大陆数字标本馆数据基础、整合的条件和设想
—以 CVH 为例

覃海宁　许哲平　李　奕　包伯坚　赵莉娜

中科院 植物研究所

【摘要】经过数年建设，「中国数字植物标本馆」（CVH）凭借标准统一、数额巨大的标本、志书和彩照等一系列关联数据库已经成为广大用户搜寻中国植物学信息的主要网站之一。新版网站（3.0）系统更重视国际标准的应用及数据整合，以提高专业服务水平，和增强网站建设的参与性和模块化管理。下阶段的主要任务是在数据更新的前提下，扩大各类数据的标准化规范化整理，以提高数据关联准确度，为生物多样性信息应用提供多层次数据服务，并实现与国际主流信息系统的接轨库。

【Abstract】Chinese Virtual Herbarium CVH becomes one of he major reference websites for China plant information with its rich and authorized data of digital specimens, e-flora and color-photos and so on. The CVH new version (3.0) focus on application of international data standards and provides users with multi-channel to view the website/data and to send the feedbacks. Future works on CVH include to update the data, to enhance the data-standardized, and to feed the users with more alive and fine data sets, as well as to mixture into global information systems.

【关键词】数字标本、标准化、数据集成、数据共享

【Keywords】digital specimens、data standardized、data integration、data sharing

1. 发展历程及工作基础

上世纪八、九十年代起，随着计算机技术的发展及其在生物学领域的应用，世界各地生物标本馆兴起标本数字化及数据库建设活动。大陆植物标本的数字化及共享服务的发展大致与欧美各主要标本馆同步，主要经历两个阶段：第一个阶段是奠基时期。上世纪九十年代初中期，中科院系统的生物标本馆和教育系统主要标本馆在中科院、科技部和国家基金委资助下分别启动生物物种数据库及标本数字化工作，建立了包括数十万份标本信息和物种名称的计算机数据库，并提供初步的共享。经过这一时期的建设，大陆标本馆系统积累了数据、初步取得数字化工作经验和锻炼了专门人才；第二阶段为快速发展时期和网络化时期。主要是二零零三年起，在实施「国家科技基础条件平台」项目中，中科院、教育部和其他部委系统及地方科学院数十家标本馆按照统一部署开展标本规范化整理、数字化表达和网络化共享等一系列活动，短短几年时间内完成了五、六百万份植物标本数字化工作，并构建了一系列相关的植物学数据库，建立了「中国数字植物标本馆（Chinese Virtual Herbarium, CVH）」（www.cvh.org.cn）和「教学标本标准化整理整合与资源共享平台」（http://mnh.scu.edu.cn）当大型门户网站，取得了良好的生物标本信息共享和知识传播效果。

CVH 自 2006 年初正式对外发布以来，经历了 1.0、2.0 和 3.0 三个版本。网站的核心数据库为 330 万份数字化标本信息，此外还有全国及各地植物志（PDF 文件）和野外生态照片，以及各类植物学数据库（表 1）。这些数据库的设计和建设均考虑了专业用户研究标本时的实际需求，且大部分数据库为课题组的数字化成果。至今为止，CVH 累计访问量为 1,320 余万 PV，其中每日独立 IP 稳定在 1.5 万左右（具体的内容访问比例为—标本查询：彩色照片：植物志=4：2：1）。该网站在用户中具有较高的知名度，成为查询中国植物标本和物种信息的专业门户网站。

刚刚发布的 CVH 新版（3.0）以加强网站功能和提高专业性信息服务水平为宗旨，着力打造参与式及可持续性发展的网站系统。并在实施过程中广泛采用国际主流信息系统的结构模式及数据标准。以标本信息共享模式为例，CVH 2.0 同时尝试了集中式（Central）和分布式（Distributed）两种形式，其中集中式共享

