NeoBiota 34: 1–20 (2017)
doi: 10.3897/neobiota.34.11139
http://neobiota.pensoft.net

RESEARCH ARTICLE

A peer-reviewed open-access journal
NeoBiota
Advancing research on alien species and biological invasions

Taxonomic perils and pitfalls of dataset assembly in ecology: a case study of the naturalized Asteraceae in Australia

Brad R. Murray[1], Leigh J. Martin[1], Megan L. Phillips[1], Petr Pyšek[2,3]

1 *School of Life Sciences, University of Technology Sydney, PO Box 123, NSW 2007, Australia* **2** *Institute of Botany, Department of Invasion Ecology, The Czech Academy of Sciences, CZ-252 43 Průhonice, Czech Republic* **3** *Department of Ecology, Faculty of Science, Charles University, Viničná 7, CZ-128 44 Prague, Czech Republic*

Corresponding author: *Brad R. Murray* (Brad.Murray@uts.edu.au)

Academic editor: *I. Kühn* | Received 10 November 2016 | Accepted 13 January 2017 | Published 3 February 2017

Citation: Murray BR, Martin LJ, Phillips ML, Pyšek P (2017) Taxonomic perils and pitfalls of dataset assembly in ecology: a case study of the naturalized Asteraceae in Australia. NeoBiota 34: 1–20. https://doi.org/10.3897/neobiota.34.11139

Abstract

The value of plant ecological datasets with hundreds or thousands of species is principally determined by the taxonomic accuracy of their plant names. However, combining existing lists of species to assemble a harmonized dataset that is clean of taxonomic errors can be a difficult task for non-taxonomists. Here, we describe the range of taxonomic difficulties likely to be encountered during dataset assembly and present an easy-to-use taxonomic cleaning protocol aimed at assisting researchers not familiar with the finer details of taxonomic cleaning. The protocol produces a final dataset (FD) linked to a companion dataset (CD), providing clear details of the path from existing lists to the FD taken by each cleaned taxon. Taxa are checked off against ten categories in the CD that succinctly summarize all taxonomic modifications required. Two older, publicly-available lists of naturalized Asteraceae in Australia were merged into a harmonized dataset as a case study to quantify the impacts of ignoring the critical process of taxonomic cleaning in invasion ecology. Our FD of naturalized Asteraceae contained 257 species and infra-species. Without implementation of the full cleaning protocol, the dataset would have contained 328 taxa, a 28% overestimate of taxon richness by 71 taxa. Our naturalized Asteraceae CD described the exclusion of 88 names due to nomenclatural issues (e.g. synonymy), the inclusion of 26 updated currently accepted names and four taxa newly naturalized since the production of the source datasets, and the exclusion of 13 taxa that were either found not to be in Australia or were in fact doubtfully naturalized. This study also supports the notion that automated processes alone will not be enough to ensure taxonomically clean datasets, and that manual scrutiny of data is essential. In the long term, this will best be supported by increased investment in taxonomy and botany in university curricula.

Keywords
Big Data, comparative ecology, conservation, harmonized dataset, macroecology, taxonomic cleaning

Introduction

Large datasets in plant ecology, composed of hundreds or thousands of species, are increasingly being assembled by combining existing lists of species (van Kleunen et al. 2015). The value of such datasets for addressing research questions is first and foremost determined by the quality of taxonomic accuracy underpinning their plant names. The task of merging multiple source datasets into one plant ecological dataset that is clean of taxonomic errors is seldom straightforward because lists that have not been actively maintained become outdated and riddled with incorrect or obsolete taxonomy (Soberón and Peterson 2004, Hulme and Weser 2011). This can lead to a lack of taxonomic congruence among existing lists and ultimately the assembly of a taxonomically unreliable dataset (Jansen and Dengler 2010). The use of unreliable datasets is of concern as they increase the risk of reaching questionable ecological conclusions and making poorly informed conservation and management decisions (Pyšek et al. 2013).

