eFloras: New directions for online floras exemplified by the Flora of China Project

Anthony R. Brach[1] & Hong Song[2]

[1] *Missouri Botanical Garden c/o Harvard University Herbaria, 22 Divinity Avenue, Cambridge, Massachusetts 02138-2094, U.S.A. brach@oeb.harvard.edu (author for correspondence)*
[2] *Saint Louis University, DuBourg Hall, 221 North Grand Boulevard, St. Louis, Missouri 63103, U.S.A. hongsong2k@yahoo.com*

Online floras provide research botanists with the opportunity to work on floristic treatments dynamically, and enable users to browse and search these treatments. A web-based program called *eFloras* (URL: http://www.efloras.org/) was developed to enable access to online "electronic" floras. Through a web interface to the data, users can browse online floristic treatments by volume, family, and genus, and can search by name, distributional data, and text. With the use of web forms, editors and authors with permissions can correct and update the data.

KEYWORDS: bracketed keys, eFlora, Flora of China, floristics, online flora.

INTRODUCTION

Since the time that medicinal plants were first noted in herbals, people have been compiling lists (checklists) of plant names. Often the names have been accompanied by illustrations, descriptions, and distributional information. Published volumes of this literature, sometimes consisting of only a few copies, had been restricted to libraries. Now, through the World Wide Web, botanists are able to instantaneously provide checklists and floras to users worldwide and update them as the taxonomies of the groups are revised and further data are gathered. Several current flora projects provide online treatments: the *Flora of Australia* (Orchard & Thompson, 1999–), *Flora Europaea* (Tutin & al., 1993–), *Flora Zambesiaca* (Exell & Wild, 1960–), *Flora Mesoamericana* (Davidse & al., 1994–), *Flora of China* (Wu & Raven, 1994–), and the *Flora of North America* (Flora of North America Editorial Committee, 1993–). More than 100 entries can be found searching for the word "flora" in the Internet Directory for Botany (URL: http://www.botany.net/IDB/) or by using a search engine such as Google (URL: http://www.google.com/). This paper will present a new look at online treatments from the vantage point of the *Flora of China* Project.

THE *FLORA OF CHINA* PROJECT

The *Flora of China* (*FOC*; Wu & Raven, 1994–) is a collaborative international project to publish the first modern English-language account of the vascular plants of China (nearly 12% of the world's plants). All taxonomic treatments are jointly co-authored by Chinese and non-Chinese botanists. Nearly 150 Chinese and 500 non-Chinese botanists are participating in the FOC Project. The project will result in the printed and online publication of 25 volumes of text of all of China's approximately 30,000 species of vascular plants, accompanied by 25 volumes of illustrations (of > 60% of China's flora) representing all genera. Published volumes are available online at www.eFloras.org. This web-based program called *eFloras* (URL: http://www .efloras.org/) was developed to enable access to online "electronic" floras. Through a web interface to the data, users can browse online floristic treatments by volume, family, and genus, and can search by name, distributional data, and text. With the use of web forms, editors and authors with permissions can correct and update the data of the *Flora of China Checklists*.

Online checklists provide an invaluable source of plant names and publication data at local, regional, and global scales (Miller & Arriagada, 2000). The *Flora of China Checklist* is a database searchable via a web interface (URL:http://mobot.mobot.org/W3T/Search/FOC/projsfoc.html) at Missouri Botanical Garden. It is a systematic reference that will contain all of the scientific names that have been published for China.

The Checklist contains all of the scientific names of species, combined with their distributions in China (at the provincial level) and adjacent, bordering countries, the elevations at which the plants grow, botanical synonyms, bibliographic citations, and endemism. The scientific names are dynamically linked to other available data (i.e., volume: page in the *Flora Reipublicae Popularis Sinicae* (FRPS) and FOC, illustrations, distri-

bution maps, type information and images) via TROPICOS (URL: http://mobot.mobot.org/W3T/Search/vast.html).

PROBLEM OF "GRAY" LITERATURE

The term "gray literature" has frequently been applied to works of limited availability because of limited numbers of copies, restricted publication regions, and language barriers. Access to this literature has been a continuing challenge, and efforts are now being made as part of several projects including the International Plant Names Index or IPNI (URL: http://www.ipni.org/) and the IOPI Global Plant Checklist (URL: http://www.bgbm.fu-berlin.de/iopi/gpc/default.asp) to database all taxonomic names.

