

植物标本汉英双语数据库管理系统 的概念与实践*

李鸣光¹⁾ Zhaoran Xu²⁾ 关朵霏¹⁾ Robert R. Haynes²⁾

张宏达¹⁾ 任善相¹⁾ Xuemei Du²⁾ 谢庆建¹⁾ 石涌岭¹⁾

(1) 中山大学生物学系/计算中心/管理学院, 广州 510275;

2) Department of Biological Sciences, University of Alabama, Tuscaloosa AL 35487, USA)

摘要 汉语有一音多字的特点, 因此汉字拉丁化后不能保持其原有的信息. 人名与地名是植物标本信息的重要组成部分, 把用汉字记录的标本资料以汉语拼音表达后, 未能完全承载原来用汉字所表达的信息. 植物标本汉英双语数据库管理系统既真实记录原有的汉语信息又同时记录其英译文, 并妥善处理了植物标本相关信息间复杂的关系. 中山大学植物标本数据库的部分汉英双语资料已进入国际计算机网, 可在世界各地国际计算机网终端上查询.

关键词 汉英双语数据库, 信息处理, 植物标本, 数据库管理系统, 计算机网络

分类号 Q949, TP392

世界上约有2亿份植物标本分藏于约2千个标本馆, 其中约4百万号为模式标本^[1]. 标本的积累, 为各项研究提供系统化、全面化资料有了潜在的可能. 如何高效率地管理如此庞大数目的植物标本是一个富于挑战性的课题. 在目前的信息时代, 资料及其寻求者之间的时空距离已不是障碍. 重要的是可资利用的信息存在与否及如何高效率地获取. 正在起步的生物信息管理学的目标就是要分析、整理、综合数量庞大的生物学资料, 使之成为科学研究、应用研究、教学和公共教育所利用.

1 植物标本信息的现代化管理

植物标本是植物学研究的依据. 植物标本包含与其相关的采集人、采集地理位置、采集时间、分类鉴定、生态特点等信息.

植物标本信息结构复杂, 植物名称具有明显的时空特征. 同一个采集号的多份同号标本(复份标本)常分别存于国内外不同的标本馆内, 每份标本可能各自被鉴定成不同的种类, 而在同一份标本上, 其鉴定结论也常随时间的推移而变动, 包括误定和订正, 有

收稿日期: 1995-04-11

* 中国国家自然科学基金会(第39410121144号)及美国国家科学基金会(National Science Foundation)(US-NSF-INT9312146)资助项目

些则因鉴定者持不同观点而异。一份标本还牵涉到具体采集人和具体的地理位置,它蕴涵的信息又与地理变迁,行政区划以及人为因素密切相关。

植物标本的信息是研究植物系统分类、区系、进化、种群、群落等的基本资料,现代生物多样性保护的正确决策也有赖于对已知植物标本信息的综合概括能力。传统的研究方法是逐份查阅标本。例如,为分析某个地区(如省、县)的植物区系状况,就要把采于该地区的标本从数十万甚至上百万份标本中逐一挑出摘记,然后才能进行分析。又例如研究一个广布于世界各地具有重要经济价值或理论意义的种的分布,要把散布于各国数百个标本馆的这个种的所有标本都调集到某个研究人员手上,即使最终能做到,亦需要大量人力物力;而研究人员把这个种在世界各地的具体分布录入并作出分布图,还要耗去数月甚至数年的不懈努力。如此艰巨的工作,完全应该利用高效率的计算机来解决。当植物标本资料存于计算机数据库,利用国际计算机网在任一终端就可以调集并分析上述资料,再在地理系统上打印出分布图,可能只需数十分钟时间。

