

基于动物志的凤蝶和弄蝶形态特征数据集

王江宁^{1,3}, 陶秋杰², 李文杰^{1,4}, 卜翠萍^{2*}

1. 中国科学院动物研究所, 北京 100101
2. 淮阴师范学院, 江苏淮安 223300
3. 国家基础学科公共科学数据中心, 北京 100190
4. 中国科学院大学, 北京 100049

摘要: 凤蝶和弄蝶形态特征数据集是在对《中国动物志 昆虫纲 第二十五卷 鳞翅目 凤蝶科 凤蝶亚科 锯凤蝶亚科 绢蝶亚科》和《中国动物志 昆虫纲 第五十五卷 鳞翅目 弄蝶科》相关文字记录数字化处理后, 利用标注工具对内容进行半自动提取得到的。本数据集包含书中记录的鳞翅目凤蝶科凤蝶亚科、锯凤蝶亚科、绢蝶亚科共计 21 属 287 种 (373 种/亚种/型), 弄蝶科共计 83 属 370 种 (640 种/亚种), 从中提取的 20654 条记录以形状、颜色、纹理为主, 是对各物种/亚种不同性别形态描述数据中单一形态特征的提取。本数据集为凤蝶、弄蝶这些类群系统进化、生物地理等生物学研究提供了基础数据, 也为自然语言处理和理解以及大语言模型等研究提供了珍贵的语料资源。

关键词: 形态特征; 凤蝶; 弄蝶; 自然语言处理; 大语言模型

数据库(集)基本信息简介

数据库(集)名称	基于动物志的凤蝶和弄蝶形态特征数据集
数据作者	王江宁, 卜翠萍
数据通信作者	王江宁 (wangjn@ioz.ac.cn)
数据时间范围	2001 年 (凤蝶科), 2015 年 (弄蝶科)
地理区域	中国
数据量	1.44 MB
数据格式	*.xlsx
数据服务系统网址	https://doi.org/10.57760/sciencedb.13653
基金项目	第三次新疆综合科学考察项目 (2021xjkk0600); 科技基础资源调查专项 (2019FY202403)。
数据库(集)组成	本数据集由 1 个数据文件组成, 收集了《中国动物志 昆虫纲 第二十五卷 鳞翅目 凤蝶科》中 287 种凤蝶、《中国动物志、昆虫纲 第五十五卷 鳞翅目 弄蝶科》中 370 种弄蝶的形态学描述数据, 以及从中人工提取的形状、颜色、纹理等形态特征数据。

ISSN 2096-2223

CN 11-6035/N



文献 CSTR:

32001.14.11-6035.csd.2023.0171.zh



文献 DOI:

10.11922/11-6035.csd.2023.0171.zh

数据 DOI:

10.57760/sciencedb.13653

文献分类: 生物学

收稿日期: 2023-12-26

开放同评: 2024-04-11

录用日期: 2024-11-22

发表日期: 2024-12-25

* 论文通信作者

卜翠萍: bucuiping3703@163.com

引言

蝴蝶一直以来备受博物学家、科学家和公众的关注，相关研究涉及物种形成、群落生态学、植物与昆虫互作、拟态、生物多样性保护等多领域，其中蝴蝶分类学研究历史最为悠久，存有大量的形态分类图书资料，如何完成纸质信息的数字化是我们亟待解决的问题。蝴蝶属鳞翅目（Lepidoptera）昆虫，全球蝴蝶分为 17 科，中国蝴蝶有 12 科、434 属、超 2000 种^[1]。我国蝴蝶分类的权威资料主要来源于《中国动物志》。目前《中国动物志》已出版《中国动物志 昆虫纲 第二十五卷 鳞翅目 凤蝶科》^[2]、《中国动物志 昆虫纲 第五十二卷 鳞翅目 粉蝶科》^[3]和《中国动物志 昆虫纲 第五十五卷 鳞翅目 弄蝶科》^[4]。其中，粉蝶科作为标记数据范例，已完成形态特征描述信息的提取工作。为构建更加完善的蝴蝶形态描述模型，我们继续补充提取凤蝶科和弄蝶科种类的形态特征描述信息。

