Quantifying Archival Silences: Phylogenetic Diversity Analysis of Controlled Vocabulary Utilization

Melvin Wevers[1], Thomas Smits[1], and Folgert Karsdorp[2]

[1] Department of History, University of Amsterdam, Amsterdam, the Netherlands
[2] Meertens Institute, Amsterdam, the Netherlands

Abstract

This study adapts Faith's Phylogenetic Diversity metric from ecology to measure controlled vocabulary utilization in archival collections, addressing limitations of traditional diversity measures that ignore hierarchical term relationships. We introduce three diagnostic ratios—Coverage, Completeness, and Cataloging Intensity—and apply them to 878,046 photographs across 16 Dutch National Archives collections cataloged with the hierarchical GTAA vocabulary. The framework provides quantitative tools for assessing how vocabulary utilization patterns influence cultural heritage accessibility while highlighting the tension between cataloging intensity and comprehensive research utility. The findings suggest that the interaction between collection content characteristics and institutional cataloging practices creates different pathways for cultural heritage discovery, revealing substantial variation in both the scope of conceptual domains (coverage ratio) addressed and the thoroughness (completeness ratio) of description within those domains. This framework provides empirical benchmarks for evidence-based collection assessment and metadata evaluation.

Keywords: phylogenetic diversity, controlled vocabularies, cultural heritage, audiovisual archives

1 Introduction

Archives mediate access to cultural heritage through controlled vocabularies that structure how materials are described and discovered. These vocabularies organize terms into hierarchical relationships, enabling archivists to describe materials consistently while helping researchers navigate collections, whether searching broadly within subject areas or locating specific types of objects. While these organizational systems may appear technically neutral, contemporary archival theory reveals how such seemingly objective practices carry implications for historical understanding and the distribution of power. This theoretical shift recognizes that archives actively shape what can be known about the past through the accumulated effects of countless classificatory decisions. Jacques Derrida's concept of "archival violence" reveals how this shaping operates structurally—archives commonly exercise power not through overt censorship but through the foundational choices embedded in classification systems, finding aids, and digitization priorities [5]. Building on this insight, Michel-Rolph Trouillot's analysis of "retrieval mechanisms" reveals the concrete pathways through which archival violence manifests in practice: historical silences emerge not only from what is excluded from preservation but from the organizational pathways that render certain materials effectively invisible to researchers [17]. These retrieval mechanisms represent the practical

Melvin Wevers, Thomas Smits, and Folgert Karsdorp. "Quantifying Archival Silences: Phylogenetic Diversity Analysis of Controlled Vocabulary Utilization." In: *Computational Humanities Research 2025*, ed. by Taylor Arnold, Margherita Fantoli, and Ruben Ros. Vol. 3. Anthology of Computers and the Humanities. 2025, 266–279. https://doi.org/10.63744/H9LnUCR9Mxmx.

instantiation of Derrida's more abstract concept, demonstrating how the very vocabularies and finding systems that enable discovery simultaneously constrain it, creating effects that extend far beyond the initial moment of classification to shape every subsequent encounter between researcher and archive.

Vocabularies represent how these theoretical dynamics play out in the practical work of archival organization. They impose a "double constraint" on archival accessibility: their internal structure may contain representational biases, and cataloging practices can further shape how thoroughly and equitably those vocabularies are applied [12]. Even when relevant materials exist in collections, this double constraint can render them effectively undiscoverable, enacting the kind of structural control that Derrida theorizes. Archival silences can thus arise from two sources: what vocabularies cannot express (vocabulary bias) and how vocabularies are applied in practice (utilization bias). Both patterns contribute to what Trouillot identifies as structural silences. Conversely, these same mechanisms can produce amplification effects when certain materials become highly discoverable through extensive vocabulary application—a process known as archival shouting [19].

Recent scholarship has emphasized the need for "critical digital archives" that actively address these structural silences [1]. To identify and measure these silences, scholars have applied various metrics for assessing collection characteristics, including entropy-based measures and frequency distributions [14; 15; 16; 20]. Others have used topic modeling and natural language processing to analyze the descriptions of items in cultural heritage archives [11], while recent work has applied knowledge graphs and deep learning to extract structured knowledge from unstructured texts [7]. While these quantitative approaches show promise, they often ignore the hierarchical nature of vocabularies, treating all terms as flat categories and thereby misrepresenting how knowledge is organized and accessed in cultural heritage institutions.

Recent work in computational humanities has begun addressing this limitation by introducing metrics from the field of ecology. One such metric is Functional Diversity (FD), which incorporates similarities and distances between elements based on their distinct functions [8]; another one is Chao1, a species estimation method that determines lower bounds of missing elements and identifies gaps in historical collections [9; 10; 13; 18].

This study extends this line of research by adapting Faith's Phylogenetic Diversity (PD) [6], a metric that quantifies biodiversity by incorporating evolutionary relationships between species. We apply this metric to measure vocabulary utilization patterns in archival records. Traditional diversity measures that count unique terms fundamentally misrepresent how vocabularies organize knowledge. Vocabularies are structured as hierarchical trees where broader terms encompass narrower ones—"transportation" branches into "automobiles," "bicycles," and "ships." Treating "transportation" and "bicycles" as equidistant concepts ignores the hierarchical conceptual structure. As a result, these measures fail to take into account how users actually navigate collections. PD addresses this limitation by explicitly accounting for hierarchical relationships. In ecology, the metric assigns greater weight to evolutionarily distinct species. For example, preserving ten closely related beetles represents less diversity than preserving five species from different evolutionary branches. PD calculates the total evolutionary history represented by a set of species by summing all branch lengths in the minimal subtree connecting them. Applied to vocabularies, this translates to measuring the total hierarchical path length required to connect all utilized terms in a collection, quantifying terminological diversity through cumulative hierarchical distance rather than simple concept counts.

