野生植物物种编目数据库数据标准和质量控制规范

Standard and Quality Control for Catalogue Database of Wild Plant Species

关键词：物种编目数据库，数据标准，质量控制

编写说明

物种编目数据库是所有物种信息系统建设的核心，如国际知名的名录数据库 Species 2000 已成为所有与物种相关的其它信息系统扩展的基础。中国是世界上植物物种最丰富的国家之一，但一直以来，由于缺乏统一的物种数据描述标准、数据基础信息内容不完整使得我国植物物种编目数据库一直没能形成。中国植物物种信息数据是一种事实数据，是数百年来科学家辛勤工作的宝贵知识的结晶，在过去它一直是存在于专著和出版物中，而今天这种数据库的形式将得以永久保存和广泛传播，并以之为核心而得以快速扩展和增值。更为迫切的任务是，为了满足国家重大科学工程“中国西南野生生物种质资源库”对我国重要野生植物种质资源的迅速收集与保存的国家需求，即基于编目数据库建设积累形成的原始科学文献数据进行科学的谋划与决策，对野生植物种质资源进行科学采集。因此，为了使植物物种编目数据库更完整、快速地建设起来，同时能够使用科学的方法控制并评价其数据质量，通过借鉴参考“中国科学院科学数据库核心元数据标准”，研究制定了植物物种编目数据库建设相应的标准与规范，形成一套适用的标准规范体系。

起草单位：中国科学院昆明植物研究所

Kunming Institute of Botany, Chinese Academy of Sciences

起草人：周兵，李拓径，何延彪，王雨华

Written By Zhou Bing，Li Tuojing, He Yanbiao, WangYuhua

野生植物物种编目数据库数据标准和质量控制规范

1. 适用范围

本规范的研制旨在为“中国植物物种编目数据库”的建设提供一个全面、科学、权威的数据标准和数据质量保证。中国植物物种编目数据库建设将基于“Flora of China”、“中国植物志”、地方植物志、公开出版学报、重要考察报告及专著等，它的建成将会是植物学最重要、最基础的本底数据资源，使植物学研究从此有一个标准依据和扩展核心。通过本规范的研制，一方面可以促进中国植物物种基础信息数据标准的统一和完善，通过物种编目数据库的建设，使数百年来存在于文献里的中国植物物种数据以数据库的形式得于永久保存和广泛传播，并以之为核心得以快速扩展和增值，使这些宝贵知识得以进一步升华；另一方面，也为国家大科学工程“中国西南野生生物种质资源库”科学收集保存我国重要野生植物种质资源工作的顺利开展提供保障，为整合全国野生植物种质资源，规范野生植物种质资源的收集、保存、鉴定、评价、研究和利用，实现野生植物种质资源的充分共享，为可持续利用奠定良好的基础。

2. 引用标准

本规范的研制参考了中国科学院“十五”信息化建设重大项目“科学数据库及其应用系统”（项目编号为 INF105-SDB）的研究成果“中国科学院科学数据库核心元数据标准” （Scientific Database Core Metadata，简称 SDBCM）。

3. 定义和术语

- 元数据（Metadata）：关于数据的数据。
- 复合元素（Compound data element）：一个复合元素是由若干数据元素、或者数据元素与其它复合元素、或者若干其它复合元素共同组成的。通常用来表示较高层次的概念。
- 数据元素(Data element)：数据元素是元数据最基本的信息单元。
- 实体（Entity）：按一定结构组织起来的数据的集合，其结构可以用一组属性来刻画。例如，关系数据库中的数据表就是一个典型的实体代表。

- 数据类型(Data type)：对数据的有效值域及对该值域中的值所允许的操作的规定。例如，整型、实型、布尔型、日期类型、字符串类型等。对于复合元素，其数据类型用“复合类型”来标识。
- 模块(Module)：本规范按照层次结构组织元数据元素，不同的数据元素和复合元素组成一个模块。该层次结构的最高起始点为复合元素“数据集元数据”，该复合元素由其它表示数据集不同方面特征的复合元素构成，即本标准中的 6 个模块——物种分类等级信息模块、物种生物学信息模块、物种异名及文献信息模块、物种生境与分布信息模块、物种珍稀特有性信息模块以及物种利用价值信息模块。模块是本标准中一个最大的组织单位。

