NATURAL HISTORY COLLECTIONS AND MUSEOMICS

Natural History Collections and Museomics 3: e195668 (2026)
DOI: 10.3897/nhcm.3.195668

Research Article

From collection to sharing: Strategies for standardising and publishing occurrence data in biodiversity platforms in the UFMG-AC Acarological Collection

Mario Henrique Assis-Silva[1], Aristóteles Góes-Neto[1], Almir Rogério Pepato[2]

1 *Programa Interunidades de Pós-Graduação em Bioinformática, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil*
2 *Departamento de Zoologia, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil*
Corresponding author: Mario Henrique Assis-Silva (marioassis1996@ufmg.br)

Abstract

Biological collections are essential for biodiversity research and conservation. However, these collections face challenges, including limited financial, human and spatial resources, inadequate infrastructure and competing curatorial responsibilities. The UFMG-AC Acarological Collection, housed at the Centro de Coleções Taxonômicas (CCT) at the Universidade Federal de Minas Gerais (UFMG), serves as a significant repository of mite diversity, containing over 16,000 catalogued specimens. This study aimed to standardise and publish the occurrence data from the UFMG-AC collection using the Darwin Core (DwC) standard to facilitate integration with global biodiversity databases. The workflow included a manual review of data to correct common errors, followed by the implementation of automated Python scripts to clean, standardise and restructure the data. A taxonomic validation pipeline was also developed to verify hierarchical consistency using the NCBI Taxonomy database and Taxallnomy API. The result is a comprehensive, high-quality dataset, compliant with DwC standards and ready for publication on platforms such as the Brazilian Biodiversity Information System (SiBBr) and the Global Biodiversity Information Facility (GBIF). By improving data quality and accessibility, our study ensures the preservation and global dissemination of the UFMG-AC's invaluable records. This effort highlights the importance of metadata digitisation, automation and standardisation in ensuring data reliability and accessibility. The integration of the UFMG-AC data into public platforms enhances its visibility and utility for biodiversity research.

Key words: Acari, biodiversity informatics, biological collections, Brazil, Darwin Core, data curation, data standardisation, GBIF, SiBBr

OPEN ACCESS

Academic editor: Pablo Dellapé
Received: 27 April 2026
Accepted: 31 July 2026
Published: 12 August 2026

Citation: Assis-Silva MH, Góes-Neto A, Pepato AR (2026) From collection to sharing: Strategies for standardising and publishing occurrence data in biodiversity platforms in the UFMG-AC Acarological Collection. Natural History Collections and Museomics 3: e195668. https://doi.org/10.3897/nhcm.3.195668

Copyright: © Mario Henrique Assis-Silva et al. This is an open access article distributed under terms of the Creative Commons Attribution License (Attribution 4.0 International – CC BY 4.0).

Introduction

Biological collections are foundational to scientific research and biodiversity conservation, serving as critical repositories of data on species distributions, morphological diversity and ecological changes across space and time (Suarez and Tsutsui 2004; Wen et al. 2015; Castillo-Figueroa 2018). These collections enable studies in taxonomy, systematics, ecology and evolutionary biology, while also serving as historical records that help track changes in biodiversity and support conservation planning (Zaher and Young 2003; De Vivo et al.

2014). Amidst the current biodiversity crisis, the relevance of these repositories has intensified, especially as new advances in bioinformatics enhance their accessibility, interoperability and analytical potential (Wen et al. 2015).

Despite their scientific value, taxonomic collections worldwide face numerous challenges. Many suffer from chronic underfunding, limited personnel and inadequate infrastructure (Snow 2005; Rafael et al. 2009). In Brazil and across Latin America, additional obstacles include bureaucratic delays, restrictive legislation and unstable public investment, which hamper both taxonomy and biodiversity research (Glienke et al. 2024). The physical vulnerability of collections, tragically illustrated by the 2018 fire at the National Museum of Brazil, which destroyed irreplaceable specimens and type material, further exposes the urgent need for digitisation and data sharing (Zamudio et al. 2018; Miller et al. 2020). Compounding these issues is the "shelf-life" problem — long delays between specimen collection and formal description — which hinders our understanding of global biodiversity (Fontaine et al. 2012; Cavallin et al. 2016; Luján et al. 2024).

In response to these challenges, a series of technological and institutional advances has emerged. The adoption of the Darwin Core (DwC) standard has enabled the consistent formatting and sharing of biodiversity data (Wieczorek et al. 2012). Initiatives such as the Global Biodiversity Information Facility (GBIF) and the Brazilian Biodiversity Information System (SiBBr) facilitate the dissemination of global and national data (GBIF 2025; SiBBr 2025). Tools such as the Integrated Publishing Toolkit (IPT) have simplified the process of publishing datasets in a standardised format, allowing researchers to contribute to open-access biodiversity databases with greater ease and accuracy (Wieczorek et al. 2012; GBIF 2025). Collectively, these efforts have improved the discoverability, preservation and scientific impact of biological collections (Heberling et al. 2021).

Standardisation not only enables data interoperability, but also reinforces the principles of open science and scientific reproducibility. By adhering to global standards such as Darwin Core, biological collections align their data practices with international initiatives promoting transparency, long-term data usability and collaborative research across disciplines and borders (Wieczorek et al. 2012; Wilkinson et al. 2016; Hardisty et al. 2019). This is particularly crucial in biodiversity science, where integrated datasets fuel conservation strategies, provide information for public policy and support environmental monitoring under rapidly changing global conditions.

Regional and local collections are especially important in this context. Collections, such as the Acarological Collection of the Federal University of Minas Gerais (UFMG-AC), uniquely help fill biogeographical and taxonomic gaps in global data repositories. Such institutions are instrumental in advancing biodiversity knowledge through localised expertise, long-term specimen curation and integration of molecular and morphological data (Pepato and Klimov 2015; Marinoni et al. 2024). Their digitisation and documentation efforts not only strengthen national biodiversity infrastructure, but also align with international goals for open science and conservation (Nelson and Ellis 2019; Hardisty et al. 2019).

Founded in 2012 and housed within UFMG's Taxonomic Collections Center (CCT), the UFMG-AC is one of Brazil's most significant repositories of mite diversity. With over 16,000 catalogued specimens from 18 countries and nearly all Brazilian States, the collection integrates classical taxonomy with the generation of genetic data. It also maintains an extensive data system that records

detailed information at each processing stage — collection, preparation, identification and sequencing — stored collaboratively via cloud-based spreadsheets.

Our study aims to document the recent efforts by the UFMG-AC to standardise, organise and publish its data in accordance with international biodiversity informatics protocols. By describing the workflow, data management and the use of tools such as IPT and Darwin Core, we highlight a replicable model that may guide other biological collections seeking to increase their scientific visibility and contribution to global biodiversity databases.

Collection description

The Acarological Collection of the Federal University of Minas Gerais (UFMG-AC) is housed within the Taxonomic Collections Center (CCT-UFMG), institutionalised in 2015 and affiliated with the Institute of Biological Sciences (ICB) at UFMG (CCT-UFMG 2025). The CCT currently encompasses 26 biological collections and is managed by a Technical-Scientific Council and a team of curators supported by technical and administrative staff.

The collection was formally established in 2012, initially based on the private mite collection of Prof. Dr. Almir Pepato. Over time, the collection expanded through targeted fieldwork, contributions from collaborators and private companies. As of 2023, the collection holds 16,767 catalogued items, including mounted specimens on slides, material preserved in alcohol and fossil specimens. It also houses 14 holotypes and 316 paratypes, as well as 3,711 voucher specimens used in molecular studies. Over 1,500 DNA sequences derived from the collection have been deposited in international genetic databases such as GenBank and BOLD.

Geographically, the collection represents broad coverage: approximately 90% of specimens were collected from nearly all 26 Brazilian States, except Amapá, the Federal District, Goiás, Roraima and Tocantins, while the remaining 10% originate from 18 other countries, including the USA, Russia, Spain, Australia and New Zealand. It also includes samples from extreme environments, such as deep-sea hydrothermal vents in the Indian Ocean and sites in Antarctica. UFMG-AC spans a wide range of habitats, from marine and freshwater systems to soil and forest litter environments, with some specimens associated with specific hosts.

Specimens are stored using appropriate preservation methods: slides are kept in labelled drawers inside metal cabinets, while alcohol-preserved material is housed in sealed containers. DNA extracts are stored at –80 °C in ultra-low temperature freezers. The physical organisation is mirrored by a digital system that ensures efficient access and curation.

The collection follows a standardised curatorial workflow. During fieldwork, data such as locality, date, habitat, host association and collector are recorded in field notebooks. In the laboratory, specimens are sorted, mounted and identified — sometimes followed by DNA extraction and sequencing. Data from each stage are documented in shared spreadsheets hosted in Google Drive, which enable simultaneous editing and access by all lab members. These records are used for sample labelling, organisation and sequencing tracking.

To address common issues such as inconsistent formatting, misspellings and misplaced data entries, the entire dataset was revised and standardised. The final data structure follows the Darwin Core (DwC) format, ensuring interoperability with global biodiversity repositories such as GBIF and SiBBr (Wieczorek et al. 2012;

GBIF 2025; TDWG 2025). This transformation has significantly improved data quality, discoverability and scientific usability (Pereira et al. 2013; Hardisty et al. 2019).