Figure 1 illustrates this principle across three vocabulary utilization scenarios with increasing diversity scores. The left panel shows low PD (4.2) for two closely related concepts requiring minimal hierarchical structure to connect them. The center panel demonstrates medium PD (5.8) for three concepts with moderate hierarchical distances. The right panel achieves maximum PD (10.049) through four concepts spanning major hierarchical divisions, representing the broadest ter-

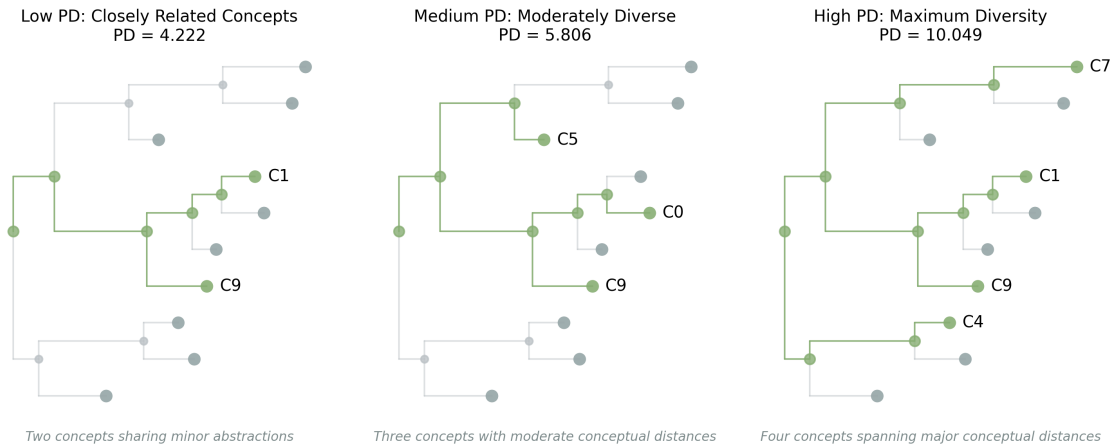

Figure 1: Phylogenetic Diversity calculation examples showing how hierarchical distance, not concept count, determines terminological diversity. Green paths represent the minimal subtree connecting selected concepts, with PD scores reflecting cumulative branch lengths.

minological coverage. The green highlighting indicates the cumulative hierarchical path lengths that are necessary to calculate PD. This figure demonstrates how the metric captures conceptual breadth rather than mere concept count, i.e., concepts spanning major hierarchical divisions contribute more to diversity than clusters of closely related terms.

The statistical estimators developed for phylogenetic diversity are universally applicable to any hierarchically structured count data, making them naturally suited for archival collections organized through vocabularies. Our approach treats these vocabularies as tree structures with uniform branch lengths, focusing on conceptual coverage rather than evolutionary distance. Additionally, we incorporate PD_{Chao1} estimators [3], which adapt the Chao1 abundance estimator to phylogenetic diversity contexts. These estimators provide lower-bound estimates of "unseen" vocabulary terms that could plausibly be applied within already-covered subject domains but remain unused in the collection. This enables assessment of completeness: how thoroughly institutions utilize available GTAA vocabulary within their established topical scope. This enables measurement of both current conceptual breadth and potential coverage gaps across entire collections.

To operationalize these insights about coverage and completeness, we introduce three diagnostic ratios that measure vocabulary use across collections: (1) the *Coverage Ratio* measures what fraction of all possible vocabulary terms a collection uses; (2) the *Completeness Ratio* assesses how thoroughly a collection applies the available vocabulary within the subject areas it covers, estimating potential cataloging gaps; (3) the *Cataloging Intensity Ratio* evaluates the terminological breadth of a collection relative to its size.

We apply these ratios to the Dutch National Archives' photograph collections (approximately 1M images) cataloged using the *Gemeenschappelijke Thesaurus voor Audiovisuele Archieven* (GTAA), a hierarchical vocabulary of over 4,000 terms.[1]

2 Methodology

We adapt Faith's phylogenetic diversity metric to quantify vocabulary utilization patterns across archival collections. Our approach generates three diagnostic ratios that quantify different aspects of conceptual representation through a three-step process of tree construction, term mapping, and diversity calculation.

[1] For more on the vocabulary see: `https://data.beeldengeluid.nl/datasets/gtaa`

2.1 Conceptual Framework

We treat archival vocabularies as phylogenetic trees where hierarchical relationships mirror evolutionary structures. In this framework, *nodes* represent individual subject terms (e.g.,"world war, 1939-1945"), *edges* denote the hierarchical relationships between broader and narrower terms (e.g., "wars" connects to "world War, 1939-1945"), *branch lengths* represent conceptual distances, with each hierarchical level constituting one conceptual "step", and *collections* refers subsets of the available vocabulary terms used to describe the archival materials they contain.