4. 目标

本规范的研制旨在为“中国植物物种编目数据库”的建设提供一个全面、科学、权威的数据标准和数据质量保证，研究、制定该系统相应的标准与规范，形成一套适用的标准规范体系，用以规范植物物种编目数据库数据资源的建设、管理和共享，进而保证数据资源的质量，建成植物学最重要、最基础的本底数据资源，使植物学研究有一个标准依据和扩展核心。通过数据标准的研制，从而保证其数据内容的全面性和科学性；在建设过程中，通过借鉴中国科学院科学数据库、生物多样性信息系统和标本馆信息化三个重大项目成熟的建库经验，对数据进行严格质量控制，从而保证其高质量的数据内容。

5. 植物物种编目数据库的数据标准规范

5.1 植物物种编目数据库的核心内容

植物物种编目数据库包括以下主要核心内容：物种分类等级信息、物种生物学信息、物种异名及文献信息、物种生境与分布信息、物种珍稀特有信息和物种利用价值信息等六个主要复合元素模块。

- 物种分类等级信息：科，属，种名（以上皆包括中文和拉丁）等；
- 物种生物学信息：形态特征描述，模式标本号，模式标本产地，馆藏地，资料来源；表型，生命周期，优势种群，基本特征描述，丰富度，盖度，

栽培状况等；

- 物种异名及文献信息：正名、异名、文献来源等；
- 物种生境与分布信息：海拔，土壤，生境类型，分布范围（国内、国外），气候类型；
- 物种珍稀特有性信息：保护级别，濒危程度，分布频度（国内、省内、地区的多少），特有状况，种型状况（所在属和科内含种的数量），古老残遗状况（发生地质年代）；
- 物种利用价值信息：用途，利用部位，利用民族，利用程度，价值意义。

5.2 植物物种编目数据库表结构及字段内容

1）分类名称信息表（主表）

数据内容	**类型**	**说明**
科拉丁名	字符	必填字段，使用科名，不加文献
科中文名	字符	
属拉丁名	字符	必填字段，使用属名，不加文献
属中文名	字符	
种中文名	字符	必填字段
种加名	字符	必填字段
定名人	字符	
种下等级	字符	
其它俗名	备注	
模式标本	字符	
文献来源	字符	必填
定名时间	日期	必填，为年
关联物种	字符	用属名+种加名表示
关联类型	字符	
绝对种编号（物种 ID 号）	字符	由系统生成，方法为属名+种加名+流水号，为关系 ID

2）异名与文献信息表

数据内容	**类型**	**说明**
物种 ID	字符	关系 ID
属拉丁名	字符	必填字段，使用属名，不加文献
种加名	字符	必填字段
定名人	字符	
种下等级	字符	
定名时间	日期	必填
文献来源	备注	必填

3）基本信息表

数据内容	**类型**	**说明**
物种 ID	**字符**	关系 ID
形态描述	备注	必填
生境与分布	备注	必填
功用	备注	
其它说明	备注	
图版	字符	外部链接图片

4）生境与分布详细信息表

数据内容	**类型**	**说明**
物种 ID	字符	关系 ID
生境类型	字符	必填
国外分布	备注	到地区或国家
国内分布	备注	必填，到县，以“省：县；”表示
海拔上限	数字	单位为米
海拔下限	数字	单位为米
北纬	字符	精确到分

东经	字符	精确到分

5）珍稀濒危特有信息表

数据内容	类型	说明
物种 ID	字符	关系 ID
所属类型	字符	必填，珍稀濒危、特有
特有类型	字符	国内特有、某省特有、某地特有
保护级别	字符	
标准出处	字符	
保护措施	备注	已有及建议保护措施

6）物种利用价值信息表

数据内容	类型	说明
物种 ID	字符	关系 ID
用途	字符	必填
使用部位	字符	必填
加工处理方法	备注	
有效成份	字符	
利用民族	字符	

7）字段约束

分类系统的选择和标注：尽管植物学学科已发展得相当成熟，但由于习惯或标引等问题，往往会出现不统一的分类标准，为了能够让植物学物种编目数据库真正成为一个标准、一个规范，必须对分类系统的选择在一个数据库中进行约束或分别标注。

表结构关系约束：表为一对多的关系，以分类名称信息表为主表，通过物种的在系统内部给定的 ID 值作为关系键。

必填字段：在数据库表和结构中为了能够充分体现植物学物种编目数据库的

内容，规定了必填字段，详细内容见数据库结构表。

字段内容约束：为了保证数据内容的正确性，必须对一些字段的内容进行某些限制，如海拔、日期、温度等，使之在一个有效的范围内取值。

字段格式约束：为了保证数据内容的正确性和一致性，在数据库结构定义时，对其格式做一定的限制。多项之间以“；”隔开。

关联文件约束：有些数据内容，如图片、多媒体文件等，不能直接进入数据库，而是以一种关链的形式建立链接，为了保证一定的质量，而做一个统一的定义。对图片文件来说应在 300-600dpi。文件命名以该物种的拉丁学名加数字编号。