与《基于动物志的鸟类形态性状特征数据集》^[5]（下文简称《鸟集》）和《基于动物志的粉蝶形态性状特征数据集》^[6]（下文简称《粉蝶集》）的产生背景一致，凤蝶、弄蝶的形态特征数据集是为了便于自然语言处理蝴蝶形态信息应用而产生，区别见表 1。本数据集基于收集整理《中国动物志 昆虫纲 第二十五卷 鳞翅目 凤蝶科》《中国动物志、昆虫纲 第五十五卷 鳞翅目 弄蝶科》中的形态描述数据，主要包括外部形态（翅脉、翅形、色彩、斑纹、体型大小、触角形状、口器构造、前足发育程度等）、外生殖器特征和幼期、卵特征，获取结构化的凤蝶、弄蝶形态描述特征数据集。另外数据集中还提取了一些寄主植物和分布的非形态数据。我们采集的数据信息可用于凤蝶科、弄蝶科的形态分类、生物地理、系统发育等研究，更是给自然语言处理、大语言模型标记等工作提供了良好的标注数据。

表 1 本数据集与《粉蝶集》、《鸟集》比较

Table 1 Comparison of this dataset with the *Pierid Dataset* and the *Bird Dataset*

比较项目	本数据集	《粉蝶集》	《鸟集》
数据采集和处理方法	相同		
数据结构组成	凤蝶和弄蝶	粉蝶	鸟类
数据质量控制和评估	相同		
数据价值	对于生物学研究来说，几个数据集应用价值相似。对于自然语言理解等研究工作来说，几个数据集有同样的研究价值。		
数据使用方法	相似，本数据集和《粉蝶集》可以联合使用		

由于本数据集和《粉蝶集》都采用了《鸟集》的采集、质量评估和控制、使用方法，结构组成基本一致，因此这些部分请参考《鸟集》的相关描述，本文此处仅对凤蝶和弄蝶数据集的一些变化详述。

1 数据采集和处理方法

1.1 原始语料数据收集和处理

本数据集的原始语料数据选自《中国动物志 昆虫纲 第二十五卷 鳞翅目 凤蝶科》和《中国动物志、昆虫纲 第五十五卷 鳞翅目 弄蝶科》。《中国动物主题数据库》中的“中国动物志数据库”

(<http://www.zoology.csdb.cn/>)^[7]对两本专著的扫描和文字识别,后整理加工形成数字化语料库,存储并共享了这些原始描述数据。本数据集的原始数据来自《中国动物主题数据库》。

