstrate high precision in the tags they generate, they consistently struggle with recall, indicating that they produce accurate but incomplete metadata. The qualitative analysis further highlights the AI services' limitations in capturing the nuanced contextual information and cultural sensitivity required for effective metadata in cultural heritage institutions.

The study underscores the need for a nuanced, category-specific approach to AI integration in digital asset management, leveraging the strengths of different services while maintaining human oversight. It also emphasizes the importance of continued development of AI models trained on diverse, culturally sensitive datasets, as well as the critical role of human expertise in ensuring the accuracy, completeness, and cultural relevance of metadata. The findings have significant implications for the integration of AI technologies in digital asset management practices, particularly for organizations dealing with large and diverse image collections.

Keywords: Metadata, AI, Computer Vision, Digital Asset Management, Image Collections,

Cultural Heritage

Introduction

In the rapidly evolving landscape of digital asset management, metadata tagging has emerged as a crucial element for the effective organization, retrieval, and management of digital assets across various sectors. This dissertation investigates the effectiveness of three leading cloudbased AI services—Google Cloud Vision, Amazon Rekognition, and Microsoft Azure Computer Vision—in generating accurate and comprehensive metadata tags for a diverse set of images. The study specifically focuses on evaluating the performance of these services in aligning with customized controlled vocabularies, which are essential for maintaining consistency and interoperability in metadata management.

Metadata, often described as 'data about data,' serves a multitude of functions, including facilitating resource discovery, providing organizational context, ensuring interoperability, supporting digital preservation, and managing rights (Gilliland, 2019). There are several types of metadata: descriptive metadata helps in identifying and discovering resources, structural metadata provides information on the organization and relationships within resources, and administrative metadata aids in the management and preservation of digital collections. The use of controlled vocabularies is essential in this context to standardize the terms used for metadata, ensuring consistency and accuracy. Controlled vocabularies are organized lists of terms that help categorize and describe digital assets uniformly, making them easier to find and navigate.

The exponential growth of digital content has necessitated the development of efficient and accurate methods for metadata enrichment, and the advent of AI-driven computer vision services has presented a promising solution to this challenge. These services employ advanced machine learning algorithms to analyse images and automatically generate descriptive tags, potentially revolutionizing the process of metadata tagging. Computer vision, a field of artificial intelligence that focuses on enabling computers to interpret and understand visual information from around the world, such as images and videos, forms the backbone of these services (Szeliski, R., 2010).

The importance of this research lies in the critical role that effective metadata management plays in enhancing the usability and value of digital collections. Metadata acts as a vital navigational tool, enabling users to find resources quickly and efficiently, thus saving time and effort. Effective metadata management enhances the usability and value of digital collections by providing critical information about the content, context, and provenance of digital assets, supporting various organizational functions such as cataloguing, searchability, and long-term digital preservation (Baca, 2016). However, traditional manual tagging methods are not only labour-intensive and time-consuming but also prone to inconsistencies and human error. AIdriven computer vision services offer a scalable and efficient alternative, potentially improving the accuracy and comprehensiveness of the tags while significantly reducing the time and effort required for metadata creation (NISO, 2004).

The primary aim of this research is to evaluate and compare the effectiveness of Google Cloud Vision, Amazon Rekognition, and Microsoft Azure Computer Vision in generating accurate and comprehensive

metadata tags for a broad set of images. These services were chosen due to their prominence in the industry, widespread adoption across various sectors, and robust feature sets specifically tailored for image analysis and metadata generation (Google, 2023; Amazon, 2023; Microsoft, 2023). Google Cloud Vision is renowned for its robust image recognition capabilities, including object detection, facial recognition, and landmark recognition. Amazon Rekognition offers a comprehensive suite of image and video analysis tools that can detect objects, scenes, and activities, as well as perform facial analysis and text recognition. Microsoft Azure Computer Vision provides a range of features such as optical character recognition (OCR), image categorization, and spatial analysis.

This overarching aim is supported by several specific objectives. First, the study will assess the accuracy and comprehensiveness of metadata tags generated by each AI service across a diverse range of image types and content. This objective is crucial as it will provide insights into the overall performance of each service across various image categories, helping to identify strengths and weaknesses in their tagging capabilities. Second, it will evaluate the alignment of AI-generated tags with customized controlled vocabularies, including the UKATArchival Thesaurus, Getty Art & Architecture Thesaurus (AAT), and Library of Congress Subject Headings. This objective is particularly important as it addresses the practical applicability of these services in real-world scenarios where specific vocabularies are often used to maintain consistency and interoperability in metadata management.

Third, the research will compare the performance of the three AI services in terms of precision, recall, and F1 score for metadata tag generation. These metrics are standard in the field of information retrieval and will provide a quantitative basis for comparing the services' performance. Precision measures the proportion of relevant tags among the retrieved tags, recall measures the proportion of relevant tags that are successfully retrieved, and the F1 score provides a balanced measure of both precision and recall. By analysing these metrics, the study aims to provide a comprehensive evaluation of each service's performance in generating accurate and relevant metadata tags.

To achieve these objectives, this study will employ a mixed-methods approach that integrates both quantitative and qualitative assessments. A convenience sample of 210 images covering a wide range of categories will be used, including people, objects, nature, cultural and historical contexts, urban and rural settings, text and documents, art and creative works, and daily life scenes. This diverse dataset will provide a comprehensive testbed for evaluating the performance of the AI services. The methodology involves processing each selected image through the three AI services to generate metadata tags, comparing the AI-generated tags to manually create a ground truth of controlled vocabularies, and calculate various performance metrics.

The quantitative analysis will involve calculating accuracy, precision, recall, and F1 scores for each service across different image categories. This will provide a clear picture of how well each service performs in generating relevant and accurate tags. The qualitative assessment will involve a detailed analysis of the

generated tags, evaluating their relevance, comprehensiveness, and alignment with the controlled vocabularies. This dual approach will offer a nuanced understanding of each service's strengths and limitations, going beyond mere numerical comparisons to provide insights into the practical usefulness of the generated tags in real-world scenarios.

Furthermore, this study recognizes the ethical implications of using AI technologies for metadata tagging, particularly concerning data privacy and the representation of culturally sensitive materials. Measures will be taken to ensure the ethical conduct of the research, including secure data storage, anonymizing sensitive information, and obtaining informed consent from relevant stakeholders. The study will also address potential biases in the AI models by diversifying the training datasets and implementing robust quality control mechanisms to ensure that the generated metadata accurately reflects the content and context of the assets. Addressing these ethical issues is essential to maintaining the integrity and trustworthiness of AI-driven metadata tagging systems (Floridi, 2019).

The findings of this study will contribute to the growing body of knowledge on the application of AI technologies in digital asset management, providing valuable insights into the strengths and limitations of different approaches to AI-driven image analysis and metadata generation.

By comparing the performance of Google Cloud Vision, Amazon Rekognition, and Microsoft Azure Computer Vision, this research aims to inform the development of more effective and efficient metadata tagging systems, ultimately enhancing the organization, retrieval, and preservation of digital assets across various sectors.

The results of this study will be of particular interest to researchers, practitioners, and organizations seeking to leverage AI technologies to enrich metadata in large-scale image collections, offering guidance on the selection and implementation of appropriate AI services based on their specific needs and requirements. Libraries, archives, museums, and other institutions managing large digital collections stand to benefit significantly from the insights provided by this research, as it will help them make informed decisions about integrating AIdriven metadata tagging into their workflows.

Moreover, this study will contribute to the ongoing dialogue about the role of AI in cultural heritage and information management fields. By critically examining the performance of these AI services, the research will shed light on the current state of the technology and its potential for future development. This information will be valuable not only for practitioners but also for policymakers and technology developers working on improving AI systems for metadata generation and digital asset management.

In conclusion, this dissertation aims to provide a comprehensive evaluation of the effectiveness of Google Cloud Vision, Amazon Rekognition, and Microsoft Azure Computer Vision in generating metadata tags for diverse image collections. By employing a rigorous mixedmethods approach and addressing both performance metrics and ethical considerations, the study seeks to offer valuable insights that will advance

the field of AI-driven metadata management and contribute to more efficient and accurate digital asset organization and retrieval systems.

This dissertation is structured into six chapters. Chapter 1, the current chapter, introduces the research problem, objectives, and significance of the study, providing an overview of the key concepts and technologies involved in AI-driven metadata tagging for image collections. Chapter 2 presents a comprehensive review of the relevant literature, covering studies on the performance of AI services in metadata tagging, challenges related to bias and accuracy, ethical considerations, and the implications for digital asset management practices. The literature review identifies research gaps and provides a foundation for the current study. Chapter 3 describes the research methodology employed, explaining the mixed-methods approach, data collection procedures, and the development and application of controlled vocabularies, as well as discussing the data analysis techniques used to evaluate the performance of the AI services. Chapter 4 presents the findings of the study, organized by image category, providing both quantitative and qualitative results. The quantitative findings include accuracy, precision, recall, and F1 scores for each AI service across different image categories, while the qualitative analysis offers insights into the contextual relevance, semantic depth, cultural sensitivity, and alignment with controlled vocabularies of the AI-generated tags. Chapter 5 discusses the findings in relation to the research objectives and the broader literature, exploring the implications for digital asset management practices and highlighting the need for human-AI collaboration, as well as addressing the limitations of the study and suggesting future research directions. Finally, Chapter 6 concludes the dissertation by summarizing the key findings, contributions, and implications of the research, emphasizing the potential of AI technologies in enhancing metadata tagging while acknowledging the challenges and the importance of human expertise in the process.

Literature review

The integration of AI technologies in metadata tagging has ushered in a paradigm shift in digital asset management, particularly in the realm of image-based collections. This literature review critically examines the current state of research on three prominent AI-driven computer vision services: Google Cloud Vision, Amazon Rekognition, and Microsoft Azure's Computer Vision. These technologies have garnered significant attention for their potential to automate and enrich metadata generation for large-scale image collections.

This review is structured around six key themes: (1) the foundational role of metadata in DAM;

(2) comparative analyses of the aforementioned AI services' performance in metadata tagging; (3) institutional case studies and real-world applications; (4) challenges inherent in AI-driven metadata generation, including issues of bias and accuracy; (5) methodological approaches to evaluating these systems and associated ethical considerations; and (6) emerging applications and future directions in the field.

By synthesizing findings from recent empirical studies, theoretical frameworks, and practitioner insights, this review aims to provide a comprehensive understanding of the capabilities, limitations, and implications of AI-driven metadata tagging in cultural heritage and broader DAM contexts. Particular attention is paid to the trade-offs between automation and human expertise, the contextual nuances that challenge AI systems, and the ethical considerations that arise from deploying these technologies in sensitive cultural domains.

Through this critical examination, the review seeks to identify gaps in current knowledge, highlight promising avenues for future research, and provide a nuanced perspective on the role of AI in shaping the future of metadata management for large-scale image collections.

The role of Metadata in Digital Asset Management

Metadata is crucial in DAM, serving as the backbone for organizing, retrieving, and preserving digital content. High-quality metadata enhances searchability, accessibility, and the overall utility of digital assets (Austerberry, 2005). The absence of detailed and rich descriptive metadata significantly impairs the discoverability and usability of cultural heritage collections, particularly in large-scale aggregation platforms like Europeana (Kaldeli et al., 2021). Kortemeyer et al. (2014) echo this observation in the context of educational resources, arguing that comprehensive metadata and quality measures are essential features that distinguish digital library assets from the vast array of resources available on the open web.

In educational settings, high-quality metadata serves multiple crucial functions: guiding users to appropriate assets, providing reliability information, and establishing meaningful connections between resources. The consequences of poor metadata quality are severe, potentially rendering resources invisible within a repository and drastically limiting their potential use and value. This problem becomes even more pronounced in large-scale systems housing hundreds of thousands or millions of resources.

The integration of AI technologies into metadata tagging aims to automate and improve this process, offering the potential for more detailed, accurate, and contextually rich metadata. However, there is limited research on how AI-generated metadata compares to traditional human-generated metadata in terms of depth, context, and usability across different types of digital assets. This highlights a significant gap in the current literature that needs to be addressed to fully understand the potential and limitations of AI in metadata generation. While the importance of metadata in DAM is well-established, there are a lack of comprehensive studies on how AI-generated metadata compares to traditional human-generated metadata in terms of depth, context, and usability across different types of digital assets. The existing literature often focuses on general metadata principles, but there is limited research on the specific challenges and opportunities presented by AI in metadata generation for large-scale, diverse image collections. While the importance of high-quality metadata is clear, the advent of AI technologies has introduced new possibilities and challenges in metadata generation. To understand these developments, we must examine the performance of leading AI services in this domain.