数据来源约束：植物物种编目数据库的建成本身将要成为一个标准，因此其数据的来源必须科学、权威。数据的标引应是来源于《Flora of China》、《中国植物志》以及公开发表的学术论文。

6. 植物物种编目数据库的质量控制规范

6.1 植物物种编目数据库的质量控制方法

植物物种编目数据库是一种简单的关系型事实数据库，但由于这是一个基础数据库，是将来其它数据比对的标准和扩展的基础，因此其内容的正确性、科学性是极其重要的。数据的内容控制是一个过程，对于植物物种编目数据库来说，应包括以下几个过程：数据库结构的设计、数据标准的制定、数据采集源及标引内容的确定、数据的著录、数据的验收与抽查。

(1) 数据库结构设计中的质量控制

建议采用 3NF 范式进行数据标准化和规范化。数据的标准化有助于消除数据库中的数据冗余。标准化有好几种形式，在植物物种编目数据库的建设中建议采用 3NF（Third Normal Form）范式，因为 3NF 在性能、扩展性和数据完整性方面达到了最好平衡。遵守 3NF 标准的数据库某个表只包括其本身基本的属性，当不是它们本身所具有的属性时需进行分解。表之间的关系通过外键相连接。它具有有一组表专门存放通过键连接起来的关联数据的特点。

使用系统生成的主键。设计数据库的时候采用系统生成的键作为主键，那么实际控制了数据库的索引完整性。这样，数据库和非人工机制就有效地控制了对存储数据中每一行的访问。采用系统生成键作为主键还有一个优点：当拥有一致

的键结构时，找到逻辑缺陷很容易。

用约束而非商务规则强制数据完整性。采用数据库系统实现数据的完整性。这不但包括通过标准化实现的完整性而且还包括数据的功能性。在写数据的时候还可以增加触发器来保证数据的正确性。不要依赖于商务层保证数据完整性；它不能保证表之间（外键）的完整性所以不能强加于其他完整性规则之上。

(2) 数据标准制定中的质量控制

需求分析是数据库设计的第一步，也是最困难、最耗时间的一步。需求分析就是要准确了解并分析用户对系统的需要和要求，弄清系统要达到的目标和实现的功能。需求分析是否做得充分与准确，决定着在其上构建数据库大厦的速度与质量。需求分析做得不好，会影响整个系统的性能，其至会导致整个数据库设计返工重做。需求分析主要包括数据字典、全系统中数据项、数据流、数据存储的描述调查和构建。通过近 3 个月与各个专题项目成员的研讨及征求相关专家的意见和需求，从各个方面了解了数据信息系统的需求和功能。

为了能够科学、全面、系统地反映出植物物种编目数据库中有关信息和数据，重点突出植物物种的特性和规范，通过需求分析，应用大型关系数据库设计的原理和方法，定制数据库结构、标准和规范。这样，一方面使结构相对简单、明了，易于有效地管理、存储和更新，方便用户的查询、统计、和分析，另一方面照顾到兼容性与可定制性，有利于其它各专题的数据采集和扩展。

(3) 数据采集源及标引中的质量控制

统一组织，专家实施。数据的来源致关重要，必须具有一定的权威性和科学性，必须是没有争议的数据标准，因此，植物物种编目数据库应由统一的组织机构进行领导建设，在顶层设计上做关键把关，同时成立专家委员会，由植物分类学专家组成。由专家委员推荐和提供数据采集源，并根据专家委员会的专业知识和用户的需求调研，对数据源进行科学的标引。该标引可由植物学分类专家来做，也可由受过培训的专业人员来做，但必须满足植物物种编目数据库的建设规范。

分级标引，逐步推进。中国植物志已基本成为中国植物学界的标准，因此建设的第 1 步应以中国植物志进行数据标引。然后，应以已出版的地方志作为第 2 梯队，最后以最新发表的学术论文作为数据的更新与补充。

清查摸底，防止遗漏。物种的全面性也是数据质量的一个重要指标，为了尽

可能地将目前我国已发现的植物物种都能够全面地记载，不仅要充分地利用现有方便的文献数据库进行系统的收集，而且应在国内外各种交流平台上广泛地收集信息，防止遗漏任何一个物种。

