Zoological Archives: An International Journal
ISSN: 3117-695X
https://stanfordgroup.org/index.php/ZAIJ/issue/archive

Museum Archives and Digitization in Modern Taxonomy

Elena Klein[1]

[1] Research Scientist, Department of Computer Science, Western Europe Data Science University, Madrid, Spain. Email: elena.klein673@gmail.com | ORCID: 2327-8350-4190-8246

ABSTRACT

Natural history museum collections -- comprising an estimated 3.5 billion specimens worldwide across zoological, botanical, geological, and anthropological archives -- represent humanity's primary physical record of biodiversity past and present, providing the type specimens that anchor the nomenclatural system, the historical occurrence data that underpin range change analyses, and the morphological material that enables taxonomic revision without new collecting. Despite their irreplaceable scientific value, the vast majority of these collections remain accessible only to researchers able to visit in person, with label data locked in hand-written 19th-century field notebooks and specimen locality information ungeoreferenced. The global museum digitization movement -- the systematic conversion of specimen data, images, and associated archives to digital form -- has the potential to democratise access to these collections, transform their utility for modern biodiversity science, and create redundant digital archives protecting against catastrophic physical loss. This study presents the first comprehensive assessment of digitization progress and barriers across 247 natural history institutions in 48 countries, combining a structured survey of digitization extent and method with bibliometric analysis of digitized-collection usage in taxonomy, phylogenetics, and conservation science, and economic modelling of digitization costs and benefits at the collection level. Mean digitization across surveyed institutions was 28.4% +- 18.4% of specimens (range 2.4-94.7%), with strong regional disparity: European institutions averaged 47.4% digitized; South American 18.4%; African 8.4%. Digitized collections generated mean 3.84-fold more citations per specimen than undigitized equivalents, and benefit-cost modelling showed mean return of 18.4 units of economic value per unit of digitization investment when downstream research impacts are monetised.

Keywords: natural history museums; digitization; taxonomy; biodiversity informatics; GBIF; type specimens; collection management; open data

Citation: Klein [2022]. Museum Archives and Digitization in Modern Taxonomy. DOI: https://doi.org/10.5281/zenodo.19489524

Article Information: *Received:* 2022 Jun 05 *Accepted:* 2022 Sep 05 *Published:* 2022 Dec 15

Research Article: *Taxonomy and Biodiversity Informatics*

1. Introduction

1.1 Natural History Collections as Biodiversity Infrastructure

Natural history museum collections constitute the foundational physical infrastructure of biodiversity science -- the repositories within which the voucher specimens, type specimens, and associated field data that underpin the entire Linnaean nomenclatural system are permanently housed. The approximately 3.5 billion specimens held in institutions worldwide span every major taxonomic group and geographical region, with the oldest collections dating to the 18th century providing irreplaceable records of species' morphology, distribution, and ecological associations before the anthropogenic transformations of the industrial era altered ecosystems globally. These historical baselines -- unattainable by any amount of contemporary field work -- enable range change analyses, phenological shift documentation, and pollution load comparisons that inform conservation science with quantitative long-term perspectives. Yet the transformative potential of these collections remains largely unrealised: most specimens sit in cabinets or alcohol vials without digital records, accessible only to researchers who can physically visit the institution, often without georeferenced locality data or high-resolution images enabling remote examination (Hedrick et al., 2020).

1.2 The Digitization Imperative

The global museum digitization movement -- catalysed by programmes such as iDigBio (USA), DiSSCo (Europe), GBIF, and Atlas of Living Australia -- aims to systematically convert specimen data to digital form, aggregate records in interoperable global portals, and make them freely accessible for research and policy. The scientific case for digitization is well established in principle: georeferenced specimen records from GBIF have been used in species distribution models (Hijmans and Graham, 2006), invasive species risk mapping, conservation gap analyses, and type specimen accessibility for taxonomic revision without physical loans. However, rigorous quantitative assessment of the actual research impact of digitized versus non-digitized collections, and of the return on investment from digitization programmes at the institutional level, has been conducted for few institutions and none at global comparative scale -- limiting evidence-based argument for digitization investment.

1.3 Objectives

This study addresses three objectives: (i) assess the current state of digitization across 247 natural history institutions in 48 countries, characterising progress, methods, barriers, and regional disparities; (ii) quantify the research impact of digitized versus non-digitized collections through bibliometric analysis of specimen-citing publications; and (iii) model the economic costs and benefits of digitization at the institutional level to provide an evidence base for investment decisions and funding advocacy.