1.2 特征描述标记

本数据集的标注工作依旧沿用编制《粉蝶集》和《鸟集》时用的标注工具(如图1)。

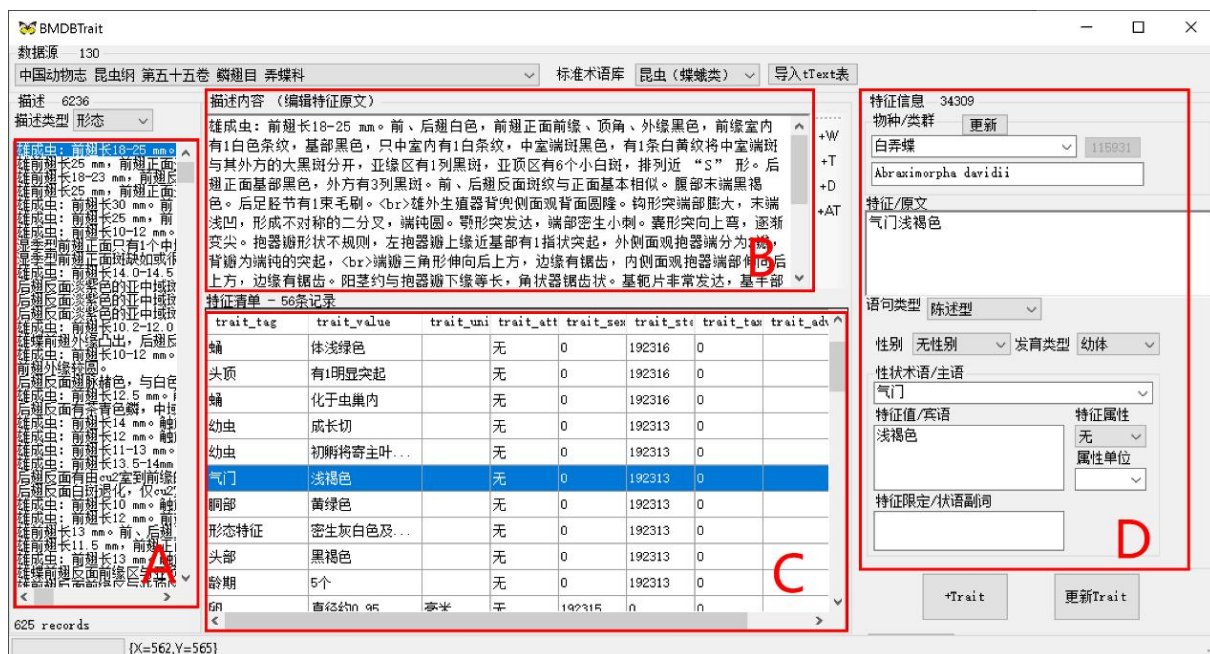


图1 生物文字描述标记工具的界面。A区为原文列表,B区是原文标记区,C区是提取列表,D区是数据编辑区。

Figure 1 User interface of the tool for labeling biology morphology description. A: the original text list; B: the original text for labeling; C: result list; D: data editing area.

工具中“特征术语/主语”项(图1中D区域)与《粉蝶集》类似,自动加载了鳞翅目专业术语,更适合蝶蛾类群的标注工作。

1.3 数据抽取原则

同《粉蝶集》和《鸟集》。

1.4 数据集设计方案

同《粉蝶集》和《鸟集》。

2 数据样本描述

数据集含有 tSpecies (物种信息)、tText (描述信息) 和 tTrait (特征数据) 三个表,每个数据表的字段与《粉蝶集》和《鸟集》基本相同,尽管略有区别,但含义基本一致,以 Excel 格式存储。这里便于读者理解,表2将元数据表格再赘列一遍,并以一组弄蝶数据(包括物种、描述和一条特征数据)作为样例,其中一些元数据的细节规范请参考《粉蝶集》。

表 2 各数据表中字段的元数据描述和示范数据

Table 2 Metadata description and demonstration records in data tables

表名	字段名	数据类型	含义	示范数据
tSpecies	tag_id	数值	物种编号	1159327
	tag_pid	数值	物种上级阶元编号	1159325
	tag_tag	文字	物种中文名称	河伯锷弄蝶指名亚种
	tag_note	文字	备注信息*	-
	tag_latin	文字	物种拉丁名称	<i>Aeromachus inachus</i> subsp. <i>inachus</i>
	tag_level	文字	物种阶元*	亚种
tText	txt_id	数值	原始语料编号	6253
	txt_content	文字	原始语料*	后翅反面翅脉赭色，与白色斑纹反差明显；雄虫体型大，前翅长 13 mm。
	type_Taxa	数值	物种编号	1159327
tTrait	trait_id	数值	特征数据编号	35024
	trait_taxa	数值	物种编号	0
	trait_sex	文字	性别	192314
	trait_tag	文字	性状	前翅长
	trait_pred	文字	语句类型	0
	trait_value	文字	特征取值	13 mm
	trait_stage	文字	发育阶段	长度
	trait_attr	文字	特征属性	毫米
	trait_unit	文字	特征属性取值单位	前翅长 13 mm
	txt_id	数值	原始语料编号	6253

注：表 2 中*是本数据集相对于《粉蝶集》增加的字段，用于存储一些特殊信息描述，多数情况下是默认值（“-”字符）。

3 数据质量控制和评估

本数据集获取方法与《粉蝶集》、《鸟集》类似，但是考虑到两本专著的出版时间差距大，写作风格复杂，除了通过一次人工全部数据校验外，做了两次抽验（增加了一次人工抽样审核工作），尽可能保证数据可靠。同样，由于自然语言中描述的复杂性，特别是类群和作者的写作风格差异构成的文风特征偏移，再加上标记工作人员对语言的理解差异，抽取后的数据描述多样性增加，但在科学性无误，例如表 2 中的 txt_id=6253 示例数据，对 txt_context 的析取出的 trait_id=35022 号记录中 trait_tag 和 trait_value 分别为“反面翅脉”和“赭色”，这个结果并没有错误，但如果析取成“后翅反面翅脉”和“赭色”或许更精准。这类情况主观性较强，在较准中，采用的是“没有错误就不修改”的原则，数量无法精确统计。

4 数据使用方法和建议

本数据集与《粉蝶集》和《鸟集》一脉相承，用法相似，这里提出一些值得特别注意的用法：

(1) 本数据集可与《粉蝶集》联合作为蝴蝶形态特征数据集使用，两个数据集的合集覆盖了已出版的中国动物志中所有蝴蝶类群的形态特征数据，可以构建蝶类形态的知识图谱，用于相关物种鉴定识别、谱系关系分析等研究。本数据集增加了 `tag_level`、`txt_content` 这 2 个字段，在使用时需注意字段的对应。