The term "subject" in "subject indexing" refers specifically to topical subjects (*trefwoordAlgemeen*), excluding other categories like persons, locations, genre, provenance, and format. Consequently, the analysis measures bias in topical coverage only, not comprehensive archival subject representation across all GTAA dimensions.

Our implementation differs from ecological applications in three key ways. First, unlike ecologists who cannot determine total species diversity in an ecosystem, we work with vocabularies that define finite, known sets of terms. The bounded context allows us to calculate the maximum possible phylogenetic diversity across the complete vocabulary tree, providing a baseline impossible in ecological studies. Importantly, however, this approach remains bounded by the vocabulary's own representational scope, measuring utilization within that scope rather than across all possible conceptual representations.

Second, while evolutionary trees use branch lengths proportional to genetic divergence, we assign uniform branch lengths (1) to all hierarchical relationships. Such uniform weighting is standard practice when only hierarchical structure is available without empirical distance measurements [3; 6]. Uniform weighting ensures that diversity metrics reflect the number of conceptual boundaries crossed rather than assumed semantic proximity.[2] Consequently, transitions from "transportation" to "automobiles" and from "automobiles" to "sedans" contribute equally to phylogenetic diversity, reflecting coverage of the vocabulary's hierarchical organization.

Third, for vocabulary branches represented in a collection, we apply Chao1-based estimation to predict the minimum number of additional terms likely to exist within those same semantic categories but remain unused in cataloging. This approach provides a lower bound estimate of incomplete documentation within covered conceptual territories. Rather than estimating missing species across entire ecosystems, our unseen diversity estimation reveals documentation gaps within specific conceptual areas of a collection.

2.2 Diagnostic Ratios

We operationalize concepts from archival theory through three complementary diagnostic ratios:

1. Coverage Ratio $= \frac{\text{PD}_{\text{collection}}}{\text{PD}_{\text{GTAA}}}$ measures the fraction of total available conceptual space utilized by a collection, with values ranging from 0 to 1. Higher values indicate broader conceptual scope across vocabulary domains.

2. Completeness Ratio $= \frac{\text{PD}_{\text{collection}}}{\text{PD}_{\text{collection}}+\text{PD}_{\text{unseen}}}$ measures documentation thoroughness within covered conceptual territory by incorporating estimated unseen diversity. Values near 1 indicate comprehensive sampling of covered conceptual areas, identifying potential archival silences.

3. Cataloging Intensity Ratio $= \frac{\text{Coverage Ratio}}{\log(\text{Collection Size})}$ measures the conceptual work invested per archival item, revealing institutional resource allocation patterns in descriptive cataloging. High intensity indicates strategic curatorial investment that enhances discoverability, while

[2] Sensitivity analyses with alternative weighting schemes (uniform, random, depth-related) confirmed that while absolute PD values change, the relative ranking and clustering of collections remain stable. Thus, uniform branch lengths provide a principled and reproducible baseline for this study.

low intensity may signal under-resourced cataloging that renders materials conceptually impoverished and harder to discover.

2.3 Methodological Constraints

A fundamental limitation is that lower diagnostic scores could reflect either (1) incomplete application of relevant GTAA terms by catalogers, (2) inadequacy of the GTAA vocabulary for describing collection content, or (3) a combination of both. Our method operates entirely within the GTAA framework and cannot distinguish between these scenarios. Consequently, our diagnostic ratios measure GTAA utilization patterns rather than absolute cataloging quality.

This framework was developed and tested specifically for subject-based indexing of visual collections in well-resourced institutional contexts. Two important scope limitations should be acknowledged. First, many archival institutions employ different approaches to description and access, including hierarchical indexing models where terms applied at higher aggregation levels implicitly extend to components within. Second, subject indexing represents only one dimension of archival description. Comprehensive archival description typically includes genre, creator, provenance, geographic location, documented functions, and other facets—often using multiple authority records and vocabularies beyond hierarchical subject thesauri. These additional descriptive dimensions frequently employ different relationship types than the hierarchical structures analyzed in this study. Our framework specifically examines GTAA subject term utilization patterns rather than evaluating the comprehensive descriptive completeness of archival records across all possible dimensions.

2.4 Data Processing Pipeline

We construct a tree-like representation of the GTAA vocabulary by processing hierarchical term relationships. We load broader/narrower term relationships from the GTAA structured data, construct a directed acyclic graph representing semantic relationships, and then add a synthetic root node to connect multiple independent hierarchical trees, creating a unified structure for phylogenetic analysis.

A key challenge involves resolving multiple parent relationships, as GTAA allows terms to belong to multiple semantic categories (e.g., “war photography” relating to both “photography” and “warfare” hierarchies). Since phylogenetic diversity calculations require strict tree structures with single parent nodes, we eliminate these multiple parent relationships while preserving hierarchical information. To address this, we implement three deduplication strategies: frequency-based (retaining the parent with the highest occurrence in metadata), depth-based (keeping the deepest branch parent), and order-based (maintaining the first encountered parent). Testing these approaches across 16 archival collections reveals remarkable stability, with Spearman correlations exceeding 0.99 between all strategies. This demonstrates that our approach captures fundamental diversity patterns rather than being sensitive to implementation details. Based on these results, we employ the frequency-based strategy to ensure our tree structure reflects empirical usage patterns rather than theoretical conceptual relationships.

