中国生物物种名录及物种数据整合

纪力强[1] 王利松[2] 乔慧捷[1] 覃海宁[2] 马克平[2]

[1]中科院 动物研究所 [2]中科院 植物研究所

【摘要】中国生物物种名录的最终目标是收集所有在中国分布的所有动物、植物、微生物和菌物物种的分类学基本信息。它的最新版本是 2009 版，包括了高等植物（被子植物、裸子植物、蕨类、苔藓）、脊椎动物（哺乳动物、鸟类、鱼类、两栖动物、爬行动物）、蜘蛛和跳小蜂等生物类群的 44,185 个物种的名录信息。中国生物物种名录的物种信息来自于数个中国物种数据库，如中国高等植物名录数据库、中国蜜蜂数据库、中国蜘蛛数据库和中国蝴蝶数据库等，数据标准遵从国际物种 2000 项目的标准数据集标准。使用者可以通过网站和光盘两种途径获得中国生物物种名录的信息。

在编辑整合中国生物物种名录的过程中，面临着诸多问题，如分类系统的选择、数据来源的规范化、编辑数据手段的多样化、中名的选择、拉丁名的核对、分布区的标准化、中文冷僻字的存储和显示等。在过去三年中，这些问题逐步得到了解决。

【Abstract】The ultimate goal of the Catalogue of Life China (CoL China) is to collect the basic data of all species of animal, plant, microbe and fungus distributed in China. The newest version of CoL China, 2009 Annual Checklist, contains data of 44,185 species from higher plant, vertebrate, spider and Encyrtid. The data of Col China are integrated from several China Species Databases, such as Catalogue of Life: Higher Plants in China, China Bee Species Database, China Spider Species Database and China Butterfly Species Database. The data structure is same with the standard dataset of global Catalogue of Life. Users could access Col China data from CD or website.

In the process of compiling the CoL China, many issues were raised and got resolved in last 3 years, such as choice of taxonomic system, standardization of data

source, choice of methods for species data editing and checking, choice of Chinese name, checking up the scientific name, standardization of distribution, storing and display the rare characters in Chinese.

【关键词】物种名录、数据标准、数据整合

【**Keywords**】species checklist、data standard、data integration

生物物种是生物多样性最重要的组成单元，直接展示着色彩纷呈的大自然的神奇。物种名录是生物物种的核心信息，不仅在微观上描述了每个物种所在的分类学位置，而且从整体上揭示了全球或一个地区的生物多样性丰富度水平。

从上世纪 90 年代起，一些从事生物多样性研究的科学家意识到物种名录的重要性。1996 年，几名科学家聚集在菲律宾的鹰角（Eagle Point）海边，一面享受着阳光海水的遐意，一面探讨着建立全球生物物种名录的可能性。他们的讨论，为紧接着举办的研讨会确定了基调，促成了国际物种 2000 组织（Species 2000）的诞生，为全球生物物种名录（Catalogue of Life, CoL）的编制奠定了组织基础。2003 年，国际物种 2000 组织与综合分类信息系统（Integrated Taxonomic Information System, ITIS）签署备忘录，合作编制并发布分年度的全球生物物种名录（Annual Checklist of CoL）和动态的全球生物物种名录（Dynamic Checklist of CoL）。

2006 年，物种 2000 中国节点成立，并分别在 2008 年和 2009 年正式发布了中国生物物种名录（Catalogue of Life China）2008 版和 2009 版年度名录。

1、 中国生物物种名录

物种 2000 中国节点的目标是收集在中国分布的所有动物、植物、微生物和菌物物种的名录信息，形成中国生物物种名录，并向全球用户提供电子化形式的共享名录数据。其中，物种名录数据主要是指所有生物物种的科学名及其它们所构成的一个分类系统等分类学基本信息，通常还包括相关的原始文献。