The FOC Project verifies the original citation of each name, and records the publication data according to recognized taxonomic standards. Many collaborators on the FOC project, and other botanists, who do not have access to all of the relevant literature, have found the checklist valuable for their work. Verification provides scientists with reliable citation information as to whether or not a name is validly published. It is estimated that the checklist will contain a total of about 135,000 botanical names, including synonyms.

HU CARD INDEX

Infraspecific names are especially difficult to track down since the *Index Kewensis* did not begin including these until 1971. To locate many of these infraspecific names, the Hu Card Index, a file of 158,844 cards for Chinese plant names, produced by Dr. Hu Shiu-ying (Professor Emeritus, Arnold Arboretum of Harvard University) is available. The index was prepared in the early 1950s when the Arnold Arboretum undertook a project to prepare a flora of China. Dr. Hu, working with a staff of four or five persons, searched all botanical literature published between 1753 and 1955 to locate all names (including infraspecific names) used for the plants of China. The cards were filed systematically by family, in a basically Englerian sequence, and alphabetically by genus, species epithet, and infraspecific epithet in three cabinets in the Harvard University Herbaria. Until recently, they were only available to staff and visitors.

In 1997, the cards were scanned as digital images at 300 dpi, 4 bit grayscale by Adaptive Computer Solutions in Blountville, Tennessee, USA. During the processing, the beginning (i.e., first card) of every family and genus was recorded for later indexing. The GIF images of the cards were then uploaded to a web server. The second author developed a web application in Java and MySQL to enable browsing, searching, and annotating the digital records.

The Hu Card Index can be browsed online (URL: http://flora.huh.harvard.edu/HuCards/). The cards are indexed by family and genus; they can be browsed as in a physical file cabinet by using "next" and "previous" buttons. The card index can also be searched by name, including "%" wildcards, in particular, "%var.%, %subsp.%, and %f.% to find infraspecific names.

FLORA OF CHINA WEB

The *Flora of China* Web (URL: http://flora.huh.havard.edu/china/) provides a regularly updated newsletter, introductory information, floristic treatments (database descriptions in HTML and PDF formats, and illustrations), interactive keys for identification (see Brach & Song, 2005), botanical papers pertaining to the FOC published in the journals *Novon*, *Annals of Missouri Botanical Garden*, and *Harvard Papers in Botany*, related searchable data (e.g., the FOC Checklist, the Hu Card Index), images, links to the FOC illustrations, guidelines for contributors, and information on editorial centers and the people involved in the Project.

ONLINE TAXONOMIC KEYS

Indented (or yoked) keys are characteristic of printed floras (Brach & Song, 2005). In both printed publications and in HTML, the sometimes long intervals between the first and second halves of each couplet are problematic. Turning to a subsequent printed page, or scrolling to a distant, second lead obscures the view of the first lead, thus creating a barrier to identification. This occurs more commonly on computer screens because of a comparatively lower resolution.

Bracketed (or parallel) keys with the alternate leads of a couplet placed adjacently resolve the problems of indented keys in HTML. Thus, it is preferable to convert indented keys to bracketed keys for use on the web (and/or to create interactive keys, particularly for large taxonomic groups, (Brach & Song, 2005).

In the brief history of web (HTML) keys, names of taxa at the terminal leads of a key were first linked to their respective descriptions manually. Later, programs were written to parse manuscripts into linked description and key pages for the internet (e.g., in Perl). Having thousands of HTML pages posed space and management problems. Databasing the descriptions and keys offered a solution to this problem. Additionally, dynamically generating name-lookups from the database was more efficient than manually inserting links. Furthermore, keys to

large genera with 50–100 or more species can become cumbersome.

Web-based interactive identification keys such as DELTA INTKEY (Dallwitz, 1980; Dallwitz & al., 1993–, 2002–); and ActKey (Brach & Song, 2005) present a simple alternative to lengthy, indented or bracketed keys. An online interface to interactive identification keys should enable users to select easily observable and readily available characteristics to identify a specimen.

EFLORAS: PARSING TREATMENTS AND TAXONOMIC KEYS INTO THE DATABASE

At the start of the FOC Web in 1996, treatments of the then 2000+ taxa were provided to the online community. Microsoft Word files were converted to HTML. An individual web page was created for each taxon. For taxonomic keys, hyperlinks were manually added to the HTML pages for the corresponding taxa.

Since the thousands of individual files posed a long-term management problem, an earlier associate (Noel Cross) wrote a program in Perl to transfer the HTML data into a database (Microsoft Access). The database was then used to dynamically create web pages. Users could then query the database by taxon name, elevation, and province.