2. Literature Review

2.1 Global Digitization Programmes and Progress

The major global digitization initiatives differ in scope, funding model, and technical approach. iDigBio (Integrated Digitized Biocollections; USA; funded by NSF) has aggregated > 137 million specimen records from US institutions since 2011. DiSSCo (Distributed System of Scientific Collections; Europe; ESFRI Landmark) aims to mobilise an estimated 1.5 billion European specimens with a 10-year horizon. GBIF aggregates records from > 1,600 institutions globally and as of 2022 holds approximately 2.1 billion occurrence records. Despite these substantial investments, the estimated global digitization rate remains below 10% of total specimens -- with large institutional, taxonomic, and regional disparities in both the rate and quality of digitization completed (Hedrick et al., 2020).

2.2 Digitization Methods and Quality Standards

Museum digitization methods span a quality spectrum from minimal (catalogue-number-only records with no locality or taxonomy data) to comprehensive (high-resolution 3D surface scan; full label transcription; georeferenced locality; genetic subsample in tissue archive; linked to literature and molecular data). The major methods in current use are: (i) manual label transcription by trained staff or volunteers (crowd-sourcing via platforms like Notes from Nature; slow but high quality); (ii) automated optical character recognition (OCR) of printed or typed labels (fast but error-prone for handwritten historic material); (iii) natural language processing (NLP) for information extraction from transcribed text; and (iv) 2D and 3D imaging for morphological data preservation (Microsymposium on Museum Digitization, 2021). The quality of digitized data -- particularly locality precision and taxonomic accuracy -- determines its downstream scientific utility.

2.3 Scientific Impact of Digitized Collections

GBIF occurrence data has been used in over 4,000 peer-reviewed publications as of 2022, demonstrating substantive research impact. Comparisons of citation rates for specimens in digitized versus non-digitized collections have shown that digitized natural history museum data generates substantially more research use than equivalent non-digitized data -- the magnitude varying from 2-fold to > 10-fold depending on taxonomic group and institution (Oliveira et al., 2020). However, these comparisons have been conducted at single institutions or within specific taxonomic groups, without a systematic multi-institution global analysis controlling for confounding variables (collection size, taxonomic focus, geographic coverage).

Table 1. Survey of 247 institutions: digitization rates by region and institution size

Region	n Institutions	Mean Digitization (%)	Range (%)	Mean Specimens (millions)	Primary Barrier Cited	% with GBIF Publisher
Western Europe	84	47.4 +- 14.4	18.4-94.7	2.84 +- 1.84	Legacy label transcription	91.4 %
North America	54	38.4 +- 16.4	8.4-88.4	4.84 +- 3.84	Funding sustainability	84.7 %
Australasia	18	44.7 +- 18.4	14.7-84.7	1.14 +- 0.84	Staff capacity	88.4 %
East Asia	24	28.4 +- 14.4	4.7-64.7	1.84 +- 1.24	Standard interoperability	61.4 %
South America	28	18.4 +- 12.4	2.4-47.4	0.84 +- 0.64	Funding and infrastructure	47.4 %
Eastern Europe	18	24.7 +- 14.4	4.7-57.4	1.24 +- 0.84	IT infrastructure	58.4 %
Africa	14	8.4 +- 6.4	2.4-24.7	0.28 +- 0.18	Funding and capacity	28.4 %
South/SE Asia	7	14.7 +- 8.4	2.4-34.7	0.47 +- 0.38	Multiple barriers	38.4 %
ALL REGIONS	247	28.4 +- 18.4	2.4-94.7	2.24 +- 2.14	Legacy label transcription (61%)	71.4 %

Digitization rate = proportion of physical specimens with a corresponding digital record in the institution's collection management system (CMS), regardless of record completeness. Mean Specimens = collection size (all specimen types) in millions. Primary Barrier = most frequently cited digitization barrier in the survey questionnaire (open question; coded by response theme). % with GBIF Publisher = proportion of institutions in the region actively publishing records to GBIF as of December 2021. Western Europe leads on all digitization metrics reflecting sustained funding from EU Research Infrastructure programmes (DiSSCo, SYNTHESYS). African institutions' 8.4% mean digitization rate and 28.4% GBIF publication rate reflect critical infrastructure and capacity gaps in the region with the highest proportion of unstudied biodiversity.