植物学界早就已考虑改善植物标本信息的管理和获取。约60年前,Grassel^[2]就设想用机械检索打孔卡的方法使植物标本管理自动化。但由于植物标本信息结构复杂,设计出易于管理、使用和维护的数据库管理系统要解决许多理论及技术难关。近20多年来,将资料数字化并正确地处理多种信息的关系来建立数据库成为植物学界的重要研究议题之一。国际植物学同行专门成立了植物分类学数据工作组(Taxonomic Data Working Group),每年召开会议讨论植物信息管理计算机化标准的理论课题,美国阿拉巴马大学(The University of Alabama)是其长期会员;近10年来,美国科学基金会每年至少资助一次全美国的植物信息现代化管理专题研讨会;本项植物标本汉英双语数据库的研制,由美国科学基金会与中国自然科学基金会共同资助。

目前,对植物信息管理有影响的包括美国加州大学柏克莱分校(The University of California at Berkeley)的“加州植物信息系统(SMASCH)”、史密森研究院(Smithsonian Institution)的植物模式标本信息库^[3]、密苏里植物园(Missouri Botanical Garden)的植物名称数据库和阿拉巴马大学的“美国东南地区标本馆联网数据库(SERFIS)”^[4];南非的PRECIS^[5]等。由于生物标本数量巨大、信息管理复杂,走在世界各国前列的美国植物标本信息管理至今也只完成了其拥有的6千万份标本的4%的录入工作;其中史密森研究院始于1968年^[3],现完成其400万号的1/5强(Russel, G. F. 个人通讯),而阿拉巴马大学则走在中小标本馆的前列,已基本完成它全部馆藏标本的录入工作^[4]。

2 双语数据库的建立

每张标本所含的采集记录、鉴定历史、馆藏记录等的大量资料,大多数是以采集人的母语写的,如中国的标本绝大多数以汉字记录。

英语是国际信息交流最通用的语言,但汉语是世界上使用人口最多的语言,也是中国学者使用的主要语言。而且唯有汉语延读广泛使用四千年以上,其运载的丰富信息没有任何其他语言能代替。外国人使用汉字困难,遂通行把中国人名与地名拉丁化,早期多用威妥玛式拼音法拼写汉字,近年则采用汉语拼音字母。

普通话的一音多字及由各种方言来的一字多音是汉语的固有特点,一个不标上声调

符的拼音字节平均对应于 16 个常用字及 20 个次常用字和不常用字^[6], 因此汉字拉丁化后不能保持其原有的信息. 唯有汉字既准确记录普通话同音字的差异, 又通过汉字把不同方言联结起来.

人名与地名是植物标本信息的重要组成部分, 把用汉字记录的标本资料以汉语拼音表达, 只能给外国人提供一个模糊的信息, 也令中国人费解. 如地名 xiàng shān qū 或为安徽马鞍山的向山区、淮北的相山区, 或为广西桂林的象山区; wèi xiàn 为河北邯郸地区的魏县或张家口地区的蔚县. 若不标上普通话的四声符, wei xian 还可以是邢台地区的威县(区、县名及所属区划根据国家技术监督局 1991)^[7].

西方通行取名字的头一个字母和完整的姓氏作为姓名的缩写, 例如 Daniel Henry Nicolson 缩写成 D. H. Nicolson. 将汉字姓名改以拼音书写再缩写, 可能再也无法辨认原来所表达的信息. 例如, 我国植物学界王姓者逾 1 千人, 缩写后出现 C. W. Wang, Y. Q. Wang 者分别有近 20 人^[6], 这种缺损的信息反馈回中国人时, 自己也变得摸不着头脑.

信息时代的来临大大缩短了人们相互间的时空距离, 不同的语言对信息交流的障碍变得异常突出, 迫使人们正视. 数据库以真实记录原有信息为目标. 双语(以至多语)数据库能完整记录原有的信息并记录其译文, 既能保留原有信息的准确性, 又有助于国际上不同民族的信息交流. 近 10 余年来, 中外生物科学交流倍增, 生物标本信息的双语处理日渐迫切, 已经提到了中美两国植物学界与科学基金会的议事日程上.

