Date of Publication:
30/07/2012

# Cunninghamia

A journal of plant ecology for eastern Australia

The Royal Botanic Gardens & Domain Trust

ISSN 0727-9620 (print) • ISSN 2200-405X (Online)

SHORT COMMUNICATION

# Using Australian Virtual Herbarium data to find all the woody rain forest plants in Australia

*[1,2]Robert Kooyman, [1]Maurizio Rossetto and [3]Shawn Laffan*

[1]National Herbarium of New South Wales, Royal Botanic Gardens & Domain Trust, Mrs Macquaries Road, Sydney, NSW 2000 AUSTRALIA;
[2]Department of Biological Sciences, Macquarie University, NSW 2109 AUSTRALIA;
[3]School of Biological, Earth and Environmental Sciences, University of New South Wales, NSW 2052 AUSTRALIA.
Corresponding author Robert Kooyman Email: robert@ecodingo.com.au; Rob.Kooyman@environment.nsw.gov.au

***Abstract***: Data bases that provide continental and global scale information about species distributions provide a valuable resource for environmental, ecological and evolutionary research. However to bring a large dataset to a standard that is suitable for quantitative analysis, data quality needed to be checked. Here we provide a worked example using a large dataset (c. 320,000 records) from Australia's Virtual Herbarium (AVH) database, based on an initial data request for full distribution data for c. 2600 woody rain forest species known to occur in Australia. To reconcile inconsistencies around taxonomic identity prior to merging with our trait data-base, and resolve issues around spatial resolution and accuracy, we implemented extensive data filtering using a 'cloud-based' solution (Google Refine). This systematic process resulted in 1) the removal of close to 45% of the records originally downloaded, and 2) a clean and powerful data set based on herbarium backed distribution records for Australia's woody rain forest species. Such resources can contribute significantly to improving research outcomes related to understanding Australia's vegetation.

Key words: Australian Virtual Herbarium, Atlas of Living Australia, databases, data cleaning, rain forest woody plants, species distributions

***Cunninghamia*** **(2012) 12(3): 177-180**
doi: 10.7551/cunninghamia.2012.12.014

## Introduction

Collections record databases such as Australia's Virtual Herbarium (AVH: www.chah.gov.au/avh) and the Australian Faunal Directory (AFD) delivered through the Atlas of Living Australia (ALA) (www.ala.org.au) provide a valuable resource describing the spatial distribution of Australian flora and fauna species. These sorts of data are increasingly used in environmental, ecological and evolutionary research, for example, 1) as part of global biodiversity informatics compilations, 2) to examine the spatial distribution of species in relation to climate change susceptibility (Gallagher et al. 2009), 3) to examine the distribution of species across geographic space and environmental gradients (Crisp et al. 2001; Mellick et al. 2011), and 4) in relation to quantified measures of evolutionary signals from phylogeny, taxonomy, endemism and genetic diversity (Bickford et al. 2004; Laffan & Crisp 2003; Rosauer et al. 2009).

The purpose of this short note is to highlight some of the data quality issues that might need to be dealt with in order to bring these large datasets to a standard that is suitable for quantitative ecological research. The spatial distribution issues of these data are well documented (e.g. Chapman 1998; Crisp et al. 2001; Newbold 2010), but there remain other issues of database accuracy, from factors such as taxonomic revisions, and data entry errors and ambiguities. A strategy for cleaning a data set of these issues is rarely examined in detail. Here we list the general issues and major problems relating to such datasets, and illustrate how we systematically addressed them in an example from our own research. Our research questions required location and distribution data for all Australian woody rain forest species (trees, shrubs and vines). These data needed to be reconciled with a trait data base and a background tree (phylogenetic) for detailed evolutionary-ecology analyses using updated information from the Angiosperm Phylogeny Group III website (APGIII, http://www.mobot.org/mobot/research/apweb/). While the data base (which contains c. 6 M collection records) is available to the public, some limitations are in place related to accessing full location data. Consequently full access to the dataset was obtained through the National Herbarium of NSW.

*Methods and progressive outcomes of data cleaning*

Our research required the compilation of distributional records (latitude and longitude) for c. 2600 woody Australian rain forest plant taxa (trees, shrubs and vines). This compilation was intended to cover the full distribution of rain forest on the Australian continent and the whole eastern sea-board from Tasmania to Cape York, and westward to the Northern Territory and the Kimberley region. Records for some taxa extended into drier non-rain forest areas of the continent.

