专题 中国生物多样性

基于动物志的鸟类形态特征数据集

薛延韬[1,2,3]，王江宁[1,3]，林聪田[1,2,3]，韩艳[1,3]，丁晓庆[1,3]，纪力强[1,3*]

1. 中国科学院动物研究所，北京 100101
2. 中国科学院大学，北京 100049
3. 国家基础学科公共科学数据中心，北京 100190

摘要：基于动物志的鸟类形态特征数据集是在对《中国动物志 鸟纲 第七卷》相关文字记录数字化处理后，利用专门工具对内容进行提取得到的。本数据集包含书中记录的夜鹰目、雨燕目、咬鹃目、佛法僧目、啄木鸟目（鴷形目）共计5目13科85种鸟类，其每条记录都是对各物种/亚种不同性别各发育阶段的形态、生物学、分布、量衡度等形态描述数据中单一形态特征的提取，以形状、颜色、纹理为主。本数据集为这些类群鸟类生物地理、系统进化等生物学研究提供了基础数据，也为自然语言处理和理解研究提供了珍贵的语料资源。

关键词：形态特征；形态学；鸟纲；自然语言处理

ISSN 2096-2223
CN 11-6035/N

文献 DOI：
10.11922/csdata.2020.0085.zh
数据 DOI：
10.11922/sciencedb.j00001.00117
文献分类：生物学

收稿日期：2020-08-29
开放同评：2021-01-27
录用日期：2021-08-26
发表日期：2021-09-29

数据库（集）基本信息简介

数据库（集）名称	基于动物志的鸟类形态特征数据集
数据作者	薛延韬，王江宁，林聪田，韩艳，丁晓庆，纪力强
数据通信作者	纪力强（ji@ioz.ac.cn）
数据时间范围	2003年
地理区域	中国
数据量	361 KB
数据格式	*.xlsx
数据服务系统网址	http://www.dx.doi.org/10.11922/sciencedb.j00001.00117
基金项目	中国科学院战略性先导科技专项（XDA19050203）；科技基础资源调查专项（2019FY202403）；中国科学院“十三五”信息化建设专项（XXH13505-03-102）。
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 鸟纲 第七卷》中5目13科85种鸟类的形态学描述数据，以及从中人工提取的颜色、形状、纹理等形态特征数据。

引 言

生物形态特征提取工作是将志书、图谱、论文等文献资料中有关物种特点的文字描述进行信息抽取，形成结构化数据的过程。结构化数据具有可量化或可类化的特点，便于计算机处理。可量化的数据指的是能以数值衡量的数据，可进行数值计算；可类化的数据指的是离散的定性描述型数据，一般都有明确的取值类

* 论文通信作者
纪力强：ji@ioz.ac.cn

型，类似枚举型数据结构的概念。另外还有一些特殊类型的数据：时间型数据是在时间维度上的数据，通常和日期、季节、时间、发育阶段等有关，作为描述的状语或限定词出现；比较型数据是针对两个以上描述对象进行特征比较的特殊数据类型，这类数据主要是针对同一个性状或特征在不同的时间或类群之间进行比较的结果，是相对值。

形态特征对于生物学研究非常重要，是生物地理、系统进化等许多生物学科的基础；此外，它也是生物学和计算机方法交流的基础[1]。在当前生物学各分支学科逐渐由定性向定量研究发展的时代，形态特征数据也逐渐从记录走向应用。提取中文形态特征数据，不仅是对全球生物性状数据的补充完善，还能对与计算机、人工智能等学科融合起一定推波助澜作用。

由于我国的生物学起步较迟，许多基础生物学数据，尤其是形态特征，常以中文文字描述的自由文本的形式记录于各类志书、图谱、论文等文献资料中，数据数字化程度不高，可用的数据库则更少。为此，我们以动物志中鸟类相关资料为对象，尝试从历史文献中提取形态特征描述信息，构建鸟类形态特征数据库。

1 数据采集和处理方法

1.1 原始语料数据收集和处理

原始语料数据来自对《中国动物志 鸟纲 第七卷》[2]的扫描和文字部分识别，将包含有形态、生物学、分布、量衡度等内容的文本信息进行数字化处理，最终抽提各物种/亚种形态学的描述部分，形成原始语料库。这些数字化的原始描述数据可以从《中国动物主题数据库》中的“中国动物志数据库”（http://www.zoology.csdb.cn/）获得。