2009 年完成的中国生物物种名录 2009 版，收录了高等植物（被子植物、裸子植物、蕨类、苔藓）、脊椎动物（哺乳动物、鸟类、鱼类、两栖动物、爬行动物）、蜘蛛和跳小蜂等生物类群的 44,185 个物种的名录信息。各个类群的统计数据见表一。

表一 中国生物物种名录 2009 版各类群物种数统计

生物类群	纲	目	科	属	种	种下	异名	别名
Bryophyta - 苔藓	4	21	116	520	2,571	178	3,958	88
Pteridophyta - 蕨类	6	11	63	221	2,267	166	6,494	388
Gymnospermae - 裸子植物	4	8	12	42	244	72	783	334
Angiospermae - 被子植物	2	61	242	3,158	29,583	4,793	53,762	19,098
Araneae - 蜘蛛目	1	1	66	618	3,300	0	1,853	0
Encyrtidae - 跳小蜂科	1	1	1	129	405	0	80	0
Agnatha -无颌纲	1	2	2	3	8	0	13	17
Pisces - 鱼纲	1	42	294	1,169	3,225	18	4,823	4,623
Amphibia - 两栖纲	1	3	12	57	346	0	1,752	115
Reptilia - 爬行纲	1	3	25	125	403	19	1,734	117
Aves - 鸟纲	1	21	83	401	1,269	0	2,374	3,358
Mammalia - 哺乳	1	14	53	236	564	3	689	1,473
Total - 合计	24	188	969	6,679	44,185	5,249	78,315	29,611

与 2008 版相同，2009 版中国生物物种名录的物种信息主要来自于中国高等植物名录数据库（Catalogue of Life: Higher Plants in China，简称：CNPC）（www.cnpc.ac.cn）和中国动物信息网（Chinese Animal Information System）两个数据库系统。中国高等植物名录数据库收录了在中国分布的被子植物、裸子植物、蕨类和苔藓等植物类群共 34,465 种植物的名称、分类系统、分布、文献和经济用途等信息。中科院植物研究所是 CNPC 的主编单位，在编撰过程中，充分利用自身优势，最大限度地依靠专家。近 3 年来，共邀请 100 余位专家对数据进行了 3 次审核。2009 年度共有 79 名专家参加审核，其中 65 位是《中国植物志》作者或《Flora of China》作者，或是（中方）作者的学生，或作者所在课题组的成员，并得到作者的指导。审核专家来自中科院植物研究所、中科院华南植物园和中科院昆明植物研究所等 26 家单位。CNPC 的最终目标是建设权威的中国高等植物物种综合信息系统，这是一个持续更新和维护的分类学信息标准。中国动物信息网由中科院动物研究所牵头建立，收集了在中国分布的 31,000 种（亚种）脊椎动物、无脊椎动物和昆虫的物种信息，包括科学名、异名、别名、分布、特征描述、生境、模式标本、文献等方面的综合信息，是一个学科综合数据库系统（Disciplinary Integrated Databases/Systems）。而此学科综合数据库系统的信息则来源于数个中国物种数据库，如中国蜜蜂数据库、中国蜘蛛数据库和中国蝴蝶数据库等。先后有 40 多位动物（昆虫）分类学专家为中国动物信息网提供了科学数据。

中国生物物种名录的数据标准遵从国际物种 2000 项目的标准数据集标准，每个物种的数据项包括：

- 接受的科学名（accepted scientific name）和相关的文献
- 异名（synonym）和相关的文献
- 别名（common name）和相关的文献
- 最终分类学审核人和审核时间（latest taxonomic scrutiny）
- 源数据库（source database）及其网址
- 科名（family to which species belongs）

- 科以上分类阶元（classification above family, and highest taxon in database）
- 分布区（distribution）
- 参考文献（reference）
- 审核人信息

