生物多样性 2015, **23** (2): 252–253 doi: 10.17520/biods.2015055
Biodiversity Science http: //www.biodiversity-science.net

植物学名校订工具

王钧杰 陈国科 马克平*

(中国科学院植物研究所植被与环境变化国家重点实验室, 北京 100093)

Tools for standardization of plant names

Junjie Wang, Guoke Chen, Keping Ma*

State Key Laboratory of Vegetation and Environmental Change, Institute of Botany, Chinese Academy of Sciences, Beijing 100093

近年来, 随着植物信息数字化的普及, 植物学领域出现了众多名称索引系统, 如英国皇家植物园(邱园)和密苏里植物园共同创办并维护的The Plant List (http://www.theplantlist.org/), 密苏里植物园维护的Tropicos (http://www.tropicos.org/), 邱园、哈佛大学标本馆和澳大利亚国家标本馆共同维护的The International Plant Names Index (IPNI, http://www.ipni.org/), 中国科学院植物研究所维护的中国高等植物信息系统(http://www.etaxonomy.ac.cn/)等, 这些索引系统涉及数以百万计的植物名称。与此同时, 许多基于上述名称索引网站提供的应用程序编程接口(API)进行名称批量校订的软件和工具包也被开发出来, 使得批量处理植物学名成为可能。对研究人员而言, 通过电脑录入植物名称可能会产生拼写错误(Rees, 2014)。此外，志书中也存在拼写或印刷错误, 如《中国植物志》中就有1,000多个学名存在拼写问题(Liu *et al*., 2013)。面对大量的植物名称, 手动校订必然花费大量精力, 而使用一些应用程序或软件包进行批量校订, 一方面节省时间, 另一方面校订的标准统一。在此我们介绍几款实用的植物学名校订工具。

1 Taxonomic Name Resolution Service (分类学名称解析系统, TNRS)

TNRS是由iPlant Collaborative开发的开源应用程序, 可以批量校订植物名录, 校正拼写错误, 判断是接受名还是异名, 并给出相应的接受名(Boyle *et al*., 2013)。TNRS目前的版本为3.2, 依据的数据源为Tropicos, Global Compositae Checklist (http://dixon.iplantcollaborative.org/compositaeweb/), USDA Plants (http://plants.usda.gov/java/) 和 NCBI Taxonomy (http://www.ncbi.nlm.nih.gov/Taxonomy/)。

登录网站http://tnrs.iplantcollaborative.org/TNRSapp.html/即可使用TNRS, 这里作简单介绍(附图1):

第一步, 输入植物学名(种名、属名、科名等)。TNRS目前支持两种输入方式: 直接复制名录(5,000条以内)到网页上的文本框, 输入的名录不需加表头, 不要有逗号等标点符号; 或者点击“Upload and Submit List”, 通过对话框上传文本文件(txt格式), 同时输入邮箱地址来接收验证码。

第二步, 在屏幕右侧的Name processing settings栏进行必要的设置(通常情况下使用默认设置), 点击“Submit List” (文本框右下角), 程序就会在后台自动进行名称校订, 校订完后会在另一个文本框中显示结果。如果输入的是文本文件, 网站会将查询结果的验证码发至上一步输入的邮箱, 在Retrieve Results栏的文本框内输入该验证码, 点击“Retrieve”即可查看校订结果。

第三步, 点击“Download results”, 在Download as栏输入输出文件的名称, 点击“ok”即可下载匹配结果(txt格式)。

第四步, 检查输出结果。建议将文本格式文件的数据导入Excel, 查看Overall_score得分, 分值低于0.99的名称建议人工核查。

收稿日期: 2015-03-05; 接受日期: 2015-03-16
基金项目: 科技部植物标本标准化整理、整合及共享平台建设项目(2005DKA21401)
* 通讯作者 Author for correspondence. E-mail: kpma@ibcas.ac.cn

2　Plantminer (植物名称校对者)

Plantminer是Gustavo Carvalho开发的批量校订种子植物名称的网络工具, 可以自动检查并校正常见的拼写错误, 给出相应的接受名。它目前依据的数据源为Tropicos, World Checklist of Selected Plant Families (WCSP) (http://apps.kew.org/wcsp/) 和The Plant List (Carvalho *et al*., 2010)。

Plantminer的使用方法非常简单, 在浏览器中打开http://plantminer.com/, 在文本框中输入接收结果的邮箱地址, 点击“Go”, 跳转到SUBMIT A LIST页面, 点击“Choose File”上传植物名录(文本文件), 点击“Submit”, 校订完成后, Plantminer会把结果(csv格式)发送到预先输入的邮箱中。打开该文件, 用Excel进行分列就会得到类似TNRS的结果。