The downloaded data set contained a range of inconsistencies across various fields that complicated the automation of follow-up analyses. For example: 1) collection locations by states given in full and abbreviated form, with and without capitalization (NT, Northern Territory, northern territory), and missing entirely (nulls); 2) plant names and authorities in multiple forms, both within a species and between species; 3) 'subspecies' and 'variety' abbreviated in different ways (e.g. subsp. and ssp.); 4) multiple acronyms for each institution; 5) null values for many geographical coordinates; 6) inconsistent geocode precision, with numerous nulls; 7) collection dates given in different formats; and 8) replications of the same collection(s). A range of issues emerged:-

*Taxonomic precision, reconciling names (part I)*

A total of c. 320 000 records were initially downloaded for the c. 2600 name queries. These data were loaded into Google Refine (www.google.com/p/google-refine/) as a comma separated variable (csv) file. We used a text facet feature to identify and group all taxa. At this stage the c. 2600 species were represented by 6742 names. A text clustering algorithm was then used to group names by (text) similarity. This process identified 672 clusters that represented names with close equivalents such as subspecies, variety, or the same name with different taxonomic authorities. As a first step these were reconciled and merged, where appropriate. The clustering process resulted in the retention of c. 210 names from the 672 clusters, leaving 6280 names.

*Taxonomic precision, updating taxonomy (part II)*

Most studies will need a systematic process of reconciling and updating species names to an accepted standard. Here we used APGIII and a process of reconciling with the underlying phylogenetic file of all Australian rain forest taxa we had previously created for this purpose (Kooyman personal data). This involved working through the remaining 6280 names in the data set one at a time using the text facet function (in Google Refine). Where appropriate, identities were merged, allocated to subspecies or variety consistent with the literature, or, very occasionally, deleted where infra-

**Table 1 Steps undertaken in cleaning data downloaded from Australia's Virtual Herbarium data base. Initial data request was for full distribution data for c. 2600 woody rain forest species known to occur in Australia.**

| STEPS | Records (c. number) | Species ID's (c. number) |
|---|---|---|
| initial download | 320000 | 6742 |
| *Taxonomic precision (part 1)* | | |
| first filter (clustering) | 320000 | 6280 |
| *Taxonomic precision (part 2)* | | |
| filtering by species | 320000 | 2600 |
| conspecific merge | 320000 | 2560 |
| *Defining and delimiting the data set* | | |
| remove marginal taxa | 280000 | 2360 |
| final taxonomic clean | 270000 | 2300 |
| *Data quality and precision, geocode precision* | | |
| geocode precision | 260000 | 2300 |
| duplications (distribution records) | 230000 | 2300 |
| *Data quality and precision, cultivated material* | | |
| cultivated specimens | 220000 | 2300 |
| *Data quality and precision, latitude and longitude* | | |
| spatial filtering (coordinates) | 200000 | 2300 |
| *Final spatial filtering and corrections* | | |
| spatial filtering (map and coordinates) | 180000 | 2300 |

specific species could not be reconciled with extant taxa. This reduced the number of names to c. 2600 which were then checked against the base phylogenetic file that included all of Australia's woody rain forest plant species. At this stage we identified 17 taxa missing from the original data request due to spelling errors or different spellings in our original search. These were corrected and the data was downloaded and merged with the partly cleaned data, resulting in all species being present in the data set. At this point we identified a number of 'unnamed species' conspecific with other 'species' in the data. Merging these reduced the final number of species by c. 40 taxa.

*Defining and delimiting the dataset*

The focus of this study was on woody obligate rain forest species. We deliberately excluded mangroves, wet and dry sclerophyll species, herbs, and some desert taxa that occur close to monsoon vine forests in the tropics. This filtering process resulted in the removal of c. 200 species (in total) from those originally requested. Palms, ferns and cordylines were not included in the original data request because trait compilations were focused on woody taxa that had comparable and equivalent traits (refer to Kooyman et al., 2010, 2011, 2012).

