# Trap of Time: Historical Common Names to Modern Taxonomy Mapping using LLMs

Jan Fillies[1,2,*], Naouel Karam[1], Alois Wieshuber[3], Giada Matheisen[3], Malte Rehbein[4], Belen Escobari[6], Sarah Fischer[5] and Adrian Paschke[1,2,7]

[1]Institute for Applied Informatics (InfAI), University of Leipzig, Germany
[2]Freie Universität Berlin, Berlin, Germany
[3]Generaldirektion der Staatlichen Archive Bayerns, Munich, Germany
[4]Chair of Computational Humanities, University of Passau, Germany
[5]Research Institute for Farm Animal Biology (FBN) Dummerstorf, Germany
[6]Botanic Garden and Botanical Museum Berlin, Germany
[8]Fraunhofer FOKUS, Berlin, Germany

**Abstract**

As society and scientific research evolve, so does the language used to express concepts, names of species, and descriptions of objects. This research addresses the challenge of mapping historic terms—specifically, common names for species in the field of biodiversity— to a modern taxonomy. Historic biodiversity collections, already sparse, are further complicated by the use of common names rather than scientific names, making exact alignment with modern taxonomies highly challenging. Changes in spelling, along with species being merged, split, or renamed, further add to this complexity. This research explores the use of a large language model (LLM), GPT-4o, to assist in this alignment process. Results show that, when provided with context, the LLM can accurately generate modern equivalents of historic instances, demonstrating an embedded understanding of historical semantic shifts in biodiversity terminology. In a test set, the LLM successfully matched (91% of cases) both unchanged and altered common names to their correct scientific names (with the inclusion of minimal context) and modern common name counterparts (68% with minimal context), underscoring its potential to standardize historical datasets and support human annotation in the future.

**Keywords**

Large Language Models (LLMs), Language Standardization, Historic Data, Semantic Annotation, Taxonomies



## 1. Introduction

Humans have always produced evidence about their observations and reflections of nature. Properly processed, these historical records can become important data for modern science. However, the language used to refer to species, as well as baselines, objects, concepts, and their interrelations, have evolved significantly since the time such historical sources originated. Preserving these changes is fundamental for subsequent analysis [1]. As scientific knowledge especially in biodiversity advances, it is becoming more and more important to align historical data with standardized and curated taxonomies, ontologies, and authority files. The advantages of controlled taxonomies are numerous, but mainly they guarantee the homogeneous use of a certain vocabulary within their field [2]. While taxonomies often carry the common and scientific name of a species, it is difficult to account for evolving language. This can be looked at from the scientific side, where species are periodically reclassified, merged, split, or renamed based on a novel scientific observation [3], or from a societal perspective, where common names can vary widely across regions, evolve over time, or become obsolete.

Even more problematic is the fact that historical observations often lack a firm and standardized nomenclature and scientific background. For example, only local common names were recorded when biodiversity data was collected in the past. In order to make this data accessible for current research, a manual matching between these historical common names and modern taxonomies is required. This matching requires a high level of domain expertise with an understanding of current biodiversity

*SWAT4HCLS 2025, Februar 24–27, 2025, Barcelona, ES*

taxonomy as well as knowledge of historical varieties of languages used to describe specific species. This complicates the process of integrating the already scarce data collections into modern biodiversity studies and creates high barriers to consistency and usability of biodiversity data.

We address these challenges within the framework of semantic annotation, leveraging our ontology repository and service, BiodivPortal [4]. BiodivPortal serves as a terminology service, enabling unified access, development, and maintenance of terminologies within biodiversity and environmental sciences. In this work, the objective is to develop a tool that supports the semantic annotation of historical data using BiodivPortal terminologies, facilitating the analysis of data despite linguistic and conceptual shifts in biodiversity vocabulary over time. By enabling this annotation with modern standards, we aim to enhance data accessibility and utility, ultimately advancing scientific research and informing policy-making in biodiversity and environmental studies.

This research specifically examines the issue of matching common names used in a historical biodiversity dataset (from 1845) to their counterparts in a standardized modern taxonomy. Recent research has shown large language models (LLMs) hold much potential for the field of biodiversity [5]. First studies have appeared demonstrating that GPT-3.5 can provide biodiversity related information about species descriptions, occurrences, and taxonomy [6]. In the conducted experiments, we build upon this existing research by evaluating the potential of a LLM, specifically GPT-4o, to assist in this alignment. It demonstrates the ability of the LLM to interpret historical language within a scientific context, and offers a new approach to detecting shifts in historical species names and classifications over time. This is becoming more relevant for research in biodiversity as more historical sources are discovered and mobilized (e.g., [7, 8, 9]).