Performance Analysis of Major AI Services in Metadata Tagging

Recent studies have explored the capabilities and limitations of various AI services in metadata tagging, providing valuable insights into their performance across different contexts and asset types. Google Cloud Vision, known for its robust image analysis capabilities, has shown promise in generating descriptive metadata for digital assets. Villaespesa and Crider (2023) conducted an in-depth analysis of its performance in tagging art collections, revealing that while Google Cloud Vision could accurately tag many objects, it faced significant challenges with specificity and contextual accuracy. This suggests that AI-generated tags, while broad in scope, sometimes lack the depth and nuanced context that human annotators can provide. The researchers found that Google Cloud Vision excelled in identifying general objects, colours, and basic compositional elements within artworks. However, it struggled with more nuanced aspects such as artistic style, historical context, and cultural significance. For instance, while the AI could identify a painting as containing a 'person' or 'face,' it often failed to recognize specific historical figures or artistic techniques that a human expert would readily identify.

Similarly, Amazon Rekognition, recognized for its object and scene detection capabilities, has demonstrated strengths in object recognition and automating metadata tagging processes (Sharma, 2022). Sharma's study highlighted Amazon Rekognition's ability to quickly process large volumes of images and generate tags for common objects, scenes, and actions. However, the research also identified limitations that need to be addressed to enhance its applicability in cultural heritage contexts, such as the range of supported image formats and the maximum image size it could effectively process. Additionally, like Google Cloud Vision, Amazon Rekognition showed limitations in identifying culturally specific elements and subtle artistic features.

Microsoft Azure's Computer Vision API offers comprehensive image analysis but exhibits performance variability based on environmental factors and object orientation (Temel, Lee, & AlRegib, 2019). Temel et al. highlighted the need for more refined models to handle diverse real-world conditions effectively, as factors such as lighting conditions, image quality, and the angle at which objects were presented could significantly impact the accuracy of the AI's analysis.

These findings collectively underscore the current limitations of AI in metadata tagging, particularly in contexts that require deep domain knowledge or cultural sensitivity. While AI services can process vast amounts of data quickly, they often lack the nuanced understanding that human experts bring to the task of metadata creation. Current research lacks a systematic comparison of the performance of AI services across a wide range of image categories, particularly in the context of cultural heritage collections. There is insufficient investigation into how these AI services perform when dealing with culturally diverse and historically significant images, which are common in many institutional collections. Having examined individual AI services, it is crucial to consider how these technologies compare to each other and how they are being applied in real-world institutional settings. Comparative studies and case applications provide valuable insights into the practical implications of AI-driven metadata tagging.

Comparative Studies and Institutional Applications

Comparative studies have provided valuable insights into the relative strengths and weaknesses of AI technologies in metadata tagging. Villaespesa and Crider (2023) conducted a comprehensive comparison of Google Cloud Vision, Amazon Rekognition, and IBM Watson in the context of art collections, revealing significant variability in the number and types of tags produced by each system. While Google Cloud Vision and Amazon Rekognition showed higher overall accuracy than IBM Watson, all systems exhibited challenges related to specificity and contextual relevance. The researchers found that Google Cloud Vision generally produced the highest number of tags per image, followed by Amazon Rekognition, with IBM Watson generating the fewest. However, quantity did not always correlate with quality. Google Cloud Vision's tags, while numerous, often included irrelevant or overly generic terms. Amazon Rekognition showed a better balance between quantity and relevance, but still struggled with highly specific or culturally nuanced content. IBM Watson, while producing fewer tags, sometimes captured unique aspects of artworks that the other systems missed. These findings underscore the need for combining AI with human expertise to achieve the best results. The variability in performance across different AI services suggests that no single system can currently meet all the metadata needs of cultural heritage institutions. Instead, a hybrid approach that leverages the strengths of multiple AI systems and incorporates human validation and enrichment may be the most effective strategy. The study also highlighted the importance of considering the specific needs and contexts of different collections when choosing AI services for metadata tagging, emphasizing the need for customized approaches and thorough evaluation before implementing AI solutions.

The potential of AI in enhancing metadata quality and interconnectivity between collections has been demonstrated in various institutional contexts. Storch (2023) conducted an in-depth study of the National Archives of Estonia, focusing on their efforts to incorporate AI into their metadata creation processes. The study emphasized the importance of combining human expertise with AI to achieve more nuanced and contextually accurate metadata, ensuring that the generated metadata is not only accurate but also contextually rich, enhancing the accessibility and usability of digital collections. The Estonian case study revealed several key findings. First, AI tools were particularly effective in processing large volumes of textual documents, extracting key information such as dates, names, and locations with high accuracy. For visual materials, AI analysis provided a good starting point, but human experts were crucial for verifying and enriching the AI-generated metadata, especially for culturally specific or historically significant elements. The integration of AI tools led to a significant increase in the speed of metadata creation, allowing the archive to process and make accessible a larger number of documents than was previously possible with manual methods alone. The combined AI-human approach resulted in more consistent metadata across the collection, as the AI provided a standardized base that human experts could then refine and enhance.

However, the study also highlighted that the optimal balance between automation and human intervention remains an area requiring further research. Questions remain about how to most effectively allocate human resources in an AI-assisted workflow and how to ensure that the efficiency gains of AI do not come at the cost of metadata quality or depth.

Other researchers have explored the potential of AI in enhancing interconnectivity between collections. For instance, Alliata, Hou, and Kenderdine (2024) presented a computational framework to enhance access to embodied knowledge archives through posture recognition and movement computing. Their work emphasized the importance of interdisciplinary collaboration among archivists, computists, artists, and knowledge holders in developing AI systems that can effectively capture and represent complex, multidimensional cultural knowledge. Their research highlights the potential of AI not just in generating descriptive metadata but in creating new ways of connecting and accessing cultural heritage materials. By recognizing patterns and relationships that might not be immediately apparent to human observers, AI has the potential to reveal new insights and connections within and across collections. While some comparative studies exist, there is a gap in research that comprehensively evaluates multiple AI services across diverse institutional contexts and collection types. The literature lacks in-depth case studies that examine the long-term impact of AI integration on metadata quality and management practices in cultural institutions. Despite the promising applications of AI in metadata generation, several significant challenges have emerged. These issues range from technical limitations to ethical concerns, highlighting the complexity of implementing AI solutions in cultural heritage contexts.

Challenges in AI-Driven Metadata Generation

A critical challenge identified across multiple studies is the issue of bias and accuracy in AIgenerated metadata. These biases often stem from the training datasets used to develop AI models, which may not represent diverse cultural and contextual nuances. Fornaro and Chiquet (2021) conducted a comprehensive study on the biases present in AI-generated metadata for cultural heritage collections, highlighting that AI systems could perpetuate biases present in their training data, leading to inaccurate or culturally insensitive metadata. This issue is particularly relevant in cultural heritage institutions, where the accuracy and sensitivity of metadata are paramount. The study identified several types of biases, including cultural bias, where AI systems trained primarily on Western art collections struggled to accurately identify and describe objects from non-Western cultures; temporal bias, where models showed better performance on contemporary art compared to historical artifacts; gender and racial bias, where AI systems often misidentified or used biased language when describing individuals of different genders or racial backgrounds; and contextual bias, where the AI frequently missed important historical or cultural contexts, leading to superficial or misleading descriptions. These findings underscore the importance of developing AI models with diverse, representative training data and implementing robust quality control measures. The researchers emphasized the need for ongoing efforts to diversify training datasets, involve experts from various cultural backgrounds in the development and testing of AI systems, and implement rigorous human oversight in the metadata creation process.

Addressing these biases requires a multifaceted approach, including diversifying training data, developing culturally aware AI, implementing bias detection tools, enhancing human-AI collaboration, and establishing ethical guidelines. This area demands further scholarly attention, as addressing these biases is crucial for

ensuring that AI-driven metadata tagging systems serve the diverse needs of cultural heritage institutions and their users.

The scalability and integration of AI technologies into existing digital asset management systems also poses significant challenges. Huddart (2022) conducted a comprehensive survey of AI integration in DAM systems across various industries, including cultural heritage institutions, highlighting that while AI has the potential to revolutionize digital asset management, its integration varies greatly among vendors. This leads to fragmented and inconsistent implementations. Key findings from Huddart's research include variability in AI capabilities, with different DAM systems offering widely varying AI functionalities; integration challenges, where many institutions faced difficulties in integrating AI tools with their existing workflows; scalability issues, with some systems struggling to maintain performance as collections grew; inconsistent metadata standards, leading to challenges in interoperability and data exchange between systems; and training and adaptation, where institutions often underestimated the time and resources required to train staff on new AIenhanced systems and to adapt workflows to make the best use of AI capabilities. These findings highlight the need for developing standardized integration frameworks to ensure consistent and effective implementation of AI technologies across different institutional contexts.

Paramount Film Production Company, one of the oldest major film studios in Hollywood, has been at the forefront of exploring the integration of AI technologies into its content management and production workflows. The case study by West, Denny, and Ruud (2021) on Paramount's DAM system provided a concrete example of both the potential and challenges of AI integration, illustrating the mixed results of AI implementation, with notable efficiency gains in areas like automatic tagging and content categorization, but also persistent accuracy issues, particularly for specialized or context-dependent content. Key lessons from the Paramount case study include the importance of a phased approach to AI integration, the need for ongoing human oversight and quality control, the value of customizing and fine-tuning AI models to the specific needs and content types of the institution, and the importance of clear communication and expectation management regarding the capabilities and limitations of AI tools. These challenges highlight the need for further research and development in several areas, such as developing more flexible and adaptable AI integration frameworks that can work with a variety of existing DAM systems; creating industry-wide standards for AI-generated metadata to improve interoperability and consistency across platforms; investigating scalable architectures that can maintain performance as collections grow and diversify; exploring user friendly interfaces and tools that make AI capabilities accessible to non-technical staff in cultural heritage institutions; and developing best practices for AI integration that address both technical and organizational challenges.

Addressing these scalability and integration issues is crucial for realizing the full potential of AI in enhancing metadata quality and management in cultural heritage contexts. Although biases in AI systems have been identified, there is limited research on effective strategies to mitigate these biases in the context of metadata generation for cultural heritage materials. The scalability of AI solutions for large, diverse collections remains

understudied, particularly in terms of maintaining performance and accuracy as collection sizes grow. Addressing these challenges requires careful consideration of both methodological approaches and ethical implications. The following section explores the current state of research methodologies and the ethical frameworks being developed to guide AI implementation in cultural heritage settings.

Methodological Approaches and Ethical Considerations

The evaluation of AI-driven metadata tagging systems presents unique challenges due to the diverse applications and contexts in which these systems are used. Current research employs a variety of methodological approaches and evaluation metrics, reflecting the complexity of the task and the need for multifaceted assessment. Many studies employ controlled experiments to evaluate the performance of AI systems under specific conditions, typically involving testing AI capabilities on pre-defined datasets and measuring metrics such as precision, recall, and F1 score (Sharma, 2022; Temel et al., 2019). While these quantitative measures provide valuable insights into the accuracy and completeness of AI-generated metadata, they may not fully capture the nuanced requirements of cultural heritage contexts.

Comparative analysis is also common, as seen in the works of Villaespesa and Crider (2023) and Storch (2023). These studies often involve side-by-side comparisons of different AI services or of AI-generated metadata against human-created metadata, revealing the relative strengths and weaknesses of different approaches and helping identify areas where AI performs well and where human expertise remains crucial.

However, there is a lack of standardized methodologies for evaluating AI performance across different institutional contexts and asset types. This presents an opportunity for methodological innovation in future research, with potential areas for development including context-sensitive evaluation frameworks that take into account the specific needs and standards of different types of cultural heritage institutions; multidimensional assessment models that consider not only accuracy but also factors like cultural sensitivity, historical context, and user relevance; longterm performance studies that track the effectiveness of AI systems over time; user-centric evaluation approaches that assess the impact of AI-generated metadata on end-user experiences; and ethical and bias evaluation frameworks that systematically assess AI systems for potential biases or culturally insensitive outputs. Developing more comprehensive and standardized evaluation methodologies is crucial for advancing the field and ensuring that AI systems meet the complex needs of cultural heritage institutions.

The ethical implications of AI in metadata tagging have gained increasing attention in recent literature, reflecting growing awareness of the potential impacts and risks associated with these technologies. Jaillant and Aske (2024) explored the ethical challenges in providing responsible access to historical medical illustrations using AI, highlighting the need for appropriate metadata to ensure discoverability and ethical access to sensitive images while balancing protection for vulnerable users and preventing misuse. Their study underscores the importance of developing ethical frameworks for AI implementation in cultural heritage contexts, an area that requires further investigation.

Key ethical considerations identified in the literature include privacy and consent, ensuring that AI-generated metadata does not inadvertently reveal sensitive information about individuals depicted in cultural heritage materials; cultural sensitivity, developing AI systems that can recognize and respect cultural taboos, sacred objects, or sensitive historical contexts; representation and bias, addressing the potential for AI systems to perpetuate or amplify existing biases in how different cultures, ethnicities, or historical periods are represented and described; accountability and transparency, establishing clear processes for how decisions are made in AI-assisted metadata creation, and ensuring that these processes are open to scrutiny and correction; intellectual property rights, navigating the complex landscape of copyright and ownership, particularly when AI systems are used to generate descriptive metadata for creative works; and access and equity, ensuring that the benefits of AI-enhanced metadata systems are equitably distributed and do not exacerbate existing digital divides.