14 家，分布式共享 15 家。两者都存在一定的弊端。新版（3.0）则借鉴 GBIF 和 EOL 的运作模式，尝试性地引入联邦式（Federated）共享模式。该模式的特点是一方面数据实体在物理上是分布式的，另一方面，要求分馆（参加单位／数据源点）的数据周期性地更新和汇总到主馆中，并通过统一的 Portal 对用户提供一站式服务，提高用户体验。实践证明，这种互动式模式既能形成集群优势和资源优势，更好的为用户服务，也可以兼顾主、分馆积极性，同时，我们完善和建立了网站管理机构及信息共享规则，包括成立 CVH 理事会、IT 技术组、规范组和用户组等组织机构，以及实施数据共享实施细则，从组织上和制度上保证了 CVH 的正常运维和可持续发展。

表 1.CVH 网站主要数据库

序号		数据量	作者及信息来源
1	一般标本查询	331 万份、167 幅图像	成员单位
2	模式标本查询	7,240 笔（仅 PE）	林祁 PE 小组
3	新种发表原始文献及模式	30,705 笔	CVH 小组编制
4	植物名称作者	3,481 笔	CVH 小组编制
5	采集地新旧地名对照	2,048 笔	CVH 小组编制
6	采集史	4 期（电子杂志）	来金朋编制
7	中国植物标本馆	325	CVH 小组编制
8	分类学文献（1949－1990）	6,879	陈心启主编，1993
9	中国种子植物名称和分布	34,056 笔	CVH 小组编制
10	中国种子植物科属词典		引自华南园网站
11	中国植物志（80 卷 125 册）	45,016 笔	植物所 BHL 小组
12	Flora of China		引自 FOC 网站
13	中国高等植物图鉴（7 卷本）	9,057 笔	植物所 BHL 小组
14	地方植物志（18 种 100 卷本）	5 万页、6 万条目	植物所 BHL 小组
15	中国植物彩色图库	46,756 幅、7,000 余种	李敏团队
16	种子植物电子检索表		文香英编制
17	植物鉴定和描述术语图解	1,133 笔	王宇飞等 2001 翻译

18	植物名称知识		CVH 小组编制
19	标本采集与装订		CVH 小组编制
20	如何鉴定植物		CVH 小组编制
21	苔藓植物板块	名称－彩照－志书－作者介绍等	贾渝编制
22	蕨类植物板块	名称－作者介绍－研究动态等	张宪春编制

CVH 3.0 通过引入 LSID 标准，基本实现物种名称数据、标本数据和文献数据的站内关联，并建立了与 GBIF、uBio、BHL、IPNI 等国际主流数据库的关联，使用户能够全面地获取相关数据（实例见 http://beta.cvh.org.cn/lsid/index.php?lsid=urn:lsid:cvh.org.cn:names:cnpc_29624）。新版 CVH 还加强了用户互动功能，不仅提倡标本数据共享，包括馆际新闻动态也已经通过 Web Service 对外共享。我们还在 Google Code（http://code.google.com/p/chinese-virtual-herbarium/）和 Flickr（http://www.flickr.com/groups/cvh/pool/）等国际知名社区站点上建立专业群组，使用户可以以多种方式参与到 CVH 的讨论和建设工作中来。

在 3.0 改版过程中，我们通过广泛调研，参考 GBIF 等国际主流网站做法，在系统中引入国际数据标准 Darwin Core 进行数据规范化和标准化整理工作，主要是从地名配准和分类名称规范化两个方面进行整理。考虑到整理的难度和实际的数据应用需求，地名配准暂时将整理精度定位县级，具体做法是以国家标准为准，找到对应地名的标准县名编码，然后再得到对应的经纬度数据和地标等级（县级），同时标记整理日期。标本地标化工作的完成大大提高网站的服务功能。目前 CVH 的 LSID 物种页面上，用户在得到标本统计数据的同时，可以直接在 Google Map 上查看这些标本数据在全国的县级分布图（表 2）。