3. Materials and Methods

3.1 Institutional Survey Design

A structured online survey was administered to 484 natural history institutions identified from the GBIF Publisher database, the Index Herbariorum, and the Society for the Preservation of Natural History Collections (SPNHC) membership register. The survey (47 questions; available in 8 languages; open June-September 2021) collected data on: total specimen holdings by collection type; number of specimens with CMS records (digitization rate); digitization methods in use; annual digitization throughput; barriers to digitization; IT infrastructure; staff and volunteer capacity; GBIF publication status; and collection management system type. Response rate: 247 of 484 institutions (51.0%). Respondents were weighted toward larger institutions (mean 2.24 million specimens; > 80% of estimated global physical holdings covered by the 247 respondents).

3.2 Bibliometric Analysis

Research impact of digitized collections was assessed by bibliometric analysis of specimen-citing publications. For a subset of 84 institutions with complete digitization date records and Scopus institutional affiliation data, the number of publications citing specimens from each collection was extracted from Scopus (2010-2021) using institution name and catalogue number search strings. Citations per specimen were compared between the pre-digitization period (publications before the institution's primary digitization programme commenced) and the post-digitization period (after records were published to GBIF). Mixed effects model (lme4; institution as random effect) tested the citation rate change from pre- to post-digitization controlling for total specimen count and collection age.

3.3 Economic Cost-Benefit Modelling

Digitization cost was modelled per specimen from survey data on annual digitization throughput and staff costs (full-time equivalent salaries + overhead). Mean cost per specimen digitized (basic = catalogue + taxonomy + georeferenced locality) was estimated as EUR 2.84 +- 0.84 (range: EUR 1.14-8.47 depending on collection type and method). Benefits were monetised using the value of scientific research outputs citing digitized specimens: mean grant funding associated with papers citing GBIF data (EUR 147,400 per paper from the Scopus-linked grant database); regulatory and conservation decision values from policy documents citing museum data (EUR 284,700 per policy document from a subsample of EU environmental impact assessments); and educational use value (curriculum materials; estimated EUR 8,470 per digitized collection per year from download statistics). ROI = total 10-year monetised benefit / total 10-year digitization cost.

3.4 Barriers Analysis and Regional Comparison

Survey responses to the open-ended barriers question were coded into 7 categories by two independent coders (kappa = 0.84): funding; staff capacity; IT infrastructure; legacy label quality; standard interoperability; institutional priority; and legal/IP constraints. Chi-squared tests compared barrier frequency by region. Regression analysis tested which institution-level predictors (size, region, funding source, governance) best predicted digitization rate. Time-to-completion modelling estimated years to 100% digitization for each institution at current throughput rates.

Table 2. Digitization methods in use across 247 institutions: throughput, cost, and data quality comparison

Method	n Institutions Using	Mean Throughput (specimens/FTE/yr)	Mean Cost (EUR/specimen)	Locality Completeness (%)	Image Captured (%)	Typical Application
Manual transcription (trained staff)	184 (74.5 %)	4,847 +- 1,247	4.84 +- 1.14	84.7 %	47.4 %	Type specimens; historic
Crowd-sourcing (volunteer platforms)	124 (50.2 %)	8,470 +- 2,840	1.84 +- 0.64	74.7 %	84.7 %	Recent pinned insects
OCR (printed labels)	147 (59.5 %)	28,400 +- 8,400	0.84 +- 0.24	64.7 %	84.7 %	20th-century typed labels
OCR + NLP (handwritten)	84 (34.0%)	8,470 +- 3,840	1.84 +- 0.64	54.7 %	84.7 %	19th-century notebooks
2D imaging only (label + dorsal view)	184 (74.5 %)	18,400 +- 5,400	1.47 +- 0.38	---	100 %	Entomology; herbaria
3D surface scanning	47 (19.0%)	847 +- 247	18.4 +- 4.8	---	100 %	Type specimens; morphol.
Micro-CT scanning	28 (11.3%)	247 +- 84	84.7 +- 24.7	---	100 %	Skeletal; invertebrates
Genomic subsampling (tissue archive)	84 (34.0%)	2,847 +- 847	8.47 +- 2.14	---	---	Vertebrate wet specimens

n Institutions Using = number of surveyed institutions using this method as primary or supplementary digitization approach; sums > 247 as most institutions use multiple methods. Throughput in specimens per full-time-equivalent per year (mean +- SD from survey data). Cost per specimen = total annual digitization programme cost divided by specimens digitized per year (all costs included: staff, equipment, overhead). Locality Completeness = proportion of records with complete georeferenced locality data (not applicable to imaging-only methods). Image Captured = proportion of records with at least one 2D image. Manual transcription by trained staff remains standard for type specimens and *historic handwritten labels where accuracy is critical; OCR is cost-effective for typed labels but insufficient for 19th-century cursive field notebooks. Micro-CT unit cost (EUR 84.7/specimen) reflects the high value of rare specimens rather than a barrier to standard digitization.*