These ethical considerations highlight the need for interdisciplinary collaboration in developing AI systems for cultural heritage contexts, involving ethicists, cultural experts, legal scholars, and technologists working together to create guidelines and frameworks that address these complex issues. Future research in this area could focus on developing ethical guidelines specifically tailored to AI use in cultural heritage metadata creation; creating mechanisms for community consultation and involvement in AI-driven metadata projects, particularly for culturally sensitive materials; exploring technical solutions for embedding ethical considerations into AI algorithms and workflows; and investigating the long-term societal impacts of AI-generated metadata on cultural memory and historical understanding. Addressing these ethical considerations is crucial for ensuring that AI-driven metadata tagging systems not only enhance the accessibility and usability of digital archives but also do so in a way that is responsible, culturally sensitive, and ethically sound.

Interdisciplinary collaboration has emerged as a crucial factor in developing and refining AIdriven metadata systems. The complex nature of cultural heritage materials and the diverse requirements of different institutions necessitate a collaborative approach that brings together expertise from various fields. Alliata, Hou, and Kenderdine (2024) presented a computational framework to enhance access to embodied knowledge archives through posture recognition and movement computing, emphasizing the importance of collaboration among archivists, computists, artists, and knowledge holders. This interdisciplinary approach ensures that AI systems are informed by a diverse range of expertise, leading to more robust and contextually accurate metadata. Their study highlighted several key benefits of interdisciplinary collaboration, including enhanced contextual understanding, improved accuracy and relevance, innovative applications, and proactive addressing of ethical considerations. However, the practical challenges and best practices for fostering such collaboration remain understudied. Future research could focus on developing frameworks for effective interdisciplinary collaboration in AI-driven metadata projects, including strategies for communication, project management, and knowledge sharing across disciplines. Alongside the benefits of interdisciplinary collaboration, the ethical implications of using AI for metadata generation in cultural heritage

contexts, particularly regarding sensitive or culturally significant materials, require further exploration. As the field of AI-driven metadata tagging continues to evolve, new applications and research directions are emerging. These developments not only address current limitations but also open up new possibilities for enhancing access to and understanding of cultural heritage materials

Future Directions and Emerging Applications

Recent research has also explored novel applications of AI in enhancing access to specialized collections. Thomas and Testini (2024) investigated the use of AI to identify and analyse image captions in historical book illustrations, demonstrating the effectiveness of AI techniques such as layout parsing and optical character recognition (OCR) in improving the searchability and interpretative value of digitized archives. Key findings from their research include AI's ability to accurately identify and extract captions from a wide range of historical book layouts and typographic styles, the potential for AI to link caption text with corresponding images, enhancing the discoverability of visual materials, and challenges in dealing with variability in caption styles and the presence of foreign languages. This research highlights the potential of AI in unlocking new ways of accessing and understanding historical materials, but also reveals the ongoing challenges in dealing with the complexity and variability of cultural heritage materials.

The application of AI in sensitive archival contexts presents unique challenges and opportunities, particularly when dealing with materials that have complex historical, cultural, or ethical implications. Dentler et al. (2024) explored the use of machine learning to improve access to and navigation of sensitive colonial visual archives, emphasizing the need for critical and transparent multimodal AI to enhance accessibility while addressing ethical concerns. The researchers developed experimental computer vision tools to analyse a large database of sensitive visual materials from colonial conflicts, demonstrating how AI can enrich archival metadata and improve search results. Key aspects of their approach included the development of culturally aware AI models, the implementation of ethical safeguards, collaboration with stakeholder communities, and transparency in AI decision-making. However, their research also highlighted the ethical challenges in balancing open access with the sensitive nature of historical materials, emphasizing the need for human oversight and contextualization. The study underscored several important considerations for future research in this area, including the development of AI systems that can recognize and appropriately handle sensitive content across diverse cultural contexts; the creation of ethical guidelines for the use of AI in processing and providing access to sensitive archival materials; the exploration of user interface designs that can effectively communicate the complex historical and ethical contexts of sensitive materials; and the investigation of the long-term impacts of AI-enhanced access to sensitive archives on historical understanding and reconciliation processes.

These ethical challenges are compounded by the socio-economic structures underpinning AI technologies, particularly the influence of platform capitalism on metadata design and deployment. One such critical issue is platform capitalism, which shapes the design and deployment of metadata systems. Platform capitalism,

as defined by Nick Srnicek in *Platform Capitalism* (2017), refers to the economic model where companies dominate markets through the ownership and exploitation of digital platforms, data, and proprietary technologies. These platforms, including those run by Google, Amazon, and Microsoft, consolidate market dominance by commodifying data through proprietary AI systems. In this context, metadata systems are not just tools for organization but are also transformed into products for profit.

These ecosystems often result in vendor lock-in, where institutions become reliant on proprietary standards, limiting flexibility and raising concerns about data migration and interoperability. Furthermore, the homogenization of metadata—optimized for efficiency rather than cultural nuance—threatens the diversity and contextual richness essential for cultural heritage collections.

For cultural institutions, these trends present far-reaching implications. Profit-driven AI tools often prioritize monetization over equitable access and cultural sensitivity. To counter these challenges, institutions should advocate for open-source AI systems that promote interoperability, transparency, and community-driven frameworks. Collaborative efforts across sectors can mitigate dependencies on proprietary systems, ensuring metadata systems preserve cultural specificity and remain accessible to diverse stakeholders. These socio-economic challenges underscore the urgency of adopting measures like paradata and collaborative frameworks to safeguard cultural heritage metadata from commercialization and homogenization.

The introduction of paradata to the archival field represents a significant development in the ongoing evolution of archival practices, particularly in relation to transparency, accountability, and record integrity in AI-driven processes. Davet, Hamidzadeh, and Franks (2023) proposed a framework for collecting paradata for AI systems in archives, suggesting three key elements: training, testing, and validation data; performance information; and comprehensive versioning information. The researchers argue that paradata is crucial for enhancing transparency, providing a basis for evaluating and improving AI systems over time, allowing for the reproduction and verification of AI outputs, and providing a valuable historical context as AI systems evolve.

However, the concept of paradata in AI-driven archival processes is still in its early stages, and several significant gaps in current research and practice remain to be addressed. These include the need for standardized methodologies for collecting, organizing, and presenting paradata across different institutional contexts; research on how to effectively integrate paradata concepts with existing archival standards and metadata schemas; comprehensive ethical guidelines for the collection and use of paradata; studies addressing the practical challenges of implementing paradata collection in real-world archival settings; research to inform the development of effective presentation and interaction models for paradata; and strategies for ensuring the long-term preservation and accessibility of paradata itself. Addressing these gaps in paradata research and practice is crucial for enhancing the transparency, accountability, and overall trustworthiness of AI-driven archival processes.

More comprehensive comparative performance studies are needed to guide institutional decision-making on AI adoption for metadata generation. These studies should evaluate the efficacy of AI systems across diverse collections and asset types, providing robust, contextually informed guidance for cultural heritage institutions.

Research into effective models for combining human expertise with AI capabilities in metadata creation workflows is a critical priority. This includes investigating strategies for leveraging the complementary strengths of human curators and intelligent automation to produce high quality, ethically-sound metadata.

Further investigation into methods for identifying and mitigating biases in AI-generated metadata is essential. Given the potential for AI systems to perpetuate or amplify societal biases, there is a critical need to develop robust bias detection and mitigation frameworks tailored to cultural heritage contexts.

Addressing the challenges of scaling AI solutions to handle large and diverse collections is another key area for future research. As institutions grapple with ever-growing digital holdings, studies are needed to explore approaches for deploying and maintaining AI-driven metadata systems at scale.

The development of comprehensive ethical guidelines and frameworks specifically tailored to AI use in cultural heritage contexts is a pressing concern. These should address issues of privacy, consent, intellectual property, and the stewardship of sensitive or culturally significant materials.

Understanding how AI-enhanced metadata impacts end-user experiences in discovering and engaging with cultural heritage materials is crucial. Research is needed to ensure that the deployment of AI-driven access and presentation mechanisms leads to equitable and meaningful interactions with these valuable resources.

Examining the long-term effects of AI-driven metadata practices on cultural memory, historical interpretation, and archival integrity is a vital area of study. As AI becomes more pervasive in the cultural heritage sector, it is essential to investigate the potential implications for the preservation and interpretation of the historical record.

Effective models for fostering collaboration between technologists, cultural heritage professionals, ethicists, and other stakeholders in AI projects must be developed. Interdisciplinary cooperation is vital for ensuring that AI systems are developed and deployed in alignment with the needs and values of the communities they serve.

Exploration of innovative applications of AI in cultural heritage contexts beyond basic metadata tagging, such as the use of computer vision and natural language processing for enhanced discovery, analysis, and storytelling, represents another promising avenue for future research.

Finally, further research and development of practical approaches to implementing paradata collection and use in AI-driven archival processes is needed. Paradata, or data about the data creation process, can provide crucial context for understanding and validating AI-generated metadata.

While novel applications of AI in specialized collections have been explored, there is a need for more research on how these applications can be scaled and adapted for broader use in diverse institutional settings. The potential of AI to enhance interconnectivity between collections through metadata generation is an area that requires further investigation, particularly in terms of cross-institutional and cross-cultural applications.

Methodology

This study employs a comprehensive multi-system approach to evaluate the efficacy of automated tagging for digital collections in organisations dealing with a large number of images. Drawing inspiration from the seminal work of Villaespesa and Crider (2020) on computer vision tagging of the Metropolitan Museum of Art's collection, this research extends their methodology to address the unique challenges faced by diverse cultural institutions. The study integrates advanced quantitative metrics with in-depth qualitative assessments, utilizing three cloud-based computer vision services: Google Cloud Vision, Microsoft Azure Computer Vision, and Amazon Rekognition.

Approach and Rationale

The research design incorporates a robust mixed-methods approach, synthesizing quantitative metrics with qualitative analysis to provide a nuanced and comprehensive evaluation of AI generated tags. This methodology is specifically tailored to address the multifaceted complexities inherent in applying digital tools to heritage collections. It acknowledges and explores the challenges posed by legacy systems, inconsistent data quality, and varying digitization standards across cultural institutions.

The rationale for this approach is grounded in the understanding that while quantitative metrics provide essential insights into the accuracy and efficiency of AI tagging systems, they alone cannot capture the nuanced cultural, historical, and contextual aspects of heritage collections. By combining these metrics with qualitative analysis, the study aims to provide a holistic assessment that considers both technical performance and cultural relevance.

Data Collection and Preparation

Initially, the study intended to utilize rich and diverse image datasets from two prestigious institutions: the Special Collection Unit at the Institute of Ismaili Studies, London, and the British Cultural Archives, London. These collections were chosen for their historical significance, cultural diversity, and potential to challenge AI systems with complex, nuanced imagery. However, during the preliminary stages of the research, both institutions took longer than expected to respond, and one withdrew their participation, citing privacy concerns and ethical considerations related to the use of cloud-based API tools for image analysis.

In response to these challenges, I had to develop a publicly available, carefully curated image dataset scraped from the web, designed to replicate the diversity and complexity of real-world organizations. This dataset encompasses a wide range of categories to ensure comprehensive evaluation:

People: This category includes a diverse array of portraits representing various ethnicities, ages, and cultural backgrounds. It also features group photos in different settings, from formal gatherings to casual social interactions. Special attention was given to including images of children engaged in play and educational activities, as well as elderly individuals participating in various cultural practices.

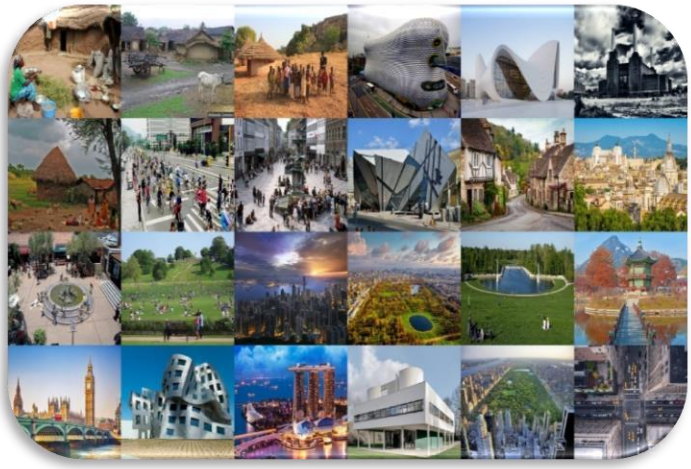
Objects: This section comprises a wide spectrum of items, from everyday household objects to rare historical artifacts. It includes technological devices spanning different eras, various types of vehicles (both modern and historical), and objects of cultural significance from diverse global traditions.

