

BIOINFORMATICS: MANAGEMENT AND PUBLICATION OF INSTITUTIONAL DESCRIPTIVE DATASETS AT THE WESTERN AUSTRALIAN HERBARIUM

Alex R. Chapman and Terry D. Macfarlane

Western Australian Herbarium, Department of Conservation and Land Management, Locked Bag 104, Bentley Delivery Centre, WA 6983, AUSTRALIA.



Abstract. Computerised methods for scoring, managing and presenting descriptive information have most commonly been used by individual workers. Some institutions are adopting descriptive data-basing, where standard character lists are developed using Descriptive Language for Taxonomy (DELTA) methodologies to provide comparable yet versatile descriptions for families, genera and species. At the WA Herbarium two state-wide datasets have been developed: *WAGENERA* - comprising descriptions of all families and genera, and the *WA Flora Catalogue* with small standard descriptions for all 12,000 species and infraspecies. Databased descriptions can be regularly maintained and the production of interactive keys and descriptions automated. They can then be integrated with related taxonomic datasets into a comprehensive flora information resource on the web, such as Western Australia's *FloraBase*.

Introduction

Institutional datasets are defined as those for which a scientific institution claims on-going custodianship and are usually related to its regional/political area of responsibility rather than a specific systematic group, which are more commonly the primary responsibility of individuals or small teams.

In the last thirty years institutional information such as census, specimen (eg. herbarium, living collection, DNA extracts and sequences) have become computerised datasets maintained centrally by the custodial organisation. Yet institutional descriptive information still commonly takes the form of floras or monographs with only secondary transformation into an accessible electronic format such as a database. The same can be said for descriptive projects at the national or international level, recent examples of which include *Flora of Australia*, *Flora of China*, or the recently published *Species Plantarum* volume. All have focused on a printed product and subsequently found different routes to capture the information with an electronic approach of some sort.

The capture of descriptive information from the outset using a rigorous method for character definition and comparability has most commonly been the domain of individual or small teams of scientists working on a specific taxonomic group. We believe this is largely because agreement on the definition, management and dissemination of descriptive character data becomes more difficult to achieve as the number of taxonomic groups (and therefore specialists) increases. Nevertheless, we see that the advantages to an institution in applying the same approach to the management of descriptive data as it does to other fundamental datasets to be equally significant.

This poster outlines the approach taken by researchers at one institution to capture, manage and disseminate descriptive data on the state's flora in a coordinated manner using database methodologies.

Objectives and Methods

The state of Western Australia comprises 2,525,500 km², or one-third of the Australian continent and is home to over 12,000 vascular plant taxa. Of the three major phytogeographic areas it is the Southwest Botanical Province (SWBP) which has the greatest diversity with almost half of the species and 79.2% species endemism (Beard, Chapman and Gioia, in prep.). The SWBP is second only to the Cape Floristic Province as a centre of biodiversity for regions with a mediterranean climate. The primary role of the Western Australian Herbarium is to be the centre for inventory and systematic research into the state's diverse and unique flora.

Institutional databasing began at the WA Herbarium in 1985 with the development of a specimen database. Completion of the backlog of specimen databasing occurred in December 1994 and since that time approximately 30,000 specimens have been added each year to take the current total to 460,000 records. Design of a centralised database of plant names relevant to WA began in 1990 and currently contains circa 17,000 accepted, synonymous and misapplied names (see Chapman and Gioia, 1995 for further details).

With the experience of successfully developing these two fundamental datasets, investigations began in 1991 into the feasibility of developing the framework, methods and tools necessary for the systematic long-term capture, maintenance and presentation of descriptive taxonomic information on the state's flora using current computer technology. This project set out to identify:

- a standard for capturing and storing descriptive data,
- mechanisms for integrating descriptive data from various projects,
- flexible methods for interrogating and disseminating this information, and
- strategies for the on-going maintenance of the data once captured.

Additionally, a mechanism for making the descriptive data available seamlessly alongside the other taxonomic datasets under the herbarium's custodianship was required. Rapid changes in computer technology made data capture in open systems a priority.

Results

Previous positive experiences on various systematic projects using DELTA - the Description Language for Taxonomy, a general system for coding taxonomic descriptions (Dallwitz, 1980, Dallwitz *et al.*, 1993) and a broad sampling of the various software which implemented the standard made its choice as a codification standard simple. The DELTA system provided the rigour of a standardised coding syntax and an explicit character definition which enabled the capture of comparable data for the target taxa.

