Retrieval of the Data from the Conservation Atlas

Group Members: Specht A.[1], Specht R.L.[2], Bolton, M.P.[3], Belbin, L.[4] and Kingsford B.[5]

[1]Australian Centre for Ecological Analysis and Synthesis; [2]University of Queensland, St Lucia, Queensland; [3]Corymbia Ecospatial Consultants, Canberra, ACT ,[4]The Atlas of Living Australia, Canberra, ACT; [5]Data Structured, Hawker, ACT.

Project Objectives

The conversion of collated species lists prepared for the *Conservation Atlas of Plant Communities in Australia* (Specht *et al.* 1995) from paper and digital text formats (species codes x locations x publications) into a collection of species observation records using the Darwin Core international standard. The resulting records will be discoverable through the Atlas of Living Australia and the Terrestrial Ecosystem Research Network providing a sustainable open-access resource to the research community and the public.

Project Background

In the first decades of European settlement observations of flora and fauna ranged from ad hoc descriptions by explorers and new settlers, artists' drawings, to specimens sent to international herbaria and museums. The early surveys were haphazard in terms of location, survey method and species coverage with little attempt to co-ordinate the work nationally. The visit of the British Association for the Advancement of Science in 1914 encouraged more systematic observations on the vegetation (Osborne, 1914, Specht and Specht, 1962) and these formed the basis of today's more quantitative approaches.

As a contribution to the International Biological Programme (IBP), the Conservation Survey of Australia (Specht *et al*. 1974) listed the major and minor plant communities extracted from the published literature and the unpublished field experience of plant ecologists. These data were incorporated into a committee-based delineation of Australia's plant communities (Specht, 1970). The great number of tree species (20 to 140) that coexist in subtropical and tropical plant communities in Australia, however, made the definition of "homogeneous" vegetation suites a challenge compared with temperate communities with 1-10 overstorey species per hectare.

The polythetic-divisive classificatory program TWINSPAN (Hill, 1973, 1979) had recently become available, and it was determined it could be used to better define these units for conservation assessment. Complete species lists reported in 705 Australian publications (refereed literature, government surveys and theses) from the 1870s were compiled and collated according to vegetation formation (using the structural classification of Specht, 1970) and geographic division (Figure 1). These segments were analyzed by Mike Dale *et al.* using the then CSIRO Division of Computing Research CSIRONet facilities based on the presence of species to determine more objective plant communities. These communities were then linked to their conservation status[1]. A spatially-represented outcome of this work was published in 1995 (Specht *et al*., 1995).