Nature: The nature category encompasses a broad range of landscapes from various geographical regions, including forests, deserts, mountains, and coastal areas. It also features closeup images of plants and flowers, emphasizing botanical diversity. Wildlife images showcase animals

in their natural habitats, while another subset focuses on different weather conditions and natural phenomena.

Cultural and Historical Contexts: This rich category includes images of cultural festivals from around the world, traditional ceremonies and rituals, and photographs documenting significant historical events. It also features famous monuments, archaeological sites, and artifacts from different historical periods, spanning ancient civilizations to modern times.

Urban and Rural Settings: The urban subset includes cityscapes from major metropolitan areas worldwide, showcasing various architectural styles, urban planning approaches, and public spaces. The rural component features landscapes from farming communities, small villages, and remote settlements, highlighting the diversity of human habitation and land use patterns.



Text and Documents: This category is crucial for testing OCR capabilities and includes a variety of textual materials. It features handwritten letters and notes from different time periods, pages from printed books spanning various genres and eras, and documents in multiple languages and scripts. Additionally, it includes signage and posters found in public spaces, as well as scanned pages from historical newspapers and magazines.

Art and Creative Works: This diverse category contains high-quality images of paintings representing various art movements and styles, from classical to contemporary. It also includes sculptures, installations, digital art, and performance art. The performance subset features images of theatre productions, dance performances, and musical concerts, capturing both traditional and avant-garde expressions.

Daily Life Scenes: This category aims to capture the mundane yet culturally significant aspects of everyday life. It includes images of marketplaces and shopping centres from different cultures, educational environments ranging from

traditional classrooms to innovative learning spaces, diverse work environments reflecting various industries, and domestic scenes showcasing home activities across different cultural contexts.

This carefully constructed dataset aims to challenge the AI systems with a level of diversity and complexity comparable to that found in actual heritage collections. It allows for a comprehensive evaluation of the capabilities of the AI services across a broad spectrum of visual content, mirroring the heterogeneity of real-world cultural heritage materials.

Evaluation of AI Services

The evaluation process involves several key steps, each designed to rigorously assess the performance and relevance of the AI-generated tags:

Tag Generation: Each image in the curated dataset is systematically processed through Google Cloud Vision, Amazon Rekognition, and Microsoft Azure Computer Vision. These AI services were chosen for their prominence in the field of computer vision and their diverse approaches to image analysis. Google Cloud Vision is renowned for its robust image recognition capabilities, including advanced object detection, face detection, and landmark recognition. Meanwhile, Amazon Rekognition offers comprehensive image and video analysis tools, with strengths in object detection, scene analysis, and text recognition. Finally, Microsoft Azure Computer Vision provides a suite of features including optical character recognition (OCR), image categorization, and spatial analysis.

Qualitative Analysis

This emphasis on qualitative evaluation stems from the unique requirements of cultural heritage collections, where contextual and cultural dimensions are critical for meaningful metadata generation This study places significant emphasis on qualitative analysis to assess the cultural and contextual relevance of AI-generated metadata. While quantitative metrics such as accuracy, precision, and recall provide essential insights into technical performance, they cannot fully capture the nuanced cultural, historical, and artistic dimensions of heritage collections. Qualitative analysis addresses this gap by offering a deeper, interpretive evaluation of metadata in relation to its contextual richness and alignment with institutional priorities.

Key elements of the qualitative methodology include:

Semantic Depth: AI-generated metadata is reviewed for its ability to go beyond surface-level descriptions and provide meaningful insights into the subject matter. For example, a painting labelled merely as "art" lacks depth, while a label such as "Expressionist painting by Wassily Kandinsky" reflects a more nuanced understanding of style, artist, and historical context.

Cultural Sensitivity: Metadata is analysed for appropriateness and respectfulness, particularly when dealing with culturally significant or sensitive materials. This includes avoiding labels that may misrepresent or oversimplify cultural artifacts. For instance, objects from Indigenous cultures are assessed to ensure that the

AI-generated descriptions align with community-endorsed terminologies and avoid perpetuating colonial perspectives.

Contextual Relevance: The study examines how well the metadata reflects the cultural, historical, and social contexts of the images. Tags that align with the specific narratives or interpretive goals of cultural institutions are considered more effective. For example, the tag "ceremonial mask" is more contextually relevant when paired with a description of its cultural origin and ceremonial use.

Alignment with Controlled Vocabularies: The AI-generated metadata is evaluated against authoritative controlled vocabularies such as the Getty Art & Architecture Thesaurus (AAT), UK Archival Thesaurus (UKAT), and Library of Congress Subject Headings (LCSH). This ensures consistency, accuracy, and interoperability in metadata creation. Controlled vocabularies provide standardized terms, enabling institutions to maintain uniformity across collections and facilitate enhanced information retrieval.

Through this qualitative framework, the study aims to bridge the gap between technological advancements in AI and the unique needs of cultural institutions. By prioritizing semantic depth, cultural sensitivity, and contextual relevance, the methodology ensures that the AI-generated metadata enhances the accessibility, interpretability, and cultural value of heritage collections.

Controlled Vocabularies

The study employs meticulously crafted controlled vocabularies as a standardized ground truth, facilitating both the calculation of quantitative metrics and the qualitative assessment of tag relevance. These vocabularies draw from several authoritative sources:

UK Archival Thesaurus (UKAT): This comprehensive thesaurus provides a standardized vocabulary for describing archival materials, ensuring consistency in terminology across different types of heritage collections.

Getty Art & Architecture Thesaurus (AAT): This structured vocabulary is particularly valuable for describing works of art, architecture, and material culture. It provides standardized terms for styles, materials, and object types.

Library of Congress Subject Headings (LCSH): This extensive list of subject headings offers a standardized vocabulary for describing the topical content of heritage materials, spanning a wide range of disciplines and subject areas.

ICOM International Committee for Documentation (CIDOC) Conceptual Reference Model (CRM): While not directly used for tagging, this ontology informs the structuring of relationships between concepts in the controlled vocabulary, ensuring a coherent semantic framework.

The implementation of these standardized controlled vocabularies promotes consistency, accuracy, and relevance in tagging digital assets across various domains (Baca, 2016; Zeng & Qin, 2016). For example:

In creating a set of controlled vocabularies for a photograph of the following Jackson Pollock painting, terms like ‘Abstract,’ ‘Modern Art,’ ‘Painting,’ ‘Jackson Pollock,’ ‘Art,’ ‘Expressionism,’ ‘Abstract Expressionism,’ ‘Action painting,’ and ‘Mid20th century American art’ are used to accurately describe the artistic movement, technique, and historical context.

For the following image of ballet performances, terms such as ‘Ballet,’ ‘Dance Group,’ ‘Performance,’ ‘Art,’ ‘Classical Dance,’ ‘Stage,’ ‘Ballerinas,’ ‘Costumes,’ ‘Black and White,’ ‘Choreography,’ are employed.

For an image of the Amazon rainforest, terms like ‘Amazon,’ ‘rainforest,’ ‘jungle,’ ‘biodiversity,’ ‘tropical,’ ‘greenery,’ and specific geological features are utilized to depict environmental and scenic elements.

These examples underscore the essential role of controlled vocabularies in ensuring the accuracy, consistency, and contextual richness of metadata across varied subject areas, thereby facilitating enhanced information retrieval and interoperability within digital asset management systems.

Data Analysis

The data analysis phase employs a sophisticated suite of software tools and statistical methods to process and interpret the results:

Python is utilized for its robust data processing and statistical analysis capabilities. Python's NumPy library is used for numerical operations, while scikit-learn's sklearn.metrics module provides implementations of various performance metrics to evaluate the machine learning algorithm.

Custom scripts are developed to automate the process of comparing AI-generated tags with the ground truth, calculating performance metrics, and aggregating results across different categories and AI services.

Data visualization is a crucial component of the analysis, helping to communicate complex findings effectively. Libraries such as matplotlib and seaborn in Python are used to create a variety of visualizations, including

o Bar Plot showing the performance of each AI service across different image categories

o Box plots illustrating the distribution of accuracy, precision, and recall scores o Network graphs visualizing the relationships between AI-generated tags and controlled vocabulary terms

These visualizations not only aid in the interpretation of results but also serve as powerful tools for communicating findings to stakeholders in the cultural heritage sector.

Limitations and Considerations

This study acknowledges several limitations and considerations that must be accounted for when interpreting its results and understanding its broader implications. These limitations are integral to the research findings

and inform the recommendations for future work and practical applications in the field of AI-driven cultural heritage.

1. Data Quality and Representativeness

While the curated dataset aims to replicate the diversity of real-world memory institutions, it cannot fully capture the immense variety and uniqueness of items found in actual collections. As a result, the performance of AI services evaluated on this dataset may not directly translate to all real-world scenarios. Future studies should explore broader datasets and assess how variations in data quality and representativeness impact AI performance.

2. Evolving AI Technologies

The field of AI, particularly in computer vision and natural language processing, is rapidly advancing. The AI services evaluated in this study may improve or change over time, affecting the long-term applicability of the findings. Continuous re-evaluation of AI technologies and their implications will be essential to maintaining the relevance of this research.

3. Cultural and Linguistic Bias

Despite efforts to create a diverse dataset, biases inherent in AI training data or controlled vocabularies may persist. These biases can influence the accuracy and relevance of metadata for items from underrepresented cultures or languages. Such limitations risk perpetuating existing inequalities in cultural representation. Addressing this issue requires the ongoing diversification of training datasets and the involvement of culturally informed stakeholders to ensure equitable metadata practices.

4. Contextual Understanding

AI services, while sophisticated, often lack the nuanced cultural or historical understanding that human experts bring to metadata creation. This limitation is particularly significant for complex or ambiguous imagery, where contextual interpretation is crucial. Collaborative workflows combining human expertise with AI capabilities can help mitigate this limitation.

5. Ethical Considerations

The ethical implications of deploying AI systems in cultural heritage contexts are substantial. One critical challenge lies in the biases within training datasets, which can disproportionately impact underrepresented communities and cultures. Dentler et al. (2024) and Fornaro and Chiquet (2021) emphasize that such biases not only misrepresent cultural narratives but also undermine the inclusivity of AI-generated metadata.

Privacy concerns also arise when AI systems handle sensitive archival materials, particularly those tied to indigenous or marginalized communities. Open access to such data must be balanced with respect for privacy and cultural values, supported by transparent governance protocols and human oversight.

The socio-economic structures underpinning AI technologies exacerbate these ethical challenges. Practices such as vendor lock-in and metadata homogenization, driven by platform capitalism, limit institutional flexibility and threaten cultural diversity. Furthermore, Shoshana Zuboff's concept of surveillance capitalism (*The Age of Surveillance Capitalism*, 2019) highlights how AI systems prioritize data extraction and commodification, raising concerns about the exploitation of cultural heritage for profit.

To address these challenges, this study adopts a robust ethical framework focusing on:

Diversification of Training Data: Creating datasets that reflect diverse cultural, linguistic, and social contexts.

Human-AI Collaboration: Involving human oversight to ensure AI outputs are contextually accurate and culturally sensitive.

Stakeholder Engagement: Actively involving community representatives to align AI practices with cultural values and ethical standards.

This approach promotes equitable and responsible AI use in cultural heritage, preserving cultural integrity while enhancing access.

6. Resource Intensiveness

Preparing large image collections for AI analysis, including digitization and metadata standardization, is resource intensive. This methodology may not be immediately applicable to institutions with limited resources or largely undigitized collections. Scalable and cost-effective strategies are needed to support smaller organizations.

7. Interoperability Challenges

The varied standards and practices across cultural institutions create barriers to implementing universally applicable AI systems. Adapting the study's findings to different institutional contexts may require further research into flexible and interoperable metadata frameworks.

These limitations are not merely caveats but essential aspects of understanding the study's scope. By acknowledging these challenges, the research lays a foundation for addressing critical gaps and advancing the responsible use of AI in cultural heritage. This nuanced understanding will inform future studies and practical implementations, ensuring that AI technologies serve as tools for inclusion, equity, and cultural preservation.