现阶段的计算机技术已提供了处理汉语生物信息的工具. 尽管中国计算机技术起步较晚, 中文的信息处理却发展迅速, 在中文 DOS 和 WINDOWS 等支持下, 在微机上常见的 dBase, Foxbase, Foxpro 等数据库系统软件都能处理汉字, 为双语数据库的研制提供了有利条件. 生物学和计算机科学必需研究设计出既适用于国际交流, 又符合中国实际的双语数据库, 妥善地衔接汉语和英语的数据库, 使之运载准确完整的信息.

3 系统组成与特点

基于上述概念, 我们初步建成了汉英双语植物学信息系统. 系统采用 dBase 语言编写, 并以 UC DOS 为汉字平台. 系统源程序达 3.3 万个命令行.

本系统具有以下特点: ①同时存贮以汉字和罗马字形式表达的数据, 因此能真实记录原始的标本信息; ②用户可用单语或双语来查询, 既有利于国际交流, 又符合中国实际; ③主要字段的数据由系统自动从一种形式转换成另一个形式, 因此主要数据只需用单语输入, 双语则由系统自行完成.

植物标本信息的复杂关系在本系统中得到了妥善处理, 相应地建立了汉语及英语采集记录库、复份标本记录库及鉴定历史记录库.

为确保录入资料的一致性, 建立了 18 个字典库. 字典库包括了植物信息处理中的关键资料: 植物学名(科名、属名、种及种下名)、人名、地名(国名、省名、县名、地方名)、标本馆名及文献资料等. 目前已经整理出与中、美两国有关的上述字典数据 25 万余条, 为准确处理中国及美国的植物学信息提供了坚实的基础. 此外, 还建立了汉字及其拼音的数据库, 供系统把汉字转换成拼音时使用.

本系统还提供与地理信息系统(Geographic Information System, 简称 GIS)衔接的界

面,能以图像形式直观显示信息。

目前,中山大学植物标本数据库的部分汉英双语资料已借助阿拉巴马大学的SERFIS系统进入国际计算机网,世界各地用户在自己的终端上已可通过public@serfisby.ua.edu来查询这部分双语资料。

4 展 望

爆炸性扩张的信息资源,以及人们对信息获取的多样化需求,使信息已无法完全依赖口头、印刷、广播等传统手段来传递。近10余年来使用电子计算机技术来传递信息是对原有的传递手段的发展,用数据库系统来存贮并传播信息已成了目前重要的发展方向。

语言是一种历史信息,不论其应用范围是否广泛,都必需保护其承载的历史信息,在档案中准确反映出来。双语以至多语数据库成了计算机信息技术的重要课题,它的发展固然有赖于数据库研究及计算机硬件在今后的发展,但它同时也促进了数据库理论和技术进步。

多媒体技术将结合到双语库中。可以预见,查询某份标本时,可有专家的语音解说;凭借在终端屏幕上再现的标本原件图像,有关专家就能准确鉴别绝大多数的种。

中国通讯现代化、信息数字化的进程正在加速。中国馆藏约1000万号植物标本,利用丰富的劳动力资源,领先于西方发达国家实现植物标本的计算机化管理是可能的,它将为我国及各国的植物学科研、应用、教学、公共教育和政府决策部门提供重要的信息。研究人员能从任一终端屏幕上远程调集所需的信息,以前所未有大范围时空尺度分析植物的系统进化,生兴衰亡,现时分布特征及数百年来环境、自然和人为因素引起的分布变迁,以及濒危植物的历史与现状。数百年来人类积累于植物标本的巨量遗产,将以崭新的姿态和方式为人服务。