Translation software allowed the codified data to be transformed into a range of outputs, including printed descriptions and interactive identification matrices. Translation of descriptive information into multiple languages also became possible (eg. Jarvie and Ermayanti, *Tree genera of Borneo*, 1995). That a number of software programmers from around the world were actively developing a range of applications which implemented or supported the DELTA standard was considered a significant advantage.

To enable the integration or sharing of data between various descriptive projects a multi-level nested set of characters was proposed. In such a system every descriptive data project would contain a small core set of characters along with the project-specific characters. Taken together, all project characters would comprise the institutional character list (see figure 1). Additional potential for data integration could be achieved if subsequent new projects started with the core characters, then selected suitable characters from the institutional list before defining any new characters.

Character Lists

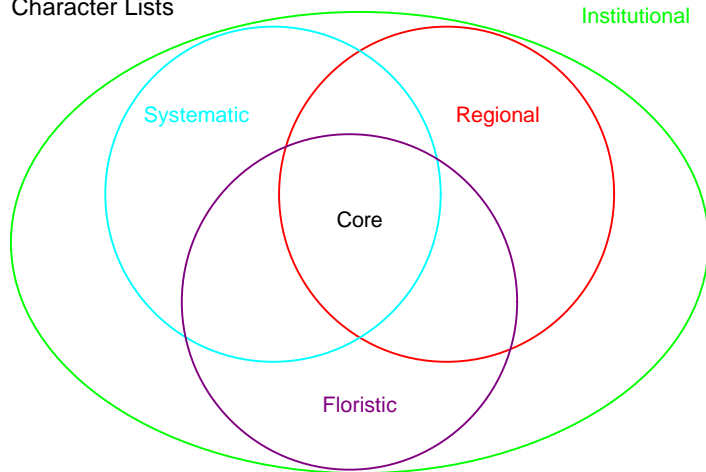


Figure 1. Illustration of the intersection of character lists from a range of project types, each possessing an agreed set of core characters and together building into the institutional character list.

To further increase the potential data compatibility we selected an existing well-developed, clearly defined and widely published character list, from *The Families of Flowering Plants* (Watson and Dallwitz, 1994) as the basis for the initial institutional character list.

Two institutional projects presented themselves for application of these proposed methods, and characters for each were drawn firstly from the newly defined institutional character list. They were:

- *The WA Flora - A Descriptive Catalogue* (Paczkowska and Chapman, in prep.) which scored a dozen basic morphological and edaphic characters for each of the 12,000 vascular taxa in the state, and
- *WAGENERA* (Macfarlane, Watson and Marchant, in prep.) which captured a much larger number of morphological characters for the 1300 genera in WA. Family data from Watson and Dallwitz (1992) was further enhanced and incorporated as part of the project.

It was also realised that the management of large character lists across an increasing number of projects would require an automated approach. In 1996 Chapman and Choo briefly outlined the functionality of a 'DELTA Integrator' which would provide a range of tools for creating, checking and coordinating DELTA-based projects. The design also allowed individual projects to work independently yet 'synchronise' their data with the institutional character list at suitable intervals.

Chapman *et al.* (1995) and Dallwitz (1996) outlined methods for enabling descriptive datasets over the web, either by the transformation of DELTA-coded data directly to hypertext markup language for browsing and linking, or through the direct download of data matrices directly into interactive identification software. A number of alternative methods were also being explored by other workers, either using web-enabled database systems or the use of java applets to remotely query descriptive data.

Discussion

As data capture for the two descriptive projects progressed and prototyping of output products began, a range of additional components were needed to increase these products' functionality. In particular, character notes, illustrations and images were required to simplify the interactive identification process in newly available software. Similarly, taxon images to aid in the verification of identification became desirable. Distribution data and maps could be drawn from the specimen database which by this time had largely been verified and the records geocoded to a given accuracy.

The creation of these 'standard components' such as taxon images and distribution maps, together with developments in integrating the specimen and census databases, all through the adoption of the NameID key as the primary computerised identifier for data elements, meant that a great deal of textual and graphical data was readily available. Concurrent testing of the web as a simplified information delivery mechanism and the movement of all datasets onto a single database platform (TEXPRESS, KE Software) facilitated the development of an integrated web-based information system, subsequently named *FloraBase*.