Conclusion

This methodology aims to provide a nuanced and multifaceted evaluation of AI-generated tags for digital image collections, combining rigorous quantitative analysis with in-depth qualitative assessment to bridge the gap between technological capabilities and the specific needs of institutions managing substantial image holdings. The research not only examines the technical performance of AI tagging services, but also explores

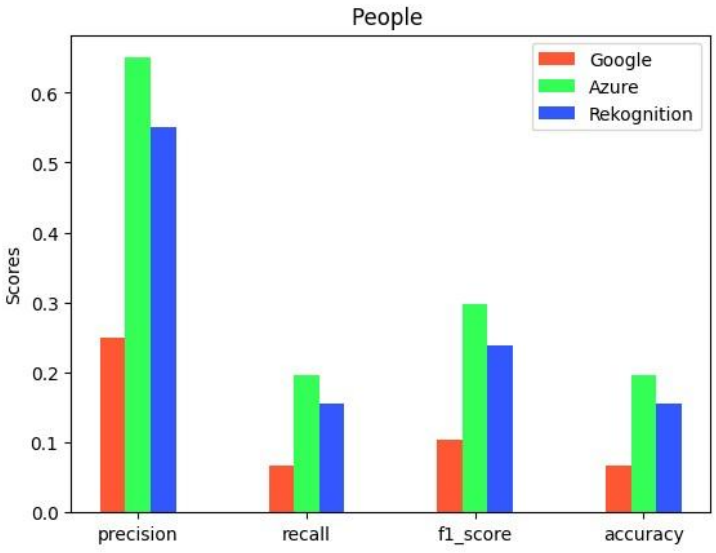
their ability to generate culturally relevant and contextually appropriate metadata, addressing both the practical requirements of digital asset management and the broader mission of organizations dealing with image-based collections. By acknowledging and examining the challenges faced in applying digital tools to large-scale image collections, including issues of data quality, ethical considerations, and resource limitations, the study contributes valuable insights to the ongoing dialogue about digital transformation in institutions with extensive image holdings. The findings have the potential to inform best practices for integrating AI technologies, guide the development of more culturally sensitive AI systems, refine metadata standards, and provide a methodological framework for future studies on AI applications in large-scale image contexts. Ultimately, this approach seeks to help institutions leverage AI technologies while preserving the richness, diversity, and nuance of human culture and history captured in their image collections, addressing critical ethical issues around data privacy, representation, and model bias.

Findings

The performance of the AI tagging services was evaluated across six image categories: People, Art and Creative Craft, Scenes of daily life, Nature, Objects, Text and Documents, and Urban and Rural Settings. The findings reveal significant variations in the performance of Google Cloud Vision, Microsoft Azure Computer Vision, and Amazon Rekognition.

People

In the “People” category, Google Cloud Vision achieved a precision score of 0.25, recall score of 0.0658, F1 score of 0.1029, and accuracy of 0.0658. Microsoft Azure Computer Vision outperformed Google Cloud Vision with a precision score of 0.65 indicating that the service identified about 19.67% of the relevant tags. It had a recall score of 0.1967, F1 score of 0.297, and accuracy of 0.1967. Amazon Rekognition had a precision score of 0.55, recall score of 0.155, F1 score of 0.239, and accuracy of 0.155.



The qualitative analysis reveals that Google Cloud Vision struggled to generate relevant and comprehensive tags, often lacking specificity and contextual accuracy. Microsoft Azure Computer Vision performed significantly better, generating more contextually relevant tags that aligned better with the controlled vocabulary. Amazon Rekognition showed reasonable performance, with generally relevant tags but slightly lower recall compared to Microsoft Azure.

In the dataset for the category of “People”, Azure demonstrated the most consistent performance, excelling in family photos and children playing sports. Rekognition showed strengths in recognizing elderly individuals and activities. Google Vision, while less consistent overall, performed exceptionally well in specific scenarios like children playing football. The varied performance across services highlights the importance of choosing the right tool for specific use cases, or potentially combining services for more comprehensive people recognition tasks.

The performance in the “People” category reflects the variability in cloud APIs’ abilities to process complex, people-centric imagery. While the services performed relatively well for generalized contexts such as “family photos” or “children playing sports,” the results reveal gaps in their ability to generate nuanced or contextually rich tags. These gaps may stem from biases in training datasets, where common, commercially significant scenarios are overrepresented at the expense of more diverse or culturally specific ones.

Additionally, the tagging inconsistencies suggest that cloud APIs may prioritize features that are visually prominent or frequently occurring in their datasets, potentially neglecting less typical or culturally specific representations of people. These findings highlight the need for broader, more inclusive training datasets that better represent the diversity of global human experiences. Improving dataset diversity and enhancing the contextual understanding of cloud APIs are critical steps for ensuring that metadata generation aligns with the complex needs of cultural heritage institutions.

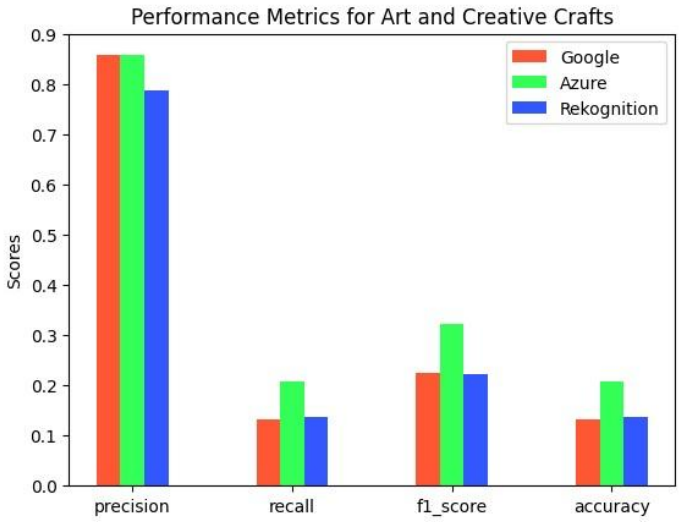
Art and Creative Craft

In the “Art and Creative Craft” category, Google Cloud Vision achieved a precision score of 0.86, recall score of 0.13, F1 score of 0.23, and accuracy of 0.13. Microsoft Azure Computer

Vision matched Google's precision with a score of 0.86 but outperformed both competitors in other metrics with a recall score of 0.21, F1 score of 0.32, and accuracy of 0.21. Amazon Rekognition had slightly lower performance with a precision score of 0.79, recall score of 0.14, F1 score of 0.22, and accuracy of 0.14.

The qualitative analysis reveals that all three services demonstrate high precision in tag generation for art and creative crafts images, indicating strong accuracy in the tags they do produce. However, the consistently low recall scores across all services suggest they struggle to capture the full range of relevant tags for these images.

Microsoft Azure Computer Vision showed the best overall performance, particularly in recall and F1 score, indicating it was able to identify about 21% of the relevant tags while maintaining high precision. Google Cloud Vision, despite its high precision, struggled with recall, suggesting it may generate fewer but highly accurate tags. Amazon Rekognition performed similarly to Google in recall and F1 score, but with slightly lower precision.



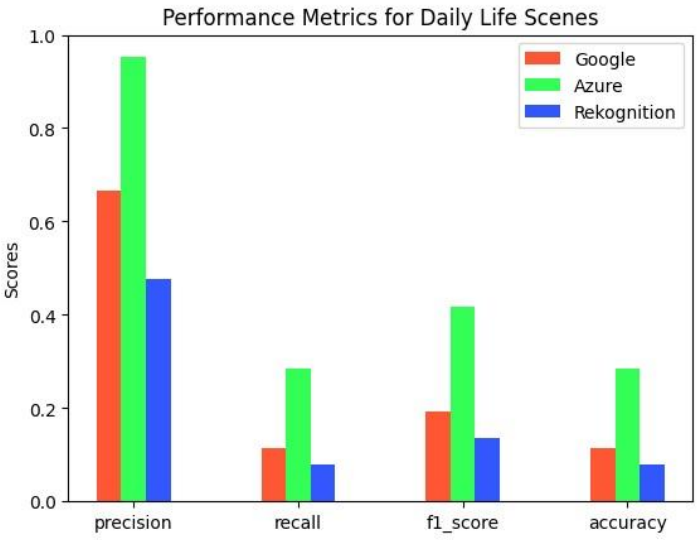
The low F1 scores across all services highlight a significant imbalance between precision and recall, indicating room for improvement in comprehensive tag generation for art and creative crafts images. This suggests that while the services are highly accurate in the tags they generate, they may be missing many relevant descriptors for these types of images.

The findings in the “Art and Creative Craft” category highlight a significant imbalance between precision and recall, with all services favouring accuracy over comprehensiveness. This imbalance may be attributed to cloud APIs being trained primarily on widely documented and commercially relevant art forms, such as Western classical art, at the expense of indigenous or underrepresented artistic traditions. While high precision scores demonstrate the APIs’ ability to recognize well-documented styles and objects, the low recall scores reflect a systemic gap in dataset diversity.

These results underscore the need for integrating marginalized art traditions into training datasets. Without addressing this gap, cloud APIs risk perpetuating the erasure of diverse artistic narratives. Enhanced collaboration between API providers and cultural heritage institutions could help ensure more equitable representation of global art in automated metadata systems.

Scenes of Daily Life

In this category, Google Cloud Vision achieved a precision score of 0.67, recall score of 0.12, F1 score of 0.19, and accuracy of 0.12. Microsoft Azure Computer Vision significantly outperformed the others with a precision score of 0.95, recall score of 0.28, F1 score of 0.42, and accuracy of 0.28. Amazon Rekognition had the lowest overall performance with a precision score of 0.48, recall score of 0.08, F1 score of 0.14, and accuracy of 0.08.



The qualitative analysis reveals that the three services demonstrate varying levels of precision in tag generation for daily life scenes, with Microsoft Azure showing exceptionally high precision. However, the consistently low recall scores across all services suggest they struggle to capture the full range of relevant tags for these images.

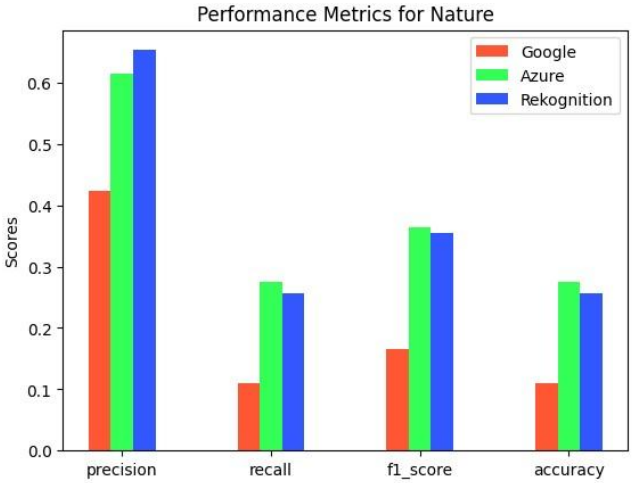
Microsoft Azure Computer Vision showed the best overall performance, particularly in precision and recall, indicating it was able to identify about 28% of the relevant tags while maintaining very high precision. Google Cloud Vision, despite its relatively high precision, struggled with recall, suggesting it may generate fewer but fairly accurate tags. Amazon Rekognition performed the poorest, with moderate precision but very low recall.

The low F1 scores for Google Cloud Vision and Amazon Rekognition highlight a significant imbalance between precision and recall, indicating room for improvement in comprehensive tag generation for images of daily life. This suggests that while these services can be accurate in the tags they generate, they may be missing many relevant descriptors for these types of images. Microsoft Azure, with its higher F1 score, demonstrates a better balance but still has considerable room for improvement in capturing the full range of relevant tags.

The tagging results for “Scenes of Daily Life” reveal the challenges cloud APIs face in balancing technical accuracy with cultural richness. Tags such as “market” or “crowd” fail to contextualize the social or cultural specificity visible in these scenes. This tendency toward generic labeling likely reflects the priorities of cloud APIs to generate universally recognizable terms rather than culturally nuanced descriptions.

These findings highlight the limitations of datasets and algorithms that prioritize generalizability over specificity. Incorporating localized data into training processes could help APIs better recognize and represent the diversity of human experiences in daily life imagery. Moreover, the integration of context-sensitive tagging frameworks could enable richer and more meaningful metadata for cultural heritage applications.

Images of Nature



In this category, Google Cloud Vision achieved a precision score of 0.42, recall score of 0.11, F1 score of 0.17, and accuracy of 0.11. Microsoft Azure Computer Vision performed better with a precision score of 0.62, recall score of 0.28, F1 score of 0.36, and accuracy of 0.28. Amazon Rekognition had the best overall performance with a precision score of 0.65, recall score of 0.26, F1 score of 0.36, and accuracy of 0.26.

The qualitative analysis reveals that the three services demonstrate varying levels of precision in tag generation for images of Nature, with Amazon Rekognition and Microsoft Azure showing higher precision than Google Cloud Vision. However, the relatively low recall scores across all services suggest they struggle to capture the full range of relevant tags for these images.

Amazon Rekognition showed the best overall performance, particularly in precision, while maintaining a recall score comparable to Microsoft Azure. This indicates it was able to identify about 26% of the relevant tags while maintaining the highest precision. Microsoft Azure performed similarly well, with slightly lower precision but a marginally higher recall. Google Cloud Vision struggled in both precision and recall, suggesting it may generate fewer and less accurate tags for nature images.