3　Taxonstand (学名标准化)软件包

Taxonstand是一个R软件包, 它依据The Plant List的数据进行名录比对, 校正错误名称, 给出名称所在的科、相应的接受名和命名人。和上述两个工具相比, Taxonstand在进行名称校订时有多个参数供手动调节, 如diffchar和abbrev等(Cayuela *et al*., 2012)。Taxonstand需要在R语言环境中运行, 因此使用前需要安装R软件。R软件可从网站http://www.r-project.org/下载, 点击download R, 选择一个镜像地址(建议选择国内的镜像如http://ftp.ctex.org/mirrors/CRAN/), 根据电脑的操作系统选择相应的程序, 之后点击base进行下载并安装。

安装后, 启动R, 电脑保持连接互联网的状态, 在R的控制台的红色“>”后输入“install.packages ("Taxonstand")”, 按回车, R会自动安装Taxonstand。安装完成后在控制台输入“library(Taxonstand)”即可调用Taxonstand提供的2个函数TPL (多个名称校订)和TPLck (单个名称校订)。在校订名录时需要用到TPL函数, 其使用方法如下(只列出主要参数):

samplelist <- TPL(splist, corr = TRUE, diffchar = 2, max.distance = 1)

samplelist用来存储结果, 可以自行定义; splist为需要校订的植物名称列表。corr这个参数用来开启或关闭模糊匹配, 默认值corr = TRUE意味着开启模糊匹配, diffchar和max.distance两个参数用于调整模糊匹配的程度, 它们的值越大, 则允许模糊匹配的程度越大, 其他的参数请在R控制台输入“?Taxonstand”, 按回车, 查看并阅读其帮助文件或者阅读Taxonstand的说明文件(Cayuela *et al*., 2014)。下面我们简单举个例子来演示这个函数的使用。

假设有5个植物名称需要校订, 它们保存在D盘根目录一个名为“testname”的文本文件里(附图2):

第一步, 用R读取这个文件。

testname <- read.csv("D:/testname.txt")

第二步, 用TPL校订名称(参数均为默认值)。

samplelist <- TPL(testname$name)

第三步, 查看结果并输出为testname_checked.txt, 保存到D盘根目录下。

samplelist

write.csv(samplelist,"D:/testname_checked.txt", row.names=FALSE)

此时, 即可在R中或者打开testname_checked.txt文件看到校订结果。

4　校订结果核查及建议

对于同一植物名称, 使用不同的校订工具可能会得到不同的结果, 如*Acer huianum* (勐海槭), TNRS和Plantminer建议的接受名为*Acer thomsonii* (巨果槭), Taxonstand建议的接受名为*Acer sterculiaceum* subsp. *franchetii* (房县槭)。这主要是因为它们使用的数据源不同, TNRS和Plantminer主要依据Tropicos, 而Taxonstand则依据TPL。因为*Flora of China* (FOC)的名录数据是Tropicos的一个子集, 对于国内的研究者而言如果想以FOC作为校订标准, 使用TNRS和Plantminer校订得到的结果具有更高的参考价值。

文中引用的参考文献及附图见附录。

致谢: 感谢陈堂婷帮助修改文稿。

(责任编委: 严岳鸿　责任编辑: 黄祥忠)

附录　Supplementary Material

附图1　Taxonomic Name Resolution Service的操作界面(http://www.biodiversity-science.net/fileup/PDF/w2015-055-1.pdf)

附图2　Taxonstand校订案例(包含5个物种名称)使用的文本文件
(http://www.biodiversity-science.net/fileup/PDF/w2015-055-2.pdf)

附文档I　参考文献(http://www.biodiversity-science.net/fileup/PDF/w2015-055-3.pdf)

附文档I 参考文献

Boyle B, Hopkins N, Lu Z, Grey JAR, Mozzherin D, Rees T, Matasci N, Narro ML, Piel WH, Mckay SJ, Lowry S, Freeland C, Peet RK, Enquist RJ (2013) The taxonomic name resolution service: an online tool for automated standardization of plant names. *BMC Bioinformatics*, **14**, 16.