Our study aims to document the recent efforts by the UFMG-AC to standardise, organise and publish its data in accordance with international biodiversity informatics protocols. By describing the workflow, data management and use of tools such as IPT and Darwin Core, we highlight a replicable model that may guide other biological collections seeking to increase their scientific visibility and contribution to global biodiversity databases.

Materials and methods

Collection management and data standardisation protocols

The methodological procedures undertaken in our study aimed to ensure the standardisation, quality control and interoperability of the metadata associated with the UFMG-AC collection. The process began with the evaluation of catalogue spreadsheets created over more than a decade by researchers affiliated with the collection. These documents, hosted on Google Drive, allowed collaborative input from laboratory members and contained detailed records of each specimen. Cataloguing followed a chronological system in which the first two digits of the specimen code represented the year of registration, starting in 2012.

Over time, however, the decentralised and multi-user nature of spreadsheet management led to inconsistencies, including typographical errors, misspellings, variations in column labels and ordering and redundant or misplaced entries. These issues made it necessary to carry out a comprehensive manual revision of all files to standardise formats and correct errors prior to data integration.

Following this manual review, a set of Python scripts was developed to automate the reorganisation of the data into the Darwin Core (DwC) standard. This included merging spreadsheets, correcting structural inconsistencies, removing extra spaces, standardising abbreviations and harmonising contributor names and taxonomic entries using dictionary-based mappings. Special attention was given to maintaining internal dependencies between DwC fields, such as those linking “taxonRank” to the appropriate classification level.

To further refine the taxonomic data and validate hierarchical structures, an additional Python pipeline was implemented. This tool queried the NCBI taxonomy database and the Taxallnomy platform (Sakamoto and Ortega 2021), enabling the verification and enrichment of taxonomic information through the retrieval of standardised ranks, unique taxon identifiers (taxonID) and parent-child relationships. This automated validation ensured a coherent and complete classification framework for the dataset.

Initial data organisation

The initial structure of the UFMG-AC catalogue consisted of yearly spreadsheets stored in a shared Google Drive folder. Each spreadsheet corresponded to a different year and the first two digits of each catalogue number indicated the year of specimen registration (e.g. “1200001” for 2012). While this system allowed decentralised and collaborative data entry by multiple users, it also led to a number of issues that compromised data quality and consistency.

One of the most frequent problems was the occurrence of typographical and orthographic errors. For example, entries such as *"Mar Bálitco"* instead of the correct *"Mar Báltico"* were common. Other errors included misspellings, omission or addition of letters and missing diacritical marks, which are especially relevant in Portuguese locality names. Additionally, the excessive use of white space — whether at the beginning, middle or end of a cell — created major complications for data processing and for the execution of scripts used in subsequent cleaning and formatting steps.

Another recurrent issue was the misplacement of data across inappropriate columns. In one example, a locality described as "Baltic Sea" was erroneously placed in the column corresponding to "Município" (municipality), when, according to the Darwin Core (DwC) standard, the correct fields would be "locality" or "waterBody". This kind of misclassification reduces semantic accuracy and complicates automated data standardisation.

The lack of data formatting consistency was also significant. Dates were represented in various formats within the same spreadsheet (e.g, "01/I/2001", "1/I/2001", "01/01/2001", "01.01.2001" and even "01.01/01"), making it difficult to parse the information programmatically. These discrepancies reflected the absence of unified curation and a review protocol.

Manual curation of data

The manual review phase was essential to identify, correct and standardise the most common inconsistencies found in the UFMG-AC data prior to the application of automated cleaning scripts. This step involved a detailed examination of all spreadsheets — organised by year since 2012 — to ensure internal coherence and compatibility across datasets. The above-mentioned issues related to typographical errors, incorrect spellings, missing diacritical marks and superfluous spaces required manual correction.

Another critical task during this stage was the unification of column names and structures. Due to the decentralised editing of annual spreadsheets by multiple users, column headers often varied in terminology, order and format. These inconsistencies were harmonised to align all spreadsheets with a single, consistent column scheme, ensuring compatibility with the Darwin Core (DwC) standard. The above-mentioned data found to be misclassified within inappropriate columns were systematically corrected during the review, as well as dates that had been all converted to a single, uniform format to facilitate parsing and interoperability. Similarly, geographic coordinates were originally expressed using different reference systems — such as Sirgas 2000, WGS84 and decimal degrees — which were unified into a single consistent geodetic reference system.

This meticulous manual correction process served as the foundation for the subsequent automation phase, allowing scripts and pipelines to be executed more efficiently and without propagation of underlying data inconsistencies.

Automated scripts for data standardisation

To enhance efficiency and ensure compliance with biodiversity data standards, a suite of automated scripts was developed in Python. These scripts, built primarily with the pandas library, addressed recurring issues identified during the

manual review stage and were designed to transform the collection's metadata into the Darwin Core (DwC) format. The scripts are available https://github.com/mariohenriique/TaxonomyDataProcessor.git.

One of the primary tasks was correcting and normalising geographic information in the "Estado" (state) column. The script was programmed to detect both abbreviations (e.g. "AC") and incorrectly spelled names (e.g. "Espirito Santos"), replacing them with standardised full names (e.g. "Acre", "Espírito Santo"). It also completed missing contextual fields such as "country" ("Brazil"), "countryCode" ("BR") and "continent" ("South America"), which are required for interoperability with platforms like GBIF and SiBBr.

Data cleaning also involved the separation of compound fields. For example, entries in the "Sexo/Estágio" (sex/life stage) column were parsed into two discrete fields: "sex" and "lifeStage", conforming to DwC specifications. This restructuring ensures that information is properly categorised and machine-readable.

To standardise person-related data, such as those found in the "recordedBy" field, dictionary-based mappings were implemented. A reference file was created listing all name variants (as keys) and their standardised versions (as values). The script applied these mappings across all datasets, effectively harmonising entries for collectors, which had previously appeared in inconsistent formats due to manual input by multiple users.

Taxonomic data fields required hierarchical logic. The script was configured to fill fields such as "taxonRank" and "scientificName", based on the most specific taxon available. For instance, a record identified as "Halacaridae" would be annotated with "family" for taxonRank and "Halacaridae" for scientificName. The "higherClassification" field was also automatically generated with the full taxonomic hierarchy: Animalia | Arthropoda | Arachnida | Trombidiformes | Prostigmata | Eupodides | Halacaroidea | Halacaridae.

The main steps of the script were as follows:

(1) **Read the data file**: Load the spreadsheet for processing;
(2) **Create missing columns**: Generate additional columns required by DwC terms;
(3) **Correct column data**: Apply corrections for common issues, including spelling, formatting and standardisation;
(4) **Update dependent columns**: Populate columns like "higherClassification" and ensure consistency in taxonomic data;
(5) **Rename columns**: Match column names to DwC standards;
(6) **Save as CSV**: Export the cleaned and standardised data to a CSV file for further use.

This automated pipeline significantly reduced the need for manual intervention, ensured structural and semantic consistency and prepared the data for seamless publication in international biodiversity data platforms.

Taxonomic validation pipeline

To address errors related to data coherence, particularly taxonomic inconsistencies and geographic mismatches, a pipeline was developed to validate and standardise taxonomic information and improve the reliability of biodiversity data. These incon-

sistencies include cases such as a mite from the family Macronyssidae incorrectly placed in the order Trombidiformes instead of Mesostigmata. Such errors required a robust and automated approach to ensure data accuracy and coherence.

The pipeline, written in Python, processes data from CSV files, preferably in Darwin Core (DwC) format, with scientific names provided in a column labelled "scientificName". It utilises external databases and tools such as NCBI Taxonomy and the Taxallnomy API (Sakamoto and Ortega 2021) to retrieve, validate and append hierarchical taxonomic data to a database, ensuring consistency across taxonomic levels.

The process begins with querying the NCBI Taxonomy database, which provides hierarchical classifications, including the scientific name, taxon rank (e.g. kingdom, phylum, order), taxonID and higher taxonomic levels. This information is appended to the database, creating a structured framework for further validation. If an error occurs during the search, such as no matches or incomplete data, the pipeline retries the query up to three times to improve reliability. After retrieving a valid taxonID, it is submitted to the Taxallnomy API, which complements the NCBI data by filling gaps in taxonomic levels and assigning temporary names to otherwise missing categories. This step ensures a complete and consistent representation of the taxonomic hierarchy.

Once the data are retrieved, the pipeline verifies whether the higher taxa exist in the local database. If not, the search process restarts. After successful validation and verification, the pipeline constructs a comprehensive hierarchical database, including relationships such as parent and child taxonIDs, to create a robust and coherent taxonomy.