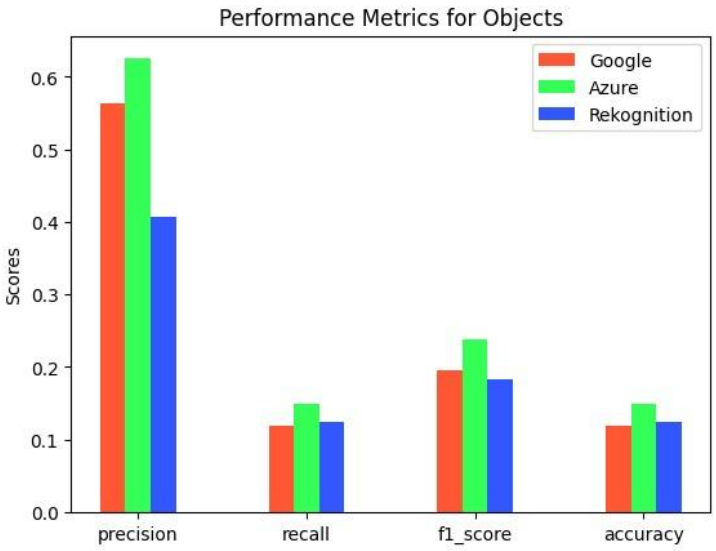
The low F1 scores across all services, particularly for Google Cloud Vision, highlight a significant imbalance between precision and recall, indicating room for improvement in comprehensive tag generation for images of nature. This suggests that while the services can generate relevant tags, they may be missing many important descriptors for these types of images. Microsoft Azure and Amazon Rekognition, with their higher F1 scores, demonstrate a better balance but still have considerable room for improvement in capturing the full range of relevant tags for nature scenes.

The findings in the “Nature” category reveal that while cloud APIs are adept at recognizing common natural elements, they struggle with ecologically unique or culturally significant natural imagery. For instance, regional flora or fauna often remain unrecognized or are tagged with overly generic terms such as “tree” or “animal.” This limitation reflects a bias in the training data, which may prioritize frequently occurring or commercially relevant natural features.

Addressing this gap requires the expansion of training datasets to include diverse ecological elements. Doing so would not only improve the accuracy of nature-related tags but also enhance the representation of culturally significant natural imagery in heritage collections.

Objects

In the “Objects” category, Google Cloud Vision achieved a precision score of 0.56, recall score of 0.12, F1 score of 0.20, and accuracy of 0.12. Microsoft Azure Computer Vision outperformed the others with a precision score of 0.63, recall score of 0.15, F1 score of 0.24, and accuracy of 0.15. Amazon Rekognition had the lowest overall performance with a precision score of 0.41, recall score of 0.13, F1 score of 0.18, and accuracy of 0.13.



The qualitative analysis reveals that all three services demonstrate moderate levels of precision in tag generation for object images, with Microsoft Azure showing the highest precision. However, the consistently low recall scores across all services suggest they struggle to capture the full range of relevant tags for these images.

Microsoft Azure Computer Vision showed the best overall performance, particularly in precision and recall, indicating it was able to identify about 15% of the relevant tags while maintaining the highest precision. Google Cloud Vision performed slightly better than Amazon Rekognition in precision and F1 score, but all three services struggled with recall, suggesting they may generate a limited number of tags for images of Objects.

The low F1 scores across all services highlight a significant imbalance between precision and recall, indicating substantial room for improvement in comprehensive tag generation for Object images. This suggests that while the services can generate some relevant tags, they are missing many important descriptors for these types of images. Even Microsoft Azure, with its slightly higher F1 score, demonstrates considerable room for improvement in capturing the full range of relevant tags for Object images.

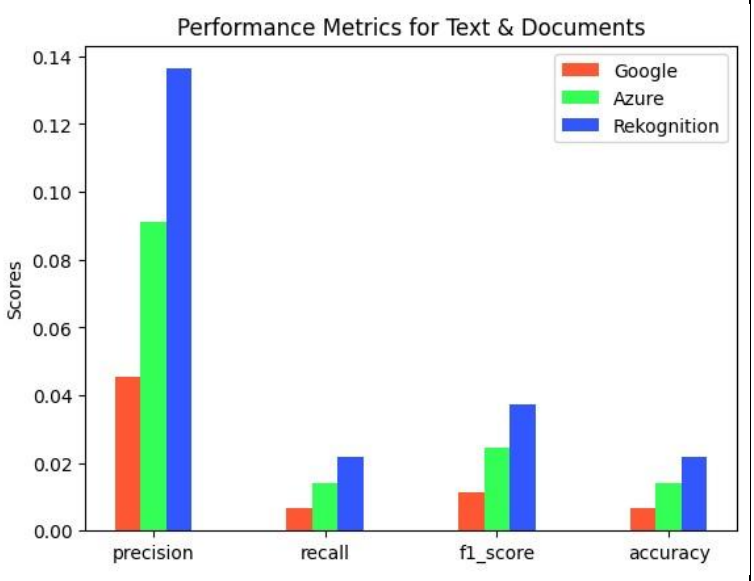
The tagging performance for “Objects” reflects a consistent trade-off between precision and recall across all cloud APIs. While precision scores suggest that APIs are effective at recognizing commonly occurring objects, low recall scores indicate a failure to capture a wider range of culturally significant items. This limitation highlights biases in the algorithms, which likely prioritize features common in commercially valuable datasets.

Improving object recognition for cultural heritage applications will require targeted dataset diversification and algorithmic adjustments to include artifacts and objects that are underrepresented in conventional training

processes. This approach would help ensure that metadata accurately reflects the unique cultural and historical value of objects in heritage collections.

Text and Documents

In this category, Google Cloud Vision achieved a precision score of 0.05, recall score of 0.01, F1 score of 0.01, and accuracy of 0.01. Microsoft Azure Computer Vision performed slightly better with a precision score of 0.09, recall score of 0.01, F1 score of 0.02, and accuracy of 0.01. Amazon Rekognition had the highest performance among the three, albeit still low, with a precision score of 0.14, recall score of 0.02, F1 score of 0.04, and accuracy of 0.02.



The qualitative analysis reveals that all three services demonstrate extremely low performance in tag generation for text and document images. This suggests a significant challenge in accurately identifying and describing content in this category.

Amazon Rekognition had the best overall performance, though still poor, with slightly higher precision and recall scores. This indicates it was able to identify about 2.2% of the relevant tags, which is marginally better than the other services but still far from satisfactory. Microsoft Azure and Google Cloud Vision both struggled significantly, with very low precision and recall scores, suggesting they generate few and largely irrelevant tags for images of text and document.

The extremely low F1 scores across all services highlight a critical imbalance between precision and recall, indicating a substantial need for improvement in tag generation images of text and document This suggests that the current capabilities of these AI services are severely limited when it comes to analysing and describing images containing text or documents. Even Amazon Rekognition, with its slightly higher scores, demonstrates that there is a considerable gap between the services' current performance and the level required for effective tagging of text and document images in digital asset management systems.

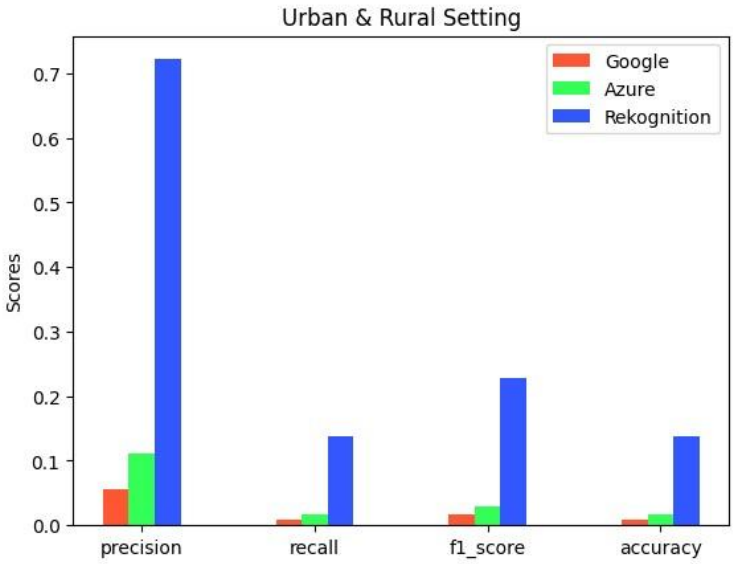
The poor performance across all cloud APIs in the “Text and Documents” category highlights significant limitations in OCR (Optical Character Recognition) capabilities when applied to heritage-focused contexts. These systems appear optimized for modern, born digital documents with standardized layouts, leaving historical, handwritten, or linguistically diverse texts largely unrecognized. For instance, manuscripts in non-

Latin scripts or with unique calligraphic styles often fail to generate meaningful tags, revealing a critical bias in the datasets and algorithms underlying these APIs.

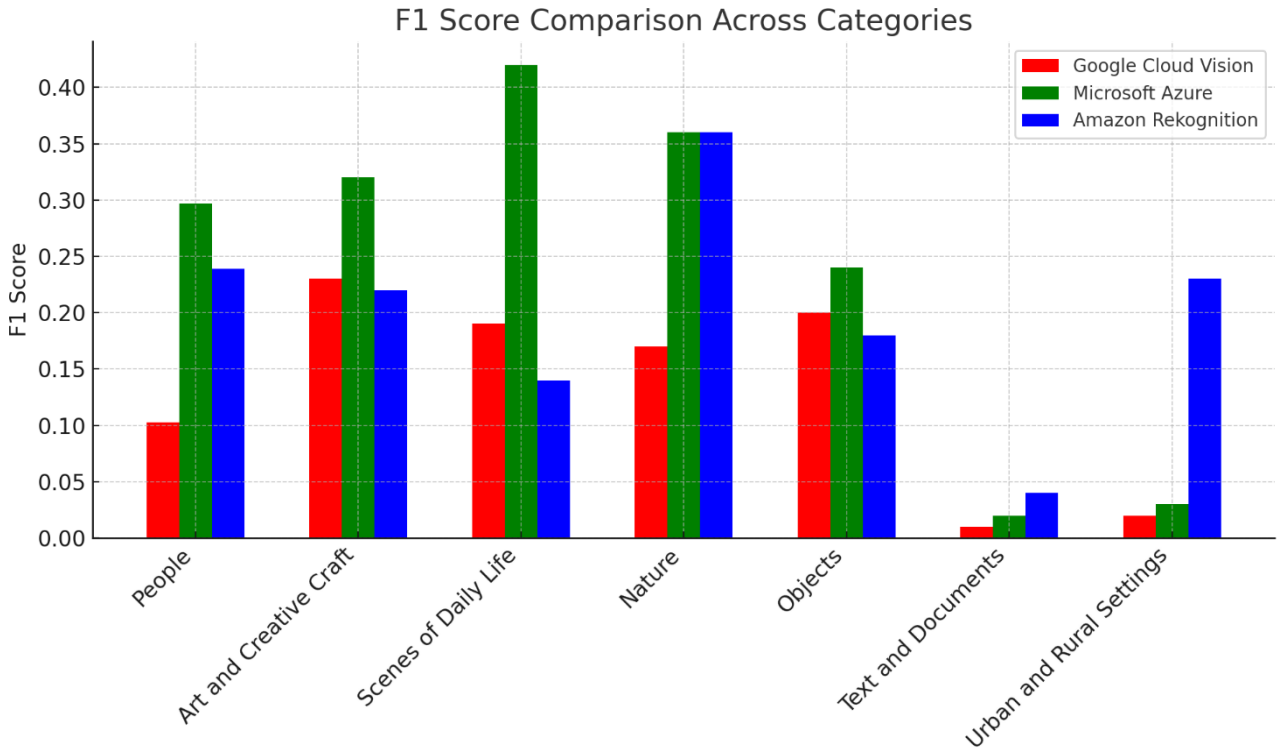
These findings suggest a commercial prioritization in the APIs' training processes, which may favor documents commonly encountered in business and consumer applications over those found in cultural heritage institutions. To address these gaps, cloud providers must expand their datasets to include a broader range of textual materials, such as historical manuscripts, multilingual archives, and culturally significant documents. Additionally, specialized OCR systems trained on heritage-specific datasets could significantly improve tagging accuracy and relevance for this category.

Urban and Rural Settings

In this category, Google Cloud Vision achieved a precision score of 0.06, recall score of 0.01, F1 score of 0.02, and accuracy of 0.01. Microsoft Azure Computer Vision performed slightly better with a precision score of 0.11, recall score of 0.02, F1 score of 0.03, and accuracy of 0.02. Amazon Rekognition significantly outperformed the others with a precision score of 0.72, recall score of 0.14, F1 score of 0.23, and accuracy of 0.14.



The qualitative analysis reveals a stark contrast in performance among the three services for images of urban and rural setting. While Google Cloud Vision and Microsoft Azure struggled significantly, Amazon Rekognition demonstrated notably higher capability in generating relevant tags.



Amazon Rekognition had the best overall performance by a considerable margin, with high precision and improved recall scores. This indicates it was able to identify about 13.8% of the relevant tags while maintaining high precision, suggesting it generates more accurate and comprehensive tags for urban and rural scenes. Google Cloud Vision and Microsoft Azure both performed poorly, with very low precision and recall scores, indicating they struggle to generate relevant tags for this category.