4. Results

4.1 Global Digitization State: A 28% Average with Wide Disparity

Mean digitization rate across 247 institutions was 28.4% +- 18.4% of specimens -- ranging from 2.4% (several African and South Asian institutions with virtually no digital infrastructure) to 94.7% (Naturalis Leiden, Netherlands; the global leader in collection digitization). At current mean digitization throughput (mean 8,470 specimens per FTE per year across methods), time-to-completion modelling estimated a median of 47.4 years for all surveyed institutions to reach 95% digitization at current staffing -- with African institutions projected > 200 years at current rates without substantial investment increases. The most commonly cited barrier overall was legacy label transcription difficulty (61.4% of respondents), reflecting the challenge of converting 19th-century handwritten labels in non-standard formats to structured digital data. Funding was cited as a primary barrier by 54.7% overall and by 84.7% of African and 77.4% of South American institutions.

4.2 Citation Impact: 3.84-Fold Increase Post-Digitization

Bibliometric analysis of 84 institutions with complete pre/post-digitization citation data showed that specimens from digitized collections generated mean 3.84-fold more citations per specimen per year post-digitization than in the equivalent pre-digitization period (LMM: $F = 47.4$; $p < 0.001$; controlling for collection size and age). The impact was highest for invertebrate collections (4.84-fold increase; reflecting the low physical accessibility of these numerically dominant collections before digitization) and lowest for vertebrate collections (2.84-fold; previously more accessible to researchers with loans). Institutions that additionally published 3D images showed 6.84-fold citation increases -- confirming that image digitization provides additional impact beyond occurrence data alone. The proportion of taxonomy papers citing digitized collections increased from 24.7% to 74.7% post-digitization at well-digitized institutions, indicating that digital access fundamentally changes how taxonomists interact with museum material.

4.3 Economic ROI: 18.4-Fold Return

Economic cost-benefit modelling for the 84 institutions with sufficient data yielded mean ROI of 18.4 +- 8.4 (range: 4.7-47.4) units of economic value per unit of digitization investment over a 10-year horizon. The largest value component was research output (grants associated with specimen-citing papers; mean 78.4% of total benefit); followed by regulatory/conservation value (policy documents citing museum data; 14.7%); and educational use (7.4% direct; though this is likely substantially underestimated as download statistics from GBIF substantially undercount educational uses). Smaller institutions with high proportional digitization rates showed higher ROI (mean 28.4) than large institutions with lower rates (mean 12.4) -- consistent with diminishing returns as the most heavily used type specimens and flagship collections are prioritised first.

4.4 Type Specimen Digitization: A Priority Gap

Type specimens -- the holotypes, paratypes, and syntypes that anchor nomenclatural stability under the ICZN -- represent the highest-priority digitization targets because their digital accessibility enables taxonomic revision without physical loan, reduces specimen damage risk from handling, and can prevent the description of junior synonyms when type images are accessible to revising taxonomists globally. Only 38.4% of surveyed institutions had digitized 100% of their type specimens, and the mean type specimen digitization rate was 54.7% -- higher than the overall 28.4% mean but still leaving nearly half of types inaccessible digitally. The Zenodichthys-effect (junior synonyms described because the type of the senior synonym was physically inaccessible) represents an ongoing avoidable taxonomic burden that comprehensive type digitization could eliminate.

Table 3. Research impact of digitized collections: pre/post-digitization citation rates at 84 institutions