The pipeline leveraged external databases, such as NCBI Taxonomy and tools such as the Taxallnomy API to ensure data accuracy and coherence (Fig. 1). The process involved the following steps:

(1) **Data Input**: The pipeline reads input from a CSV file, preferably in Darwin Core (DwC) format. The column containing the scientific name to be validated is identified (e.g. "scientificName");
(2) **Search in NCBI Taxonomy**: The scientific names are queried in the NCBI Taxonomy database, which provides a hierarchical structure of taxonomic classifications;
(3) **Appending Taxonomic Data**: The pipeline appends retrieved information to the database, including taxon rank (e.g. kingdom, phylum, order), taxonID and higher taxonomic levels;
(4) **Error Handling**: If errors occur during the search (e.g. no matches or incomplete data), the process retries up to three times to ensure data retrieval;
(5) **Integration with Taxallnomy API**: The taxonID from NCBI is submitted to the Taxallnomy API. This API complements the hierarchical data from NCBI by filling gaps in taxonomic levels and assigning placeholder names when necessary;
(6) **Taxa Verification**: The pipeline verifies whether the retrieved higher taxa exist in the local database. If not, it restarts the search process;
(7) **Hierarchical Data Construction**: After successfully retrieving and verifying the taxonomic data, the pipeline builds a hierarchical database. It includes relationships such as parent and child taxonIDs, ensuring a complete and consistent taxonomic structure.

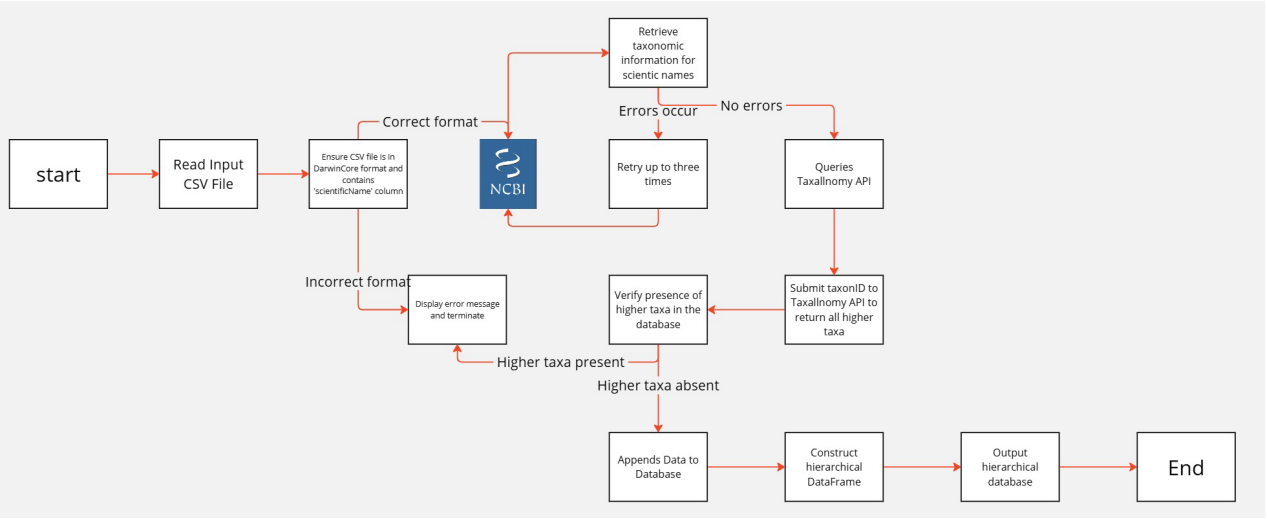

Figure 1. Taxonomic validation and hierarchical data construction workflow.

The pipeline addresses several common challenges encountered during the process. For instance, cases where multiple taxa share the same scientific name, such as the incorrect query "*Porolohmanella violaceae* (Kramer, 1879)" returning information for the plant family Violaceae instead of the intended freshwater mite species *Porolohmanella violaceae* (Kramer, 1879), are logged as errors. These errors are reviewed and corrected manually, when necessary. Additionally, the pipeline detects and corrects potential spelling or typing errors in scientific names, reducing ambiguities and enhancing data quality.

This approach not only corrects taxonomic errors, but also ensures coherence between different fields in the database, such as taxonomic ranks and geographic information. The hierarchical database produced by the pipeline provides a comprehensive framework for scrutinising taxonomic collections and facilitates integration with biodiversity platforms.

By leveraging the widespread use and hierarchical structure of NCBI Taxonomy, along with the gap-filling capabilities of the Taxallnomy API, this pipeline represents a significant advancement in automating data validation and standardisation. It effectively addresses inconsistencies, improves data coherence and provides reliable taxonomic classifications, making it an invaluable tool for researchers and biodiversity data management.

Key standardisation actions

To achieve compliance with DwC and improve semantic clarity, the dataset underwent significant restructuring, including:

- **Renaming of Columns** to DwC equivalents, for example:
  - Tombo → catalogNumber
  - Autor e ano → scientificNameAuthorship
  - Código → otherCatalogNumbers
  - Determinador → identifiedBy
  - U.F./Província → stateProvince
  - Coletor(es) → recordedBy

- **Splitting Composite Fields**:
  - Sexo/Estágio → sex and lifeStage
  - Exempl. → individualCount and preparations
  - Altitude → minimumElevationInMeters and maximumElevationInMeters
  - Coletado em → associatedTaxa and habitat
- **Translating and Mapping Taxonomic and Geographical Terms**:
  - Ordem, família, gênero, espécie → order, family, genus, species
  - país, município, localidade → country, municipality, locality

New columns introduced

To meet Darwin Core's structural and data requirements, new fields were added to the dataset. These include:

- **Core DwC fields**: basisOfRecord, modified, datasetName, institutionCode, collectionCode, occurrenceID, license, rightsHolder, amongst others.
- **Additional fields** for enhanced information: associatedSequences, recordNumber, dynamicProperties, identificationRemarks, samplingEffort, eventRemarks, waterBody, verticalDatum and more.

Auxiliary columns for internal use

Although not part of the published dataset, auxiliary columns support data entry and consistency. For instance, coordinates were initially entered in degrees, minutes and seconds and then programmatically converted into decimal degrees using custom scripts. Likewise, intermediate taxonomic categories were retained to improve internal taxonomic resolution and future classification efforts.

Final dataset and automation

The complete dataset was consolidated in a master spreadsheet encompassing all records from 2012 to 2023, formatted according to DwC. The file includes embedded formulas and a Google Apps Script that automatically updates the modified field upon any change. A separate spreadsheet for 2024 was created following the same structure and is intended for integration into the master file upon completion.

Only DwC-compliant columns are published via the Integrated Publishing Toolkit (IPT), ensuring compatibility with platforms such as SiBBr and GBIF. Auxiliary fields, while essential for internal workflow, are omitted from public datasets to maintain standard compliance.

Results

Data publication and integration with biodiversity platforms

The final standardised dataset of the UFMG-AC collection, of all records catalogued from 2012 to 2023, fully aligned with the Darwin Core (DwC) standard, was successfully published through the Integrated Publishing Toolkit

(IPT), making the collection openly available on both the Global Biodiversity Information Facility (GBIF) and the Brazilian Biodiversity Information System (SiBBr) platforms.

A separate spreadsheet was also created for the years 2024 onwards, following the same DwC structure and automation logic. Upon completion, these new records will be integrated into the master file, ensuring continued consistency and completeness of the dataset.

The IPT was configured with metadata required by GBIF and SiBBr, including dataset title, abstract, taxonomic and geographic scope, data licence, contact information and a methodological description. It allowed:

- The publication of over 16,000 occurrence records in machine-readable format;
- Hosting of the dataset in both GBIF and SiBBr, enhancing visibility and long-term accessibility;
- Automation of data updates to support continuous data curation and version control;
- Compliance with international biodiversity data standards, promoting interoperability, reuse and citation;
- The availability of the dataset under an open licence, supporting transparency and open science.

This publication increases the visibility and scientific usability of UFMG-AC, positioning the collection as a contributor to global biodiversity knowledge. It also reinforces the collection's role in national and international networks of taxonomic data sharing, conservation planning and ecological research.

Temporal coverage

The temporal profile of the UFMG-AC collection reflects both its historical depth and its dynamic institutional growth. Officially established in 2012 by Prof. Dr. Almir Rogério Pepato, the collection initially comprised primarily marine mites from his personal repository. Over time, it expanded significantly through research collaborations, academic networks and the dedicated efforts of professors and students at the university. Specimens from a wide range of environments and origins have since been incorporated, enriching both the taxonomic and temporal breadth of the collection.

Despite the oldest specimen dating back to 1984, items collected before 2010 are relatively few (7.5%), creating a noticeable gap between the earliest material and the steady growth observed after collection formal establishment. Since then, the collection has expanded at a consistent pace, averaging 1,339 new catalogue entries per year. This upward trend is visible in Fig. 2, which shows the annual number of catalogued records from 2012 to 2023.