(4) 数据著录中的质量控制

数据录入之前要求对标引目标进行人工预审，录入时要有专业人员随机跟班答疑，并建立工作档案，登记工作进度。对于已录入的资料要随机抽取 5%的数据样本与标引目标进行比对。若样本录入差错率超过 1%，要对已录入数据进行全面的对照检查，以更正数据录入中产生的差错。

数据审核控制。数据审核包括逻辑关系审核和合理性检查两部分。要求打印逻辑审核出错清单和合理性检查清单，经查明原因后才能进行更正、修改，更正、修改要有记录。

综合数据控制。利用计算机对科、属、种进行分析统计和汇总，通过专家知识对综合数据进行逻辑性和合理性检查。汇总结果有问题的，必须查明原因并对基层数据进行校正，不可直接对综合数据进行校正，校正情况要有记录。

(5) 数据验收与抽查中的质量控制

在数据库最终形成之前，必须经过数据的验收才可提供给用户使用，因此正式入库之前的数据必须经过一个审核和验收的过程。审核和验收应由专业委员会及专业组成员组成，不仅从专业知识上对物种的名称及相关信息做一个仔细认真地检查，而且需要再次从数据库的结构上对数据的结构和逻辑做一次合理性校验，这样才可最终形成数据库。为了进一步保证数据的质量，对已最终形成的数据库仍需要一次抽查工作，若抽查的错误率仍然超过 1%，则需要将数据重新返回，予以检查和修改。

6.2 植物物种编目数据库的质量评价方法

植物物种编目数据库的质量评价是对数据库的科学性和权威性的评定，主要评价内容应为数据库的描述标注和数据库规范的检验。描述标注主要指元数据的描述是否科学、合理和完整；数据库规范检验主要检验数据内容的完整性、一致性和相关性等。

(1) 质量评价流程

如图 1 所示：首先，对数据库进行描述标注评价，是否有描述标注；若有，是否符合“科学数据库核心元数据标准”。然后，进行物种编目数据库规范评价，分类系统是否一致？必填字段是否完整？物种名称是否规范？格式是否正确？内容是否一致？最后，评价记录的物种覆盖该领域的程度如何？这属于专家评价。只有都满足了以上条件，才可能说明该数据库属于高质量数据。通过逐项评价后，将分别对各种评价打印出评价报告，以便用户和建设者使用和改进。