In *FloraBase* three forms-based query interfaces to the names, specimen and descriptive data are available, but there is a common look and feel to the main report pages and the use of the standard component approach is exemplified by the sample page presented (figure 2), which is assembled on the fly at the time of record retrieval.

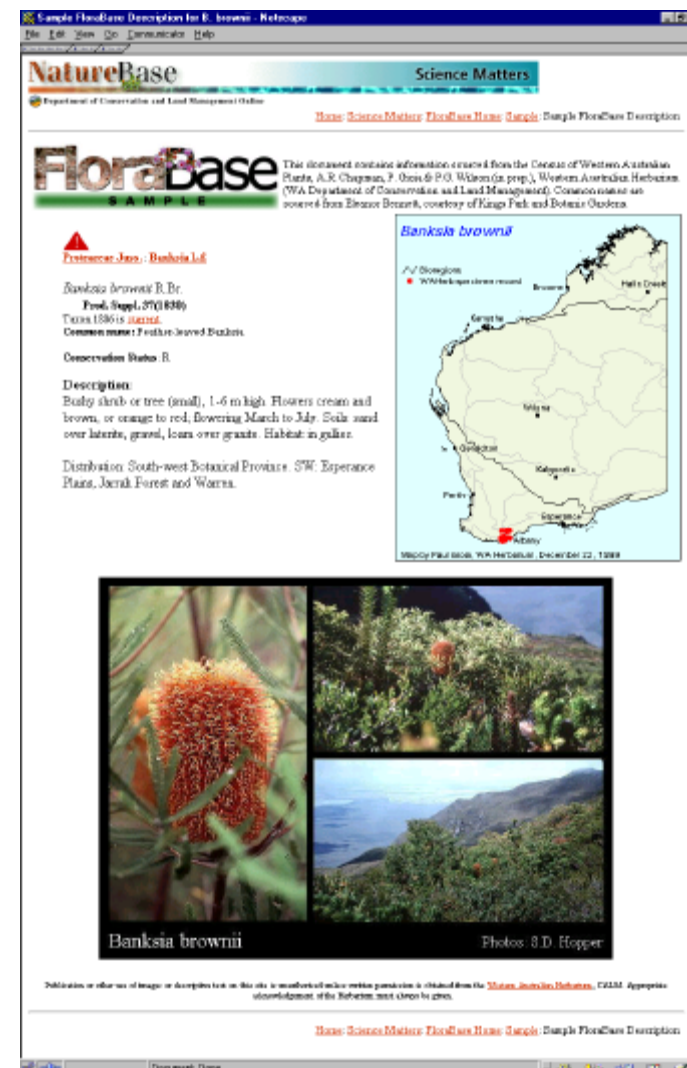


Figure 2. A screenshot of the main descriptive page from *FloraBase* showing the amalgamation of various standard components (classification, citation, status, brief description, distribution, illustration) onto a single page. Such a page can be generated for every vascular plant species or infraspecies recognised as occurring in Western Australia and with all components present in most cases, excepting the taxon image.

In the *FloraBase* page in figure 2, the classification, name, citation, and status data are drawn from the census names table. The distribution statement and map are derived from geocoded records in the specimen database and the full specimen records for the relevant taxon are available by clicking on the map itself. The standard taxon image, illustrating diagnostic features, habit and habitat, is used both for the website and in interactive keys for groups containing the species, and for related CD-ROM products.

The brief description is a warehoused version of the *WA Flora Catalogue* data transformed from its codified DELTA form to HTML and stored in atomic fields in *FloraBase*. It is the textual data that is queryable via *FloraBase*, and that is reformulated on this page as a descriptive paragraph.

The classification hierarchy provides links to comprehensive pages of descriptive data for the respective genus and family, warehoused in *FloraBase* from the *WAGENERA* DELTA database. Use of the data for identification will be enabled not by simple database query as is the case for the species-level descriptions, but utilising the advanced facilities of INTKEY (see Dallwitz, Paine, and Zurcher, 1998). It is planned that customised interactive identification datasets to genera will be linked to in each family page and, providing the end-user has already installed INTKEY, the data matrix, character and taxon images will be downloaded as required from the *FloraBase* web site. This approach is now used by a range of descriptive data projects, including Watson and Dallwitz' (1992) *The Families of Flowering Plants*.