The inclusion of new records peaked in 2017, whereas 2016 recorded the highest number of collected specimens. This lag reveals a recurring pattern: most collections have only been finalised in subsequent years, primarily due to time-intensive processes such as DNA extraction, specimen identification and particularly slide preparation and labelling. Based on collection and cataloguing dates, the average delay between specimen collection and identification

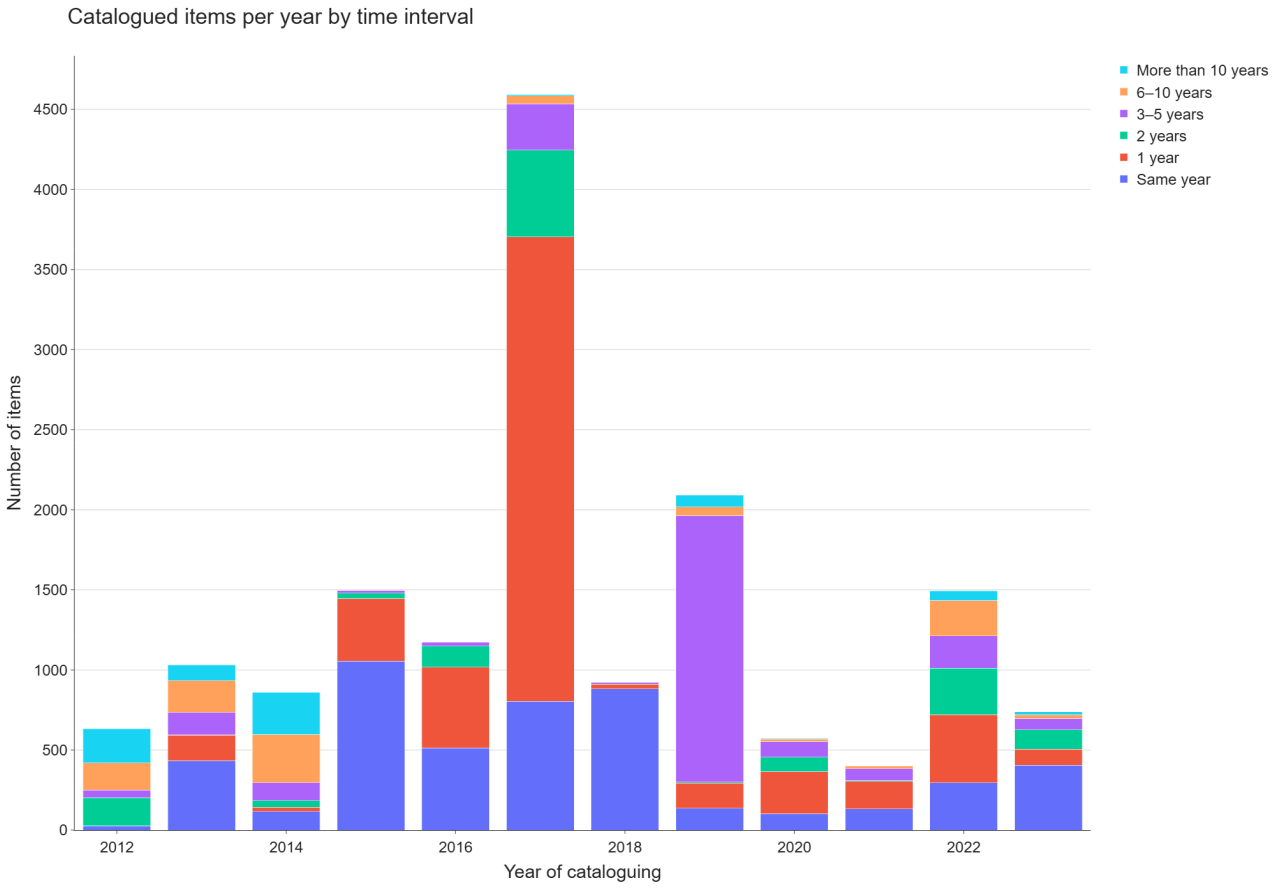

Figure 2. Number of items catalogued per year according to the time elapsed between collection and cataloguing. Stacked bars represent annual totals, subdivided into six time intervals: same year (≤ 0 years), 1 year, 2 years, 3–5 years, 6–10 years and more than 10 years between collection and cataloguing.

was estimated at 9 years, highlighting a significant temporal disconnect in the processing pipeline.

This time lag is particularly relevant in the context of type material. For holotypes and paratypes, which underpin species descriptions, the average delay between collection and official cataloguing was 1.95 years — a metric often referred to as the "shelf-life" in taxonomic workflows (Fontaine et al. 2012). Fig. 3A illustrates the number of catalogued items per year, segmented by time elapsed since collection.

An additional temporal analysis focused on weekly collection patterns. As shown in Fig. 2C, Monday had the highest number of recorded collection events. This concentration likely reflects the logistical structure of field expeditions, which often starts early in the week following preparatory activities carried out in preceding days. While weather conditions and project schedules may also influence this distribution, the data suggest that strategic planning can significantly affect collection productivity.

A deeper investigation into annual collection dynamics revealed an outlier in 2017, which presented a disproportionately high number of catalogued items. Interestingly, 2013 had the highest number of unique collectors (n = 64), but this did not translate into higher cataloguing volumes. Examination of the recordedBy column showed that contributions especially from Equipe Carste. The "Carste ciência e meio ambiente" is a private company that offers scientific

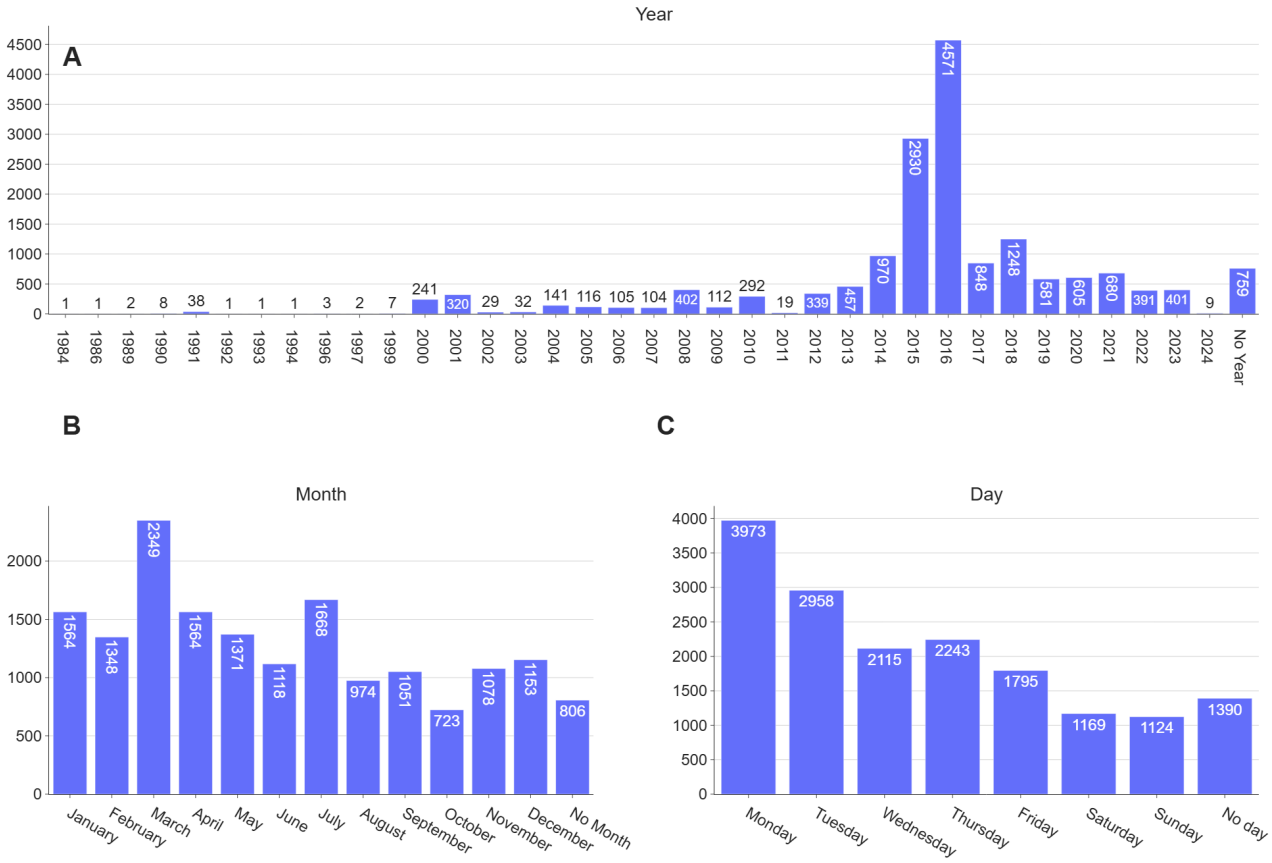

Figure 3. Temporal distribution of collected items. **A.** Annual number of collected items, including records without year information ("No year"); **B.** Monthly distribution of collected items, including records without month information ("No month"); **C.** Distribution of collected items by day of the week, including records without day information ("No day"). Values represent absolute counts of collection records.

services to mining companies that need to comply with the Brazilian environmental laws. They were instrumental in boosting collection rates in years like 2016, by introducing specially colletec mites from caves, helping to explain some of the annual fluctuations observed. Fig. 4 presents the correlation between occurrence records per year and collector identity.

Spatial coverage

The UFMG-AC collection demonstrates a broad and well-documented spatial coverage, with specimens collected from across Brazil and 18 additional countries spread in different continents (Fig. 5). This wide geographic scope reflects both local biodiversity research and international collaborations, although the collection maintains a clear emphasis on Brazilian fauna. Approximately 90% of catalogued specimens originate from Brazil, reinforcing the collection's central mission of documenting the country's mite diversity.

