



Integration of Taxonomic Descriptive Data across Multiple Database Platforms and Softwares (Weed Information Network — a Case Study)

2001

M. H. C. Choo and A. Spooner

Abstract

Taxonomic data are generally maintained in a number of formats across multiple database software platforms. Descriptive data in particular tends to be maintained in isolation and its integration with related systems is an important and inevitable issue.

This presentation describes the framework for managing DELTA data in an integrated environment. It identifies the steps and processes involved in integrating DELTA descriptive data with other descriptive data and data maintained in other corporate database systems. It describes how the issues of data integrity, currency and duplication are resolved. It uses the Western Australian Herbarium's Weed Information Network (WIN) as a case study to demonstrate how this integration is achieved.

The challenge faced by our DELTA team was to implement an institutional DELTA database with a minimum of staffing resources and, at the same time, maintain currency of plant names and distributions. A coordinated approach, using staff and volunteers to capture only data not already maintained in other corporate systems, was adopted. Critical to this approach is an automated mechanism for extracting and generating DELTA-formatted data from these other systems.

DELIA, the DELta IntegrAtor, developed at the WA Herbarium for managing institutional DELTA data, is the answer. Although it lends itself to general application, DELIA is currently being used to manipulate weed data by mapping, merging and creating DELTA datasets. Lists of weed species are extracted (by family), generated and distributed to scorers as DELTA Editor files. DELIA later merges the scored data into the WIN Master set. Finally, it automatically generates scores from data held in related non-DELTA systems (eg. distribution and family names) and interfaces directly with Confor for natural language descriptions and Intkey for interactive identification.

Aim

The main aim was to provide an integrated environment for the management of institutional DELTA data (Dallwitz 1980; Dallwitz and Paine 1993).

For the Weed Information Network (WIN) project, the aim was to develop an integrated and easy-to-maintain DELTA database with a minimum of resources and within a short timeframe.

Introduction

The Western Australian Herbarium has been managing DELTA projects for a long time, and over the years has wrestled with the issue of maintaining them in an integrated corporate environment.

Our research and development into this area has seen the implementation of a number of projects where descriptive data are scored and then coded in the DELTA format as well as extracted from related systems.

This presentation describes our recent experience at developing the WIN project in an integrated environment.

In the traditional database environment, integration of corporate data is the norm. In the arena of taxonomic descriptive data (Dallwitz, 1980) however, this integration is not readily available and maintenance of discrete databases is generally the norm. Chapman and Choo (1996) explored maintaining institutional descriptive databases and this led to the development of a system for managing institutional DELTA data (Choo, 1998), and the subsequent development of the DELTA integrator (DELIA), a software for integrating and managing DELTA databases.

In early 2001, the WIN project was initiated. It is a Natural Heritage Trust supported initiative to develop a comprehensive weed watch program and online information system for Western Australia. It aims to be the primary authoritative information resource concerning the state's weed species. WIN's DELTA team's charter was to produce an interactive key of all weed species in Western Australia to be accessible online through FloraBase. Minimal funding was available and the project had limited staffing resources. In its favour, a number of relevant corporate systems and softwares were available and WIN had ready access to taxonomic and software expertise.

The challenge faced by our WIN DELTA team was to implement an institutional DELTA database, utilising existing softwares, with a minimum of staffing resources and, at the same time, maintain currency of plant names and distributions.

A coordinated and systematic approach mobilising expertise in management, taxonomy and computer systems was adopted.

Approach

The chosen approach involved a team of volunteers and staff with varying degrees of taxonomic and computer expertise, together with a computer systems specialist. Central to the working processes is DELIA, a software for managing and integrating WIN's DELTA data across hardware and software platforms.

DELIA facilitated the processes by enabling a species template for Western Australian weeds to be generated by combining existing data with the WIN character list. To aid scoring, this data set was then separated into small manageable segments eg by family / genus. Each family was then scored before another family was processed. Scoring was done using CSIRO's DELTA Editor (Dallwitz, Paine and Zurcher 1999).

DELIA was then used to manipulate the scored data and integrate them with other corporate data to form the institutional WIN system.

Finally, data from this institutional system was passed to Confor (Dallwitz, Paine and Zurcher 1993) for syntax checking and natural language production and to Intkey (Dallwitz, Paine and Zurcher 1995) for interactive identification.

During the project the following issues were addressed:

- infrastructure
- project team
- training
- systems integration
- levels of responsibility & custodianship
- security and integrity of data
- system protocols
- change controls and conflict resolution
- integration with other databases
- data manipulation within/across projects
- institutional and core characters
- image management

Infrastructure

The two major corporate data software platforms are ORACLE and TEXPRESS. Paradox is used in-house for smaller systems. The Web is used to deliver the end product. CSIRO's DELTA system is used to input and manipulate DELTA-formatted data. DELIA is used to manipulate and manage institutional DELTA data.