1.2 形态特征标记

为便于对上述语料库进行处理，我们基于 C#语言设计开发了一套可用于 window 平台的生物形态特征标记专用工具。其主要功能是对数字化的原始描述数据中所包含的对每个物种每个部位的特征描述进行标记，并形成特征清单。该工具的工作界面如图 1 所示。标记的主要过程如下：

第一步：选择标记语料。选择区域 A 中的清单后，对应的具体语料会在区域 B 中完整展示。这些语料已经完成数字化，附有物种对应的描述类型信息。后续标记工作都基于区域 B 中的语料开展。

第二步：标记每个形态描述。在区域 B 中选择文字描述内容的“标注短语”，点击“+W”按钮自动识别特征原文，此时识别内容会自动填入区域 C 中。

第三步：编辑语料信息。在区域 C 中对识别后的特征名词/术语/主语、特征/属性值/宾语、特征/属性、属性单位等项目进行编辑。若“特征名词/术语/主语”识别不准确，可在“特征/原文”处选择特征名，单击“+T”即可重新识别；若“特征/属性值/宾语”识别不准确，可在“特征/原文”处选择特征属性值，单击“+D”即可重新识别，同时手动修正其他数据。

第四步：特征数据编辑确认无误后，点击“+Trait”按钮即可成功添加一条特征清单，显示在区域 D 中。

此外，对于个体形态特征描述比较简洁明了的内容，在第二步时可选择自动提取整段内容的快捷方式，即可快速完成对该物种特征清单的添加。此操作结束后仍需对特征清单进行逐条审核，如有误则在右侧进行逐条编辑修改，最后点击“更新 Trait”按钮即可。



图 1　本研究开发和使用的生物形态特征标记工具的工作界面

1.3　数据抽取原则

在生物形态特征的抽取过程中，依照如下操作原则进行：

原则 1：追寻本意。基于作者表达的真实意思进行提取，提取出的量化特征不与原文表达的意义相悖。

原则 2：原子提取。对语料中出现的性状描述尽可能进行量/类化和细化，每一条清单内容是对该物种某一形态的单一特征描述，即能抽象成“1 特征取 1 值”的形式。

原则 3：体现核心。形态和描述是一条数据的关键两部分组成，需要按描述重点对语料划分，通常位置、部位、方位等定位描述用于修饰性状的细节结构，一般归置到形态主语部分；其他如颜色、形状、数量等记录作为描述部分。

原则 4：保存原样。对于表达复杂且无法简单拆分的语句，则保留原始描述，待后续分析。

1.4　数据集设计方案

综上所述，本数据集包括原始语料数据和标记后的形态特征数据两类核心实体。按照数据库设计的一般性原则和实际使用情况，我们将原始语料数据设计成两张数据表，一张用于记录物种及其分类系统关系，另一张用于记录形态学的原始语料数据；而形态特征数据被设计成另一张表。各实体在 Excel 中分别用 tSpecies、tText 和 tTrait 表记录，实体之间的关系如图 2 所示。

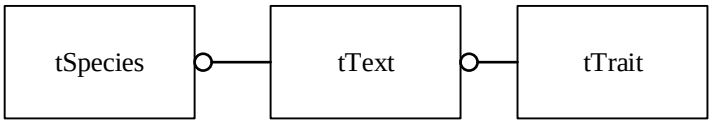


图 2　数据实体之间的关系示意图

2 数据样本描述

数据集以 Excel 表格形式进行存储，包含 3 个数据表单：第 1 个为 tSpecies 数据表，用于记录物种及其分类系统关系；第 2 个为 tText 数据表，用于记录形态学的原始语料数据；第 3 个为 tTrait 数据表，用于记录形态特征数据。3 张数据表中实体对应的元数据描述和示范数据分别见表 1–3。本数据集结构较简单，此设计便于模式识别研究人员直接使用或者稍作调整后使用。需要说明的是，tTrait 数据表中一些字段由默认值 0 填充，这些默认值一般是在原始语境中作为一种默认的上下文情景，文中没有加以说明，反而是有特定取值的情况会进行特殊说明，因此在数据表设计时候采用了大量默认值设计。各字段默认值含义在“取值说明”栏都加以说明。

表 1　tSpecies 数据表中字段的元数据描述和示范数据

字段名	sp_id	sp_pid	sp_name	sp_latin
数据类型	数值	数值	文字	文字
主/外键	主键			
含义	物种编号	物种上级阶元编号	物种名称	物种名称
取值说明	自动编号	取值为 sp_id 中的值	物种中文学名	物种拉丁学名
示范数据	1127620	0	佛法僧目	Coraciiformes
	1127625	1127620	翠鸟科	Alcedinidae
	1127637	1127625	翠鸟属	Alcedo Linnaeus, 1758