目前，除通过光盘获得全部的名录信息外，使用者还可以通过物种 2000 中国节点的网站（http://www.sp2000.cn/）查询最新版本的中国生物物种名录数据。光盘数据管理系统和网站的后台程序一样，均提供了分类树浏览和关键词搜索等查询功能。网站还提供了几个 web service 功能（http://webservice.sp2000.cn/）：

- 根据物种 ID 查询物种信息
- 根据物种科学名查询物种信息（全匹配方式）
- 根据物种科学名查询物种信息（部分匹配方式）
- 根据物种科学名查询物种信息（模糊查询方式）
- 根据物种科学名、别名和分布地查询
- 根据科名下载科的名录

使用者可以在自己的应用程序中使用这些 web service，进行数据分析、展示和综合。

2、 物种数据整合中的有关问题

中国生物物种名录由中科院生物多样性委员会提供资助，由物种 2000 中国节点负责组织分类学专家队伍收集整理信息，中科院植物研究所和动物研究所分别负责植物和动物方面数据汇总和编辑，并提供了网站和光盘系统信息技术的支持。100 多位分类学专家提供或审核了 2009 版中国生物物种名录的物种数据。图 1 显示了物种名录数据收集整理的过程。图的上半部是动物数据整理的

流程，下半部是植物数据整理的流程。二者的区别在于数据规范化与专家审核的先后次序，动物部分的数据先经过规范化处理，然后再由专家审核；植物部分的数据先由专家审核，然后再进行规范化处理。有在规范化处理时要对不规范或缺失的数据进行处理，因此，植物部分需要增加一个二次审核的环节。