Conclusion

The successful launch of *FloraBase* onto the web in November 1998, and its adoption as the primary source of authoritative data on the Western Australian flora by nearly 1,000 registered users in the following months has created a new model for the dissemination of flora information in the state.

While access to the latest accepted names for the state at the lowest level, or detailed specimen information as an aid to systematic or conservation research at the highest, will remain fundamental to the site, it is envisaged that the most popular and dynamic part of the site will be the descriptive components.

The range and scope of interactive identifications available on the site will expand. It is planned that the small list of characters currently available for species-level descriptions will grow to include the full 'core characters list' as part of the ongoing dataset maintenance. Family and generic level identifications will be added and refined with more illustrations and images, and a closer focus on significant groups can be incorporated. New hypotheses of angiosperm phylogeny may be more readily communicated. Collaborative projects will see similar methods applied to more general subjects such as propagation or weeds.

Advantages extend beyond the immediate increase in institutional profile. Faster access to the full range of authoritative data for staff, combined with moderated end-user access which empowers them to, for example, perform identifications of the state's flora, will allow a renewed focus on systematic research at the herbarium, which benefits the long-term goals for the institution and systematics generally. Comprehensive information systems concerning local flora and managed by the custodial institution should naturally 'anastomose' into national or global information systems. Cornerstones to such a network are now coming into existence such as the *International Plant Names Index* - coordination remains the key.

Literature cited

- Beard, J.S., Chapman, A.R. and Gioia, P. (in prep). Species Richness and Endemism in the Western Australian Flora - An updated assessment in comparison with South Africa and California.
- Chapman, A.R. and Choo, M. (1996). Institutional DELTA Databases: A case study. *DELTA Newsletter* 12. CSIRO Division of Entomology, Canberra.
- Chapman, A.R. and Gioia, P. (1995). The Smart Collection. *Landscape* 10 (4): 49-53. Department of CALM, Perth.
- Chapman, A.R., Lander, N.S., Macfarlane, T.D. and Dallwitz, M.J. (1995). DELTA and Hyper-Text Markup Language. *DELTA Newsletter* 11: 5-7. CSIRO Division of Entomology, Canberra.
- Dallwitz, M. J. (1980). A general system for coding taxonomic descriptions. *Taxon* 29, 41-6.
- Dallwitz, M. (1996). Using Intkey data files directly from the WWW. *DELTA Newsletter* 12. CSIRO Entomology, Canberra.
- Dallwitz, M. J., Paine, T. A., and Zurcher, E. J. (1993 onwards). *User's Guide to the DELTA System: a General System for Processing Taxonomic Descriptions*. 4th edition. <http://biodiversity.uno.edu/delta/>
- Dallwitz, M. J., Paine, T. A., and Zurcher, E. J. (1998). Interactive keys. In *Information Technology, Plant Pathology and Biodiversity*, pp. 201-212. (Eds P. Bridge, P. Jeffries, D. R. Morse, and P. R. Scott.) (CAB International)
- Jarvie, J.K. and Ermayanti (1995 onwards). *Tree Genera of Borneo*. <http://djangoharvard.edu/users/jarvie/borneo.htm>
- Macfarlane, T.D., Watson, L. & Marchant, N.G. (in prep.) *WAGENERA - A floristic database of the families and genera of Western Australian flowering plants*
- Paczkowska, G. & Chapman, A.R. (in prep.) *The WA Flora - A Descriptive Catalogue*
- Watson, L., and Dallwitz, M. J. (1992 onwards). *The Families of Flowering Plants: Descriptions, Illustrations, Identification, and Information Retrieval*. Version: 28th May 1999. <http://biodiversity.uno.edu/delta/>
- Watson, L. and Dallwitz, M.J. (1994). *The Families of Flowering Plants*. CD-ROM. Version 1.0 for MS-DOS. (CSIRO Publications: Melbourne.)
- Western Australian Herbarium (1998 onwards). *FloraBase - Information on the Western Australian flora*. Department of Conservation and Land Management. <http://www.calm.wa.gov.au/science/florabase.html>