表 2　tText 数据表中字段的元数据描述和的示范数据

字段名	txt_id	txt_content	sp_id
数据类型	数值	文字	数值
主/外键	主键		外键
含义	原始语料编号	原始语料	
取值说明	自动编号		值取自 tSpecies 表
示范数据	849	跗蹠被羽；腰部无暗色羽干；下体灰褐。翼长 122–130 mm。	193208
	858	形态 (依据指名亚种描述。) 　　成鸟 头、颈灰褐色，各羽末端具白色细的羽缘；背、腰黑褐色，各羽末端也具细的白色羽缘。尾上覆羽基部白色，形成一约 14 mm 宽的大白斑，白色斑的各羽具极细的黑色羽干纹；尾上覆羽的端部和尾羽黑褐色。翼的各羽的外羽片为黑褐色，内羽片为灰褐色。颏、喉、前颈为白色；胸、腹、尾下覆羽为灰褐色，各羽端部均有较宽的白色羽缘。 　　虹膜暗褐；嘴黑；跗蹠和趾紫黑，爪黑褐色。 	193220

注：此表数据中含有 html 标识符，使用者请自行过滤。

表 3　tTrait 数据表中字段的元数据描述和的示范数据

字段名	trait_id	sp_id	trait_sex	trait_stage	trait_tag	trait_pred	trait_value	trait_attr	trait_unit	note	txt_id
数据类型	数值	数值	文字	文字	文字	文字	文字	文字	文字	文字	数值
主/外键	主键										外键
含义	特征数据编号	物种编号	性别	发育阶段	性状	语句类型	特征取值	特征属性	特征属性的取值单位	原始性状描述	原始语料编号
取值说明	自动编号	0：默认值，指继承于原始段落的类群属性（即对应 tText 表的 sp_id 值）。其他值：一般存在于和其他鸟类进行比较的语段中。	0：默认值，指不区分性别（即两性的特征一样）。其他值：雄性，雌性	0：默认值，指成鸟/成体/成年个体。其他值：幼鸟	必填字符串，对于不同领域（或描述角度），对应的术语也不同：①分类/形态/鉴定/度量衡，对应解剖结构术语；②生态/生物学，对应行为、环境术语；③生物学，对应发育、生理等术语。	0：默认值，指陈述型 其他值：1118932 比较型，1118934 因果型	一般是数值、文字或专业术语。	包括大小、长度、宽度、颜色、形状、纹理等属性名词，不选择时为默认值“无”；对一些表达比较丰富的数据值进行区分，如长度值分为均值和范围。	一般指长度和重量单位，其他属性无单位。	必填项目	值取自 tText 表
示例数据	572	0	0	0	跗蹠	0	被羽	无		跗蹠被羽	849
	574	0	0	0	下体	0	灰褐	颜色		下体灰褐	849
	577	0	0	0	腰部	0	无暗色羽干	无		腰部无暗色羽干	849
	578	0	0	0	翼长	0	122–130	范围	毫米	翼长 122–130 mm	849

3 数据质量控制和评估

本数据集采用的生物形态特征提取方法经过了长期的探索和实践，开发并逐步完善了专用标记工具（如图 1 所示），用于在特征提取中的规范化和简洁化操作。在提取形态特征的过程中，所面临的描述语段往往非常复杂，存在诸如无关语句、比较句型、倒置句型等语句层问题，以及多主语、主语指代不明、细微结构词、复杂限定状语等句式问题，且内容与作者的写作风格、描述对象的特征、写作时的文化环境有紧密相关性。因此，在长期探索和实践的过程中，我们提出了形态特征提取原则以及问题解决方案，形成了一套独特的形态特征提取方法，最大程度保证数据库的准确性和实用性。同时，我们还设计开发并逐步完善了生物形态特征提取工具，使数据提取过程标准化、程序化和规范化，且该工具经过多人验证使用，运行稳定可靠。此外，对于提取后的形态特征值，我们仍然进行了多次人工抽样核查，以保证标记结果的准确性。