在此评价流程中，会用到利用规范及借用统计方法的评价方法，具体内容将在下一部分中介绍。

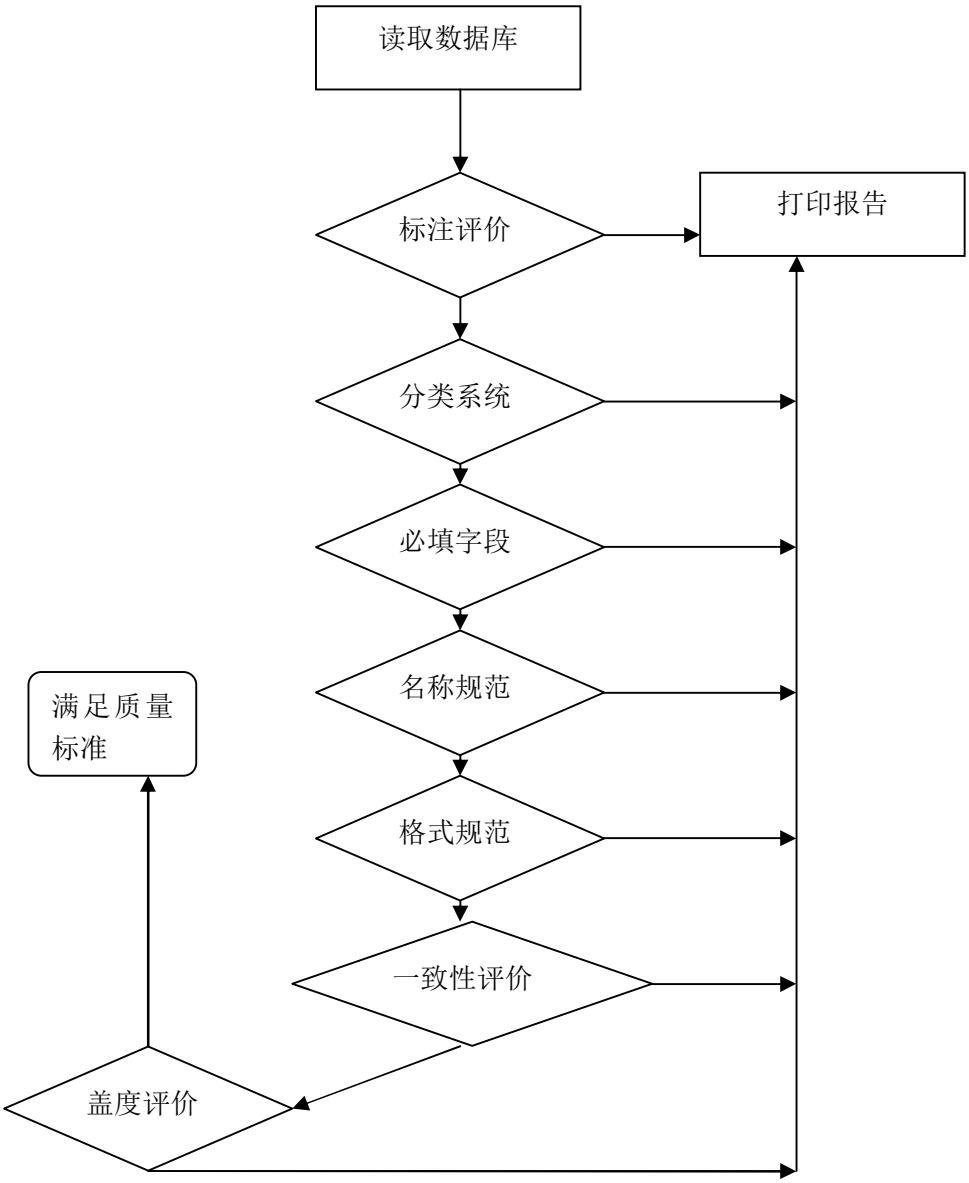

图 1

(2) 质量评价方法

- 专家评价

描述标注评价：要求编目数据库应有按“科学数据库核心元数据标准”填写的关于该数据库的相关描述，若没有或不符合元数据标准规范的，视为不够质量的数据库。

分类系统评价：由于植物学并存有不同的分类系统和标准，在编目数据库中

应在元数据中说明该库是以哪一个分类系统建设的，或者对每一条记录应有明显的字段来标明所属的分类系统。否则，视为混乱数据。

盖度评价：编目数据库的目标是尽可能全地收集中国所有的植物物种，数据库所记录的物种范围直接关系着该数据库的质量。专家或正式发表的学术书刊对某领域或某类生物都有一个大概的范围，该范围应该提供在元数据的描述中，根据该范围来评定本数据库的覆盖程度。

- 计算机自动评价

根据编目数据库核心数据规范，通过计算机编程对约束性字段进行评价，如必填字段、名称规范、格式规范和一致性评价等都可以使用计算机自动评价。即在限定的约束条件下，通过编程技术，按照数据规范要求对每条记录进行逐项检查，并打印出错误分析报告。

必填字段的评价条件：在编目数据库核心信息规范中共有 17 个必填字段，它们是名称信息表中的科拉丁名、属拉丁名、种加名、种中文名、文献来源、定名时间，异名信息表中的属拉丁名、种加名、定名时间、文献来源，基本信息表中的形态描述、生境与分布，分布信息表中的生境类型、国内分布，珍稀特有信息表的所属类型，利用价值信息表中的用途、使用部位。

名称规范的评价条件：中文名字段中不能出现英文字符；拉丁名中不能出现中文字符；在科属种的拉丁名字段中都是一个单词；图版的命名应是“属拉丁名+种加名+种下等级+_TUBAN+.JPG”；国外分布应与国际通用名称库比对；国内分布应依据国内最新的行政区划的名称规范；珍稀特有的类型只有珍稀濒危种和特有种两个类型。

格式规范的评价条件：定名时间应为只为年的格式；在模式标本、国外分布、国内分布、使用部位、利用民族字段中若出现并列多项，各项之间要用“；”隔开；海拔要使用数字表示，默认单位为米，分布范围最高不能超过 8848 m；分布范围的经纬度以度、分的格式表示，经纬度范围限制在中国境内。

一致性评价的条件：主要评价基本信息表中的“生境与分布”中的内容信息是否与国内分布的地方同属于相同的气候带，是否与分布的海拔和经纬度范围之间有矛盾冲突。例如，如果某一物种分布在热带，在国内的分布就不可能在中国东北某省县，其经纬度范围也不可能在这温带区域，海拔也不应超过 3000 m。