Project Team

The team, who all worked part-time on this project, comprised of 2 project managers, 2 coordinators, a taxonomist, a systems specialist, 12 volunteers and 2 additional staff members.

Training

A series of workshops and training sessions were conducted to familiarise the team with various aspects of the project. Particular attention was paid to the DELTA entry protocols and to consistency in interpretation and measurement.

Systems Integration

The key to maintaining data currency is the adoption of an institutional approach that integrates data from a number of different sources. The WIN project derives its data from sources across different software systems, in both DELTA and non-DELTA format. DELIA provides this integration by maintaining DELTA data in an institutional database and integrating this with associated databases (nomenclatural, specimen data and distribution) in Oracle and TEXPRESS. Within this framework, all DELTA projects are integrated and managed effectively on a corporate basis.

Principally, DELIA was used to perform the following functions:

1. Extract weed species data from a number of DELTA sets (Weed Categories data, South West Flora data and WAGenera data).
2. Generate weed templates using species data as described above and merge them with the WIN character list, ie transform weed scores across projects by mapping characters and states across the different projects.
3. Generate templates of weed species for each family by using relationships held in the master list (WACENSUS), and their conversion to WIN character list for external scoring.
4. Integrate the externally-scored weed species by family into the corporate WIN database.
5. Generate scores for auto characters (eg common name & distribution) using data from related systems.
6. Autogenerate the T-images file from images stored in the system.
7. Create the associated DELTA directive files for integration with CSIRO's DELTA system (eg Confor and Intkey).
8. Transfer to external DELTA softwares (eg. Intkey for interactive identification and Confor for natural language production).
9. Transfer to the Western Australian's FloraBase Web application for online access.

Levels of responsibility & custodianship

The WIN project identified two distinct levels of responsibility, the corporate level and the project level. At the project level, each scorer is responsible for his/her data. When scoring at this level is completed, the data are transferred to the WIN coordinator who then integrates the data into the WIN master set. The WIN coordinator is the custodian of this corporate data.

Security and Integrity of data

All data sets are backed up on the server and the server is backed up daily.

The project coordinators and the systems custodian have control access to the system.

System Protocols

The WIN coordinator maintains the corporate set of files. Any proposed change has to go through the coordinator to the project managers before it is implemented. Data are scored on a family-by-family

basis. For each family, the coordinator generates a template project for the species and loads the template on the scorers' machines for scoring. The project becomes the responsibility of the scorer, who scores and backs up scored data to the network server at the end of each day. When scoring is completed, the scored project is transferred to the coordinator and a new project (family) is initiated by the coordinator. At any time, the coordinator or scorer may run checks and other procedures against the scored data. Errors detected go to the scorer for amendment.

Change Controls and Conflict resolution

Amalgamating a number of established DELTA projects requires assembling a master character list from a number of existing character lists created independently of each other. Definitions of taxa, characters and states may differ in different projects causing data duplication and inconsistencies. The creation of an institutional character list and linking taxon descriptions to related systems will prevent these difficulties from occurring with new projects.

Integration with other databases

A unique nameID is used as a key to access species-level data across all systems. Using this nameID, non-DELTA data (eg. name, genus, family & distribution) can be accessed from the corporate databases and converted automatically to DELTA scores. These DELTA characters are called "auto characters" and are left out of the manual scoring process, avoiding data duplication. Data currency is maintained by generating the auto characters when required.

The coordinator may invoke DELIA at any time to populate auto character scores for WIN.

Data Manipulation within/across projects

Data may be moved freely within and between projects. Any character that is stored in the system may be included in a project. Scored data from any project within the system may be included in a project. Within a project, characters may be moved or re-sequenced.

New projects may be created by nominating characters from a number of projects (or alternatively nominating a subset of characters from a project). Data can then be extracted from existing projects by nominating the required taxa. Only those data scored with characters in the target project will be transferred across.

This feature is used by the coordinator to generate templates for each family ready for the scorers to input scored data.

Institutional and Core characters

The institutional character list is a comprehensive list of characters used in the system. It serves as the authoritative character list and is maintained by the custodian. New projects draw characters from this list. In addition to this is a list of core characters to be used by all projects. WIN derives its characters from these two lists.

Image Management

The NameID is also used to integrate taxon images to the system. This is done by incorporating the nameID as part of the image file name e.g. 10001ic1.jpg. Given a nameID and an algorithm for generating the name, the system can then associate all images to a specified taxon.

Results and Discussions

The results obtained by the WIN's project team were encouraging.

In five months and using a list of 117 characters, the team was able to produce over 100 fully coded and checked taxon descriptions and 87 fully coded and partially checked descriptions, documenting the Western Australian weeds belonging to 6 families. Over 1000 specimens were examined during this process. The DELTA data set thus produced is fully integrated with other related corporate data.

Initial and ongoing training of the volunteers and the botanical and technical support made available to them ensured their interest in and loyalty to the project. During the project their botanical knowledge and computing skills were extended, increasing their confidence in these fields.