2.5 Phylogenetic Diversity Calculation

We implement Faith’s Phylogenetic Diversity (PD) calculation by identifying the minimum spanning path T connecting all terms associated with a collection and summing the branch lengths e within this structure (see Figure 1).[3] PD quantifies the total “conceptual distance” spanned by a

[3] There is also a modern implementation offered by Scikit-Bio, which calculates phylogenetic diversity by identifying the minimal subtree that connects all terms associated with a given collection and summing the branch lengths within this

collection's subject annotations.

Following the theoretical definition, PD of term set s equals the sum of lengths of all branches in the corresponding minimum spanning path T:

$$PD(T) = \sum_{e \in \text{minimum\_spanning\_path}(T)} \text{length}(e)$$

We employ uniform branch lengths of 1 for computational consistency, treating each hierarchical relationship as conceptually equidistant. This standardized approach provides a framework for comparative analysis across collections, even though it does not capture varying semantic distances between terms.

2.6 Unseen Diversity Estimation

We employ the PD_{Chao1} estimator, which extends the Chao1 abundance estimator specifically for phylogenetic data [3; 4]. This method estimates the phylogenetic diversity of unobserved terms, thereby providing lower bounds for unexpressed conceptual potential.

The implementation follows three sequential steps. We begin with node-level frequency analysis across the GTAA tree, calculating the sum of branch lengths grouped by child node frequency in collection metadata. We then calculate the sum of branch lengths for singletons (terms appearing once, yielding g_1) and doubletons (terms appearing exactly twice, yielding g_2). Next, when doubletons exist we apply the following formula: $\hat{g}_0 = \frac{n-1}{n} \times \frac{g_1^2}{2g_2}$, where n represents total term observations. In cases where multiple singletons are observed but no doubletons exist, we employ the modified formula $\hat{g}_0 = \frac{n-1}{n} \times \frac{g_1(g_1-1)}{2}$ to avoid division by zero. Finally, we calculate the complete phylogenetic diversity estimate by combining observed PD with estimated unseen diversity. This yields lower bound estimates of the total conceptual coverage potential, accompanied by 95% confidence intervals derived through log-normal approximation [2; 3].

3 Data: Dutch National Archive Photographic Collection

The study analyzes metadata terms from the digitized photograph collections of the Dutch National Archives.[4] We retrieved metadata from the archive's online photographic database in May 2025 via their SPARQL endpoint. The complete corpus encompasses nearly 1 million photographs distributed across multiple collections of varying sizes and institutional contexts. These range from news photography agencies (*Anefo*) and government documentation services (*Rijksvoorlichtingsdienst*) to specialized archives such as the KNVB football collection. To ensure reliable statistical estimates, we applied a minimum threshold of 1,000 items with subject metadata per subcollection. This filtering process yielded our final dataset of 878,046 photographs spanning 16 archival collections. These collections exhibit substantial variation in size, thematic focus, and curatorial origin, as detailed in Table 1. The GTAA terms analyzed in this study were applied by Dutch National Archives catalogers after collections were acquired, rather than by the originating institutions. The patterns we observe may reflect the National Archive's decisions about how to catalog different collection types, consisting of diverse historical materials.[5]

connecting structure. For consistency between methods, both the Unseen PD and our approach rely on the traditional calculation method rather than this modern implementation.`https://scikit.bio/docs/dev/generated/skbio.diversity.alpha.faith_pd.html`

[4] `https://www.nationaalarchief.nl/onderzoeken/fotos`

[5] To ensure transparency and reproducibility, we are sharing the complete dataset used in this study. The dataset includes metadata terms from the digitized photograph collections of the Dutch National Archives, as well as the GTAA ontology CSV file. You can access the code and dataset here: `https://github.com/melvinwevers/CHR_GTAA_Diversity`

Collection	Focus	Period	Total Items
Spaarnestad Onderwerpen	Publisher	1918-1997	397,084
Anefo	News Agency	1903-1989	376,056
Elsevier	Publisher	1752-1995	42,415
Dienst voor Legercontacten Indonesië	Military	1944-1974	33,085
Rijksvoorlichtingsdienst Eigen	Government	1688-1989	33,085
Van de Poll	Individual Photographer	1925-1967	32,834
Nederlandse Heidemaatschappij	Government	1923-1984	30,171
RVD / Koninklijk Huis	Government	1880-2002	7,346
Kantoor voor Voorlichting en Radio Omroep Nederlands Nieuw Guinea	Government	1954-1962	4,558
Arbeidsinspectie	Government	1913-1953	4,476
Eerste Wereldoorlog	Historical Archive	1914-1919	3,303
Deli Maatschappij	Private Company	1897-1958	2,023
Anefo / Londen	News Agency	1937-1945	1,637
KNVB Fotocollectie	Sports Archive	1894-1940	1,455
558 Ph. C. Visser	Individual Photographer	1911-1935	1,383
Anefo / RVD Londen Positieven	News Agency	1940-1957	1,015

Table 1: Overview of analyzed archival collections showing focus, time period the collection covers, and total items with subject metadata.

4 Results

4.1 Coverage and Completeness Patterns

The research examined how 16 different photo collections at the Dutch National Archives applied the GTAA to describe their materials, calculating the coverage, completeness, and cataloging intensity ratio.

Coverage Ratio: Collections explore limited conceptual territory Each collection captures only a small fraction of the total conceptual diversity possible within GTAA, ranging from 0.1% to 62.2% of the full conceptual space. Most collections concentrate on relatively narrow conceptual domains rather than spanning the broad range of concepts available in the vocabulary.[6] This reflects how specialized collections naturally focus on specific subject areas rather than attempting to cover all possible conceptual territory.