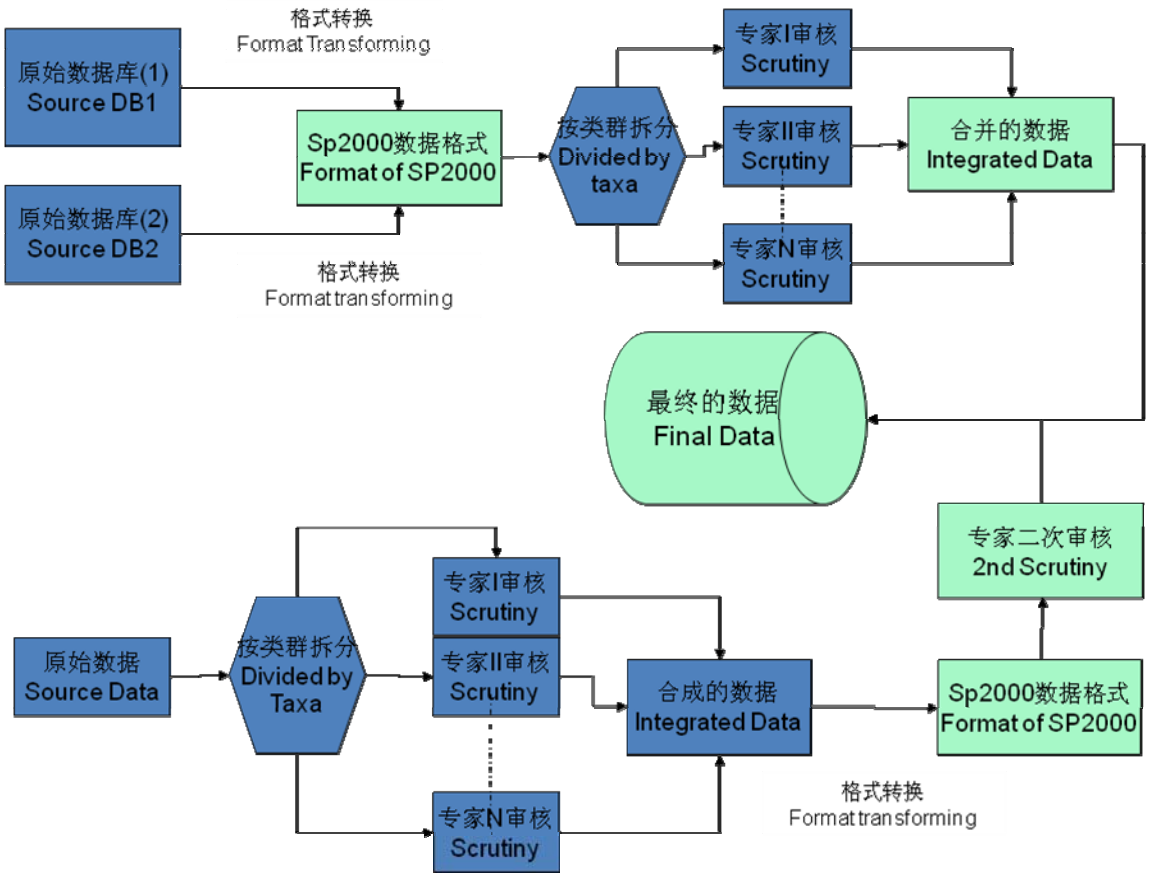

图 1 中国生物物种名录数据收集整理过程示意图

在编辑整合中国生物物种名录的过程中，面临着诸多问题，如分类系统的选择、数据来源的规范化、编辑数据手段的多样化、中名的选择、拉丁名的核对、分布区的标准化、中文冷僻字的存储和显示等。在过去三年中，这些问题逐步得到了解决。

（1） 分类系统的选择

随着分类学研究的进展，分类系统总是处于不断的变化中。由于科学家对某些生物类群的分类概念存在认识上的差异，有的生物类群同时存在两个或两个以上的分类系统。如何选择分类系统，是建立物种名录首先面临的问题。物种 2000 中国节点在当初建立的时候成立了一个工作组。工作组针对每个大的生

物类群指定一名或两名首席专家。首席专家根据工作类群的大小，可以再聘请若干专家组成类群专家组，通过讨论协商的方式确定该类群采用的分类系统。

（2） 数据来源的规范化

在以前建立与生物物种有关的数据库时，分类学专家经常把各种来源的数据都放在数据库中，包括未发表或发表在非正式出版物上的内容，且未做详细的数据来源说明，导致数据库整体的数据质量下降。在编辑中国生物物种名录时，我们严格按照国际物种 2000 的规范，要求每条记录的内容都要有公开发表的文献对应，并把数据内容与出处建立连接，从而保证数据来源的规范化。

（3） 编辑数据手段的多样化

为了规范名录数据的格式，提高数据收集和审核的效率，我们专门设计了一个名录数据采集和审核工具软件。但是在实际收集整理物种数据时，分类学专家对计算机的操作技能差异很大，有些年长的专家甚至不能自己使用计算机。为此，我们设计了多种数据编辑和核对的手段，供不同的专家选用：a.使用自行开发的名录数据采集和审核工具软件，直接将结果存入数据库中；b.使用我们设计的 Excel 数据采集表格录入数据，录入完成后统一转换成标准格式的数据库形式；c.使用我们设计的 Word 格式的数据采集单，每页一个物种，可以打印空白表手工填写，或直接在计算机上填写，然后再转换到数据库中。

（4） 中名的选择

脊椎动物和高等植物的一些类群，往往有多个中文名称，并且很多都出现在正式的出版物上。多数分类学专家建议，每个物种确定一个或两个中文名称，作为推荐的正式名称—中名，避免名字使用的混乱。经过讨论决定，尽量使用中国动物志、中国植物志等权威学术专著、期刊上使用的中文名称作为中名；如果该类群的动物志还没有出版，则由首席专家确定物种采用的中名。其他中文名称都作为别名收录。