(2) 因分类学研究的发展，本数据集的原始数据涉及的著作中的分类体系和《中国凤蝶志》^[8]、《中国蝴蝶分类与鉴定》^[9]、《中国蝶类志（修订版）》^[10]等蝶类著作分类体系可能有差异，特别是最新版《中国生物物种名录》^[11]中的分类体系更为明显，在做生物学相关研究时，需要读者根据物种数据表（即 `tSpecies` 表）中的拉丁名或者中文名自行做一些校准工作。

(3) 可以作为自然语言处理和理解研究的语料中的训练集和验证集使用，如：从本数据集 `trait_tag` 字段抽出蝶类形状特征术语模型（即术语表），从 `trait_tag` 和 `trait_value` 字段组合中构建蝶类形态描述模型。这些建构的模型可用于析取其他蝶类书籍中的形态描述数据，在图 1 展示的标记工具中，实际上已经将《粉蝶集》制作中优化形成的形状特征术语模型应用在部分辅助人工标注工作中。另外，由于鳞翅目昆虫的近缘性质，构建的模型或许可以用于蛾类的数据标注工作，这也将是我们后续将尝试的一些工作。

致 谢

感谢中国科学院动物所韩艳、林聪田、纪力强协助整理和提供《中国动物主题数据库》中相关数据。

作者分工职责

王江宁（1982—），男，江苏省淮安市人，博士，馆员，研究方向为信息资源管理。主要承担工作：文章撰写，标记工具研发运维，数据集设计、编审、维护。

陶秋杰（2001—），女，江苏省宿迁市人，本科生，研究方向为生物分类学。主要承担工作：数据标记和校审。

李文杰（1991—），女，河南省周口市人，博士研究生，研究方向为昆虫功能与形态学。主要承担工作：数据校准。

卜翠萍（1982—），女，山东省东平县人，博士，副教授，研究方向为昆虫分类。主要承担工作：数据标记和编审，文章校审。

参考文献

- [1] 寿建新. 蝴蝶分类系统及最新数据[J]. 西安文理学院学报：自然科学版, 2010, 13(3):91-102. [SHOU J X. Butterfly Classification System and the Latest Data, Journal of Xi'an University of Arts & Science[J], 2010, 13(3):91-102.]
- [2] 武春生. 中国动物志 昆虫纲 第二十五卷 鳞翅目 凤蝶科 凤蝶亚科 锯凤蝶亚科 绢蝶亚科[M]. 北京：科学出版社，2001. [WU C S. Fauna of Sinica: Insecta Vol. 25: Lepidoptera: Papilionidae: Papilioninae, Zerynthiinae, Parnassiinae [M]. Beijing: Science Press, 2001.]

- [3] 武春生. 中国动物志 昆虫纲 第五十二卷 鳞翅目 粉蝶科[M]. 北京: 科学出版社, 2010. [WU C S. Fauna of Sinica: Insecta Vol. 52: Lepidoptera: Pieridae [M]. Beijing: Science Press, 2010.]
- [4] 袁锋, 袁向群, 薛国喜. 中国动物志 昆虫纲 第五十五卷 鳞翅目 弄蝶科[M]. 北京: 科学出版社, 2015. [YUAN F, YUAN X Q, XUE G X. Fauna of Sinica: Insecta Vol. 55: Lepidoptera: Hesperidae [M]. Beijing: Science Press, 2015.]
- [5] 薛延韬, 王江宁, 林聪田, 等. 基于动物志的鸟类形态特征数据集[J/OL]. 中国科学数据, 2021, 6(3). (2021-03-24). DOI: 10.11922/csdata.2020.0085.zh. [XUE Y T, WANG J N, LIN C T, et al. A dataset of morphological traits of birds based on Fauna Sinica [J/OL]. China Scientific Data, 2021, 6(3). (2021-03-24). DOI: 10.11922/csdata.2020.0085.zh.]
- [6] 李文杰, 王江宁, 卜翠萍, 等. 基于动物志的粉蝶形态特征数据集[J/OL]. 中国科学数据, 2022, 7(1). (2021-12-21). DOI: 10.11922/11-6035.csd.2021.0063.zh. [LI W J, WANG J N, BU C P, et al. A dataset of morphological traits of pierid based on Fauna Sinica [J/OL]. China Scientific Data, 2022, 7(1). (2021-12-21). DOI: 10.11922/11-6035.csd.2021.0063.zh.]
- [7] 中国科学院动物研究所. 中国动物主题数据库[DB/OL]. 2009. <http://www.zoology.csdb.cn>. [Institute of Zoology, Chinese Academy of Sciences. Chinese Animal Theme database [DB/OL]. 2009. <http://www.zoology.csdb.cn>.]
- [8] 白九维, 王效岳. 中国凤蝶志[M]. 新北: 淑馨出版社, 1998. [BAI J W, WANG X Y, Swallowtail Butterflies in China[M]. Xinbei: Shuxin Press, 1998.]
- [9] 周尧. 中国蝴蝶分类与鉴定[M]. 郑州: 河南科学技术出版社, 1998. [ZHOU Y. Classification and identification of Chinese butterflies[M]. Zhengzhou: Henan Science and Technology Press, 1998.]
- [10] 周尧. 中国蝶类志(修订版)[M]. 郑州: 河南科学技术出版社, 2000. [ZHOU Y. Butterfly of China [M]. Zhengzhou: Henan Science and Technology Press, 2000.]
- [11] The Biodiversity Committee of Chinese Academy of Sciences. Catalogue of Life China: 2023 Annual Checklist[DB/OL]. 2023. <http://www.sp2000.org.cn>.