Completeness Ratio: Collections thoroughly describe their chosen conceptual areas Collections consistently capture 67.9% to 100% of the possible conceptual diversity within their chosen domains. This means that when a collection focuses on a particular subject area, it tends to comprehensively explore the full range of conceptual nuances and variations available for that topic, rather than leaving gaps or blind spots. The KNVB Fotocollectie exemplifies this coverage-completeness contrast: despite achieving only 0.1% coverage by focusing exclusively on sports-related vocabulary, it achieves 100% completeness within the sports hierarchy.

Cataloging Strategy: Selective domain choice with systematic term application These patterns reveal how the Dutch National Archives systematically describe newly acquired collections using the GTAA vocabulary system. The wide variation in coverage (0.1-62.2%) reflects catalogers making decisions about which conceptual domains from GTAA best capture each collection's subject matter. A sports photography collection receives sports-focused vocabulary application, while a news agency photographic collection gets described using political, cultural, and societal concepts. The remarkably consistent completeness rates (67.9-100%) demonstrate that once catalogers identify the GTAA domains for a collection, they systematically apply nearly all relevant concepts within those domains. This thoroughness ensures that collections are comprehensively described within their identified subject areas. Such structured approaches reflect deliberate cataloging decisions about scope and depth for different collection types that influence how cultural materials become discoverable and accessible to researchers and the public. Six collections achieving greater than 90% completeness. The distribution suggests modest improvement potential across collections: three could potentially add 5-10% additional terms, three demonstrate near-perfect utilization (>95%), and one shows potential for 10-15% enhancement. However, these apparent "improvement opportunities" require careful interpretation. As noted in our methodological constraints, lower completeness scores may reflect either incomplete term application or inadequate GTAA coverage of specific collection subject matter. Our analysis cannot distinguish between these scenarios, making direct recommendations for improvement problematic without additional investigation.

4.2 Cataloging Practice Typology and Collection Archetypes

The distribution of individual collections and their coverage and completeness ratios reveals four distinct collection archetypes, each reflecting different cataloging approaches for varying subject

[6] The distribution has a pronounced right skewness (1.47), indicating that most collection fall within the left part of the range

matter and collection characteristics. Using median thresholds for coverage (0.059) and completeness (0.838), we can classify collections according to their positioning within this two-dimensional space.[7]

Archetype	Coverage	Completeness	n	Representative Collections
Comprehensive	Above Median	Above Median	7	Anefo (0.621, 0.971), Rijksvoorlichtingsdienst Eigen (0.353, 0.909), Elsevier (0.348, 0.938)
Broad but Selective	Above Median	Below Median	1	Dienst voor Legercontacten Indonesië (0.104, 0.832)
Focused and Thorough	Below Median	Above Median	1	KNVB Fotocollectie (0.001, 1.000)
Limited Scope	Below Median	Below Median	7	Eerste Wereldoorlog (0.052, 0.792), Collectie 558 Ph. C. Visser (0.036, 0.719), Deli Maatschappij (0.024, 0.798), and 4 others

Table 2: Collection Performance Archetypes

Comprehensive collections achieve both extensive coverage and high completeness, representing collections with a diverse scope and comprehensive utilization of the GTAA vocabulary. *Anefo* exemplifies this archetype with 62.1% coverage and 97.1% completeness across its 346 thousand images. As the former General Dutch Photo Bureau documenting newsworthy events from 1945 to 1989, it is not surprising that *Anefo* exhibited broad terminological coverage. This pattern typically characterizes major news agencies and national documentation services. *Broad but selective* collections maintain above-median coverage while exhibiting completeness levels just below the median threshold. The *Dienst voor Legercontacten Indonesië* exemplifies this pattern with 10.4% coverage but 83.2% completeness. Produced by the Dutch military during the Indonesian War of Independence (1945-1949), the collection contains diverse geographical locations, political events, and scenes. Catalogers applied terminology selectively rather than exhaustively, reflecting the challenge of comprehensively cataloging broad collections that span multiple subject domains. *Focused and thorough* collections achieve near-perfect completeness within carefully defined domains. The *KNVB fotocollectie* demonstrates this approach through systematic documentation of 1,300 photographs of the Dutch national football team from 1894-1946. Originally compiled from photobooks with precise match metadata, this collection's bounded subject matter enables exhaustive term application within a focused branch of the hierarchy. *Limited scope* collections exhibit both restricted coverage and lower completeness, potentially reflecting either subject matter poorly aligned with GTAA's representational framework or inadequate use of available terms to the materials. The *Deli Maatschappij* collection, containing photographs from a tobacco company operating in the former Dutch East Indies, illustrates this challenge where corporate and colonial contexts may require specialized vocabularies beyond GTAA's standardized terminology.

[7] We use the median because the data is skewed.