（5） 拉丁名的核对

从中文专著上采集的物种拉丁学名，是否还需要与全球名称索引（Global Names Index）核对，以保证它是一个正确的学名？经过讨论认为，应该记录文

献发表的原始状态，因此，不必一一查对。在一个类群的名录基本完成后，可以用程序批量核对。核对的结果提交给首席专家，由他对发现的差异进行标注或修改。

（6） 分布区划分和名称的标准化

不同年代和不同作者的文献对于分布区的描述方式可能有所不同。为了能够与全球生物物种名录中的分布区数据项相衔接，也为了以后用地理信息系统图形化地显示各生物类群的分布区，经过讨论，推荐各个类群都采用规范的地理分布描述方式，即采用国家行政区划代码（国家标准 GB2260-2007）、中国河流名称代码（SL249-1999）、中国湖泊名称代码（SL261-98）、中国山脉山峰名称代码（GB/T 22483-2008）或国际生物分类学数据库工作组世界地理标准等级四（Level 4 of the TDWG World Geographical Scheme）等国家和全球的数据标准之一的地理名称描述，作为分布区的规范化描述方式。如果原始文献使用了非规范化的描述语言，在保留原始描述的前提下，将其转换成上述五个标准之一的描述，做为规范化描述的内容。

（7） 中文冷僻字的存储和显示

中国生物物种名录的汉字内码采用的是 UNICODE 编码。在 2009 版名录中，大约有 50 个字在目前任何电子化字符集中都找不到，包括 UNICODE 编码标准的 UTF-32 字符集中也没有。因此，为这 50 个汉字设计了从 16x16 到 128x128 的汉字字模，并使用 UNICODE 没有使用的内码段为他们赋予内码值。为了避免中文字库安装与否而带来的问题，所有汉字均使用图像的方式显示，因此，这些中文冷僻字的字模也以图像的方式保存在光盘和网站上。不论使用者的计算机是否有中文字库，均可以正常显示汉字。

3、 结束语

生物物种名录看似结构简单，但实际上涉及的面很广，即使是通常被分类学家看起来没有大问题的脊椎动物部分，也需要花相当的人力和时间才能达到规范化的要求。由于中国是一个生物多样性丰富的国家，很多生物类群的分类

学信息还有待专家研究、整理和发表，因此可以预计，全面完成中国生物物种名录还需要时日。

自 2008 年开始，全球生物物种名录增加了一项新的内容—生命科学标识符（LSID）。它为每个生物类群的分类学概念和名称提供了一个全球唯一的代码，为追溯分类学研究过程、查询引证和分类系统比较，提供了一个便捷的工具。中国生物物种名录也将采用这个设计，在 2010 年或 2011 年的年度名录中增加 LSID 的内容。

2010 版的名录编制工作正在进行，预计将增加部分无脊椎动物和昆虫类群的数据。

在名录编制的过程中，仍然存在很多的专业和技术难题，比如分类学信息的更新、物种数据库的建设和维护、历史文献的信息采集、汉字的处理、多源信息的搜索和整合等。相信台湾同行也已经遇到相类似的问题，或许这些问题已经有了较好的解决方案。我们愿意与台湾同行开展深入的交流合作，共同提高，实现双赢。

【作者简介】

纪力强 男

职　　称：中科院 动物研究所 研究员

研究领域：生物多样性信息学

个人简介：1990 年在中科院动物研究所获博士学位。目前的研究方向是生物多样性信息学，主要研究生物多样性信息采集、整理、存储、处理和共享过程中的关键技术和手段，探讨生物多样性评价的方法并开发评价工具，研究制订生物多样性数据规范和标准，规划、设计并实施生物多样性信息数据库系统建设。先后主持建立了中国生物多样性信息系统和信息中心动物学分部、中国生物物种名录（Catalogue of Life China），以及相关的生

物多样性数据库。目前任国际物种 2000 项目全球工作组成员、国际 DIVERSITAS 计划中国国家委员会副秘书长、中科院生物多样性委员会委员兼秘书、中科院科学资料委员会委员。

联络电话： 86-10-64807129

联络邮箱： ji@ioz.ac.cn