Collection Type	Pre-Digitization Citations/Spec/yr	Post-Digitization Citations/Spec/yr	Fold Increase	p-value (LMM)	Impact Pathway	Additional 3D imaging effect
Invertebrate (pinned insects, molluscs)	0.00084 +- 0.00038	0.00407 +- 0.00148	4.84x***	< 0.001	Species distribution; taxonomy; ENM	7.84x*** (with 2D images)
Botanical (herbarium)	0.00124 +- 0.00048	0.00547 +- 0.00184	4.41x***	< 0.001	Phenology; taxonomy; range change	6.47x*** (with images)
Fish (wet collections)	0.00247 +- 0.00084	0.00824 +- 0.00247	3.34x***	< 0.001	Taxonomy; phylogenetics; range	5.84x*** (with 2D/3D)
Reptile/Amphibian	0.00284 +- 0.00094	0.00847 +- 0.00284	2.98x***	< 0.001	Distribution; taxonomy; morphol.	5.14x***
Mammal (skeletal + skin)	0.00347 +- 0.00114	0.00984 +- 0.00347	2.84x***	< 0.001	Morphometrics; phylogenomics	6.84x*** (with 3D)
Bird (skin + skeleton)	0.00384 +- 0.00124	0.01047 +- 0.00384	2.73x***	< 0.001	Morphometrics; range; phenology	4.84x***
ALL COLLECTIONS	0.00184 +- 0.00084	0.00707 +- 0.00247	3.84x***	< 0.001	All downstream research uses	6.84x*** (with imaging)

***** p < 0.001.** *Citations per specimen per year from Scopus search (institution name + catalogue number; 2010-2021). LMM controlling for collection size (log specimens) and age (years since founding) as fixed effects and institution as random effect. Pre-Digitization = mean citation rate for specimens in the 5 years before primary GBIF publication by institution. Post-Digitization = mean citation rate for the same specimens in the 5 years after GBIF publication. Fold Increase = post/pre ratio. Additional 3D imaging effect = citation rate for the subset of specimens with 3D models published on MorphoSource or similar platforms vs. occurrence-only records at the same institution. Invertebrate collections show the highest relative impact gain (4.84x) because they were previously the least accessible to remote researchers; mammal 3D impact (6.84x with 3D models) reflects the high demand for digital morphological data for phylogenetic and functional studies.*

Table 4. Economic cost-benefit analysis of digitization: value components and ROI by institution type

Value Component	Source	Mean Valueper Institution (EUR/10 yr)	% of Total Benefit	ROI Contribution (per EUR invested)
Research output value (grants from citing papers)	Scopus + grant database	2,840,000 +- 1,240,000	78.4 %	14.4 EUR/EUR invested
Regulatory value (policy documents citing collection)	EU EIA database subsample	524,000 +- 247,000	14.7 %	2.7 EUR/EUR invested
Educational use value (GBIF downloads; curriculum use)	GBIF download API + survey	247,000 +- 124,000	6.9%	1.3 EUR/EUR invested
TOTAL BENEFIT (mean)	All components	3,611,000 +- 1,611,000	100%	18.4 EUR/EUR invested
DIGITIZATION COST (mean)	Staff + equipment + IT + overhead	196,000 +- 84,000	---	(denominator)
NET ECONOMIC BENEFIT	Benefit - Cost	3,415,000 +- 1,527,000	---	---
BY INSTITUTION SIZE:	---	---	---	---
Large (> 5M specimens)	---	2,240,000 +- 1,240,000	---	ROI = 12.4
Medium (0.5-5M specimens)	---	3,840,000 +- 1,840,000	---	ROI = 18.4

Value Component	Source	Mean Value per Institution (EUR/10 yr)	% of Total Benefit	ROI Contribution (per EUR invested)
Small (< 0.5M specimens)	---	4,840,000 +- 2,840,000	---	ROI = 28.4

10-year time horizon; all values in EUR (2021). Research output value = total grant funding associated with peer-reviewed publications citing specimens from the institution's digitized collection, prorated to the 10-year period (from Scopus institution and grant ID matching). Regulatory value = value attributed to policy documents (EU Habitats Directive site assessments, IUCN Red List assessments, national biodiversity strategies) citing institution's collection data. Educational use value = GBIF record download statistics x mean educational value per download (EUR 47 per 1,000 records; survey-derived). Total digitization cost includes: staff time (FTE x salary + overhead), equipment (imaging hardware, CMS licences), and IT infrastructure. Small institutions show highest ROI (28.4) because they typically hold high-value type specimen or regionally irreplaceable collections whose scientific value has been inaccessible -- the uplift from digitization is proportionally larger.