论文引用格式

王江宁, 陶秋杰, 李文杰, 等. 基于动物志的凤蝶和弄蝶形态特征数据集[J/OL]. 中国科学数据, 2024, 9(4). (2024-12-25). DOI: 10.11922/11-6035.csd.2023.0171.zh.

数据引用格式

王江宁, 卜翠萍. 基于动物志的凤蝶和弄蝶形态特征数据集[DS/OL]. V1. Science Data Bank, 2024. (2024-04-11). DOI:10.57760/sciencedb.13653.

A dataset of morphological traits of Papilionidae and Hesperidae based on *Fauna Sinica*

WANG Jiangning^{1,3}, TAO Qiujie², LI Wenjie^{1,4}, BU Cuiping^{2*}

1. Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, P.R. China

- 2. Huaiyin Normal University, Huai'an 223300, P.R. China
- 3. National Basic Science Data Center, Beijing 100190, P.R. China
- 4. University of Chinese Academy of Sciences, Beijing 100049, P.R. China

*Email: bucuiping3703@163.com

Abstract: The dataset of morphological traits of papilionid and hesperid based on Fauna Sinica is extracted by using a semi-automatic labeling tool after digitizing the records of the INSECTA Vol. 25 Lepidoptera Papilionidae, and INSECTA Vol. 52 Lepidoptera Hesperidae both in the book *Fauna Sinica*. This dataset contains 287 species (373 species/subspecies/types) from 21 genera in Papilioninae, Zerynthiinae and Parnassiinae of Papilionidae, and 370 species (640 species/subspecies/types) from 83 genera in Hesperidae. The 20,654 records are extracted primarily in terms of shape, color and texture, representing the extraction of single morphological features from the morphological descriptions of different species/subspecies, such as different gender forms and quantitative measures. This dataset can provide basic data for biological studies on phylogeny and biogeography, and serve as valuable corpus resource for information fields like natural language processing and understanding as well as large language model.

Keywords: morphological trait; Papilionidae; Hesperidae; natural language processing; large language model

Dataset Profile

Title	A dataset of morphological trait of Papilionidae and Hesperidae based on <i>Fauna Sinica</i>
Data corresponding author	WANG Jiangning (wangjn@ioz.ac.cn)
Data authors	WANG Jiangning, BU Cuiping
Time range	2001 (Papilionidae), 2015 (Hesperidae)
Geographical scope	China
Data volume	1.44 MB
Data format	*.xlsx
Data service system	< https://doi.org/10.57760/sciencedb.13653 >
Sources of funding	The Third Xinjiang Scientific Expedition Program (2021xjkk0600); Basic Resources Investigation Project of Science and Technology (2019FY202403).
Dataset composition	The dataset consists of one data file, collecting the morphological description data of 287 species of swallowtail butterflies from the INSECTA Vol. 25 Lepidoptera Papilionidae and 370 species of hesperids from the INSECTA Vol. 55 Lepidoptera Hesperidae both in the book <i>Fauna Sinica</i> , along with artificially-extracted morphological feature data (e.g. shape, color, and texture).