Taxonomic cleaning during the assembly of plant ecological datasets can be an especially difficult process for non-taxonomists, not only because of the inherent complexities of taxonomy and the ongoing nature of taxonomic change (Chapman 2005), but also given the recent decline of taxonomic expertise and resources (Wheeler 2014, Halme et al. 2015) that would normally be the first point of contact for taxonomic assistance (Gotelli 2004). The sorts of problems that need to be overcome during the assembly of plant ecological datasets include, among others, locating scientifically reliable source datasets, resolving issues of synonymy so that species' names are correct and currently accepted, and, where relevant, assigning the correct ecological status (e.g. rare, naturalized, invasive) to each species' name. For instance, a status of common may have switched to a status of rare by the time of dataset assembly (e.g. Murray and Hose 2005). Despite these problems, the increasing global availability of large volumes of ecological data and the growing reliance on Big Data to address the world's environmental problems mean that efforts must continue to assemble taxonomically clean and reliable datasets.

In an effort to assist ecologists not familiar with the finer details of taxonomic cleaning and who may not have previously assembled an ecological dataset, our first aim in the present study is to describe the range of taxonomic difficulties likely to be encountered when combining existing lists of plant species into a harmonized dataset. To facilitate this, we present a systematic taxonomic cleaning protocol for merging multiple source datasets into a single plant ecological dataset. The protocol draws partly on established knowledge and procedures for taxonomic cleaning (e.g. Chapman 2005, Chavan 2007, Kooyman et al. 2012, Pyšek et al. 2013, Mathew et al. 2014) and expands upon these in a systematic way to include searches for taxa new to a study

region since the production of source datasets, confirmation of the occurrence of taxa in the region through manual inspection of distribution records, and verification of the ecological status of taxa. Our second aim is to present a case study that assembles a dataset of naturalized species and infra-species in the Asteraceae in Australia by merging two publicly available source datasets (Groves et al. 2003, Randall 2007). Importantly, we use this case study to quantify the impacts and to highlight the ramifications of ignoring the critical process of taxonomic cleaning.

Methods

Dataset design

Data cleaning identifies inaccurate and incomplete data and improves the quality of a dataset through correction of detected errors and omissions (Chapman 2005). We describe an eight step protocol for taxonomic cleaning (Fig. 1) that produces a final dataset (FD) linked to a companion dataset (CD). The FD is the cleaned dataset of species and infra-species (together referred to as taxa) ready for use in ecological studies (Suppl. material 1). The CD provides clear details of the path from source dataset to the FD taken by each taxon that has required some form of cleaning (Suppl. material 2).

The first four columns in both datasets contain genus, species, infra-species marker, and infra-species names while the fifth column contains the title(s) of the source dataset(s) in which taxon names occur. Central to the construction of the CD is checking off each taxon name against one or more of 10 categories listed in the CD. Each category, which has its own column in the CD for noting whether a taxon meets the requirements of the category, is described in full detail below (examples of each category are provided in Table 1). The CD is critical as it allows future studies to trace the origin of taxa in the FD exactly in the taxonomic form that they were collected and revisit them if need be. A comment column is included in both the FD and CD to ensure clearly articulated pathways of communication about the cleaning process between the two datasets. The 10 categories and comments columns transparently summarize all taxonomic modifications and updates, additions of new taxa, and taxon exclusions.

Taxonomic cleaning protocol

The eight step protocol presented here can be used to integrate any number of source lists, ranging from two to hundreds, into a single dataset from which taxonomic uncertainties and inaccuracies have been removed. The protocol is applicable to any taxonomic clade and in a consistent manner both to the assembly of datasets that target one or more geographic regions (from local plant communities to continental or global floras). The protocol can also be used to assemble comparative datasets that require large numbers of taxa to test ecological and evolutionary hypotheses which may not