To centralize the keys for improved management of the pages, editorial corrections and updates, the original static HTML key files for more than 4000 taxa were transferred into an Access database. A program was written in Perl to parse the HTML files into delimited ASCII text. The components of each lead (i.e., lead_id, lead_number, lead_half, description, taxon_id, and taxon_name) were used to parse each lead into corresponding fields. Cardinality maintained order and identification within and among keys.

Final manuscripts are automatically parsed into delimited ASCII text files with the use of a Perl program. The text files are then imported into a database. The relational database contains all of the treatments and keys from the published volumes as well as new records for taxa published after the production of a particular FOC volume. As new volumes are finalized, the manuscripts are parsed and appended to the database. Treatments are dynamically linked to taxonomic databases and taxon images (e.g., the FOC Illustrations in TROPICOS, URL: http://mobot.mobot.org/W3T/Search/image/imagefr.html, and photos from the Biodiversity of the Hengduan Mountains Region website, URL: http://maen.huh.havard.edu:8080/china/).

The current version of eFloras.org stores data in a Microsoft SQL Server database hosted at Missouri Botanical Garden. The web site is programmed in C#, and the web server is Microsoft's Internet Information Server. The databased taxonomic keys were dynamically converted into bracketed keys, and each taxon name within the keys is linked to the database record for that taxon (Table 1).

A web interface was developed for new additions (e.g., new names published after the printing of a particular volume) and corrections to the database (Fig. 1). Earlier programs written in Java for browsing and searching "Flora Online", were replaced by a new interface using ASP.Net and C#.

BROWSE AND SEARCH INTERFACE

We are using a web interface to the database to dynamically create web pages. Users can browse treatments by volume, family, and genus. Users may query the database by taxon name, province, adjacent country, and elevation (Fig. 2). Simple text searches allow queries of descriptive, habitat, and use terms. For example, based on 11 available volumes of the FOC, text queries for use terms "medicinal" and "ornamental" resulted in 813 and 198 records, respectively. Text queries for habitat types "limestone" and "mountains" resulted in 421 and 730 records, respectively. Similarly, text queries for texture states "leathery", "papery", and "membranous" resulted

Table 1. Example of a bracketed key (*Saxifraga* sect. *Irregulares* Haw.). Underlined text indicates hyperlinks.

1	Stolons arising from axils of basal leaves, filiform	20. *S. stolonifera*
+	Stolons absent.	(2)
2(1)	Leaf blade peltate or ovate or broadly so, or elliptic to oblong, abaxially sometimes brown spotted.	(3)
+	Leaf blade reniform to orbicular, abaxially usually concolorous.	(4)
3(2)	Leaf blade peltate or ovate or broadly so, abaxially brown spotted.	18. *S. mengtzeana*
+	Leaf blade elliptic to oblong.	14. *S. kwangsiensis*
4(2)	Leaf blade with foliar embryos in sinus adaxially.	19. *S. epiphylla*
+	Leaf blade without foliar embryos.	(5)
5(4)	Longest petal serrate at margin.	17. *S. fortunei*
+	Longest petal entire at margin.	(6)
6(5)	Flowering stem and inflorescence reddish long glandular villous.	15. *S. rufescens*
+	Flowering stem and inflorescence shortly glandular pubescent.	16. *S. imparilis*

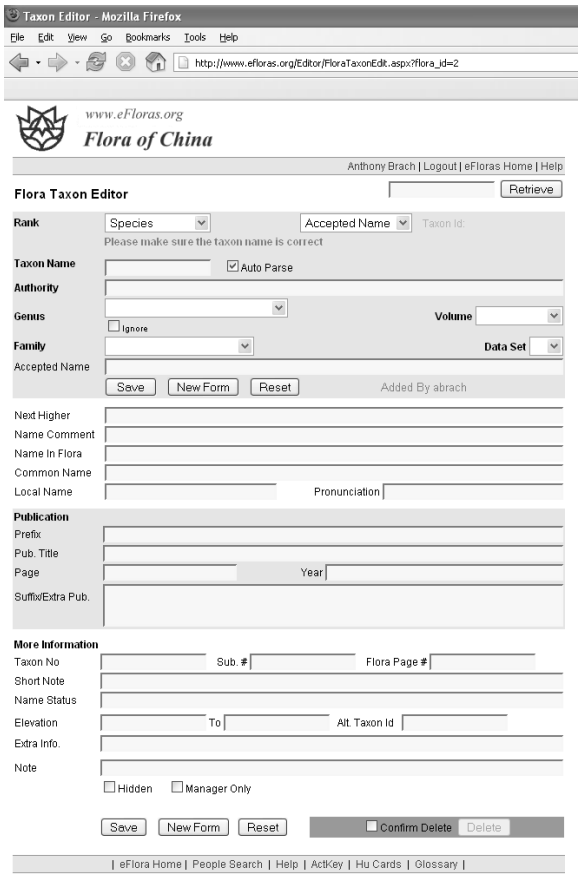

Fig. 1. Flora Taxon Editor.

in 2536, 2241, and 2836 records, respectively.