This study shows that within a given context, GPT-4o can identify and match historical common names to modern equivalents in 68% of cases with minimal context and achieve alignment with scientific standards in 91% of cases when additional context is provided. The model's ability has been demonstrated on two groups: unchanged common names and semantically shifted terms. The results suggest that LLMs can serve as good tools to support the human-guided process of standardizing historical biodiversity datasets. While some challenges remain, LLMs have the potential to enhance the efficiency of the data integration process, particularly by reducing the time of human annotation and supporting the modernization and harmonization of biological records across time.

In the following, we will describe the used dataset in Section 2 and explain the applied prompting strategies in Section 3. Section 4 gives an overview over the conducted experiment and the results. The findings are discussed in Section 5 and we conclude and give an outlook on future research in Section 6. For reproducibility, the scripts used for the experiments are shared on GitHub[1].

## 2. Dataset

This research used the dataset provided by Rehbein et al. [10], which systematically recorded and analyzed a scientific study conducted by the Bavarian government in 1845. The records of this study, consisting of 520 handwritten pages, have been preserved in the Bavarian State Archives[2]. They document the occurrence of 44 selected vertebrate species across Bavaria. The original study, led by zoologist Andreas Wagner, relied on surveys sent to the 119 forestry offices in the state. Rehbein et al. [10] not only published these surveys in a computer-readable format but also provided a deeper analysis, including human-based matching of the recorded historical species to modern-day English common names and equivalent scientific names. This enriched dataset forms the foundation for evaluating the proposed approach to create an LLM-based matching system from historical common names to modern-day scientific and common names.

An inspection of the dataset, conducted in collaboration with domain experts, revealed three primary challenges inherent in working with historical common names. These challenges, while not exhaustive,

[1] https://github.com/biodivportal/historic_biodiversity_term_matching_with_LLM/

[2] BayHStA, Zoologische Staatssammlung, 208-217: https://www.gda.bayern.de/service/findmitteldatenbank/Kapitel/0ea38d12-d425-4b3e-a497-7b6830f439e1

provide an essential baseline for evaluating machine-based matching systems:

1. Changes in spelling: The spelling of historical common names may differ from their modern-day counterparts (e.g., "Murmelt**h**ier" -> "Murmeltier"). Subclasses also include misspellings and misinterpretations.
2. Discontinued or evolved common names: Some historical common names fall out of use, are not clearly identifiable in a historic setting, or scientific advances lead the community to adopt different terms. For example, in 1845, the term "Steinkrähe" (Alpine chough) referred to what is now known as "Alpendohle" (red-billed chough).
3. Redefined scientific concepts or increased granularity: Scientific concepts may be redefined or subdivided into more specific categories. This includes cases where common names historically referred to broader groups, making it challenging to match them directly to a modern scientific equivalent. For instance, "Ente" (duck) now encompasses many modern-day subspecies.

These challenges underscore the complexity of aligning historical biodiversity data with contemporary taxonomies and highlight the importance of advanced tools, such as LLMs, to support this process.

## 3. Prompting to create the matching

This study applies prompt engineering to test the ability of the LLM GPT-4o [11] to align historical biodiversity terms, particularly common species names, with modern scientific taxonomies. GPT-4o was selected for its frequent releases, its model size, and demonstrated human-level performance in various academic benchmarks [12].

The work by Ekin [13] outlines multiple approaches regarding prompt engineering and distinguishes between two variants, basic and advanced approaches. For the basic approach, control codes, templates, iterative testing, and refinement are applied. This approach focuses on fast iterative result driven testing. It is advantageous for rapid prototype development or initial testing of an approach. Advanced techniques, on the other hand, involve strategies such as temperature and token control, prompt chaining, and adapting prompts, which are more suitable for fine-tuning mature systems. Given the exploratory nature of this study, we designed standard prompt engineering templates, emphasizing iterative testing and refinement to achieve accurate results in the majority of cases.

To transfer the common names found in historic text to modern-day common and scientific names, two classes of prompts were designed: one class to translate historic to modern common names and one for matching historic common names to modern day scientific names. Both classes follow the same structure: they start with a simple background setting followed by the task and end with specification about the expected format of the output. An example prompt for historic common names to modern day scientific names is as follows:

"You are a biodiversity expert. In a text you find the species 'species' please provide the correct modern scientific Latin name. Just answer with the scientific Latin name."

In a second step, we tested if the prediction quality of the LLM would increase with more context. In this case, the prompt was adjusted to include the section: "[...] In an old German Bavarian text from 1845 you find the species 'species'.". All example prompts can be found on GitHub[3].

## 4. Initial Experimental Results

An excerpt of the results for predicting common names is displayed in Table 1 (full results available in our GitHub). All matching results were evaluated by two human annotators and categorized into three classes: completely correct matches (e.g., "Dachs" to Badger), partially correct matches (e.g., "B"ar" to Bear instead of Brown Bear), and incorrect matches (e.g., "Trappe" matched to Truffle instead of Bustard).