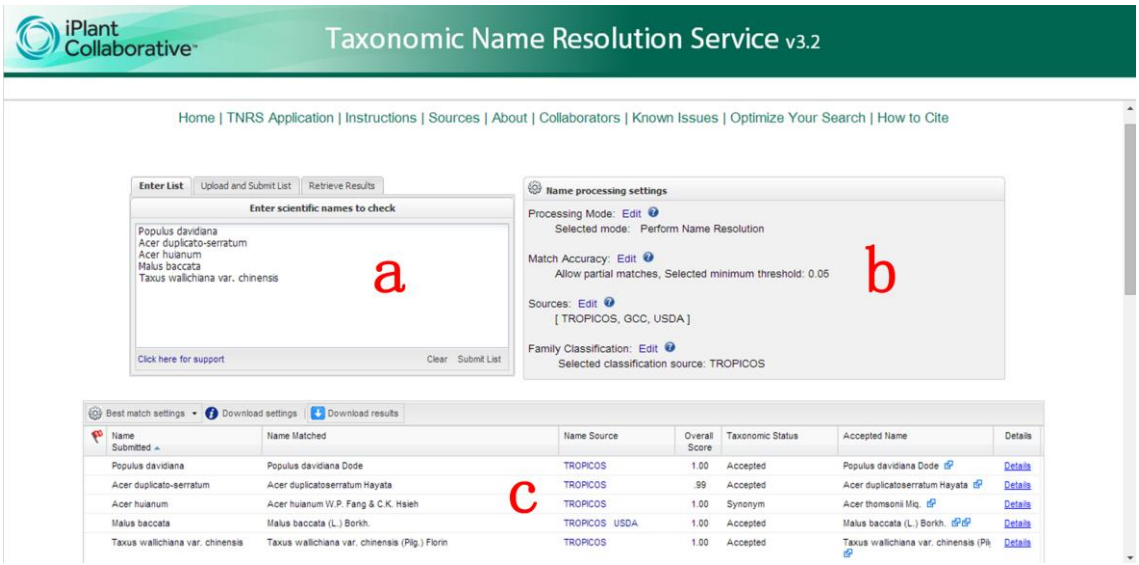
Carvalho GH, Cianciaruso MV, Batalha MA (2010) Plantminer: a web tool for checking and gathering plant species taxonomic information. *Environmental Modelling & Software*, **25**, 815–816.

Cayuela L, Granzow-de la Cerda Í, Albuquerque FS, Golicher DJ (2012) Taxonstand: an R package for species names standardisation in vegetation databases. *Methods in Ecology and Evolution*, **3**, 1078–1083.

Cayuela L, Oksanen J, Cayuela ML (2014) Package "Taxonstand". http://cran.irsn.fr/web/packages/Taxonstand/Taxonstand.pdf.

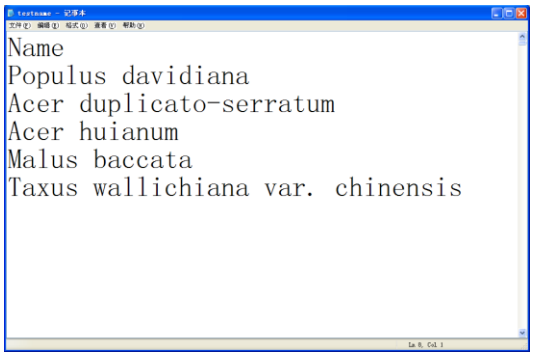
Liu S, Liu B, Zhu XY (2013) Corrections of wrongly spelled scientific names in *Flora Reipublicae Popularis Sinicae*. *Journal of Systematics and Evolution*, **51**, 231–234.

Rees T (2014) Taxamatch, an algorithm for near ('fuzzy') matching of scientific names in taxonomic databases. *PLoS ONE*, **9**, e107510.



附图1 Taxonomic Name Resolution Service的操作界面。a: 输入文本框，可以直接把一列名称复制到该文本框，点击**Upload and Submit List**可切换为导入文本文件的文本框，点击**Retrieve Results**则会显示输入验证码的文本框; b: 设置栏，有4个项供调整，分别为处理模式、是否进行模糊匹配、数据源、科名依据; c: 输出文本框。

Fig. S1 Operation interface of Taxonomic Name Resolution Service. a: input text box for inputting a list of plant names, clicking Upload and Submit List button to switch to a text box for uploading a text file, or clicking Retrieve Results button to switch to a text box for inputting identifying code; b: name processing settings, it has four settings to fit: Processing Mode, Match Accuracy, Sources, Family Classification; c: output text box.



附图2 Taxonstand校订案例使用的文本文件，这个案例包含5个物种名称。用TNRS和Plantminer校订时，需删去表头(第一行)。

Fig. S2 The txt file to be checked, which includes 5 plant names. Using TNRS or Plantminer for checking, please delete the header.