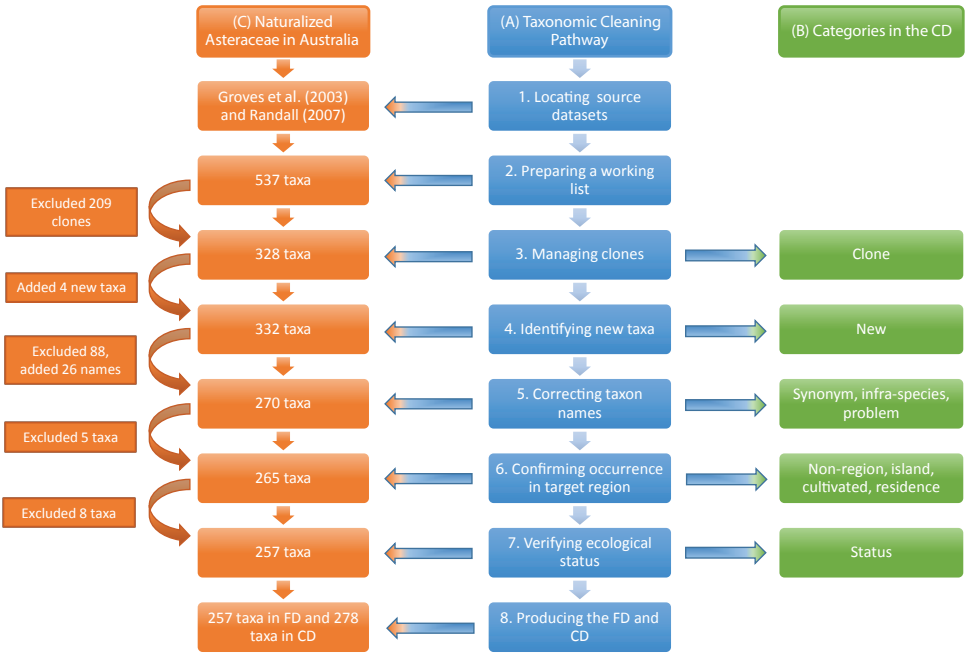

Figure 1. A Flowchart of the eight steps in the taxonomic cleaning protocol **B** Ten categories in the companion dataset that are populated with taxon names during the cleaning process, located adjacent to relevant steps in the protocol **C** A walkthrough of the case study of naturalized Asteraceae in Australia, with a numerical breakdown of the taxa in the working list at each step to the production of the final and companion datasets.

necessarily be tied to a particular geographic region. Recently-developed automated processes for various aspects cleaning (e.g. Cayuela et al. 2012, Boyle et al. 2013, Pennell et al. 2016) can be implemented while following our protocol.

We do not explore issues related to cleaning geographic coordinate records of taxa as these have been covered in detail elsewhere (e.g. Chapman 1998, Chapman 2005, Kooyman et al. 2012, Maldonado et al. 2015, Robertson et al. 2016). Our check for the occurrence of taxa in a region (step 6) is simple in that we are interested in whether a taxon is either in the study region or not. However, this step of the protocol does require careful scrutiny of available data such as inspection of comments on herbarium records and perhaps even new field surveys to ensure that specimens were collected within the study region.

Step 1: Locating source datasets

Protocol. Datasets can be obtained from a wide range of sources, including published floras, scientific papers, herbaria and museums. There is also an expanding availability of relevant data from sources such as the Global Biodiversity Information Facility

Table 1. Descriptions of the 10 categories in the companion dataset with examples of naturalized Asteraceae in Australia. FD = final dataset, CD = companion dataset, GD = Groves et al. (2003), RD = Randall (2007).