4.3 Archetype Visualization and Size Effects

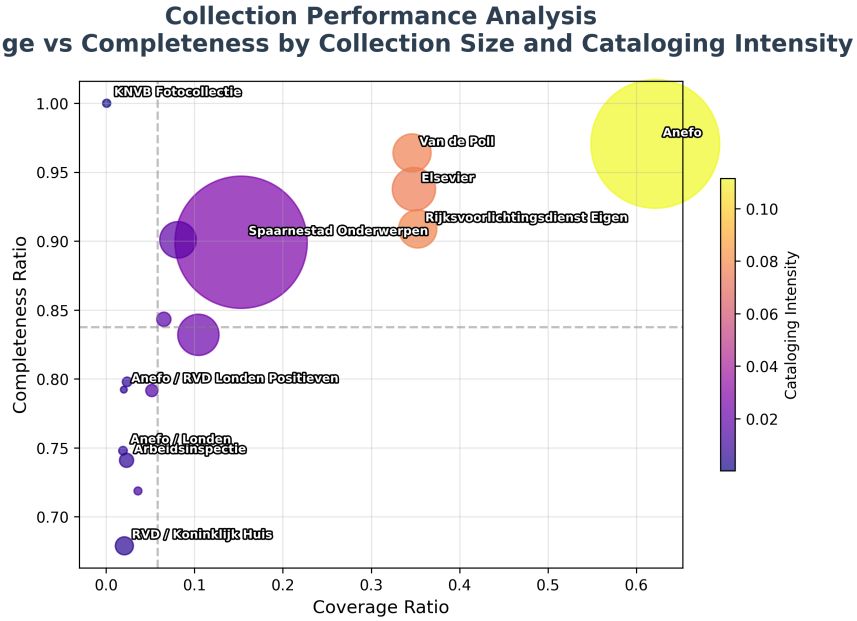

Figure 2: Coverage versus Completeness Ratios across Dutch National Archives Collections. Bubble size represents collection size, color intensity indicates cataloging cataloging intensity ratio. The median thresholds (coverage = 0.059, completeness = 0.838) are shown as dashed lines, delineating the four cataloging practice archetypes identified in Table 2.

Figure 2 visualizes the archetypes presented in Table 2 across the coverage-completeness parameter space, revealing three patterns. First, collections show pronounced concentration within the high-completeness ranges (> 0.85), confirming the consistency of thorough term application across diverse institutional contexts. Second, coverage varies dramatically, spanning 2.5 orders of magnitude from highly specialized to broadly comprehensive approaches. Third, observable relationships emerge between collection size (bubble size) and cataloging intensity (color intensity), suggesting that institutional size influences but does not determine cataloging coverage.

The cataloging intensity ratio reveals significant variation in how much conceptual work institutions invest per archival item. Cataloging intensity varies substantially across collections (range = 0.0002 to 0.1115, $\sigma = 0.0344$) with a right-skewed distribution (1.22), indicating that most collections receive modest cataloging investment while a few demonstrate exceptional descriptive depth per item. Larger collections generally achieve broader terminological coverage, as confirmed by significant correlations through both parametric (Pearson $r = 0.584$, $p = 0.018$) and non-parametric (Spearman $\rho = 0.844$, $p < 0.000$) measures. However, scale alone does not determine cataloging intensity. This pattern becomes clear when comparing collections of different sizes: while Anefo achieves exceptional intensity through substantial conceptual investment per item, the similarly large Spaarnestad collection demonstrates much lower coverage and intensity. Conversely, much smaller collections like Van de Poll, Elsevier, and Rijksvoorlichtingsdienst achieve coverage levels well below Anefo's but maintain intensity levels only slightly lower, indicating intensive cataloging work that maximizes conceptual representation per item.

This variation illuminates how subject matter characteristics interact with institutional cataloging investment. Collections whose content aligns well with GTAA's vocabulary structure can achieve high conceptual coverage with intensive cataloging work regardless of size, while collections covering topics poorly represented in GTAA's framework may require exceptionally high

cataloging intensity to achieve broad conceptual representation. This suggests that cataloging intensity patterns reflect both the match between collection content and vocabulary design, and institutional decisions about descriptive resource allocations that directly impact material discoverability.

5 Discussion

5.1 Understanding Vocabulary Patterns

The consistency in completeness ratios (mean = 84.5%, $\sigma = 0.098$) across collections with markedly different coverage levels suggests patterns in cataloging practices. This finding aligns with expectations of consistent cataloging standards applied by the Dutch National Archives, though structural constraints within GTAA itself may also contribute to this consistency. If cataloging quality varied significantly across collections, we would expect greater scatter in completeness measures. The observed clustering suggests that either cataloging practices are remarkably consistent across collections, or vocabulary-content alignment constraints dominate individual cataloging variations, or both factors operate simultaneously.

These patterns reveal variation in cataloging approaches to vocabulary breadth for different collection types, with potential implications for material discoverability. Collections with limited terminological coverage may create discovery gaps, while comprehensive vocabulary application enhances material visibility. Both approaches operate at what Trouillot identifies as the retrieval mechanisms juncture, where cataloging decisions directly influence user access to cultural heritage materials. The relationship between collection scale and cataloging intensity adds further nuance to these dynamics. While larger collections exhibit greater coverage ($r = 0.587$), cataloging intensity ratios reveal that strategic subject curation can substitute for smaller scale. Several mid-sized collections achieve cataloging intensity ratios comparable to much larger ones, indicating that collection size alone does not determine cataloging intensity and suggesting that other factors, potentially including subject matter characteristics and vocabulary-content alignment, may play important roles.

5.2 Implications for Research Discovery and Access

The observed pattern of deep but narrow GTAA vocabulary application reveals a fundamental tension in archival cataloging between cataloging intensity and comprehensive access. While this specialized approach ensures thorough description within identified subject domains, it creates significant limitations for interdisciplinary research and cross-domain discovery.