表 2.CVH3.0 技术及功能特点（部分）

类型	次级类型	CVH2	CVH3	备注
数据标准	Darwin Core	✗	✓	对照 Darwin Core 进行数据转换，整合

				了蜡叶标本、彩色照片和化石标本三类数据
	KML(Keyhole Markup Language)	✗	✓	标本县级分布实时显示，同时可以在 Google Map 或 Google Earth 上共享
	LSID(Life Sciences Identifiers)	✗	✓	分类、标本、文献、图片以及外部关联数据集成在一个页面中
数据整理	地标配准	✗	✓	按照《中华人民共和国行政区划代码（GB/T 2260-2007）》进行整理
	分类名称	✗	✓	参考 Darwin Core 标准进行规范化整理
Web Service		✗	✓	包括县级分布图、基于 Darwin Core 标准标本详细信息、动态新闻、物种简要信息（标本数、分布省份和标本馆列表）
用户社区	自建系统	✗	✓	基于 Drupal 系统，包括用户注册、论坛、评论等 Web 2.0 功能
	Flickr 群组	✗	✓	与全球专业爱好者进行数据维护和交流
外部数据集成		✗	✓	与国际主流生物多样性系统的数据关联
检索功能	拉丁名智能提示补全	✗	✓	减轻用户记忆负担，提高输入正确率
	结果按标本馆分组	✗	✓	强化分馆显示度，提高责任感
	集成 Google 站内检索	✗	✓	

使用统计	按分馆进行统计	✗	✓	分馆有权利知道自身数据的访问情况

包括标本物种名称在内的其他数据库也选择相应标准进行规范化整理。整理之后的网站数据无论在数据质量上，还是在数据的规范化程度上都有较大的提升，能够较好地与各种应用相结合，给予查询用户更为专业性的信息服务，同时也为下阶段数据整理工作提供了经验和参考。

总之，我们在实施 CVH 3.0 解决方案中，全面引入和采用国际技术和数据标准，目前已经采用的标准包括 LSID、Darwin Core、KML 等。特别是 Darwin Core 的引入，使标本数据能够与 GBIF 等数据较好地对接，同时也为下一步通过 GBIF IPT 工具进行数据集成和发布提供了良好的数据基础。目前，CVH 通过 Darwin Core 标准在同一个入口整合了蜡叶标本、彩色照片和植物化石三类数据的查询。

2. 整合条件

从数据源来看，大陆标本馆的数据库大至可分为三个类型：一是 SQL Server 2000 数据库，主要是较大型标本馆使用，结构设计基本相同。二是 Access 数据库，多为小型标本馆使用，表结构与 SQL Server 相似。此外，还有部分单位使用自己设计的数据库。如此看来，在参照 Darwin Core 数据结构的基础上，对现有数据结构进行映射，将有效地减轻现有系统负担，更好地为用户提供数据服务。

在数据标准方面，由于台湾地区已经是 GBIF 的节点，采用的是 Darwin Core 标准，而 CVH 也已开始采用该标准，同时还做了一些扩展，能够通过 web service 和 KML 进行数据共享。但在此之前，还应通过建立交互查询系统或名称对照字典等办法解决植物名称对接问题。由于分类系统和人文历史存在差异的原因，两岸在植物名称上普遍存在“同物异名”现象，同种植物在两地的拉丁学名和中文名可能都不同。只有实现标本物种名称等项目的对接整合才有可能实现对于两岸标本数据的联合鉴定、数据补充和整理，以及扩大项目应用的数据来源。

项目需求将是数据整合的一个重要推动条件。无论是台湾还是大陆的相关项目（如物种编目、珍稀濒危植物评估等）都需要有坚实的数据底库做支撑，而只有将两岸的标本数据合并来分析才能得到更加可信的结果。

此外，与标本查询和研究相关的其他数据也是数据整合的一个外围条件，如文献、图片、生态、土地变化等数据。这些数据目前也有不少积累（中国自然标本馆、植物图像库、BHL 中国节点等），如果能够实现关联共享，必然能增加用户的关注度，同时也是对标本数据的一个良好补充。

3. 未来设想

未来的数据共享和关联将建立在本体和语义分析的基础上，这也要求我们的数据更加规范化和标准化，数据质量要求更高。一方面，要启动植物专家系统校订标本物种名称，而为了使空间数据更精确，需要进一步展开县下地名的配准整理（方法包括小地名反推确定县级名称，也可通过采集人及采集号信息查找副份标本进行类推）。另一方面，在地标数据和分类学数据整理的基础上，还要对大量的人名（采集人和鉴定人）和时间（采集时间和鉴定时间）数据进行规范化整理，以利于用户进行数据的时空分析和采集事件分析。后期的工作还应考虑引入 GBIF IPT 工具，或者在此基础上加以扩展，使整理之后的数据在可视化表达、数据分组分析等方面有较好表现，便于用户的查询和选择。