The low F1 scores for Google Cloud Vision and Microsoft Azure highlight a critical imbalance between precision and recall, indicating substantial room for improvement in tag generation for images of urban and rural setting. Amazon Rekognition's higher F1 score suggests a better balance, though there is still significant potential for improvement. This stark difference in performance indicates that Amazon Rekognition may have more advanced capabilities or better-trained models for recognizing elements specific to urban and rural environments. However, even with Amazon Rekognition's superior performance, the results suggest that AI services still have considerable room for improvement in comprehensively tagging urban and rural setting images.

The results in the “Urban and Rural Settings” category highlight disparities in how cloud APIs are optimized for geographic and environmental imagery. Amazon Rekognition’s superior performance suggests that its algorithms are better tuned to recognize architectural and environmental features, particularly those common in urban settings. However, even this system demonstrates biases toward Western architectural

styles and commercially significant urban elements such as modern buildings or road networks, while rural and culturally specific landscapes are often misclassified or overlooked.

These biases likely stem from training datasets that prioritize commercially prevalent geographies, leaving less documented or culturally rich settings underrepresented. To improve tagging accuracy in this category, cloud APIs must expand their training data to include diverse architectural, environmental, and regional features. This would ensure that urban and rural imagery from non-Western contexts is appropriately recognized and tagged, enriching the metadata for global heritage collections.

The findings highlight the need for cultural heritage institutions to adopt a strategic and adaptive approach when integrating cloud APIs for metadata generation. Institutions should select APIs based on their strengths, such as leveraging Microsoft Azure Computer Vision for general-purpose tagging with balanced precision and recall, or Amazon Rekognition for ecological and geographic imagery. Combining multiple APIs may also enhance performance, allowing their complementary strengths to address individual weaknesses. Furthermore, collaboration with API providers to develop diverse training datasets, particularly for underrepresented categories like indigenous art and historical documents, is essential for improving inclusivity. Finally, incorporating human oversight alongside AI can ensure metadata generation aligns with ethical standards and reflects the cultural and contextual richness required for heritage collections. These strategies collectively empower institutions to optimize the value of AI tools while safeguarding the integrity of their collections.

Analysis

The evaluation of three leading AI-driven computer vision services—Google Cloud Vision, Microsoft Azure Computer Vision, and Amazon Rekognition—has revealed critical insights into their performance across diverse image categories. These findings illuminate both the promise and limitations of current AI technologies, particularly in their application to cultural heritage contexts.

This study evaluated the performance of three leading AI-driven computer vision services— Google Cloud Vision, Microsoft Azure Computer Vision, and Amazon Rekognition—in generating metadata tags for diverse image categories. The findings reveal significant variations in performance across different image types and services, highlighting both the potential and limitations of current AI technologies in digital asset management for organizations dealing with large image collections.

The results demonstrate that the effectiveness of AI service's in generating relevant metadata tags varies considerably depending on the image category. For instance, in the 'People' category, Microsoft Azure Computer Vision showed the best overall performance, while Amazon Rekognition excelled in 'Urban and Rural Settings'. This variability aligns with the findings of Villaespesa and Crider (2023), who observed similar inconsistencies in their study of AI tagging for art collections. The literature review highlighted the importance

of metadata in digital asset management, with high-quality metadata enhancing searchability, accessibility, and the overall utility of digital assets (Austerberry, 2005; Kaldeli et al., 2021; Kortemeyer et al., 2014). The varying performance of AI services across image categories underscores the need for a nuanced approach to metadata generation, as relying on a single AI service may not yield optimal results for all image types.

This variability in performance across categories highlights the systemic influence of "platform capitalism" on AI development. As Srnicek (2017) observes, this economic model prioritizes data commodification, incentivizing AI systems to excel in commercially profitable domains. For cultural heritage institutions, this translates into inadequate support for underrepresented narratives, underscoring the need for alternative approaches to metadata generation.

It is important to note that this study encountered significant challenges in its initial stages, which shed light on broader issues facing the cultural heritage sector in the digital age. Initially, the research aimed to utilize rich and diverse image datasets from two prestigious institutions: the Special Collection Unit at the Institute of Ismaili Studies, London, and the British Cultural Archives, London. However, during the preliminary stages of the research, both institutions took longer than expected to respond, and one withdrew their participation, citing privacy concerns and ethical considerations related to the use of cloud-based API tools for image analysis.

These challenges reflect broader concerns articulated in Zuboff's "surveillance capitalism" (2019), which critiques the unchecked commodification of data by technology platforms. For cultural heritage institutions, this dynamic exacerbates fears of losing control over sensitive content, particularly as AI-driven tools embed themselves into institutional workflows.

This unexpected development brought into sharp focus a constellation of issues that numerous cultural heritage institutions are grappling with in the digital age, including data privacy, ethical considerations surrounding AI deployment, complex legacies of collection practices and ownership rights, inadequate digital infrastructure, and varying levels of digitization and metadata standardization across collections. Collectively, these issues paint a picture of a sector in transition, striving to balance the preservation of cultural heritage with the opportunities and challenges presented by digital technologies.

In response to these challenges, the study pivoted to developing a publicly available, carefully curated image dataset scraped from the web, designed to replicate the diversity and complexity of real-world organizations. While this approach allowed the research to proceed, it is important to acknowledge that this dataset may not fully capture the nuances and complexities present in institutional collections.

The “Art and Creative Craft” and “Daily Life Scenes” categories saw better performance across all services, with higher precision scores but consistently low recall. This indicates that while the AI services can generate accurate tags for these categories, they struggle to provide comprehensive coverage of relevant descriptors. This limitation echoes the observations of Fornaro and Chiquet (2021), who noted the challenges AI systems face in capturing nuanced contextual information in diverse image collections. These findings reveal a

systemic bias in metadata generation, where underrepresented cultural narratives risk exclusion. For cultural heritage institutions, such gaps perpetuate the erasure of less-documented traditions, limiting their ability to foster inclusive storytelling and equitable access. The literature review emphasized the importance of contextually rich metadata in cultural heritage institutions (Storch, 2023), and the findings suggest that AI services alone may not be sufficient to meet this requirement without human intervention.

Notably, all three services performed poorly in the "Text and Documents" category, suggesting a significant gap in their ability to process and tag images containing textual content. This finding aligns with the need for specialized OCR capabilities when dealing with documentheavy collections, as highlighted by Storch (2023) in their study of AI applications in archival contexts.

A consistent pattern observed across all image categories was the trade-off between precision and recall. Generally, the AI services demonstrated higher precision scores but struggled with recall, indicating that they generate accurate but incomplete sets of tags. This precision-recall trade-off reflects broader limitations inherent in commercial AI models driven by platform capitalism. Optimizing for efficiency and marketability often leads to the exclusion of cultural specificity and nuanced interpretations, critical for heritage applications. This imbalance was particularly pronounced in categories like "Nature" and "Objects", where even the best performing services identified only a fraction of the relevant tags. This precision-recall tradeoff aligns with the findings of Sharma (2022), who observed similar patterns in object detection using Amazon Rekognition. The implication for organizations dealing with large image datasets is that while AI-generated tags can be trusted for their accuracy, they should not be relied upon for comprehensive description without human oversight, particularly for internal users who require detailed and context-rich metadata for their work.

The qualitative analysis revealed that even when AI services generated accurate tags, they often lacked the contextual depth and nuance necessary for meaningful description in various professional contexts. This limitation was evident across all image categories but was particularly noticeable in "Art and Creative Craft" and "Cultural and Historical Contexts". This finding underscores the observations of Bakker and Castro (2024), who emphasized the importance of combining AI capabilities with human expertise to ensure sensitive and contextually rich metadata. The literature review also highlighted the issue of bias and accuracy in AI-generated metadata (Fornaro & Chiquet, 2021), and the inability of AI services to consistently capture nuanced contexts reinforces the need for human curation and interpretation in digital asset management for organizations dealing with diverse image collections. To address these systemic concerns, cultural heritage institutions should advocate for the development of open-source AI tools that prioritize inclusivity and transparency. Collaborative frameworks involving domain experts and AI developers could ensure that metadata systems reflect diverse cultural contexts, aligning technological innovation with ethical stewardship.

It is important to acknowledge that the use of customized controlled vocabularies as ground truth to calculate metrics may introduce bias, and the results could be influenced by the researcher's positionality. Future research should consider the potential impact of researcher positionality and the choice of ground truth on the evaluation of AI-driven metadata tagging systems.

The varied performance of AI services across different image categories has significant implications for their integration into digital asset management systems in organizations handling large image datasets. While these technologies show promise in certain areas, their inconsistent performance suggests that a one-size-fits-all approach to AI-driven metadata generation is not feasible. Instead, organizations might consider a hybrid approach, leveraging the strengths of different AI services for specific image categories while maintaining human oversight. This aligns with the recommendations of West, Denny, and Ruud (2021), who advocated for a balanced integration of AI technologies in DAM systems.

The findings of this study highlight the need for continued development of AI models specifically trained on diverse image collections relevant to different industries and sectors. As Christophe et al. (2024) suggested, collaboration between AI developers and domain experts could lead to more nuanced and effective metadata tagging systems. The literature review also emphasized the importance of interdisciplinary collaboration in developing AI-driven metadata systems (Alliata, Hou, & Kenderdine, 2024; Fornaro & Chiquet, 2021), and the findings reinforce the need for such collaboration to address the limitations of current AI services.

The limitations inherent in the cloud-based nature of the AI services evaluated in this study - including data privacy and security concerns, limited control over training data, dependency on external infrastructure, potential for bias, and lack of customization - align with the ethical considerations discussed in the literature review (Dalal-Clayton & Rutherford, 2024; Jaillant & Aske, 2024). These limitations underscore the importance of a critical and cautious approach to implementing cloud-based AI tools in organizational workflows, especially for internal users dealing with large and diverse image datasets.

While this study provides valuable insights into the current capabilities of AI-driven metadata tagging for large image collections, it has several limitations. Future research could benefit from partnerships with multiple organizations across different sectors to access more diverse and representative datasets and should explore the performance of these AI services across different languages and cultural frameworks. By addressing the socio-economic structures underpinning AI development—namely, platform capitalism and surveillance capitalism—future research can better align AI technologies with the ethical and practical needs of cultural heritage institutions. This shift requires interdisciplinary collaboration and a rethinking of AI’s role as not just a tool for efficiency but a medium for cultural inclusivity and integrity.

Discussion

The findings of this study provide critical insights into the capabilities and limitations of AI-driven metadata generation services—Google Cloud Vision, Amazon Rekognition, and Microsoft Azure Computer Vision—when applied to diverse image categories. This discussion synthesizes these findings to address their

practical, theoretical, and ethical implications for cultural heritage institutions and the broader field of digital asset management (DAM).

5.1 Practical Implications

The variability in performance across image categories underscores the necessity of adopting a strategic, hybrid approach to AI integration in DAM workflows. Microsoft Azure Computer Vision demonstrated consistent performance in categories requiring contextual accuracy, such as “People” and “Daily Life Scenes,” making it a suitable choice for general-purpose tagging. Conversely, Amazon Rekognition excelled in categories like “Urban and Rural Settings,” suggesting its strength in geographical and architectural recognition. However, no single service proved universally effective, emphasizing the need to combine AI services for complementary strengths while incorporating human oversight to fill gaps in recall and contextual depth.

Practical recommendations include:

Hybrid Approaches: Combining multiple AI services to leverage their unique strengths and address individual weaknesses.

Enhanced Training Datasets: Encouraging collaboration with AI providers to expand training datasets, particularly for underrepresented cultural and ecological contexts.

Human-AI Collaboration: Establishing workflows where AI handles initial tagging, followed by human validation to ensure cultural and contextual accuracy.

5.2 Theoretical Contributions

This study enriches the theoretical discourse on metadata generation by highlighting systemic biases embedded in AI algorithms. The analysis revealed a persistent trade-off between precision and recall across all services, reflecting the economic prioritization of efficiency under "platform capitalism" (Srnicek, 2017). AI systems, driven by commercial imperatives, prioritize datasets optimized for high-frequency, commercially valuable tags at the expense of nuanced, culturally sensitive representations.

These finding challenge current theories of AI neutrality and underscores the need to incorporate socio-technical perspectives in metadata research. By aligning with the theories of Zuboff’s "surveillance capitalism" (2019), this study highlights how AI development, influenced by profit-driven imperatives, often neglects the ethical and cultural dimensions critical for heritage applications.

5.3 Ethical Implications

The study surfaces critical ethical challenges in deploying AI for cultural heritage metadata:

Bias in Training Data: The exclusion of marginalized cultural narratives from AI training datasets perpetuates systemic erasures, particularly evident in the pretrained models of cloud-based APIs like Google Cloud

Vision, Amazon Rekognition, and Microsoft Azure Computer Vision. These pretrained models rely heavily on datasets curated for global scalability, often prioritizing high-frequency, commercially significant tags that align with Western cultural contexts. For example, Google's vast image dataset has shown a propensity for identifying globally prevalent objects and concepts but often fails to recognize Indigenous artifacts, such as specific ceremonial masks or regionally significant botanical elements. Similarly, Amazon Rekognition's training datasets appear optimized for consumer-focused applications, such as retail product identification or security surveillance, rather than the nuanced recognition required for cultural heritage contexts. Microsoft Azure Computer Vision demonstrates slightly better adaptability due to its emphasis on enterprise use cases, but its performance still reflects biases in favor of mainstream categories, as evidenced by its struggles with linguistically diverse texts and historical documents.