At the international level, specimens have been acquired from nearly all continents, with the exception of Africa. The most represented countries outside Brazil include Russia (219 specimens), Spain (200), Australia (117), New Zealand (94), the United States (72) and Chile (62) (Table 1). Other countries with smaller representations include Germany, Azerbaijan, Iran, Bolivia and several

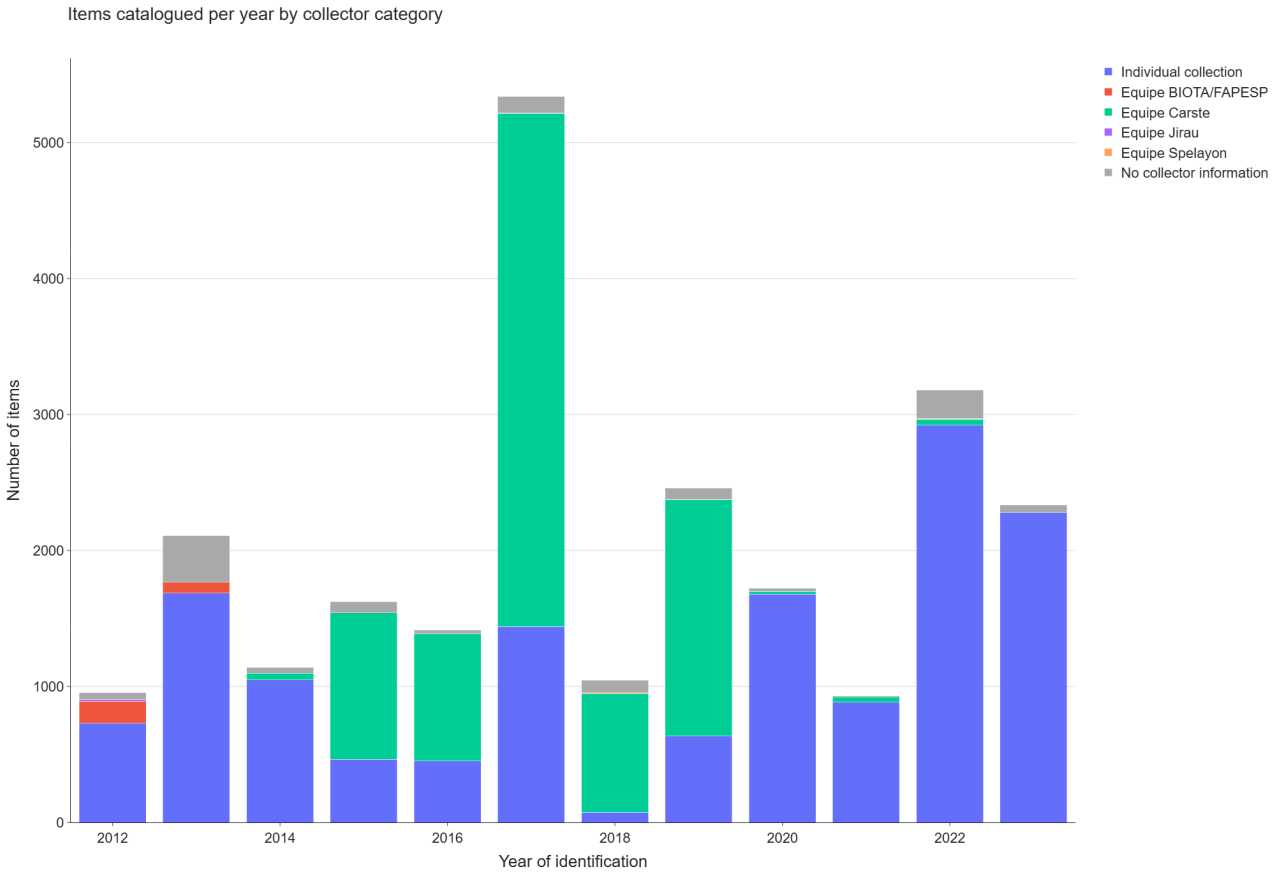

Figure 4. Annual cataloguing output by collector category. Number of catalogued items per year grouped by collector category. Stacked bars represent annual totals subdivided into individual collections, institutional teams (Equipe BIOTA/FAPESP, Equipe Carste, Equipe Jirau and Equipe Spelayon) and records lacking collector information. Values indicate the proportional contribution of each category to the yearly total.

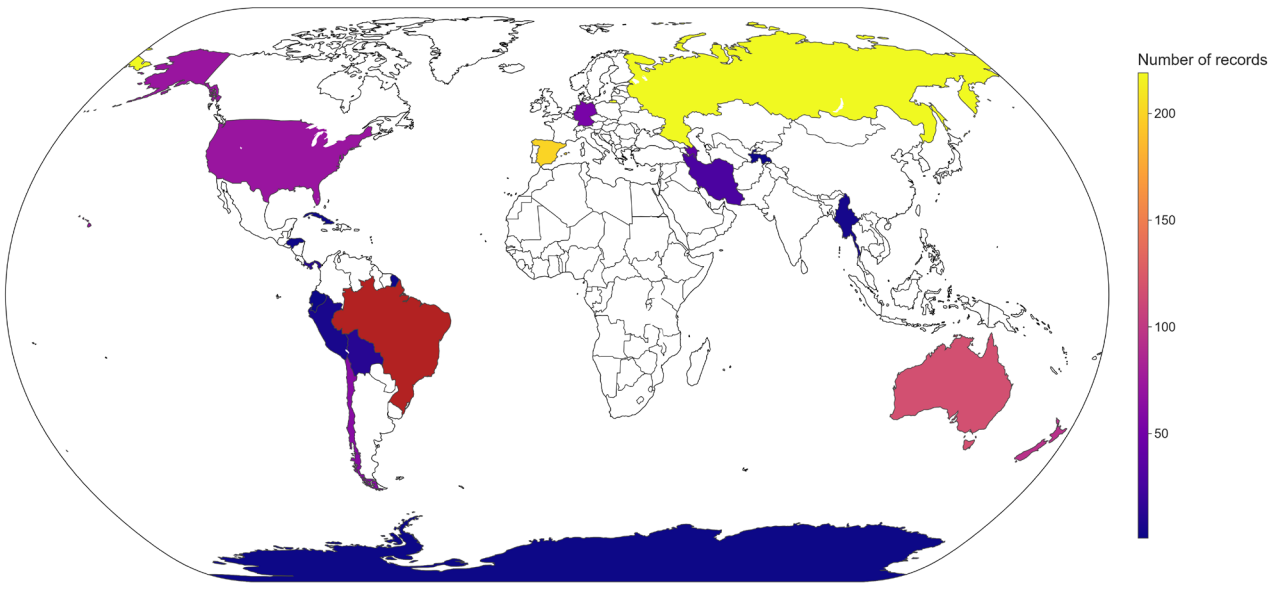

Figure 5. Geographic distribution of records by country. Countries are shaded according to the total number of records using a continuous colour scale. Brazil is highlighted in red and excluded from the scale due to its substantially higher number of records relative to other countries.

Table 1. Summarizes the number of catalogued items per country and continent, including both absolute and relative percentages.

Continent	Country	Quantity	Absolute percentage (%)	Relative percentage (%)
South America	Brazil	15372	91.68	—
	Chile	62	0.37	6.47
	Bolivia	12	0.07	1.29
	Peru	6	0..04	0.65
	French Guiana	2	0.01	0.22
	Ecuador	1	0.01	0.11
Central America	Panama	12	0.07	1.29
	Cuba	2	0.01	0.22
	Honduras	2	0.01	0.22
North America	USA	72	0.43	7.76
Europe	Spain	200	1.19	21.57
	Germany	53	0.32	5.72
Asia	Russia	219	1.31	23.62
	Azerbaijan	39	0.23	4.21
	Iran	28	0.17	3.02
	Myanmar	5	0.03	0.54
	Tajikistan	2	0.01	0.22
Oceania	Australia	117	0.70	12.62
	New Zealand	94	0.56	10.14
Antarctica		1	0.01	0.11
Missing country information		465	2.77	–
Total		16766	100.00	100.00

nations in Central America and Asia. The spatial distribution of individual occurrence records further illustrates the geographic dispersion and sampling intensity of the collection (Fig. 6). Notably, marine mites collected near hydrothermal vents in the Indian Ocean and specimens from Antarctica highlight the taxonomic and ecological breadth of the collection.

Within Brazil, the Southeast Region dominates the dataset, accounting for 88.17% of all catalogued records, followed by the Northeast (4.72%), North (4.26%), South (1.85%) and Centre-West (0.40%) regions (Table 2). At the state level, Minas Gerais alone accounts for 76.76% of all Brazilian records, a pattern likely influenced by the logistical advantages of proximity to the collection headquarters. Other States with notable contributions include São Paulo (8.44%), Pará (3.15%) and Espírito Santo (2.40%) (Fig. 7).

This strong regional bias is due the Carste Ciência e Meio Ambiente company activities and convenience for the researchers at UFMG, since sampling in the State allows for lower-cost and more frequent fieldwork, a factor that must be considered when interpreting geographic biases in biological repositories. The uneven representation across Brazil also underscores the importance of promoting fieldwork in under-represented regions — particularly in the Centre-West and certain northern and north-eastern states — as well the need for the establishment of other taxonomical collections that would promote a more intense sampling in their neighbourhoods.

The high spatial resolution of UFMG-AC's records enables valuable insights into species distributions, biogeographic patterns and sampling biases.