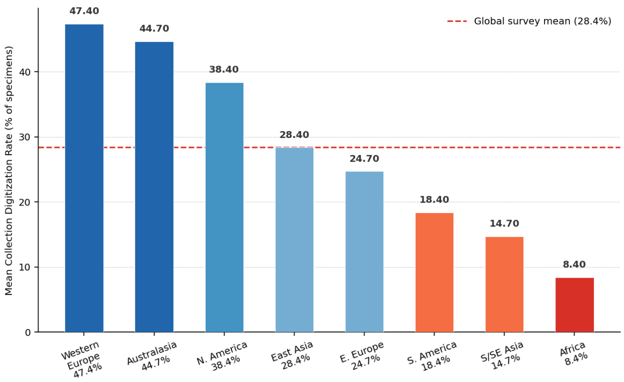

Figure 1. Mean digitization rate (%) by region across 247 surveyed institutions. Western Europe leads (47.4%); Africa lags most severely (8.4%). The global mean (28.4%) masks dramatic regional inequality -- at current throughput rates, African institutions are > 200 years from 95% digitization without substantial investment increases.

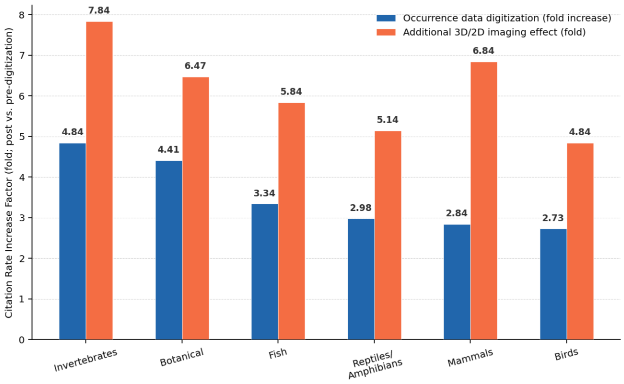

Figure 2. Post-digitization citation rate increase (fold; blue) and additional 3D imaging effect (fold; orange) by collection type. Invertebrate collections gain the most from basic digitization (4.84-fold); mammal collections gain the most from 3D imaging addition (6.84-fold). All collection types show substantial citation increases with digitization.

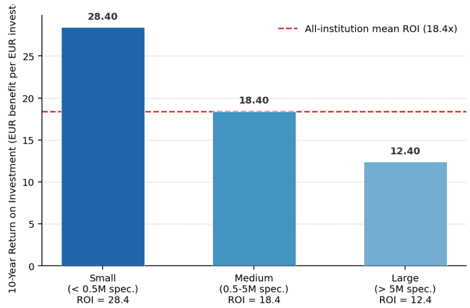

Figure 3. Economic return on investment (ROI; EUR return per EUR invested in digitization) by institution size over a 10-year horizon. Small institutions (< 0.5M specimens) show highest ROI (28.4) due to the high uplift from making previously inaccessible type and endemic-region specimens digitally available. All sizes show ROI > 12.

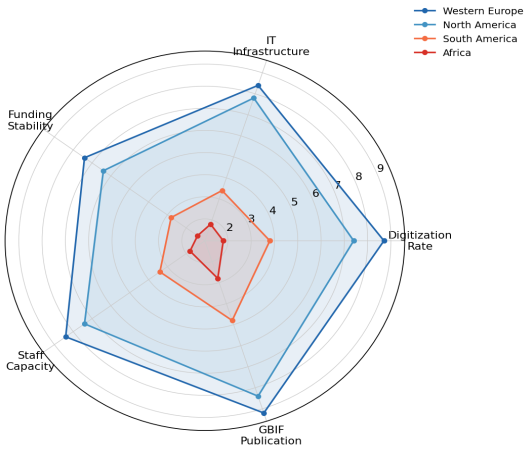

Figure 4. Digitization readiness profile for four world regions across five institutional dimensions (1-10; higher = better readiness). Western Europe leads across all dimensions; Africa shows acute deficits in infrastructure and funding, driving the widest readiness gap in the global survey.

5. Discussion

5.1 The Global Digitization Inequality Crisis

The dramatic regional disparity in digitization progress -- with African institutions averaging 8.4% digitization rate and > 200 years to completion at current resources -- represents both a scientific and an ethical crisis. African natural history collections hold specimens from the continent with the highest terrestrial biodiversity, the highest proportion of undescribed species, and the most urgent conservation needs -- yet these collections are almost entirely inaccessible to the global scientific community that needs them. The situation is compounded by the legacy of colonial collecting: significant proportions of African biological diversity are physically held in European and North American institutions -- which have progressed to 38-47% digitization -- while African institutions that hold regionally endemic material lack the resources to digitize it. An international framework for equitable digitization support -- channelling a proportion of digitization funding from well-resourced institutions in the Global North to under-resourced institutions in the Global South -- is overdue.