Category	Descriptions and taxa examples
1. Clone	A taxon with an identical entry of its name in more than one source dataset. *Facelis retusa* has the same name in GD and RD. *Facelis retusa* is placed in the FD and in the CD checked off against the clone category.
2. New	A taxon found to occur either within a study region or in clades that are the focus of a study, since the time when the source datasets were originally constructed. *Bidens aurea* has become naturalized in Australia since the preparation of GD and RD. *Bidens aurea* is placed in the FD and in the CD checked off against the new category.
3. Synonym	A taxon with an old, no longer accepted scientific name listed in a source dataset, and that is now recognized by a new, currently accepted scientific name. *Cnicus benedictus* in GD and RD is a synonym of the currently accepted name *Centaurea benedicta*. *Centaurea benedicta* is placed in the FD and *Cnicus benedictus* is placed in the CD checked off against the synonym category.
4. Infra-species	A taxon whose [genus + species] and [genus + species + infra-species] names in source datasets are taxonomically valid. *Centauria nigrescens* ssp. *nigrescens* in GD and *Centauria nigrescens* in RD are both valid names. We placed *Centaurea nigrescens* ssp. *nigrescens* in the FD and *Centaurea nigrescens* in the CD checked off against the infra-species category, as we chose to include [genus + species + infra-species] names in the FD over [genus + species] names.
5. Problem	A taxon in a source dataset for which there is either current uncertainty regarding the correct name that should be used or whose name cannot be officially verified. *Palafoxia rosea* cannot be taxonomically verified and is excluded from the FD and placed in the CD checked off against the problem category.
6. Non-region	A taxon in a source dataset that is found on close inspection not to occur in the study region. *Brachylaena discolor* does not occur in Australia (both known herbarium records are from overseas) and is excluded from the FD and placed in the CD checked off against the non-region category.
7. Island	A taxon in a source dataset that is found on a nearby island, not on the mainland study region. *Picris hieracioides* is not on mainland Australia but has possibly been recorded on nearby Norfolk Island. *Picris hieracioides* is excluded from the FD and placed in the CD and checked off against the island category.
8. Cultivated	A taxon in a source dataset that is found in the study region, but only in cultivated form. There are no examples of naturalized Asteraceae in the source datasets that are only in Australia in cultivation.
9. Residence	A taxon in a source dataset that is native when the focus of the study is on exotic taxa, or a taxon that is exotic when the focus of the study is on native taxa. There are no examples of naturalized Asteraceae in the source datasets excluded from the FD because they are native to Australia.
10. Status	A taxon whose ecological status in the source dataset does not match the required status. *Anacyclus radiatus* is excluded from the FD and placed in the CD checked off against status because it is doubtfully naturalized in Australia.

(GBIF, www.gbif.org), the Global Invasive Species Dataset (GISD, www.issg.org) and the TRY Plant Trait Dataset (TRY, www.try-db.org). Each source dataset used during dataset assembly is given a unique title to keep track of the origin of taxon names throughout the cleaning process.

Confidence that source datasets are scientifically reliable and have been produced carefully is an essential requirement for dataset assembly. No matter how much a source dataset is cleaned, if the underlying compilation of taxa in the source dataset is questionable, then use of the dataset will subsequently lead to the assembly of an unreliable dataset. The best-case scenario is found in regions with a long history of botanical work and record-keeping. In such cases, obtaining reliable and up-to-date source datasets is straightforward. For example, the alien flora of the Czech Republic has been carefully described (Pyšek et al. 2002, Pyšek et al. 2012a), and a solid body of research which has used and refined this work provides a supportive framework for new research (e.g. Mihulka et al. 2003, Pyšek et al. 2003, Chytrý et al. 2005, Křivánek and Pyšek 2006, Chytrý et al. 2009, Phillips et al. 2010, Pyšek et al. 2012b). The strength of such source datasets is that there is usually a wealth of information about how they were built, including references, contained in peer-reviewed papers. There is an important point of distinction, in terms of confidence in a source dataset, between regions with such dataset availability and regions which have lists that are perhaps only available online and that are not attached to an institution, lacking any information about their construction or ongoing taxonomic maintenance.

Naturalized Asteraceae. Australia was permanently settled by Europeans in 1788, and even within the first 14 years of settlement, 29 exotic plant taxa that were introduced either accidentally or deliberately had started to naturalize (Groves 2002). Since then, over 2,500 plant taxa have become naturalized across the continent (Groves et al. 2005). Our case study assembled a dataset of species and infra-species in the Asteraceae that have become naturalized in the natural environment in Australia since permanent European settlement. We selected the Asteraceae for our study because a large number of taxa in the group have become naturalized in Australia and many have become invasive and problematic across the landscape (Radford and Cousens 2000, Parsons and Cuthbertson 2001, Hamilton 2005, Dodd et al. 2015).