[3]https://github.com/biodivportal/historic_biodiversity_term_matching_with_LLM/

**Table 1**
Extract of LLM Predictions of historic German Common Species Names to modern day Species Names in English

| Provided Name | LLM Prediction (Context) | LLM Prediction (No Context) | Annotated Common Name |
|---|---|---|---|
| Bär | Bear | Bear | Brown bear |
| Dachs | Badger | Badger | Badger |
| Steinmarder | European stone marten | European Stone Marten | Stone marten |
| Murmelthier | Marmot | Marmot | Marmot |
| Saatkrähe | Rook | Rook | Rook |
| Alpendohle | Alpine Chough | Alpine chough | Alpine chough |
| Steinkrähe | Hooded Crow | Hooded Crow | Stone crow |
| Trappe | Great Bustard | Truffle | Bustard |
| Schnepfe | Snipe | snipe | Sandpiper |
| Saatgans | Bean Goose | Greater White-fronted Goose | Bean goose |
| Enten | Ducks | Ducks | Ducks |
| Kupferotter | European Copper Skink | European adder | Common European adder |

**Table 2**
Evaluation of Prediction Performance in rounded Percent

| LLM Prediction | Completely right | Completely Wrong | Partially right |
|---|---|---|---|
| Common Names no Context | 0.66 | 0.09 | 0.25 |
| Common Names with Context | 0.68 | 0.05 | 0.27 |
| Scientific Names no Context | 0.84 | 0.16 | - |
| Scientific Names with Context | 0.91 | 0.09 | - |

In Table 2, rows one and two display the overall prediction performance of GPT-4o on the common names (44 names) in the historical species dataset. The data shows that most common names were correctly identified without additional context. Including the partially correct matches, 91% of all common names were accurately translated into modern language. The model effectively handled old German spelling, as seen in the "Murmelthier" example. However, it is worth noting that the outdated common name "Alpendohle" did not align with human annotations, failing to account for changes in scientific terminology. Providing a brief context improves prediction quality slightly but does not result in significant changes in error types or introduce new issues.

An excerpt of the predicted results for matched scientific names can be observed in Table 3, with full results available on GitHub. The third and fourth rows in Table 2 show the performance. It is clearly visible that the LLM is able to predict the correct scientific names with high certainty (84%) without any additional context. The prediction quality improves even further when context is provided, reaching 91%.

Notably, spelling did not affect the results in this case, and it appears that the transition in common names from "Steinkrähe" to "Alpendohle" was detected when context was included. Regarding the final challenge of handling broken-down scientific concepts (see "Ente" - duck), although the prediction with context is still incorrect—since there is no single, universally correct answer—the suggestion of "Anas platyrhynchos" is significant, as it is the most common duck species in the Bavarian region. It is clear that without context, the LLM failed to produce a scientific name for the species "Trappe" (Bustard).

## 5. Discussion and Limitations

The results indicate that, without any fine-tuning, GPT-4o can correctly identify common modern names and modern scientific names for historical texts. For the common names, this might seem less impressive, as many of these names remain unchanged. However, the correct identification of scientific names is more advanced and challenging, as this typically requires expertise from domain specialists. It is particularly interesting that GPT-4o accurately captured semantic shifts within common

**Table 3**
Extract of LLM Predictions of German Common Species Names to Scientific Names

| Provided Name | LLM Prediction (Context) | LLM Prediction (No Context) | Annotated Scientific Name |
|---|---|---|---|
| Murmelthier | Marmota marmota | Marmota marmota | Marmota marmota |
| Alpendohle | Pyrrhocorax graculus | Pyrrhocorax graculus | Pyrrhocorax graculus |
| Steinkrähe | Pyrrhocorax graculus | Corvus monedula | Pyrrhocorax graculus |
| Nachtigall | Luscinia megarhynchos | Luscinia megarhynchos | Luscinia megarhynchos |
| Mauerspecht | Jynx torquilla | Dendrocopos major | Tichodroma muraria |
| Auerhuhn | Tetrao urogallus | Tetrao urogallus | Tetrao urogallus |
| Birkhuhn | Tetrao tetrix | Lyrurus tetrix | Tetrao tetrix |
| Schneehuhn | Lagopus lagopus | Lagopus muta | Lagopus muta |
| Trappe | Otis tarda | nan | Otis tarda |
| Schnepfe | Gallinago gallinago | Scolopax rusticola | Scolopax rusticola |
| Bekassine | Gallinago gallinago | Gallinago gallinago | Gallinago gallinago |
| Enten | Anas platyrhynchos | Anas | Aves p.p. |
| Kupferotter | Vipera berus | Vipera ammodytes | Vipera berus |

names (e.g., from "Steinkrähe" (Alpine crow) to "Alpendohle" (Stone crow)) and could predict the correct modern scientific name. Similarly, it was able to account for spelling errors and historical spellings, as demonstrated in examples like “Murmelthier” becoming “Murmeltier,” and it could even make informed guesses for broad categories based on species distribution (“Enten” (ducks) to “Anas platyrhynchos,” the most common duck species in the region where the data was collected). A limitation of this study is the dataset size and scope, which also restrict the ability to test more edge cases in challenging settings.