The centralized database enables searches within and between floras within eFloras.org (see URL: http://www.efloras.org/search_page.aspx?flora_id=0). Thus, users can locate treatments and images of shared taxa in separate floras.

DISCUSSION

The centralized, relational database structure of eFloras presents opportunities to create dynamic links to online taxonomic databases (e.g., IPNI, IOPI Global Plant Checklist; botanists, publications, and specimen databases, e.g., URL: http://www.huh.harvard.edu/databases/). A potentially unlimited number of images and other objects (PDF files, maps, illustrations, web sites, etc.) can be linked to each taxon name. The taxon name in one flora is automatically linked to the other treatments and objects in other floras within the same eFloras database. The search for a taxon name can be based on one or all floras.

DELTA INTKEY (Watson, 1997), and additionally, the Prometheus Taxonomic Project (Raguenaud & al., 2002) offer possibilities to examine classification alternatives (Prometheus I: URL: http://www.dcs.napier.ac.uk/~prometheus/#prometheus1), and for writing well-defined character descriptions (Prometheus II: URL: http://www.dcs.napier.ac.uk/~prometheus/#prometheus2).

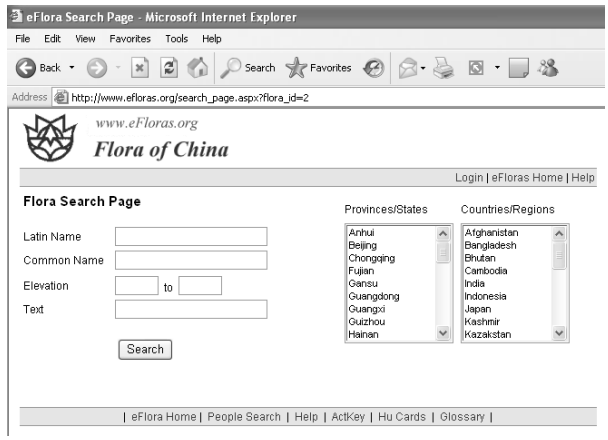

Fig. 2. eFloras: Search Page.

eFloras.org provides a tool for compiling checklists and catalogues. Each project can define its own dataset. A data entry screen will be generated dynamically based on the project's dataset (e.g., A Catalogue of the Vascular Plants of Madagascar, URL: http://www.efloras.org/key_page.aspx?set_id=10001&flora_id=12). The eFloras system provides the structure for creating a fully electronic flora from the start, whereby, the collaborating authors and editors can draft, edit, revise, and review the flora online (e.g., Digital Flora of Taiwan Project).

We are improving searches for interactive identifications (see Brach & Song, 2005). eFloras.org provides an online interactive key builder. Authorized users will be able to use this tool to create and edit character sets, build descriptions of taxa using character data, and create interactive searches of the taxa (e.g., URL: http://www.efloras.org/flora_page.aspx?flora_id=1001).

Additionally, it should be possible to improve text searches using existing technology such as proximity and frequency. Internet agents can collect data pertaining to flora accounts from related sites such as type specimen and medicinal use databases. Eventually, with the creation of formal models of botanical language (e.g., with natural language processing; Beck & al., 1994), and using techniques derived from computational linguistics, it should be possible to transform descriptions that were formerly unstructured data into globally accessible resources.

Converting floristic treatments to Extensible Markup Language (XML, URL: http://www.w3.org/XML/), Structure of Descriptive Data (SDD, see URL: http://160.45.63.11/Projects/TDWG-SDD/), and TaxonX (see URL: http://research.amnh.org/informatics/taxlit/) offers

further possibilities to database treatments (e.g., Cui, 2004) and to provide extensive access to them. However, as a cautionary note from our experiences with earlier parsing, automatic markup of treatments requires much checking and verification.