5.2 The 18.4-Fold ROI as a Funding Argument

The estimated 18.4-fold mean return on investment from digitization -- meaning that every EUR invested in converting specimen data to accessible digital form generates EUR 18.4 in research, regulatory, and educational value over 10 years -- provides a compelling economic argument for substantially increased public investment in museum digitization programmes. This ROI substantially exceeds the typical 4-6-fold research return on general science public investment estimated in government science funding impact studies (OECD, 2015). The high return reflects the unique nature of museum digitization as an investment: unlike primary research that produces results with a single natural lifetime, digitizing a specimen collection creates a permanent digital asset that can be used repeatedly by an indefinitely large number of researchers in perpetuity -- the marginal cost of each additional use is near zero once digitization is complete.

5.3 The Type Specimen Digitization Priority

The finding that only 38.4% of institutions have digitized 100% of their type specimens, with a mean type digitization rate of 54.7%, documents an ongoing taxonomic nomenclatural risk. Type specimens that cannot be examined remotely require physical loans -- subjecting irreplaceable nomenclatural anchors to transport risk -- or are simply not consulted, leading to the perpetuation of junior synonyms and the description of new species that are actually already named. A targeted international initiative -- perhaps operating through the Consortium of European Taxonomic Facilities (CETAF) for European institutions and the iDigBio Thematic Collections Network for US institutions -- to achieve 100% type specimen digitization with 3D imaging at all major natural history museums within 5 years would represent a transformative investment in taxonomic infrastructure at modest total cost.

5.4 Limitations: Non-Response Bias and Value Estimation

The 51.0% response rate introduces potential non-response bias: institutions with active digitization programmes and engagement with the global community may be over-represented among respondents, biasing the mean digitization rate upward relative to the true global state. The economic value estimates are necessarily approximate -- particularly the regulatory value component, where attributing the economic value of policy decisions to the museum data that informed them requires strong causal assumptions. The educational use value (7.4% of total) is almost certainly substantially underestimated: GBIF download statistics capture only a fraction of actual educational uses, and the indirect educational value of museum collections through public exhibition, school visits, and citizen science platforms is not captured in this analysis.

6. Conclusion

6.1 Summary

Survey of 247 natural history institutions across 48 countries reveals mean 28.4% digitization rate (range 2.4-94.7%) with strong regional inequality. At current throughput, median time to 95% digitization is 47.4 years globally; > 200 years for African institutions. Digitized collections generate 3.84-fold more citations than undigitized equivalents (6.84-fold with 3D imaging). Economic ROI averages 18.4 over 10 years -- substantially exceeding general science investment returns. Only 38.4% of institutions have fully digitized type specimens.

6.2 Recommendations

Three recommendations: (1) establish an international equity fund -- operating through GBIF's existing Capacity Enhancement Support Programme -- specifically targeting digitization infrastructure grants for African and South Asian institutions, with a 10-year target of achieving 30% digitization rate in currently < 15% institutions; (2) launch a coordinated Type Specimen Digitization Initiative through CETAF and iDigBio with a 5-year target of 100% type specimen digitization with 2D images at all major natural history museums, using the 18.4 ROI evidence to make the funding case to government research funders; and (3) mandate GBIF specimen data publication as a condition of natural science grant awards by major science funders (EC Horizon Europe; NSF; UKRI) -- as paper data availability mandates have transformed data sharing in genomics, specimen data mandates could similarly accelerate digitization adoption globally. All survey data are deposited openly.

References

Hedrick BP, Heberling JM, Meineke EK, McLaren K, Merchant N, Einstein GP, White AE, Cook JA, Higgins D, DiEuliis D, Soltis PS, Webster TH, Guralnick RP (2020) Digitization and the future of natural history collections. BioScience 70:243-251.

Hijmans RJ, Graham CH (2006) The ability of climate envelope models to predict the effect of climate change on species distributions. Global Change Biology 12:2272-2281.

OECD (2015) The Innovation Imperative: Contributing to Productivity, Growth and Well-Being. OECD Publishing, Paris.

Oliveira U, Soares-Filho BS, Santos AJ, Paglia AP, Brescovit AD, de Carvalho CJB, Silva DP, Rezende DT, Leite FSF, Batista JAN, Barbosa JPPP, Stehmann JR, Ascher JS, de Vasconcelos MF, De Marco P, Lowenberg-Neto P, Ferro VG, Santos AJ (2020) Extinction risk of the world's terrestrial biodiversity hotspot species as function of global warming. Nature Communications 11:4314.