Volunteers provided significant cost savings and their inclusion in any project is worth exploring at this time, when funding is limited.

Adding a systems specialist to the team provided the technical knowledge to automate and integrate many of the processes, a situation that is becoming increasingly important in this electronic age.

Streamlining the processes and separating the functions enabled our resources to be used efficiently. This was made possible by using DELIA to automate many of the processes:

- DELIA's integration with related corporate data ensured non-duplication of coding and allowed each custodian to maintain control of his own data.
- DELIA's ability to find intersecting records between datasets allowed the total species list to be intersected with the weed species list to automatically produce the weed list for each family. This list was then merged with the WIN character list to automatically generate templates for scoring the weed species for each family.
- DELIA's ability to relate across taxonomic hierarchies allowed species-level data to be generated from family and genus level scores (inheritance of higher level scores by lower levels). This facility was used repeatedly in the generation of template lists of species in Western Australia for nominated families.
- DELIA's ability to move items and characters between projects provided the mechanism to create tailored lists of weeds.
- Its exception reporting facility provided the mechanism to trap and correct errors.

In the WIN project we used these features to:

1. Share characters and items across projects. (DELIA's ability to do this makes it useful not only for managing but also for exploring data).
2. Extract and merge family and genus level data with WIN's species level data.
3. Generate required weed species templates using weed species data from a number of projects. This automatic generation of templates for input facilitated the keying process.
4. Mask out redundant characters thus simplifying the input process.

Therefore not only were we able to generate 'personalised' datasets for our volunteers but also to partially populate those data sets. This reduced complexity and minimised errors inherent in a project of this kind, such as the mis-keying of botanical names and the mis-interpretation of data.

DELIA's facilities to manipulate characters and items across projects allowed the team to explore and modify characters during the initial project phase. It eliminated the need for the character list to be finalised before scoring commenced.

Scoring by family proved effective. It provided the team with a systematic approach whereby each family could be dealt with before embarking on the next. It also provided the facility to engage any number of coders without the problem of data duplication. It also allowed priority species to be dealt with first. This flexibility also allowed the volunteers to choose to score a particular family in their area of interest or expertise.

DELIA's ability to generate character scores based on the scores of other characters (ie derived characters) provided the mechanism for keeping current those character scores whose scores are derived from other characters. An example is *weed category*, which is based on *distribution*. As the distribution of a weed changes, its category changes. DELIA interrogates distribution and generates the category score based on the weed's current distribution.

Integration with corporate databases allows data integrity to be maintained.

The taxon image (Timage) automatic generation facility provided automatic mapping of current images to the system. It provided the mechanism for a separate team of volunteers to continually and independently scan and update taxon images. In addition, it eliminated the time necessary to manually create and manage a file of all weed images (>1000 species).

DELIA's interface with DELTA provided the mechanism for validation and processing of scored data at will. It provided valuable feedback and control during the development phases. Its ability to generate DELTA directive files makes it easy for non-DELTA users to use the DELTA suite of programs, typically Confor and Intkey. This provided valuable feedback during the input phases (Confor checks, interactive keys and natural language outputs).

Summary

Corporate DELTA projects straddle hardware and software platforms. A software that understands the DELTA format and can communicate across the various platforms is necessary to avoid duplication and to maintain the currency and integrity of data. DELIA fills this role and provides the desired integration for the Western Australian Weed Information Network project.

The WIN project exemplifies the integration problems faced by DELTA project teams. Their resolution provides other teams with the means of achieving this integration.

The WIN project demonstrated the feasibility of maintaining an integrated, easy-to-maintain taxonomic descriptive database.

References

- Chapman A.R., and Choo, M.H.C. 1996. Institutional databases — a case study. *DELTA Newsletter*. 12: 14–15.
- Choo, M.H.C. 1998. An institutional model for managing descriptive taxonomic data (Abstract). Biodiversity, Biotechnology & Biobusiness: 2nd Asia-Pacific Conference on Biotechnology: Perth, Western Australia, 23–27 November 1998. Programme & Abstracts, pp 37. (Eds. M van Kuelen & M.A. Borrowitzka). Australian Biotechnology Association (W.A. Branch), Perth.
- Dallwitz, M. J. 1980. A general system for coding taxonomic descriptions. *Taxon* 29: 41–6.
- Dallwitz, M. J., and Paine, T.A. 1993 onwards. Definition of the DELTA format. <http://delta-intkey.com>
- 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://delta-intkey.com>
- Dallwitz, M. J., Paine, T. A., and Zurcher, E. J. 1995 onwards. User's guide to Intkey: a program for interactive identification and information retrieval. <http://delta-intkey.com>
- Dallwitz, M. J., Paine, T. A., and Zurcher, E. J. 1999 onwards. User's guide to the DELTA Editor. <http://delta-intkey.com>
- Western Australian Herbarium. 1998. FloraBase — Information on the Western Australian flora. Department of Conservation and Land Management. <http://www.calm.wa.gov.au/science/florabase.html>