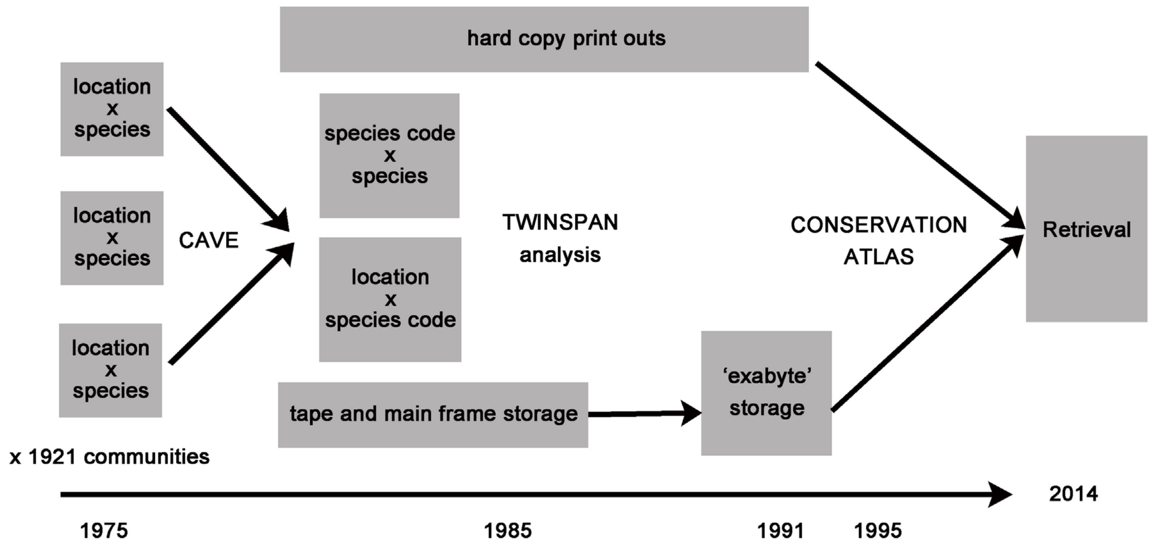

Figure 1: The path from initial collation of literature through to the retrieval project. Digital backup in 1991 used Exabyte tapes. CAVE: Classification of Australian Vegetation system developed to pre- and post-process the data using University of Queensland computing facilities.

[1]It was reasoned that if all major plant communities (and associated abiotic habitat) were conserved, then associated environments and their biodiversity would also be conserved. It was assumed that because all animal species are dependent on vegetation either for food or shelter or both, most of the resident invertebrates and vertebrates would be included in these reserves. Special considerations would need to be made for the conservation of mobile vertebrates and invertebrates.

Project Background (continued)

Original publications and hard-copy output were archived in storage boxes and in filing cabinets (Figure 2). Digital data were initially stored on hard drives, but with the high account and storage costs at the time, were saved to large reel magnetic tapes and in 1991 to 'Exabyte' tapes and stored in two separate locations (Figure 1). With the establishment of the Atlas of Living Australia (www.ala.org.au) in 2005 and the Terrestrial Ecosystem Research Network (www.tern.org.au) in 2009 and their objectives to collect, integrate and share biodiversity-related data for the better understanding and management of our ecosystems, came the realization that these data needed to be exposed publically.

The retrieval project started in June 2014.

Figure 2: File storage boxes (left) in which many original papers used for the project have been stored for three decades, and examples of print-outs of alphanumeric codes by location (right). The scale in left image is 20cm per colour band.

Methods

Existing computer printouts were assembled and the number of locations, communities and species were estimated for each vegetation formation (Table 1). The Exabyte tapes contained text files, source data and analytical results.

Table 1: Data used for the Conservation Atlas available through print-outs.

Formation+	Locations	Communities	Number of Species*
Sclerophyll vegetation SW W.A.	64	172	1,761
Sclerophyll vegetation E. Aust.	189	544	2,581
Savanna understorey	57	198	1,313
Mallee open-scrub	28	41	395
Desert Acacia	54	148	1,229
Chenopod low shrubland	25	55	410
Forested wetland (incl. brigalow)	30	34	193
Arid wetlands	20	41	642
Freshwater swamp vegetation	80	80	139
Coastal dune vegetation	47#	49	315

* Not including introduced species and species recorded in only one community, but included in the raw data

+ Some formations were analysed separately from the pathway outlined in Figure 1 and were not readily available for this project. These were the rainforests, dry scrubs (NT & south-east Qld), alpine vegetation, coastal wetlands and hummock grass understorey.

The tasks were to:

1. Recover all available data from
a. Hard copy
b. Exabyte tape
c. Other data in digital form (e.g., Excel spreadsheets)

2. Design a structure that reflected how the data should be viewed from current perspectives
a. Site data/metadata (latitude longitude x vegetation structure x comments)
b. Species alphanumeric codes and their associated scientific names
c. Sites x species codes (some with multiple communities)

3. Update the species codes/names to current nomenclature
a. Use the Atlas of Living Australia's web services (api.ala.org.au), the National Species Lists and Australian Plant Census (CHAH)) to semi-automate the current identification of species names
b. Manually check any ambiguous or missing names

4. Generate a mapping of the fields used in the Conservation Atlas/CAVES project to the Darwin Core Standard (http://rs.tdwg.org/dwc/).
a. Collate the terms used in the previous studies
b. Determine the intent of the fields
c. Find the best equivalent term in the Darwin Core standard

5. Collate and integrate the data
Produce a list of species observations using the Darwin Core at each site (defined by a consensus latitude/longitude) with metadata that includes vegetation type/structure, references, author and processing comments.

