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

1.工作基础

大陆植物标本的数字化和网络化共享经历两个主要阶段：第一个阶段是奠基时期。上世纪九十年代中期，中科院和教育系统的主要标本馆在中科院、科技部和基金委资助下启动生物物种及标本数字化工作，建立了包括数十万份标本的计算器数据库，并提供初步的共享；第二阶段为快速发展时期，主要是在实施「国家科技基础条件平台」项目中，中科院、教育部和其它部位系统及地方科学院数十家标本馆按照统一方案完成了六百万份植物标本的整理和数字化工作，并与所构建的其它关联数据结合提供网络查询服务，取得了到良好的信息共享和知识传播效果。

「中国数字植物标本馆（Chinese Virtual Herbarium, CVH）」 是「国家科技基础条件平台」植物标本项目的成果显示和信息共享平台。自 2006 年正式对外发布以来，CVH 经历了 1.0、2.0 和 3.0 三个版本。目前，网站标本资料量达 331 万份，其中带图像 167 万份。还包括标本查询和研究所必需的其它植物学数据库，主要是《中国植物志》（80 卷 125 册）、部分地方植物志和数万张彩色植物照片等。至今为止，CVH 累计访问量为 1,068 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。

CVH 标本数据共享机制包括集中式（Central）和分布式（Distributed）两个两种形式，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。目前在建的 CVH 新版（3.0）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。我们认为只有通过这种互动模式，才能兼顾主、分馆积极性，才能形成集群优势和资源优势，更好的为用户服务。此外，我们在 2009 年，还召集专门会议，讨论和建立了 CVH 理事会、IT 技术组、规范组和用户组等组织机构，以及数据共享实施细则，从组织上合机制上保证了 CVH 的正常运维和可持续发展。

CVH 3.0 还通过引入 LSID 标准，基本实现物种名称数据、标本数据、文献数据和植物园数据的站内关联，同时还建立了与 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。在与用户互动方面，新版 CVH 不仅提倡标本数据共享，包括分馆新闻的动态信息也已经通过 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/）等国际知名小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。

从 2009 年开始，我们通过广泛调研，参考 GBIF 等国际主流网站经验基础，在 CVH 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，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。在完成标本地标化工作后，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。

对于分类名称的整理，首先是与权威的「中国植物名称数据库（China Plant Catalogue, CNPC）」进行匹配，剩余记录按 Darwin Core 的推荐格式进行整理（如将命名人与学名存放在不同字段等）。在此基础上，再将整理结果与 CNPC、uBio 进行匹配。整理之后的数据无论在数据质量上，还是在资料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，同时也为进一步的数据整理工作提供了经验和参考。

可以说从 2009 年起，我们在 CVH 3.0 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像，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 也已经采用该标准，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不同的问题。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。如此的标本整合之后将利于植物学专业用户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（如红色名录、入侵种研究等），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，与标本查询和研究或者说植物分类学研究相关的数据，也是数据整合的一个外围条件，如文献、图片、生态、土地变化等资料。这些资料目前也有不少积累（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推理）。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 界面。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它数据。如果生物多样性 e-Science 平台的构建进展顺利，这些标本数据将成为整个平台的重要组成部分，为实现不同层次的数据应用提供基础素材，为生物多样性保护、控制外来入侵种、气候变化等领域的应用研究提供强有力的支撑。

关键词　数字植物标本馆、标准、整合

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