Two publicly available datasets of naturalized plants in Australia were used, Groves et al. (2003) and Randall (2007) (referred to as GD and RD respectively, here and in the FD and CD), to merge naturalized Asteraceae from these source datasets into a single dataset. These are older sources of information, but we selected these to specifically demonstrate the problems that are to be expected and the errors that can arise when combining existing lists of plant species into a harmonized dataset. The book by Groves et al. (2003), commissioned by the Department of Environment and Heritage in 2000 and the Bureau of Rural Sciences in 2001, was compiled by 14 plant specialists from all the States and Territories of Australia with high-level expertise in taxonomy and botany. Naturalized exotic plants were defined as species or infra-species that have been introduced, become established and that reproduce naturally in the wild, without human intervention, consistent with descriptions in Richardson et al. (2000). The book by Randall (2007), which not only provides a comprehensive list of all exotic plant species introduced to Australia, but also identifies those that have become naturalized somewhere in Australia, was a publication of the CRC for Australian Weed Management and represents

a development of *A Global Compendium of Weeds* (Randall 2002), a major dataset of all weedy flora of the world. Both GD and RD source datasets represent the results of years of meticulous botanical work.

Step 2: Preparing a working list

Protocol. All taxa from the source datasets are placed in an initial working list that is a precursor to the FD. Some taxa will be present more than once in the working list under exactly the same name when source datasets are merged. These repeat entries are kept in the working list at this stage with their different source titles.

Naturalized Asteraceae. There were a total of 537 taxa of naturalized Asteraceae in Australia in the working list resulting from the merging of GD and RD.

Step 3: Managing clones

Protocol. Clones are repeat, completely identical entries of a taxon name from more than one source dataset. Once all clones have been identified, their occurrence in the working list is reduced to a single-name entry for each cloned taxon. Each cloned taxon is placed in the CD and checked off against the *clone* category (Fig. 1), retaining all source titles for each taxon. This step is important for record keeping as it provides an initial evaluation of consistency among source datasets (Chapman 2005).

Naturalized Asteraceae. There were 209 clones across the 328 unique taxa derived from both source datasets. This translates to 76.6% of the 273 taxa in GD and 79.2% of the 264 taxa in RD that were initially common to both datasets, leaving 64 taxon names found only in GD and 55 taxon names found only in RD.

Step 4: Identifying new taxa

Protocol. This step ensures that the FD contains all taxa currently known to occur either within a target region (*sensu* Pyšek et al. 2004) or in clades that are the focus of study. Taxa new to a study region (e.g. newly discovered natives, recent introductions of exotics or non-endemic natives) and recently described taxa – since the time when the source datasets were originally constructed – need to be identified. Each new taxon is attached to a unique source title to keep track of the origin of the taxon name and placed in the FD. The names of these new taxa and their source titles are also placed in the CD and checked off against the *new* category in the companion dataset (Fig. 1).

Naturalized Asteraceae. To gather information about newly naturalized taxa in the Asteraceae in Australia since the compilation of the two source datasets, we conducted a literature search of publications from the Australian state herbaria and botanical gardens including *Austrobaileya*, *Cunninghamia*, *Telopea*, *Muelleria*, *Journal of the*

Adelaide Botanical Gardens and *Nuytisia*. These journals periodically publish lists and records of plants newly recorded or identified as naturalized within Australia. We located three sources documenting new naturalizations in Australia, Hosking et al. (2007), Hosking et al. (2011) and Parsons (2012).