在数据关联方面，越来越多的国际性主流信息系统如 GBIF、BHL、uBio、EOL 等都开放了 web service 接口。今后应调用这些接口与我们自身标本数据进行结合，使用户获得更多标本相关的其他数据。随着生物多样性信息学如 e-Science 平台的深入发展，这些标本数据将成为整个平台的重要组成部分，为生物多样性编目、监测和保护等不同层次的数据应用提供基础素材，并成为地区性和全球性生物多样性信息系统的重要组成部分

参考文献

1. Dublin Core 官方网站：http://dublincore.org/

2. EOL Transfer Schema:
http://services.eol.org/schema/EOL_Transfer_Schema_Documentation.pdf

3. GBIF 官方网站：http://www.gbif.org/

4. TDWG 官方网站：http://www.tdwg.org/

5. Li JJ（李健钧）, Delta system-an international standard for processing plant taxonomic descriptions. *Acta Phytotaxonomica Sinica*（植物分类学报）, 23, 447-452, 1996.（in Chinese with English abstract）

6. Zhang ML（张明理）, DELTA System, a recommendable information processing tool for taxonomic description. *Journal of Plant Resources and Environment*（植物资源与环境学报）, 18 (1), 87-90. 2009.（in Chinese with English abstract）

7. Daltio J & Medeiros CB. Aondê: An ontology Web service for interoperability across biodiversity applications. *Information Systems*, 33, 724-753. 2008.

8. Qiao HJ（乔慧捷）, Han Y（韩艳）, Li N（李诺）, Ji LQ（纪力强）. A model of biodiversity information integration. *Biodiversity Science*（生物多样性）, 12, 553-561. 2004.（in Chinese with English abstract）

9. Chapman, AD. *Uses of Primary Species-Occurrence Data*. Version 1.0. Report for GBIF, Copenhagen, 2005（陈映筠、柯智仁译，邵广昭校，2007，*物种出现原始资料之用途*。台北：中研院生物多样性研究中心）

10. Chapman, AD. *Prinsiples and Methods of Data Clearing_ Primary Species and Species-Occurrence Data*. Version 1.0. Report for GBIF, Copenhagen, 2005.（陈映筠、柯智仁译，邵广昭校，2008，*资料清理原则与方法*。台北：中研院生物多样性研究中心）

11. Chapman, AD. *Principales of Data Quality.* Version 1.0. Report for GBIF, Copenhagen, 2005（陈映筠、柯智仁译，邵广昭校，2008，*资料品质原则*。台北：中研院生物多样性研究中心）

【作者简介】

覃海宁　男

职　　称：中科院 植物研究所 研究员

职　　务：中科院 植物研究所「生物多样性信息学重点实验室」常务副主任

研究领域：植物分类学、生物多样性保护、科学数据库

个人简介：1987 年 9 月获中科院植物研究所植物分类学专硕士学位，同年于该所工作至今。1995 获博士学位，1997 年被聘为副研究员，2008 年为研究员（资格）。1991～2002 年担任分类室及标本馆副主任、主任；2004～2008 年担任网络信心中心副主任、2008 年～担任文献信息中心副主任、2010 年～担任生物多样性信息学重点实验室常务副主任；2004 年 10 月～CNC-DIVERSTAS 副秘书长、2002 年～担任 IUNC/SSC 中国植物专家组组长。上世纪八、九十年代完成世界性木通科植物进行分类修订，建立了新的分类系统；近年来，在项目负责人（马克平研究员）指导下，组织所内外相关专业人员完成「中国数字植物标本馆」网络信息共享平台的搭建，和中国野生高等植物编目工作；利用世界自然保护联盟（IUCN）红色名录标准，初步完成对中国野生高等植物绝灭危险的评估。

联络电话：86-10-62836023

联络邮箱：hainingqin@ibcas.ac.cn