These limitations are deeply rooted in the commercial imperatives driving cloud-based API development, which prioritize datasets that maximize generalizability and efficiency over inclusivity. As a result, heritage institutions relying on these services often encounter gaps in the representation of culturally specific or historically nuanced materials. For instance:

Google Cloud Vision tends to misclassify or overlook Indigenous artifacts due to the lack of regional and ethnographic datasets in its training pipeline.

Amazon Rekognition often fails to tag culturally significant features in non-Western rural settings, highlighting a lack of diversity in its training data.

Microsoft Azure Computer Vision has shown limited capability in recognizing non-Latin scripts or historically significant calligraphy, further evidencing a gap in linguistic and cultural representation.

These biases not only impact the accuracy of metadata but also risk perpetuating systemic erasures by underrepresenting marginalized narratives. Addressing these issues requires:

Collaborative Dataset Development: Cloud API providers must collaborate with cultural heritage institutions to incorporate diverse, underrepresented cultural artifacts, languages, and symbols into their training datasets.

Localized Pretraining Models: Developing region-specific pretrained models could improve the recognition of Indigenous artifacts and linguistically diverse texts, bridging the gap between global scalability and localized accuracy.

Transparent Training Processes: Enhancing transparency about dataset composition and training methodologies would enable institutions to assess whether a given service aligns with their cultural heritage goals.

By focusing on these measures, cloud-based APIs can evolve beyond their current commercial biases, contributing to the equitable representation and preservation of global cultural diversity in digital asset management systems

Data Privacy Concerns: The reluctance of cultural institutions to share sensitive materials with cloud-based AI services highlights growing fears of losing control over data, exacerbated by surveillance capitalism practices.

Homogenization of Metadata: Optimizing metadata for generality over specificity risks diluting the cultural richness of heritage collections.

Addressing these issues requires:

Open-Source Development: Advocating for open AI systems that promote transparency and inclusivity.

Community Involvement: Actively engaging stakeholders, particularly from underrepresented communities, in dataset curation and algorithm development.

Ethical Frameworks: Establishing robust guidelines to navigate the trade-offs between automation and cultural sensitivity.

5.4 Limitations and Future Directions

While this study provides valuable insights, several limitations warrant consideration:

Dataset Representation: Although the curated dataset attempted to mimic institutional diversity, it may not fully reflect the complexity of actual collections.

Evolving AI Technologies: Rapid advancements in AI mean that findings could become outdated, necessitating continuous re-evaluation.

Researcher Positionality: The controlled vocabularies and metrics chosen reflect specific interpretive frameworks, which could influence the evaluation of AI performance.

Future research should:

Expand comparative studies across additional languages, cultural contexts, and underrepresented image categories.

Investigate scalable models for integrating AI and human expertise in metadata workflows.

Develop context-sensitive metrics to evaluate AI-generated metadata not only for accuracy but also for cultural and historical relevance.

5.5 Broader Implications

This study underscores the dual role of AI as both a tool for efficiency and a medium for cultural stewardship. The socio-economic structures underpinning AI development, particularly platform capitalism, limit its utility for equitable representation in cultural heritage applications. To counter these challenges, institutions must prioritize interdisciplinary collaboration and advocate for AI systems that align with ethical principles of diversity, inclusivity, and transparency.

Moreover, this research emphasizes the potential of AI to augment, rather than replace, human expertise. By reframing AI as a collaborator rather than a substitute, organizations can harness its strengths while mitigating its limitations. This hybrid approach promises to preserve the integrity of cultural narratives while unlocking new possibilities for digital asset management.

Conclusion

This study evaluated the capabilities of Google Cloud Vision, Microsoft Azure Computer Vision, and Amazon Rekognition in generating metadata tags for diverse image categories, specifically within the context of digital asset management (DAM) and cultural heritage preservation. The findings illuminate the dual potential and limitations of these AI-driven tools, offering critical insights for organizations navigating the challenges of managing large-scale, diverse digital image collections.

Key Insights

Variability in Performance: The AI services demonstrated significant differences in precision, recall, and contextual sensitivity across image categories. For example, while Microsoft Azure excelled in "People" and "Daily Life Scenes," Amazon Rekognition outperformed in "Urban and Rural Settings," revealing that no single service can meet all DAM needs comprehensively.

Bias and Exclusion: Systemic biases in pretrained models, shaped by commercially-driven datasets, limit the representation of marginalized cultural narratives. These biases perpetuate erasures, particularly of Indigenous artifacts, linguistically diverse texts, and regionally specific contexts, as evidenced by all three services.

Human-AI Synergy: Despite advancements in AI, human expertise remains indispensable. The inability of these tools to consistently generate nuanced, contextually rich metadata underscores the need for workflows where human validation complements automated tagging.

Ethical and Practical Challenges: Ethical concerns, such as data privacy, metadata homogenization, and the dominance of platform capitalism, highlight the need for culturally sensitive and ethically grounded approaches to AI integration.

Implications for Practice

Hybrid Strategies: Organizations should adopt hybrid approaches, combining multiple AI services to leverage their strengths while employing human oversight to address gaps in recall and cultural nuance.

Training Data Diversity: Collaboration with AI providers is crucial to expand training datasets, emphasizing culturally and linguistically diverse content to ensure equitable representation in metadata generation.

Human-Centric Interfaces: Developing intuitive interfaces that facilitate human-AI collaboration can enhance the accuracy and relevance of metadata tagging, ensuring that human expertise enriches AI-generated outputs.

Advocacy for Open Standards: Cultural institutions must advocate for open-source, transparent AI frameworks that prioritize inclusivity over commercial imperatives, aligning with the mission of heritage preservation.

Directions for Future Research

Culturally Specific AI Models: Developing AI systems tailored to the unique needs of cultural heritage organizations can address the limitations of generic, commercially oriented tools.

Ethical Framework Development: Further studies should establish robust guidelines for addressing biases, preserving data privacy, and ensuring cultural sensitivity in AI applications.

Longitudinal Impact Studies: Investigating the long-term effects of AI-generated metadata on user engagement, cultural heritage access, and historical interpretation will provide deeper insights into its broader implications.

Scalable Collaborative Models: Future research should explore scalable frameworks for human-AI collaboration, ensuring that metadata workflows balance efficiency with cultural and contextual richness.

Broader Implications

This study underscores the transformative potential of AI in DAM while emphasizing the critical need for ethical, equitable, and inclusive practices. The socio-economic dynamics of platform capitalism and surveillance capitalism impose systemic constraints on AI's utility for cultural heritage, prioritizing efficiency and marketability over diversity and depth. To counter these challenges, institutions must champion interdisciplinary collaborations and advocate for AI systems that uphold the principles of cultural stewardship.

Ultimately, AI's role in digital asset management should be reimagined as a collaborator rather than a substitute for human expertise. By fostering synergies between machine intelligence and human judgment, organizations can unlock new possibilities for preserving and accessing digital collections while safeguarding the cultural richness and historical significance that define our shared heritage.

Appendices

Appendix A: Code Repository

All the code used in this dissertation is available in the GitHub repository at the following link: https://github.com/khalidxansari/dissertation-code

Bibliography

Alliata, A., Hou, S., & Kenderdine, S. (2024). Enhancing access to embodied knowledge archives through posture recognition and movement computing. Digital Humanities Quarterly, 18(1), 1-20.

Austerberry, D. (2005). The components of a digital asset management system. Journal of Digital Asset Management, 1(2), 131-145.

Baca, M. (2016). Introduction to metadata. Getty Publications.

Bakker, R., & Castro, M. (2024). From pixels to Python: When digital collections befriend artificial intelligence. Journal of Digital Media Management, 12(2), 157-166.

Black Cultural Archives. (n.d.). Adamah Papers. Retrieved July 27, 2024, from

https://collections.blackculturalarchives.org/repositories/2/resources/54

Christophe, A., Solomon, N. H., Kneubuhl, H., Donnellan, V., & Caldeira, L. (2024). Hosting and integrating a Hawaiian language taxonomy in the British Museum's collection database. Collections: A Journal for Museum and Archives Professionals, 00(0), 1-26. https://doi.org/10.1177/15501906241234945

Dalal-Clayton, A., & Rutherford, A. (2024). Provisional semantics: Addressing the challenges of representing multiple perspectives within public collections. Collections: A Journal for

Museum and Archives Professionals, 00(0), 1-20. https://doi.org/10.1177/15501906241232427

Davet, J., Hamidzadeh, B., & Franks, P. (2023). Archivist in the machine: Paradata for AI-

based automation in the archives. Archival Science, 23(3), 275-295. https://doi.org/10.1007/s10502-023-09408-8

Field, A. (2013). Discovering statistics using IBM SPSS statistics. Sage.

Floridi, L. (2019). Establishing the rules for building trustworthy AI. Nature Machine Intelligence, 1(6), 261-262.

Fornaro, P., & Chiquet, V. (2021). Artificial intelligence for content and context metadata retrieval in photographs and image groups. Journal of Digital Media Management, 9(4), 297304.

Getty Research Institute. (2024). Getty Art & Architecture Thesaurus (AAT). Retrieved July

27, 2024, from https://www.getty.edu/research/tools/vocabularies/aat/

Gilliland, A. J. (2019). Conceptualizing 21st-century archives. Society of American Archivists.

Huddart, K. (2022). Artificial intelligence powered digital asset management: Current state and future potential. Journal of Digital Media Management, 11(1), 6-17.

Ismaili Special Collections Unit. (2024). The Institute of British Studies. Retrieved July 27, 2024, from https://www.iis.ac.uk/library-and-special-collections/special-collections/

Jaillant, L., & Aske, A. (2024). Responsible access to historical medical illustrations: Ethical challenges and the use of AI. Journal of Medical Ethics, 50(2), 123-129.

Kaldeli, E., Mitzias, P., Kokkinakis, D., Kapasakis, D., Kontopoulos, E., Kompatsiaris, I., & Papatheodorou, C. (2021). Digital curation and metadata enrichment in aggregated cultural heritage data. Communications in Computer and Information Science, 1388, 20-32.

Kortemeyer, G., Kashy, E., Benenson, W., & Bauer, W. (2014). Experiences using the opensource learning content management and assessment system LON-CAPA in introductory physics courses. American Journal of Physics, 76(4), 438-444.

Library of Congress. (2024). Library of Congress Subject Headings (LCSH). Retrieved July 27, 2024, from https://id.loc.gov/authorities/subjects.html

Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval.

Cambridge University Press.

NISO. (2004). Understanding metadata. National Information Standards Organization.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

Perez, P. J., Escueta, J. K., & Jequinto, P. M. P. (2022). Digital cultural heritage inequality:

Philippine museums during the COVID-19 pandemic. Collections: A Journal for Museum and

Archives Professionals, 18(4), 568-585. https://doi.org/10.1177/15501906221121624

Sharma, V. (2022). Object detection and recognition using Amazon Rekognition with Boto3. Proceedings of the 6th International Conference on Trends in Electronics and Informatics (ICOEI 2022), 727-732. https://doi.org/10.1109/ICOEI53556.2022.9776884

Storch, H. (2023). Collection insight and interconnectivity through artificial intelligence image analysis: A collaboration with the National Archives of Estonia. Journal of Digital Media Management, 12(1), 35-45.

Szeliski, R. (2010). Computer Vision: Algorithms and Applications (1st ed.). Springer Nature. https://doi.org/10.1007/978-1-84882-935-0

Temel, D., Lee, J., & AlRegib, G. (2019). Object recognition under multifarious conditions: A reliability analysis and a feature similarity-based performance estimation. Proceedings of the IEEE International Conference on Image Processing (ICIP 2019), 3033-3037.

https://doi.org/10.1109/ICIP.2019.8803317

Thomas, D., & Testini, A. (2024). AI for book illustrations: Identifying and analyzing image captions in historical books. Digital Scholarship in the Humanities, 39(1), 78-95.

UK Archival Thesaurus. (2024). UKAT - UK Archival Thesaurus. Retrieved July 27, 2024, from http://ukat.aim25.com/

Villaespesa, E., & Crider, S. (2023). Computer vision tagging the Metropolitan Museum of

Art's collection: A comparison of three systems. Journal of Information Technology & Museums, 12(3), 45-67.

West, D., Denny, C., & Ruud, R. (2021). Case study: Integrating artificial intelligence metadata within Paramount's digital asset management system. Journal of Digital Media Management, 9(3), 198-208.

Zeng, M. L., & Qin, J. (2016). Metadata (2nd ed.). Neal-Schuman.