Furthermore, by making these data openly available through platforms such as GBIF and SiBBr, the collection contributes to regional and global biodiversity assessments and conservation planning.

Table 2. Details the spatial distribution of specimens across Brazilian States, organised by geographic region.

Region	State	Quantity	Absolute percentage (%)	Relative percentage (%)	City
Sudeste	Minas Gerais	11801	76.76	–	96
	São Paulo	1298	8.44	36.91	23
	Espírito Santo	369	2.40	10.49	17
	Rio de Janeiro	125	0.81	3.55	10
Norte	Pará	485	3.15	13.79	18
	Amazonas	126	0.82	3.58	3
	Rondônia	33	0.22	0.92	1
	Acre	8	0.05	0.23	1
Centro-oeste	Mato Grosso do Sul	51	0.33	1.45	2
	Mato Grosso	11	0.07	0.31	3
Sul	Santa Catarina	122	0.79	3.47	11
	Rio Grande do Sul	119	0.77	3.38	4
	Paraná	49	0.32	1.39	4
Nordeste	Bahia	292	1.90	8.30	20
	Pernambuco	113	0.74	3.21	8
	Alagoas	61	0.40	1.73	4
	Rio Grande do Norte	52	0.34	1.48	2
	Ceará	49	0.32	1.39	2
	Paraíba	41	0.27	1.17	4
	Sergipe	14	0.09	0.40	3
	Maranhão	9	0.06	0.26	1
Missing state information		56	0.36	–	
Total		15372	100.00	100.00	237

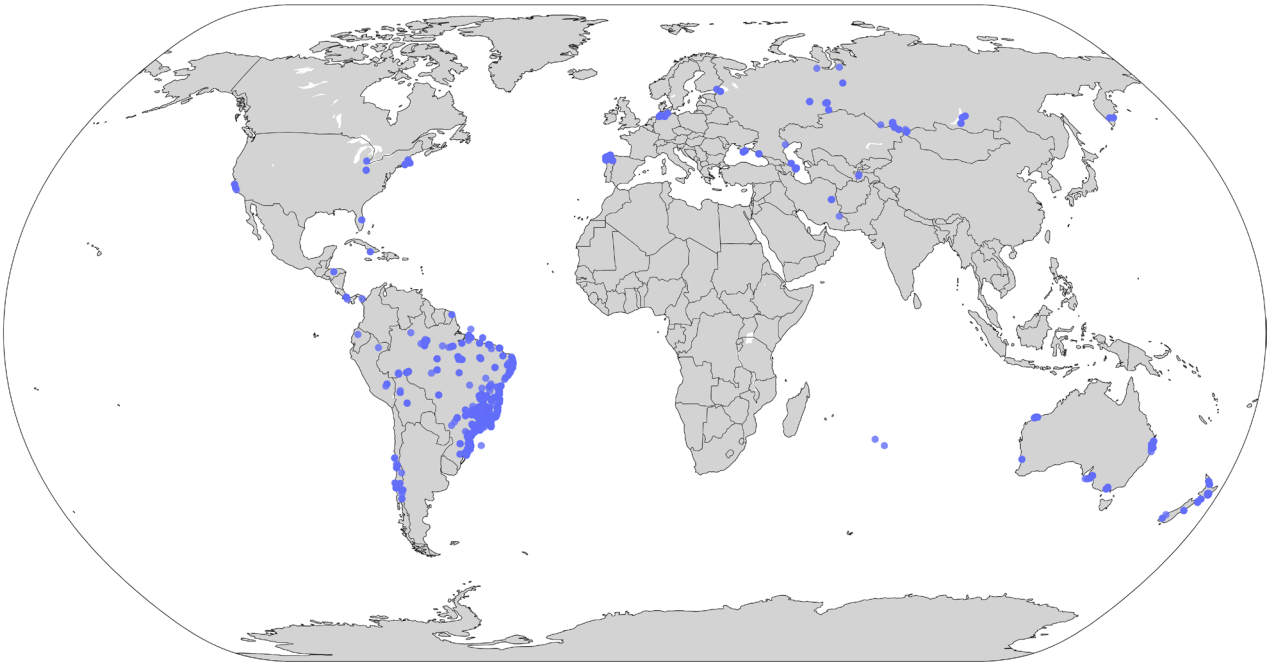

Figure 6. Spatial distribution of occurrence records showing the geographic location of sampling points in UFMG-AC. Each point represents a georeferenced occurrence record.

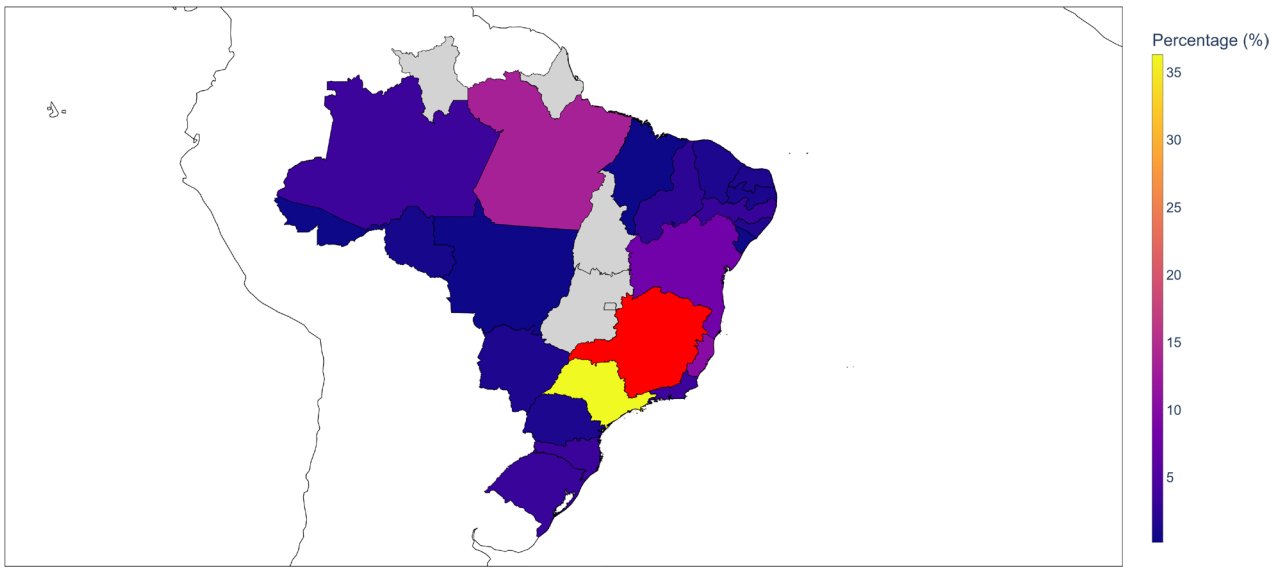

Figure 7. Percentage distribution of occurrence records by Brazilian federal state. Minas Gerais (red) is excluded from the colour scale due to its substantially higher percentage of records. States shown in grey indicate absence of occurrence records (Amapá, Distrito Federal, Goiás, Roraima and Tocantins).

Taxonomic coverage

The UFMG-AC collection exhibits substantial ecological — encompassing parasitic, marine, aquatic and cave dwelling mites — and taxonomic breadth, encompassing a wide range of hierarchical levels including order, family, genus and species. As of 2023, the collection includes specimens distributed across 14 taxonomic orders. The mite specimens fall under six orders of the subclass Acari: Trombidiformes, Mesostigmata, Opilioacariformes, Sarcoptiformes, Holothyrida and Ixodida. Nonetheless, due to the co-collection of minute organisms during field sampling, the database also includes non-mite taxa such as Araneae, Copepoda, Opiliones, Ostracoda, Palpigradi, Pseudoscorpiones, Schizomida and Tanaidacea. These are currently retained in the dataset for logistical reasons, but will be reallocated to the appropriate collections within CCT-UFMG or to partner institutions in the future. Furthermore, they represent only 0.002% of individuals included in the UFMG-AC collection.

Taxonomic identification remains an ongoing process, particularly at lower taxonomic ranks. Currently, approximately two-thirds of the specimens have been identified at the family level, which represents a key milestone in the classification workflow. This hierarchical identification process is cumulative: specimens not identified at the family level are inherently unclassified at the genus or species level, making the advancement in each rank dependent on the accuracy of higher-level identifications (Fig. 8).

An analysis of family-level representation reveals strong concentrations in groups relevant to the collection's research agenda, with the predominance of Halacaridae, a family of marine mites, reflecting its focus on marine and intertidal acarology. There are also families employed within research projects by graduate students (Erythraeidae, Spinturnicidae**),** often found in large numbers in caves (Trhypochthoniidae, Laelapidae and Macronyssidae)**.**

A histogram displaying the frequency of families with more than 100 records (Fig. 9) visually illustrates the taxonomic diversity represented in the collection.