参 考 文 献

- 1 Holmgren P K, Holmgren NH, Barnet L C. Index Herbariorum. 1990. 693
- 2 Grassel C O. Visualizing our herbaria by the application of mechanical methods of tabulation and indexing. Museums Journal 1936. 36:373~384
- 3 Shetler S G. The botanical type specimen register. In: Shetler S G. et al. (eds.) An Introduction to the Botanical Type Specimen Register. Smithsonian Contributions to Botany, 1973. 12:1~25
- 4 Haynes R R, Zhaoran Xu. SERFIS, a regional approach to herbarium database management. The ABS Bulletin, 1995, 42(2):108
- 5 Gibbs—Russell G E. The PRECIS computer system and its applications for the flora of southern Africa. In: Morin et al. (eds.), Floristics for the 21st century, 1989. 11~22
- 6 Zhaoran Xu, Nicolson D H. Don't abbreviate Chinese names. Taxon 1992. 41:499~504
- 7 国家技术监督局. 中华人民共和国国家标准: 中华人民共和国行政区划代码 GB2260-91, 1991

Conceptual Issues in the Development of English—Chinese Bilingual Botanical Databases

Li Mingguang * *Zhaoran Xu* *Guan Duoqi*

Robert R. Haynes *Chang Hong-ta* *Ren Shanxiang*

Xuemei Du *Xie Qingjian* *Shi Yongling*

Abstract In order to make the Chinese language comprehensible to Western people, earlier *Wade—Giles*, and more recently, *pinyin*, have been used to phonetically transcribe Chinese personal and geographic names into Roman alphabets. Phonetically synonymous characters are common in the Chinese language, and one *pinyin* syllable on average accounts for 16 different commonly used characters plus 20 additional characters of less common usage. This phonetic synonymy results in loss of information when Chinese personal and geographic names are transcribed into syllables because these names alone have no context to help correctly interpret the original writing. Therefore, a monolingual database that convert and store Chinese data in Roman alphabets will present ambiguous information to both Western and Chinese colleagues. A bilingual (and possibly multilingual) database is preferable because it can archive the original information and simultaneously presents a translation of the original data in English, making the data comprehensible to a larger audience.

An English—Chinese bilingual information system should possess the following functions : 1) the database can store two formats of data, i. e. Roman (English) alphabets and Chinese characters; 2) users may retrieve data in monolingual format (either English or Chinese) or bilingual format (the same piece of data being displayed in both English and Chinese); and 3) the system can automatically convert the key data from one format to the other. Therefore, a bilingual information system will not require double data entry time in comparison to a monolingual system.

We have built an English—Chinese Bilingual Botanical Information System (BBIS95). The system is compiled with the dBase language and UC DOS is taken as the Chinese platform. Our system provides all the three functions mentioned above. In addition, the system contains eighteen bilingual dictionaries with more than 250 000 records of taxonomic, geographic, personal, institutional and referential data concerning Chinese and American plants. Data of multiple annotations from a single specimen sheet and multiple duplicates of a single collection deposited in different herbaria are properly recorded. The system also include interface with Geographic Informa-

* Department of Biology, Zhongshan University, Guangzhou 510275

tion System (GIS), Image Processing and Multimedia. Some bilingual collection data from the herbarium of Sun Yatsen University (SYS) have been made available on the Southeastern Regional Floristic Information System (SERFIS) at the University of Alabama, which is accessible through the Internet computer network (Email to zxu@serfis.by.ua.edu for detail). In the future, more Chinese data are expected on-line in similar ways.

Keywords bilingual, English—Chinese, database, botany, network

附 记

植物信息管理研讨会将于 1995 年 11 月 6 日在美国阿拉巴马大学举行,中美合作研制的汉英双语植物信息系统是本次研讨的主要内容。海峡两岸、美国、欧洲、澳大利亚的植物学专家、标本信息系统设计者、计算机专家等将参加研讨。

本次研讨会得到美国科学基金会资助,中国国家自然科学基金会也向参加本次会议的中方人员提供资助。研讨会由阿拉巴马大学与中山大学共同主持。中山大学、中国科学院北京植物研究所、广西植物研究所等单位将派有关专家参加。