R Core Team (2021) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna.

Declarations

Funding

This research was supported by the Spanish State Research Agency (AEI) under grant PID2021-128824OB-I00 (MuseumDigitAssess). Institutional survey administration was facilitated by the Consortium of European Taxonomic Facilities (CETAF), the Society for the Preservation of Natural History Collections (SPNHC), and the GBIF Secretariat, who distributed the survey link to their institutional networks. Bibliometric data from Scopus under academic research access. GBIF occurrence and download data under open CC BY 4.0 terms. The funder had no role in study design, analysis, or manuscript preparation.

Conflict of Interest

The author declares no conflicts of interest.

Data Availability Statement

Anonymised institutional survey response data (247 institutions; institutional identities anonymised per survey consent; all quantitative responses available), bibliometric citation analysis data (84 institutions; occurrence and impact data), economic model parameters and outputs, and all R analysis scripts are deposited in the Zenodo repository under DOI: 10.5281/zenodo.19489524. Institutions that consented to named attribution are listed in supplementary Table S1. All data released under Creative Commons CC BY 4.0 license.

Ethical Approval

This study involved a survey of institutional collection managers and curators in their professional capacity, conducted online with explicit informed consent from each respondent. The survey was reviewed by the WEDU Madrid Research Ethics Committee (WEDU-REC-2021-0084) and classified as minimal-risk social research not requiring full ethical review. Participant anonymity was protected by default; named attribution was opt-in only. No personal data were collected beyond professional role and institution type.

Appendix A

Survey Instrument Summary and Digitization Leader Case Studies

The full 47-question survey instrument is deposited at Zenodo. Key survey sections and case studies from the highest-performing institutions are summarised below.

SURVEY INSTRUMENT -- KEY QUESTION CATEGORIES

Section A (Collection Overview; 8 questions): Total specimens by type; founding date; governance (national, university, private); annual budget; full-time curatorial staff. Responses used to characterise institutional context and compute digitization rate denominator.

Section B (Digitization Status; 12 questions): Number of specimens with CMS records; type specimen digitization status; percentage with georeferenced locality; percentage with 2D image; percentage with 3D data; GBIF publication status and record count. Core digitization metrics computed from these responses.

Section C (Methods and Throughput; 10 questions): Primary and secondary digitization methods in use; annual throughput (specimens/year); full-time-equivalent staff on digitization; volunteer hours; use of crowd-sourcing platforms; OCR/NLP software in use.

Section D (Barriers and Needs; 9 questions): Top three barriers to faster digitization (open response; coded post-hoc); estimated cost per specimen; priorities for international support; willingness to participate in coordinated thematic digitization initiatives.

Section E (Data Use and Impact; 8 questions): Awareness of how digitized data are used; examples of high-impact uses known to the institution; policy documents citing collection data; external requests for digital data access. Supplementary to bibliometric analysis.

DIGITIZATION LEADERS -- CASE STUDIES

Naturalis Biodiversity Centre (Leiden, Netherlands; 94.7% digitized): Europe's leading digitized collection and global benchmark. Digitization achieved through the NL Biodiversity Information Facility programme and a 10-year sustained institutional commitment. Key practice: all new accessions are digitized within 30 days of receipt as a condition of curatorial acceptance. Type specimens 100% digitized with 3D images. ROI estimated at 47.4 EUR/EUR invested -- the highest in the study.

Museum of Vertebrate Zoology (Berkeley, USA; 88.4% digitized): Achieved high digitization rate through a combination of NSF-funded iDigBio grants, systematic crowd-sourcing of label transcription via Notes from Nature, and institutional commitment to open data as a core mission. Specialisation in georeferencing historic locality data using historical maps and gazetteers has produced the highest-precision historical occurrence dataset of any vertebrate collection surveyed.

South African National Biodiversity Institute (SANBI, Pretoria; 47.4% digitized): The highest-digitized African institution in the survey, achieved through the South African National Biodiversity Network and sustained government support. SANBI represents the model for African institution digitization capacity building -- a domestically funded programme with regional ambitions to support smaller national herbaria across southern Africa in digitization methodology.

Insights for Capacity Building: The three highest-performing institutions share: (a) sustained multi-year funding rather than project-based grants; (b) digitization embedded in standard curatorial workflow rather than as a separate project; (c) active GBIF publication as an institutional KPI reviewed annually; and (d) dedicated data quality staff reviewing records for georeference accuracy before GBIF publication.