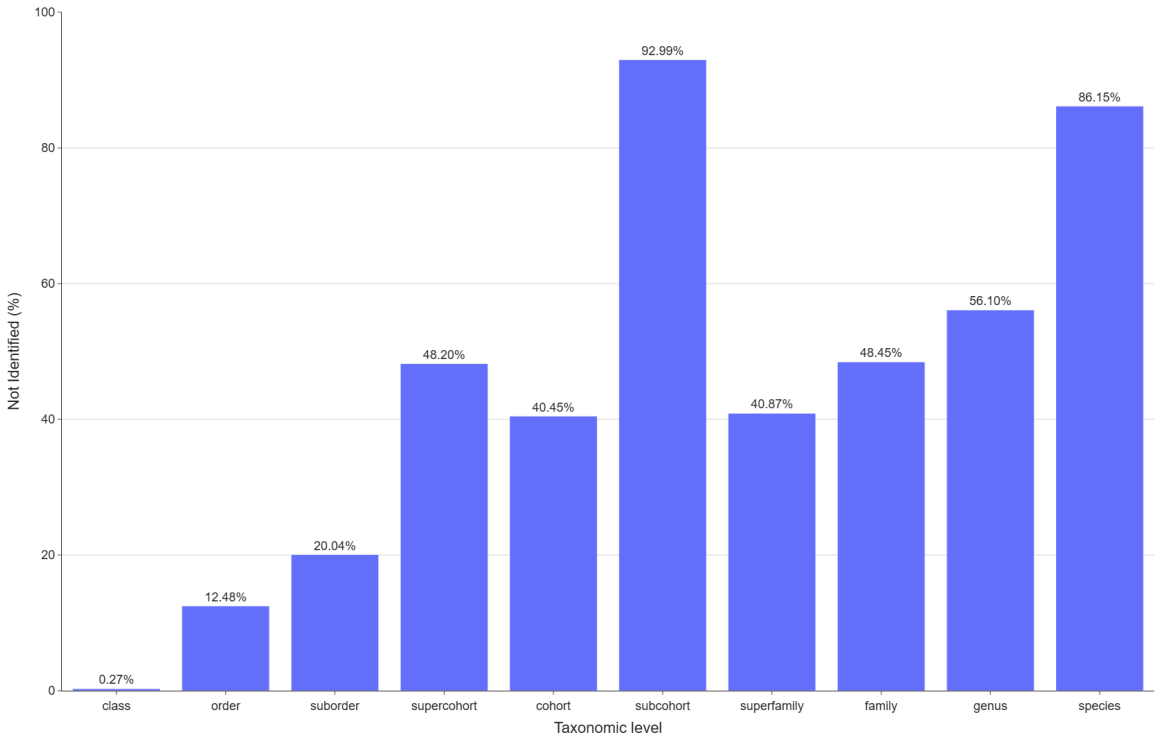

Figure 8. Percentage of missing values across taxonomic levels. Taxonomic level names are presented as recorded in the database. The dataset contains a mixture of Darwin Core (DwC) standardised terms and legacy taxonomic rank names that do not have direct DwC equivalents (e.g. Subordem) and, therefore, some labels appear in Portuguese, while others follow DwC conventions. Not all specimens contain complete hierarchical classifications; some records lack lower ranks (e.g. Subcoorte) while retaining higher ranks (e.g. Coorte).

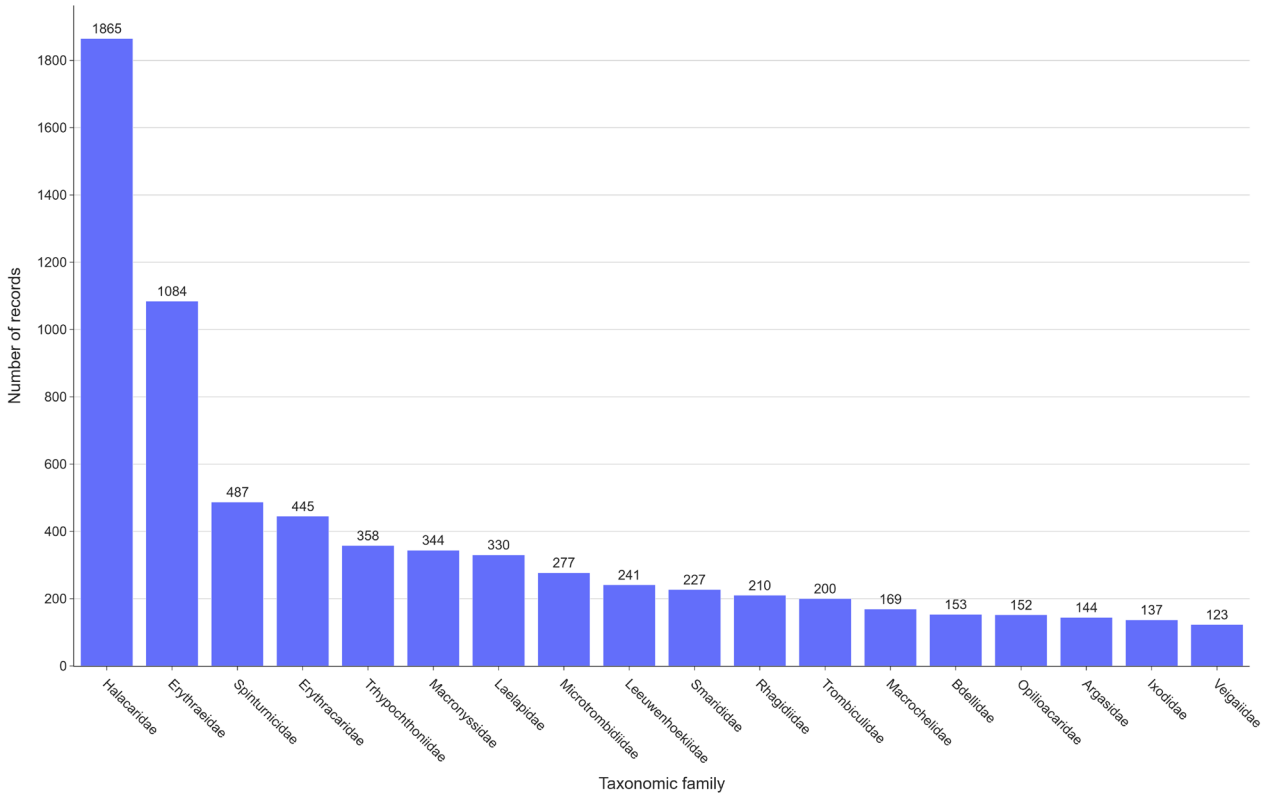

Figure 9. Frequency distribution of taxonomic families, based on the number of occurrence records. Bars represent the total number of records per family.

Standardisation

The data standardisation of the UFMG-AC collection was a pivotal step in transforming heterogeneous, multi-format metadata into a structured and interoperable database. The resulting dataset, fully aligned with the Darwin Core (DwC) standard (Wieczorek et al. 2012), supports both internal curation and external data publication via biodiversity platforms such as GBIF and SiBBr.

The final version of the dataset consists of 89 columns, of which 12 are auxiliary fields designed to aid internal processes, but excluded from public repositories due to their non-compliance with DwC. These auxiliary fields include geographic detail columns such as graus, minutos, segundos, N/S, graus.1, minutos.1, segundos.1, W/E and hierarchical taxonomic intermediates such as Subordem, Supercoorte, Coorte and Subcoorte widely employed in acarology. Despite not being exported, the information from these fields is embedded or mirrored in DwC-compliant columns like decimalLatitude, decimalLongitude, verbatimCoordinates and higherClassification.

Discussion

The key outcome of this work is the successful application of the Darwin Core (DwC) standard for organising and publishing biodiversity data. Despite its widespread adoption, the implementation of DwC in Brazilian collections remains limited due to lack of technical expertise, institutional support or awareness. The methodology employed by UFMG-AC — combining manual review, automated Python scripts and a taxonomic validation pipeline — serves as a replicable model for other institutions seeking to enhance data interoperability and visibility (Wieczorek et al. 2012). Moreover, the use of tools, such as Google Sheets and Apps Script, demonstrates that low-cost, accessible technologies can still yield rigorous and scalable data management solutions.

Nonetheless, the process also revealed persistent challenges faced by regional collections. These include limited personnel, uneven taxonomic expertise and the burden of managing large volumes of historical data without centralised curatorial oversight. As emphasised by Snow (2005) and Rafael et al. (2009), curators are often overextended, balancing teaching, research and collection maintenance — a reality also reflected in the UFMG-AC workflow. The observed average delay of 1.95 years between collection and cataloguing of type material and up to 9 years for general identification, exemplifies the issue known as “shelf life” in taxonomy (Fontaine et al. 2012). While this delay is not uncommon globally, its impact is more pronounced in regions with high biodiversity and under-resourced institutions, where undescribed species may remain hidden for years, delaying conservation assessments and ecological research.

The spatial and taxonomic coverage of UFMG-AC also provides insights into collection biases and research priorities. While the vast majority of records originate from Minas Gerais and the Southeast Region, this pattern is largely shaped by the logistical feasibility of conducting fieldwork near the institution and the inclusion of material collected by local companies. Similar trends have been reported in other Brazilian collections, reinforcing the need for targeted funding and collaboration to support regional collections in under-represented areas. Taxonomically, the over-representation of families such as Halacaridae

and Erythraeidae reflects the research focus of the institution, but also reveals gaps in groups with fewer specialists or reference materials.