The study observed that even minimal contextual information can significantly improve prediction quality. Further trials revealed that some encountered prediction errors could not be reproduced, suggesting that this knowledge is indeed encoded within the model. This inability to reproduce is a major risk in the research. As GPT-4o is a commercial model and not open source, it is impossible to assess the training data fully and understand how the predictions are produced or what they are based on. An open-source model might provide greater transparency in this regard.

It is important to note that the matches produced here are not directly linked to a taxonomic resource. However, as shown by the ground truth provided by [10] all identified scientific names were recognizable concepts within a taxonomy. Thus, this research offers valuable insights and represents a step toward automatically matching historical species names to standardized taxonomies. When working with large language models, it is essential to be aware of the risks; since these models are prone to biases and errors, they should not be used as standalone solutions. This research advocates for a “human-in-the-loop” approach, where human annotators are substantially supported by the model in mapping historical terms to modern scientific names.

## 6. Conclusion and Future Work

This research used historical texts and expert-based annotations as the foundation to evaluate whether common historical names for species in the field of biodiversity can be accurately matched to their scientific names in modern-day taxonomy. While there are still challenges, the findings demonstrate that small variations in spelling do not pose a problem, discontinued common names can be matched to modern-day scientific names with minimal context, and knowledge of species distribution is encoded, allowing for assumptions about subspecies addressed in broader common names. With such clear results from an untrained general LLM, the potential of this technology to support human annotation and machine-guided matching is clearly demonstrated.

Future work should focus on comparing different, in the best-case open-source LLMs, evaluating predictions on a broader dataset, and testing more intricate solutions such as fine-tuning or retrieval-augmented generation for suggestions.

## Declaration on Generative AI

The authors have not employed any Generative AI tools.

## References


[1] A. Kohlbecker, N. Karam, A. Paschke, A. Güntsch, Preserving taxonomic change and subsequent taxon relationships over time., in: JOWO, 2021.

[2] E. Svenonius, The intellectual foundation of information organization, 2000.

[3] G. Rouhan, M. Gaudeul, Plant taxonomy: a historical perspective, current challenges, and perspectives, Molecular plant taxonomy: Methods and protocols (2014) 1–37.

[4] N. Karam, J. Fillies, C. Jonquet, S. Bouazzouni, F. Löffler, F. Zander, B. König-Ries, A. Güntsch, M. Diepenbroek, A. Paschke, Biodivportal: Enabling semantic services for biodiversity within the german national research data infrastructure, Datenbank-Spektrum 24 (2024) 129–137. URL: https://doi.org/10.1007/s13222-024-00474-5. doi:10.1007/S13222-024-00474-5.

[5] T. Osawa, N. Tsutsumida, et al., The role of large language models in ecology and biodiversity conservation: Opportunities and challenges (2023).

[6] M. Elliott, J. Fortes, Using chatgpt with confidence for biodiversity-related information tasks, Biodiversity Information Science and Standards 7 (2023) e112926.

[7] S. Govaerts, Biodiversity in the late middle ages: Wild birds in the fourteenth-century county of holland, Environment and History 30 (2024) 241–266. doi:10.3197/096734022X16627150608122.

[8] P. Barkham, Country diary 100 years on: sheep and dogs dominate over rabbits and house martins, The Guardian (2024-09-28) 49–50.

[9] D. S. Viana, F. Blanco-Garrido, M. Delibes, M. Clavero, A 16th-century biodiversity and crop inventory, Ecology 103 (2022) e3783. doi:10.1002/ecy.3783.

[10] M. Rehbein, A. B. Escobari Vargas, S. Fischer, A. Güntsch, B. Haas, G. Matheisen, T. Perschl, A. Wieshuber, T. Engel, Historical Animal Observation Records by Bavarian Forestry Offices (1845), 2024. URL: https://doi.org/10.5281/zenodo.13899541. doi:10.5281/zenodo.13899541.

[11] OpenAI, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, R. Avila, I. Babuschkin, more, Gpt-4 technical report, 2024. arXiv:2303.08774.

[12] S. Minaee, T. Mikolov, N. Nikzad, M. Chenaghlu, R. Socher, X. Amatriain, J. Gao, Large language models: A survey, 2024. arXiv:2402.06196.

[13] S. Ekin, Prompt engineering for chatgpt: A quick guide to techniques, tips, and best practices, 2023. doi:10.36227/techrxiv.22683919.