*Data quality and precision, geocode precision*

Spatial resolution is an issue for all studies, and choices around what level of geocode precision to retain will be influenced by a range of factors, including, for example, the desirability of retaining older collection records for comparison. We allocated the AVH measures of geocode precision to ranked values. All records with geocode precision >25 000 metres were removed, resulting in c.10 000 records including most of the 'older' records being deleted. A secondary filtering process based on an assessment of geocode precision relative to other location records for the same taxon was then undertaken to test if locations with differing precisions were closely aligned (or widely divergent) for the same locations (described in the text location descriptions). Records that aligned consistently were retained. Records that did not were removed. In addition, whenever duplicate records were encountered they were removed. This resulted in c. 23 000 records (total) being removed.

*Data quality and precision, cultivated material*

The focus then shifted onto removing records of all non-wild-collected (cultivated) specimens. Cultivated specimens were defined as those recorded as having been collected from public and private botanic gardens, arboreta, and research and experimental plantings. Filtering searches required every name variation for each agency, individual, and herbarium to be identified and used (with case sensitivity removed) to identify all cultivated specimens. These were all removed, resulting in another large (c. 10 000) reduction in numbers.

*Data quality and precision, latitude and longitude*

Spatial distribution accuracy is a key component of data reliability for analyses. We began by filtering the data by species locations (latitude and longitude). Latitude and longitude fields (columns) were first duplicated and rounded down to the nearest degree (no decimal places). This allowed for another round of facet filtering by species and locations for each taxon. The known distributional extent of species (Kooyman personal data) was used to check if the AVH records fell within the broad parameters of species known distributions. The process of checking (though still reliant on prior knowledge) can also be performed in the Biodiverse software package (Laffan et al. 2010). The filtering process highlighted numerous records with incorrect coordinates (including records seaward of the coast and not corresponding to known islands with rainforest), and incorrect data entry. These were removed for all taxa one at a time, resulting in many 1000s of deletions. An alternative to this process would be a record by record comparison of geocodes (Lat. Long.) to described collection locations, followed by manual correction of latitude and longitude errors. While considerably more laborious, this process could result in the retention of many more records, but time constraints did not permit this option in this case.

Following that filtering process, the spatial mapping facet was activated (in Google Refine) using the actual coordinates. This allowed all remaining species records seaward of the coast, and other anomalous records to be highlighted. Those not occurring on islands, or on distant oceanic islands, were removed. A species by species reconciliation of known spatial distribution with the AVH data distribution records was then undertaken for all taxa in the list. This process also detected the residue of cultivated arboretum and collection (planting) records remaining in the data, and the remaining errors (incorrect data entries, poor location information and other factors). All these were removed.

*Final filtering*

At this point c. 45% of the records originally downloaded for the research project had been removed for various reasons in the data cleaning and filtering process. Several of the filtering processes were then repeated (including text facet filtering, location coordinate filtering, and location description filtering) to detect any residue of error using text identification commands with variations of key words previously found to provide the main errors. Other command lines (e.g. planted, 'name of towns', street, island) were then tested to detect any other logical errors. As a final filter the whole process of spatial reconciliation for each species (c. 2300) was repeated and double checked. This resulted in the removal of a small but significant number of errors (e.g. inland and desert locations for coastal species) that would create noise in any subsequent spatial analyses.

The outcomes of the filtering processes described are listed in Table 1. Following the final filtering the data could be described as taxonomically and spatially clean, with high accuracy relative to available data on species distributions.

## Discussion

For such a large dataset, many weeks were needed to complete data filtering, and the final outcome was made possible by a close working knowledge of the flora and the individual taxa. The objectives of other researchers relative to different projects may vary significantly, as will the time required to undertake the filtering and cleaning. However, the general process and sequence of steps in filtering and cleaning the data using the tools suggested will remain similar, but may vary relative to research objectives. The limitations of herbaria and museum data in relation to capturing complete species distributions was described by Newbold (2010) who advocated supplementing those data with distribution modelling. Depending on the scales and resolution required for any particular study question, this remains an option (e.g. Crisp et al. 2001; Mellick et al. 2011).

Alternative processes for some aspects of the data cleaning and reconciliation include bringing species binomials to a common taxonomy across datasets by matching names against the accepted names in the Plant List (www.theplantlist.org/) and / or sources such as the International Plant Name Index (IPNI; www.ipni.org/) and Tropicos (www.tropicos.org/). Binomial matching with the Plant List and IPNI can be done using a matching algorithm (in software such as R). Species that return multiple matches (using the algorithm) must still be examined and corrections will need to be made one at a time.