4 数据使用方法和建议

本数据集用途较多，可以：

（1）作为相关鸟类的形态特征数据集使用，用于相关物种的谱系关系分析等研究。

（2）用作自然语言处理研究中的语料，作为训练集和验证集使用。

本数据集数据形式简单，在使用时注意：

（1）形态特征数据应与文本编码和物种/类群编码结合使用。

（2）目前对于形态特征的提取尚无统一标准，本数据集所使用的原则和方法已在本文中说明。如有异议请参考原文，以原文表述为准。

（3）本表数据集的原始数据采集自谭耀匡和关贯勋于 2003 年编著的《中国动物志 鸟纲 第七卷》[2]，书中的分类体系和当前《中国生物物种名录》[3]中的新分类体系有差异，因此在使用物种数据表（即 tSpecies 数据表）中的数据时，请根据物种的中文名或者拉丁名自行校对。

致 谢

感谢中国科学院软件研究所马龙龙在数据审核校准中所做的工作。

数据作者分工职责

薛延韬（1992—），男，山西省孝义市人，博士研究生，研究方向为生物多样性信息学。主要承担工作：论文撰写和数据标记。

王江宁（1982—），男，江苏省淮安市人，博士，助理研究员，研究方向为昆虫图像识别。主要承担工作：数据集设计和整理，数据标记和校准。

林聪田（1982—），男，福建省漳州市人，博士研究生，工程师，研究方向为生物多样性信息学。主要承担工作：数据集原始数据整理。

韩艳（1972—），女，北京市人，学士，工程师，研究方向为生物多样性信息学。主要承担

工作：数据集原始数据采集和整理。
丁晓庆（1989—），女，河南省商丘市人，学士，科研助理。主要承担工作：数据标记。
纪力强（1961—），男，北京市人，博士，研究员，研究方向为生物多样性信息学。主要承担工作：总体规划和方案设计。

参考文献

[1] 谢强, 卜文俊, 王昕, 等. 现代动物分类学导论[M]. 北京: 科学出版社, 2012.
[2] 谭耀匡, 关贯勋. 中国动物志 鸟纲 第七卷 夜鹰目 雨燕目 咬鹃目 佛法僧目 鴷形目[M]. 北京: 科学出版社, 2003.
[3] 中国科学院生物多样性委员会. 中国生物物种名录 2020 版 [EB/OL]. (2020–05–25)[2020–08–25]. http://www.sp2000.org.cn.

论文引用格式

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.

数据引用格式

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[DB/OL]. Science Data Bank, 2021. (2021-01-27). DOI: 10.11922/sciencedb.j00001.00117.

A dataset of morphological traits of birds based on *Fauna Sinica*

XUE Yantao[1,2,3], WANG Jiangning[1,3], LIN Congtian[1,2,3], HAN Yan[1,3], DING Xiaoqing[1,3], JI Liqiang[1,3*]

1. Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, P.R. China
2. University of Chinese Academy of Sciences, Beijing 100049, P.R. China
3. National Basic Science Data Center, Beijing 100190, P.R. China

*Email: ji@ioz.ac.cn

Abstract: The dataset of morphological traits of birds based on *Fauna Sinica* is obtained through content extraction by using special tools after digitizing the related records of the book *Fauna Sinica (Aves 7)*. This dataset contains 85 species of birds, 13 families in 5 orders recorded in the book, including *Caprimulgiformes, Apodiformes, Trogoniformes, Coraciiformes, and Piciformes*. Each record in this dataset is an extraction of the single morphological characteristics from morphology, biology, distribution, measurement, and other morphological description data of different species/subspecies in different genders and various developmental stages, mainly in shape, color, and texture. This dataset provides not only basic data for the biological research on these birds groups, such as biogeography and phylogeny, but also valuable corpus resources for natural

language processing and understanding research.

Keywords: morphological trait; morphology; Aves; natural language processing

Dataset Profile

Title	A dataset of morphological traits of birds based on *Fauna Sinica*
Data corresponding author	JI Liqiang (ji@ioz.ac.cn)
Data authors	XUE Yantao, WANG Jiangning, LIN Congtian, HAN Yan, DING Xiaoqing, JI Liqiang
Time range	2003
Geographical scope	China
Data volume	361 KB
Data format	*.xlsx
Data service system	<http://www.dx.doi.org/10.11922/sciencedb.j00001.00117>
Sources of funding	Strategic Priority Research Program of the Chinese Academy of Sciences (XDA19050203); Basic Resources Investigation Project of Science and Technology (2019FY202403); 13th Five-year Informatization Plan of Chinese Academy of Sciences (XXH13505-03-102).
Dataset composition	The dataset consists of 1 data file, which collects the morphological description data of 85 species of birds, 13 families in 5 orders, as well as color, shape, texture and other characteristic data artificially extracted from the book *Fauna Sinica (Aves 7)*.