Looking forward, future efforts should aim to:

- Expand digitisation and taxonomic resolution, particularly at the species level;
- Reduce shelf life by using integrative taxonomy, including DNA barcoding and automated identification tools;
- Promote capacity-building initiatives focused on data standardisation and curation across other regional collections in Brazil;
- Encourage long-term investment in personnel, infrastructure and collaborative networks to ensure the sustainability of biological repositories.

Finally, this work offers a replicable model for other institutions aiming to modernise their biological repositories. It emphasises that investment in standardised digital infrastructure, personnel training and collaborative networks is not only feasible, but essential for the long-term sustainability of taxonomic collections in the face of ecological, institutional and environmental uncertainties.

In light of recurring threats to physical collections — such as the fires at the Museu Nacional and Instituto Butantan in Brazil — the digital preservation and open publication of biological metadata are vital measures for minimising irreversible losses. The UFMG-AC example demonstrates that, with proper planning, institutional commitment and relatively low-cost tools, regional collections can meet international data standards and actively contribute to global biodiversity science.

References

Castillo-Figueroa D (2018) Beyond specimens: linking biological collections, functional ecology and biodiversity conservation. Revista peruana de biología 25: 343–348. https://doi.org/10.15381/rpb.v25i3.14246

Cavallin EK, Munhoz CB, Harris SA, Villarroel D, Proença CE (2016) Influence of biological and social historical variables on the time taken to describe an angiosperm. American Journal of Botany 103: 2000–2012. https://doi.org/10.3732/ajb.1600120

CCT-UFMG (2025) Centro de Coleções Taxonômicas. https://www2.icb.ufmg.br/cct/

De Vivo MD, Silveira LF, Nascimento F (2014) Reflexões sobre coleções zoológicas, sua curadoria e a inserção dos Museus na estrutura universitária brasileira. Arquivos de Zoologia 45: 105–113. https://doi.org/10.11606/issn.2176-7793.v45iespp105-113

Fontaine B, Perrard A, Bouchet P (2012) 21 years of shelf life between discovery and description of new species. Current Biology 22: R943–R944. https://doi.org/10.1016/j.cub.2012.10.029

GBIF (2025) Quick guide to publishing data through GBIF.org. https://www.gbif.org/

Glienke C, Petters-Vandresen DA, Souto AD, Marinoni L, da Silva M (2024) Microbiological collections in Brazil: current status and perspectives. Diversity 16: 116. https://doi.org/10.3390/d16020116

Hardisty AR, Michener WK, Agosti D, García EA, Bastin L, Belbin L, Bowser A, Buttigieg PL, Canhos DAL, Egloff W, De Giovanni R, Figueira R, Groom Q, Guralnick RP, Hobern D, Hugo W, Koureas D, Ji L, Los W, Manuel J, Manset D, Poelen J, Saarenmaa H, Schigel D, Uhlir PF, Kissling WD (2019) The Bari Manifesto: An interoperability framework

for essential biodiversity variables. Ecological informatics 49: 22–31. https://doi.org/10.1016/j.ecoinf.2018.11.003

Heberling JM, Miller JT, Noesgaard D, Weingart SB, Schigel D (2021) Data integration enables global biodiversity synthesis. Proceedings of the National Academy of Sciences 118: e2018093118. https://doi.org/10.1073/pnas.2018093118

Luján M, Lemos RM, Lucas E, Michelangeli FA, Prance GT, Pennington TD, Rzedowski J, Santamaría-Aguilar D, Serpell E, Sothers CA, Zuntini AR (2024) Trials and tribulations of Neotropical plant taxonomy: Pace of tree species description. Plants, People, Planet 6: 515–527. https://doi.org/10.1002/ppp3.10469

Marinoni L, Gasper AL, Chiquito EA, Glienke C, Fonseca CB, Juarez KEM, Basílio DS, Henriques DK, Petters-Vandresen DAL, Morais GS, Jorge I, Pimenta AD, Dal Molin A, Souto ASS, Santos CRM, Serejo CS, Souzamotta CM, Attili-Angelis D, Werneck FP, Carvalho FR, Garbino GST, Stehmann JR, Giacomin LL, Jesus MA, Silva M, Fukuda MV, Rosa RS, Costa-araújo R, Maués MM, Thiengo SC, Silva TRS, Andrade TS, Cavalcanti TB, Dutra VF, Vicente VA (2024) Introdução e orientações às boas práticas para as Coleções Biológicas Científicas Brasileiras. Sociedade Brasileira de Zoologia. https://doi.org/10.7476/9786587590042

Miller SE, Barrow LN, Ehlman SM, Goodheart JA, Greiman SE, Lutz HL, Misiewicz TM, Smith SM, Tan M, Thawley CJ, Cook JA, Light JE (2020) Building natural history collections for the twenty-first century and beyond. BioScience 70: 674–687. https://doi.org/10.1093/biosci/biaa069

Nelson G, Ellis S (2019) The history and impact of digitization and digital data mobilization on biodiversity research. Philosophical Transactions of the Royal Society B 374: 20170391. https://doi.org/10.1098/rstb.2017.0391

Pepato AR, Klimov PB (2015) Origin and higher-level diversification of acariform mites–evidence from nuclear ribosomal genes, extensive taxon sampling, and secondary structure alignment. BMC Evolutionary Biology 15: 178. https://doi.org/10.1186/s12862-015-0458-2

Pereira HM, Ferrier S, Walters M, Geller GN, Jongman RHG, Scholes RJ, Bruford MW, Brummitt N, Butchart SHM, Cardoso AC, Coops NC, Dulloo E, Faith DP, Freyhof J, Gregory RD, Heip C, Höft R, Hurtt G, Jetz W, Karp DS, McGeoch MA, Obura D, Onoda Y, Pettorelli N, Reyers B, Sayre R, Scharlemann JPW, Stuart SN, Turak E, Walpole M, Wegmann M (2013) Essential biodiversity variables. Science 339(6117): 277–278. https://doi.org/10.1126/science.1229931

Rafael JA, Aguiar AP, Amorim DD (2009) Knowledge of insect diversity in Brazil: challenges and advances. Neotropical Entomology 38: 565–570. https://doi.org/10.1590/S1519-566X2009000500001

Sakamoto T, Ortega JM (2021) Taxallnomy: an extension of NCBI Taxonomy that produces a hierarchically complete taxonomic tree. BMC Bioinformatics 22: 1–23. https://doi.org/10.1186/s12859-021-04304-3

SiBBr (2025) O que é o SiBBr. https://sibbr.gov.br/

Snow N (2005) Successfully curating smaller herbaria and natural history collections in academic settings. BioScience 55: 771–779. https://doi.org/10.1641/0006-3568(2005)055[0771:SCSHAN]2.0.CO;2

Suarez AV, Tsutsui ND (2004) The value of museum collections for research and society. BioScience 54: 66–74. https://doi.org/10.1641/0006-3568(2004)054[0066:TVOMCF]2.0.CO;2

TDWG (2025) Darwin Core. https://dwc.tdwg.org/

Wen J, Ickert Bond SM, Appelhans MS, Dorr LJ, Funk VA (2015) Collections based systematics: Opportunities and outlook for 2050. Journal of Systematics and Evolution 53: 477–488. https://doi.org/10.1111/jse.12181

Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, Giovanni R, Robertson T, Vieglais D (2012) Darwin Core: an evolving community-developed biodiversity data standard. PLOS ONE 7: e29715. https://doi.org/10.1371/journal.pone.0029715

Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, Santos LBS, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJG, Groth P, Goble C, Grethe JS, Heringa J, Hoen PAC't, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone SA, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B (2016) The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3: 1–9. https://doi.org/10.1038/sdata.2016.18

Zaher H, Young PS (2003) As coleções zoológicas brasileiras: panorama e desafios. Ciência e Cultura 55: 24–26. http://cienciaecultura.bvs.br/scielo.php?script=sci_arttext&pid=S0009-67252003000300017

Zamudio KR, Kellner A, Serejo C, de Britto MR, Castro CB, Buckup PA, Pires DO, Couri M, Kury AB, Cardoso IA, Monné ML, Pombal J, Patiu CM, Padula V, Pimenta AD, Ventura CRR, Hajdu E, Zanol J, Bruna EM, Fitzpatrick J, Rocha LA (2018) Lack of science support fails Brazil. Science 361: 1322–1323. https://doi.org/10.1126/science.aav3296

Additional information

Conflict of interest

The authors have declared that no competing interests exist.

Ethical statement

No ethical statement was reported.

Artificial Intelligence (AI) use

The authors accept full responsibility for the content of the manuscript, including the disclosure of any use of AI.
Regarding the use of AI in the preparation of this manuscript, the authors declare the following: GPT.
Used for: Language, style and writing, Software and automation.
Artificial intelligence (AI) tools were used in a limited and supportive capacity during the preparation of this manuscript. Specifically, AI-assisted models were employed to improve language clarity, grammar, and overall readability, as well as to refine the organization of the text. AI tools also assisted in data visualization, code refinement, figure formatting, and translation of labels into English.

Funding

No funding was reported.

Author contributions

All authors have contributed equally.

Author ORCIDs

M.H. Assis-Silva https://orcid.org/0009-0005-2280-4522
A. Góes-Neto https://orcid.org/0000-0002-7692-6243
A.R. Pepato https://orcid.org/0000-0002-2944-4209

Data availability

All of the data that support the findings of this study are available in the main text.