6. Generate a collection-level metadata record.

aceas
Australian Centre for Ecological Analysis & Synthesis

Major Findings

The fundamental data in this project were plant species lists for 1,378 plant communities throughout Australia from 1879 to 1994. These data provide a unique baseline historical record of high value as such continent-wide observations cannot be repeated. Their value also lies in providing a basis for comparison with current-day observations.

The project was expected to highlight the challenges of data curation and the conversion from old formats, and it did not disappoint.

• Conversion of paper records to digital format was a challenge that was made easier by the timely recovery of significant data from Exabyte tapes. Finding a reader for the Exabyte tapes and their contents was not easy, reminding us that rapid technology changes require regular updating of stored data formats to ensure recovery.

• A master site data/metadata file aligned the many disparate original files and highlighted many inconsistencies, for example, how sites were coded. Manual alignment for consistency was required but some inconsistencies were able to be overcome by judicious programming.

• The largest workload stemmed from the many significant taxonomic changes due to re-classification either along geographical boundaries, or complete name changes due to taxonomic work in the intervening 25 years. The ALA's web services enabled sophisticated look-ups of previous to current names but a proportion of fuzzy or poor matches required significant manual validation. This highlighted the need for more extensive lists of historical changes, but even such lists could not address incorrect original names.

• There has been a significant evolution in the precision and accuracy of recording site locations. The data sources also varied from landscape to accurate site scale.

Key papers or products

These data will be made available through the public data repositories of the Atlas of Living Australia (www.ala.org.au; Belbin 2011) and the Terrestrial Ecosystem Research Network (www.tern.org.au).

One paper is in preparation about the challenges of retrieval of such archival data and the workflow developed.

How will this affect Australian ecosystem science and management?

This collection will provide a significant, extensive and complementary data collection into the public domain, greatly enhancing our understanding of past conditions and aid more effective management of Australia's natural resources.

Without an historical reference point there is no possibility of comparison into the future. Without historic data you cannot anticipate or evaluate the effect of climate and land use change.

The data released will complement other collection methods, giving a more complete picture of any particular location in space and time. The effort required in the recovery of these data demonstrate the imperative of ensuring any and all biodiversity data needs to be urgently integrated into a national repository before it is lost.

References

Belbin, L. (2011) The Atlas of Living Australia's Spatial Portal. pp. 39-43. In, *Proceedings of the Environmental Information Management Conference 2011* (*EIM 2011*) (Eds *M.B. Jones*, M.B. & C. Gries, C.), Santa Barbara, California, USA.

Hill, M.O. (1973) Reciprocal averaging: an eigen vector method of ordination. *J. Ecology* 61: 237-49.

Hill, M.O. (1979) TWINSPAN– A FORTRAN program for arranging multivariate data in an ordered two-way table by classification of the individuals and attributes. Section of Ecology and Systematics, Cornell University, Ithaca, NY, USA.

Osborne, T.G.B. (1914) Sketches of vegetation at home and abroad. VIII. Notes on the flora around Adelaide, South Australia. *New Phytologist* 13: 109-21.

Specht, M. and Specht, R.L. (1962) Bibliographia Phytosociologia Australia. Excerpta Botanica. Sectio B, 4:1-58.

Specht, R.L. (1970) Vegetation pp. 44-67. In, *The Australian Environment.* Fourth Edition (Ed. G.W. Leeper) CSIRO and Melbourne University Press, Melbourne, Vic, Australia.

Specht, R.L., Roe, E.M. and Boughton, V.H. (eds) (1974) Conservation of major plant communities in Australia and Papua New Guinea. *Aust. J. Botany Suppl. No. 7.*

Specht, R.L., Specht, A. Whelan, M.B. and Hegarty, E.E (1995) *Conservation Atlas of Plant Communities in Australia.* Southern Cross University Press in association with the Centre for Coastal Management, Lismore, NSW, Australia.