When catalogers focus primarily on the subject matter most fitting to the collection, they run the risk of systematically missing important themes embedded within the same materials. Two distinct patterns in our data illustrate different manifestations of this problem. The *Broad but Selective* archetype demonstrates high coverage with low completeness, exemplified by the *Dienst voor Legercontacten Indonesië collection*. This collection spans multiple conceptual domains but remains incompletely described within those domains, indicating substantial unseen diversity that points to archival silence. The second archetype *Limted Scope* shows low coverage with low completeness, representing collections with limited conceptual scope that are also inadequately described within their narrow focus. This contrasts with a collection such as the *KNVB Fotocollectie*, which also demonstrates low coverage but achieves high completeness within its sports-focused domain. The second archetype indicates inadequate attention to collections that may not align well with institutional cataloging priorities (utilization bias) or vocabulary strengths (vocabulary bias).

This cataloging approach in these two archetypes reinforces existing conceptual silos and limits opportunities for serendipitous discovery. Photographs often capture multiple layers of social, eco-

nomic, and cultural information that transcend their apparent primary subject matter, yet current cataloging practices may not make this richness discoverable across disciplinary boundaries.

The cataloging intensity patterns revealed in our analysis connect directly to Trouillot's concept of retrieval mechanisms as sites where archival silences are produced. Collections receiving high cataloging intensity—such as the KNVB Fotocollectie with its systematic application of sports terminology—become highly discoverable within their conceptual domains. Conversely, collections with low intensity may contain rich materials that remain effectively invisible due to sparse descriptive investment. This dynamic illustrates how institutional resource allocation decisions about cataloging depth can amplify or diminish a collection's research visibility regardless of its actual content richness, creating differential access patterns that extend beyond the initial moment of classification.

5.3 Bridging Computation and Critical Theory

The three diagnostic ratios introduced in this study align with central concerns in archival theory, offering quantitative perspectives on traditionally qualitative concepts.*Coverage* reflects the epistemic scope of a collection's representation—low coverage ratios may indicate domains that remain archivally invisible, connecting to Trouillot's structural silences at the retrieval level. *Completeness* gauges representational thoroughness within chosen domains, distinguishing between incomplete cataloging (where relevant terms exist but were not applied) and structural vocabulary gaps (where appropriate terms don't exist in GTAA). *Cataloging Intensity* reveals institutional investment patterns in descriptive work, connecting to archival silence theory by showing how resource allocation decisions create differential pathways to discovery. High intensity collections receive substantial conceptual investment per item, potentially mitigating archival silences through enhanced discoverability. Low intensity may indicate under-resourced cataloging that leaves materials conceptually impoverished, creating retrieval barriers that function as structural silences regardless of content quality.

5.4 Future Research Directions

Three research directions could advance this framework beyond its current scope. *Content-vocabulary alignment studies* would combine phylogenetic diversity analysis with multimodal content analysis to distinguish between incomplete term application and vocabulary inadequacy. This approach would move beyond vocabulary-bounded analysis to examine actual relationships between collection content and available terminology, providing a ground truth needed to interpret utilization patterns. *Cross-ontology comparative analysis* could systematically compare different vocabularies within specific domains, informing vocabulary selection strategies for cultural heritage institutions. Such studies would reveal whether observed patterns reflect characteristics of particular vocabulary systems or represent broader institutional cataloging behaviors. The framework also opens possibilities for *cross-domain applications* to other hierarchically organized data. For example, medieval manuscript studies could apply phylogenetic diversity to analyze stemmatic relationships and textual transmission patterns. Likewise, research could assess music catalogues' coverage across genre taxonomies. These applications would test the framework's adaptability beyond archival contexts and potentially reveal universal patterns in how institutions organize and represent knowledge. Additionally, *multi-faceted descriptive analysis* could extend this framework to evaluate how subject indexing interacts with other descriptive dimensions such as genre, creator, and place vocabularies. Such analysis would need to account for different relationship types beyond hierarchical structures to measure descriptive completeness across interconnected vocabularies and authority records. This expanded framework would more accurately reflect the complex descriptive ecosystems of modern archival practices.

The four cataloging archetypes identified in this study (Comprehensive, Broad but Selective, Focused and Thorough, and Limited Scope) provide humanities researchers and archivists with empirical benchmarks for evidence-based collection assessment and metadata evaluation. This framework quantifies how the interaction between collection content characteristics and institutional cataloging practices creates different pathways for cultural heritage discovery, revealing substantial variation in both the scope of conceptual domains (coverage ratio) addressed and the thoroughness (completeness ratio) of description within those domains. However, this approach also illuminates significant limitations in current cataloging practices. The focus on primary subject matter may create blind spots that limit serendipitous discovery and cross-domain research potential. Materials containing rich secondary themes may remain invisible to researchers working outside the primary cataloging domain, reinforcing disciplinary silos in cultural heritage access. Understanding these dynamics advances our knowledge of how metadata decisions influence cultural heritage accessibility while highlighting the tension between cataloging intensity and comprehensive research utility. Future work that combines phylogenetic diversity analysis with multimodal content analysis could distinguish between incomplete term application and vocabulary inadequacy, moving beyond vocabulary-bounded analysis to examine actual relationships between collection content and available terminology.