Online floras allow botanists the opportunity to work on dynamic floristic treatments. The program eFloras (URL: http://www.efloras.org/) enables users to browse treatments by flora volume, family, and genus, and to search by name, distributional data, and free text entries. Taxonomic treatments can be imported and revised online with the use of web forms.

ACKNOWLEDGEMENTS

We are thankful to Myriam Fica (MO) and Noel Cross (deceased) for programming assistance and helpful comments; Hu Shiu-ying (A) for the use of her extensive card file of Chinese plant names; the editorial committee of the Flora of China Project, especially Dave Boufford (A), Mark Watson (E), and Mike Gilbert (MO at K) for helpful comments and suggestions; Michele Funston (MO) for downloads from TROPICOS; Anne Marie Countie (GH) for her dedicated support of servers at Harvard University Herbaria, and others who have made this work possible. We are grateful to Mike Dallwitz (Giralang, Australia) and two anonymous reviewers for helpful review comments. Support of the U.S. National Science Foundation (BSR-8906215, BSR-9201378, DEB 9626806, DEB 0072682, and DBI 0343439) is gratefully acknowledged.

LITERATURE CITED

Beck, H. W., Mobini, A. M. & Kadambari, V. 1994. *A Word is Worth 1000 Pictures: Natural Language Access to Digital Libraries.* URL: http://www.ncsa.uiuc.edu/SDG/IT94/Proceedings/Searching/beck/beckmain.html

Brach, A. R. & Song, H. 2005. ActKey: a web-based interactive identification key Program. *Taxon* 55: 1041–1046.

Cui, H. 2004. *Automating Semantic Markup of Semi-Structured Text via an Induced Knowledge Base: A Case Study Using Floras.* Ph.D. Dissertation, Univ. Illinois, Urbana-Champaign.

Dallwitz, M. J. 1980. A general system for coding taxonomic information. *Taxon* 29: 41–46.

Dallwitz, M. J., Paine, T. A. & Zurcher, E. J. 1993–. User's *Guide to the DELTA System: A General System for Processing Taxonomic Descriptions*, ed. 4. URL: http://delta-intkey.com/. CSIRO Division of Entomology, Canberra.

Dallwitz, M. J., Paine, T. A. & Zurcher, E. J. 2002–. *Interactive Identification Using the Internet.* URL: http://delta-intkey.com/.

Davidse, G., Mario Sousa, S. & Chater, A. O. (eds.). 1994–. *Flora Mesoamericana*. Universidad Nacional Autónoma de México, Instituto de Biología, México, D.F.; Missouri Botanical Garden Press, St. Louis.

Exell, A. W. & Wild, H. (eds.). 1960–. *Flora Zambesiaca: Mozambique, Federation of Rhodesia and Nyasaland, Bechuanaland Protectorate.* Edited by A. W. Exell and H. Wild on behalf of the editorial board, on behalf of the Governments of Portugal, the Federation of Rhodesia and Nyasaland, and the United Kingdom by the Crown Agents for Oversea Governments and Administrations. London. URL: http://www.kew.org/efloras/

Flora of North America Editorial Committee. 1993–. *Flora of North America North of Mexico.* Oxford Univ. Press, New York.

Miller, N. G. & Arriagada, J. E. 2000. Web site and unpublished data sets for the Southeast flora. *Sida Bot. Misc.* 18: 83–96.

Orchard, A. E. & Thompson, H. S. (eds.). 1999–. *Flora of Australia*, ed. 2. ABRS/CSIRO Australia, Canberra.

Raguenaud, C., Pullan, M. R., Watson, M. F., Kennedy, J. B., Newman, M. F. & Barclay, P. J. 2002. Implementation of the Prometheus Taxonomic Model: a comparison of database models and query languages and an introduction to the Prometheus Object-Oriented Model. *Taxon* 51: 131–142.

Tutin, T. G,. Heywood, V. H., Bruges, N. A., Valentine, D. H., Moore D. M., Walters, S. M. & Webb, D. A. (eds.). 1993–. *Flora Europaea*. Cambridge Univ. Press, Cambridge.

Watson, L. 1997 *Angiosperm Families - INTKEY and WWW Packages.* Post to Taxacom Listserver. URL: http://listserv.nhm.ku.edu/cgi-bin/wa.exe?A2=ind9704&L=TAXACOM&P=R1745.

Wu, Z. Y. & Raven, P. H. (eds.). 1994–. *Flora of China.* Science Press, Beijing; Missouri Botanical Garden Press, St. Louis.

Note. – URLs (web pages) accessed on 23 January 2006.