It is becoming increasingly common for researchers to work with large continental and global scale data sets (Moles 2005; Chave et al. 2009). The challenge of these databases is that analyses must be preceded by intensive cleaning and filtering and be done with great care and attention to spatial detail (Chapman 1998). However, once those steps are taken, it is equally clear that these are incredibly valuable resources, that provide researchers with direct access to spatial data that is herbarium vouchered and validated. It would be difficult to overstate the value of such data, particularly when it is used in conjunction with analyses designed to examine and compare measures of diversity or endemism, or to look at the phylogenetic signal of diversity across geographic space (Crisp et al. 2001; Laffan & Crisp 2003; Bickford et al. 2004). Once issues around taxonomy (e.g. conspecifics, synonyms and unresolved subspecies) and spatial accuracy are resolved the data become very powerful. For the future, the potential is for the development of web-based flora lists that include similarly accessible tabulated details on a range of informative traits (generally already listed in the text of existing floras). That addition would assist local, continental and global scale studies even further. What is evident is that even in cases where only general patterns are being searched for, such resources provide the potential to uncover some of the most significant secrets of Australia's vegetation.

## References

Atlas of Living Australian (ALA) www.ala.org.au/

Australian Virtual Herbarium (AVH) www.chah.gov.au/avh

Bickford, S.A., Laffan, S.W., de Kok, R.P.J. and Orthia, L.A. (2004) Spatial analysis of taxonomic and genetic patterns and their potential for understanding evolutionary histories. *Journal of Biogeography* 31: 1715–1733.

Chapman, A.D. (1998) Quality control and validation of point-sourced environmental resource data. Third International Symposium on Spatial Accuracy Assessment: Land Information Uncertainty in Natural Resources, Quebec (ed. by K. Lowell and A. Jaton), Ann Arbor Press, Chelsea, MI.

Chave, J., Coomes, D., Jansen, S., Lewis, S.L., Swenson, N.G. and Zanne, A.E. (2009) Towards a worldwide wood economics spectrum. *Ecology Letters* 12: 351-366.

Crisp, M.D., Laffan, S., Linder, H.P. and Munro, A. (2001) Endemism in the Australian Flora. *Journal of Biogeography* 28: 183-198.

Gallagher, R.V., Hughes, L. A and Leishman, M.R. (2009) Phenological trends among Australian alpine species: using herbarium records to identify climate-change indicators. *Australian Journal of Botany* 57: 1–9.

Google Refine. http://code.google.com/p/google-refine/

Kooyman, R.M., Cornwell, W. and Westoby, M. (2010) Plant functional traits in Australian sub-tropical rain forest: partitioning within community from cross-landscape variation. *Journal of Ecology* 98: 517-525.

Kooyman, R.M., Rossetto, M., Cornwell, W. and Westoby, M. (2011) Phylogenetic tests of community assembly across regional to continental scales in tropical and sub-tropical rainforests. *Global Ecology and Biogeography* 20: 707-716.

Kooyman, R.M., Rossetto, M., Allen, C. and Cornwell, W. (2012) Australian tropical and sub-tropical rainforest: phylogeny, functional biogeography, and environmental gradients. *Biotropica* 10.1111/j.1744-7429.2012.00861.x

Laffan, S.W. and Crisp, M.D. (2003) Assessing endemism at multiple spatial scales, with an example from the Australian vascular flora. *Journal of Biogeography* 30: 511-520.

Laffan, S.W., Lubarsky, E. and Rosauer, D.F. (2010) Biodiverse, a tool for the spatial analysis of biological and related diversity. *Ecography* 33: 643-647.

Mellick, R., Lowe, A. Rossetto, M. (2011) Consequences of long- and short-term fragmentation on the genetic diversity and differentiation of a late successional rainforest conifer. *Australian Journal of Botany* 59: 351-362.

Moles, A.T., Ackerly, D.D., Webb, C.O., Tweddle, J.C., Dickie, J.B. and Westoby, M. (2005) A brief history of seed size. *Science* 307: 576-580.

Newbold, T. (2010) Applications and limitations of museum data for conservation and ecology, with particular attention to species distribution models. *Progress in Physical Geography* 34: 3-22.

Rosauer, D., Laffan, S.W., Crisp, M.D., Donnellan, S.C. and Cook, L.G. (2009) Phylogenetic endemism: a new approach for identifying geographical concentrations of evolutionary history. *Molecular Ecology* 18: 4061–4072.

Manuscript accepted 4 June 2012