5.5 Acknowledgments

We thank the Dutch National Archives for providing access to their digitized collections and metadata. We also acknowledge the developers of the GTAA vocabulary for creating the hierarchical structure that made this analysis possible. We thank Leon van Wissen for his help with SPARQL.

References

[1] Caswell, Michelle, Punzalan, Ricardo, and Sangwand, T.-Kay. “Critical Archival Studies: An Introduction”. In: *Journal of Critical Library and Information Studies 1* (2 2017).

[2] Chao, Anne. “Estimating the Population Size for Capture-Recapture Data with Unequal Catchability”. In: *Biometrics 43*, no. 4 (1987), pp. 783–791.

[3] Chao, Anne, Chiu, Chun-Huo, Colwell, Robert K., Magnago, Luiz Fernando S., Chazdon, Robin L., and Gotelli, Nicholas J. “Deciphering the Enigma of Undetected Species, Phylogenetic, and Functional Diversity Based on Good-Turing Theory”. In: *Ecology 98*, no. 11 (2017), pp. 2914–2929.

[4] Chiu, Chun-Huo, Wang, Yi-Ting, Walther, Bruno A., and Chao, Anne. “An Improved Nonparametric Lower Bound of Species Richness via a Modified Good–Turing Frequency Formula”. In: *Biometrics 70*, no. 3 (2014), pp. 671–682.

[5] Derrida, Jacques. *Archive Fever: A Freudian Impression*. Trans. by Eric Prenowitz. Chicago: University of Chicago Press, 1996.

[6] Faith, Daniel P. “Conservation Evaluation and Phylogenetic Diversity”. In: *Biological Conservation 61*, no. 1 (1992), pp. 1–10.

[7] Huang, Y. Yuexin, Yu, S. Suihuai, Chu, J. Jianjie, Fan, H. Hao, and Du, B. Bin. “Using Knowledge Graphs and Deep Learning Algorithms to Enhance Digital Cultural Heritage Management”. In: *Heritage Science 11*, no. 1 (2023), p. 204.

[8] Karsdorp, F. B., Manjavacas, Enrique, and Fonteyn, Lauren. “Introducing Functional Diversity: A Novel Approach to Lexical Diversity in (Historical) Corpora”. In: *Proceedings of the Computational Humanities Research Conference 2022 1613* (2022), p. 0073.

[9] Karsdorp, Folgert, Kestemont, Mike, and de Koster, Margo. “Beyond the Register: Demographic Modeling of Arrest Patterns in 1879-1880 Brussels”. In: (2024).

[10] Kestemont, Mike, Karsdorp, Folgert, de Bruijn, Elisabeth, Driscoll, Matthew, Kapitan, Katarzyna A., Ó Macháin, Pádraig, Sawyer, Daniel, Sleiderink, Remco, and Chao, Anne. “Forgotten Books: The Application of Unseen Species Models to the Survival of Culture”. In: *Science (New York, N.Y.) 375*, no. 6582 (2022), pp. 765–769.

[11] Klein, Lauren F., Eisenstein, Jacob, and Sun, Iris. “Exploratory Thematic Analysis for Digitized Archival Collections”. In: *Digital Scholarship in the Humanities 30* (suppl_1 2015), pp. i130–i141.

[12] Knowlton, Steven A. “Three Decades Since Prejudices and Antipathies: A Study of Changes in the Library of Congress Subject Headings”. In: *Cataloging & Classification Quarterly 40*, no. 2 (2005), pp. 123–145.

[13] Koeser, Rebecca Sutton and LeBlanc, Zoe. “Missing Data, Speculative Reading”. In: *Journal of Cultural Analytics 9*, no. 2 (2024).

[14] Langer, Lars, Burghardt, Manuel, Borgards, Roland, Böhning-Gaese, Katrin, Seppelt, Ralf, and Wirth, Christian. “The Rise and Fall of Biodiversity in Literature: A Comprehensive Quantification of Historical Changes in the Use of Vernacular Labels for Biological Taxa in Western Creative Literature”. In: *People and Nature 3*, no. 5 (2021), pp. 1093–1109.

[15] Quinn, Frank. “Measuring Diversity of Opinion in Public Library Collections”. In: *The Library Quarterly 65*, no. 1 (1995), pp. 1–38.

[16] Topaz, Chad M., Klingenberg, Bernhard, Turek, Daniel, Heggeseth, Brianna, Harris, Pamela E., Blackwood, Julie C., Chavoya, C. Ondine, Nelson, Steven, and Murphy, Kevin M. “Diversity of Artists in Major U.S. Museums”. In: *PLOS ONE 14*, no. 3 (2019), e0212852.

[17] Trouillot, Michel-Rolph and Carby, Hazel V. *Silencing the Past: Power and the Production of History*. Boston, Massachusetts: Beacon Press, 2015.

[18] Wevers, Melvin, Karsdorp, Folgert, and van Lottum, Jelle. “What Shall We Do With the Unseen Sailor? Estimating the Size of the Dutch East India Company Using an Unseen Species Model”. In: *Proceedings of the Computational Humanities Research Conference 2022 1613* (2022), p. 0073.

[19] Wulf, Karin. “Archival Shouting – AHA”. Perspectives on History. Apr. 10, 2024.

[20] Yang, Le, Wei, Fuyi, and Chen, Enci. “Developing an Assessment Index for Collection-User Suitability: Application of Information Entropy in Library Science”. In: *The Journal of Academic Librarianship 48*, no. 1 (2022), p. 102477.