Step 5: Correcting taxon names

Protocol. This step requires careful scrutiny of taxon names in the working list to ensure that taxa are represented with their currently accepted and correct names. How difficult a task this is will ultimately depend on the availability of up-to-date taxonomic information via sources such as publications, online datasets and tools, detailed herbarium records, and taxonomists and their expertise. The guiding principle when updating taxa with their currently accepted names is to adopt a taxonomic system that provides an accepted, current authority in the jurisdiction of interest. Where no single authoritative source is available and competing taxonomies exist, researchers will need to make a choice and be explicitly clear about their taxonomic choices. This step in the process also corrects misspellings and lexical variants (i.e. different ways of writing the same name), and misapplications (where an incorrect name has mistakenly been given to a taxon), with any corrected taxon names checked in case they are clones of taxa already in the working list (step 3), to ensure that clones are limited to single-name entries. In some cases, it might be helpful to make use of automated recognition and correction tools for plant taxonomy, such as TaxonStand (Cayuela et al. 2012), the TNRS (Boyle et al. 2013) and taxonlookup (Pennell et al. 2016). If such tools are implemented, the version used must be carefully documented as these tools are also reliant on their underpinning sources of taxonomic information being maintained and kept up-to-date.

One of the most difficult taxonomic cleaning issues is dealing with the complex issue of synonymy. In taxonomy, a synonym is an old, no longer accepted scientific name that applies to a taxon that is now recognized by a new, currently accepted scientific name. Homotypic synonyms are problematic when assembling a dataset from multiple source datasets, as the inclusion of two or more names that refer to the same taxon (i.e. two or more names given to the same type specimen) leads to pseudo-replication in the dataset and thus problems with subsequent analyses and conclusions. Heterotypic synonyms consist of different names for different type specimens, which were all at one point considered distinct taxa, but which have now been lumped into the one taxon. Heterotypic synonymy needs to be resolved not only because the single, up-to-date taxon could have a broader geographic range than its constituent synonyms (an important distinction for macroecological studies of range size variation), but also because variation in life-history and ecological traits will probably be greater for the wider ranging up-to-date taxon (an important detail for comparative studies of life-history variation). It is also important to identify and correct any homonyms in the working list, which refer to a name for a taxon that is identical in spelling to another such name, that belongs to a different taxon, as well as any misapplications (i.e. where a taxon has been incorrectly

identified). Once all issues of synonymy have been identified, the single currently accepted name of a taxon is retained in the working list and non-current or misapplied names are excluded from the working list and placed in the CD and checked off against the *synonym* category (Fig. 1). Source titles are retained for each taxon with specific notes kept on the link that each synonymous taxon has to its currently accepted name in the working list, remembering that the working list becomes the FD at the end of the process.

It may become apparent that source datasets have chosen a different approach in relation to infra-species epithets. For example, a taxon might be represented with a [genus + species] name in one source dataset, but represented with [genus + species + infra-species] name in another (and in some cases both might be included). Sometimes, in checking the up-to-date names of such taxa, both names are considered to be current. An approach for dealing with infra-species in dataset assembly is to decide at the outset whether to include infra-species epithets across the whole working list, or if not, to pool infra-species into a [genus + species] name where appropriate. The latter approach can perhaps be used to deal with 'difficult' taxonomic groups where there are unresolved taxonomic issues. This pooling approach, however, can have disadvantages. Pooling infra-species into one larger taxon ignores potentially important differences among infra-species in their geographic distribution, life history, physiology and ecology. We suggest that where possible, infra-species are included in the working list. In such cases, the [genus + species] name that is not used is placed in the CD and checked off against the *infra-species* category and only the [genus + species + infra-species] name is retained in the working list with the relevant source title (Fig. 1). Where infra-species are not recognized, then [genus + species + infra-species] names are placed in the CD and checked off against the *infra-species* category, and the [genus + species] names are placed on the working list. The *infra-species* category provides the opportunity to contrast patterns emerging from the FD in analyses with and without infra-species if desired.

Some taxa may need to be removed from the working list, placed in the CD and checked off against a *problem* category (Fig. 1). These are either taxa for which there is current uncertainty regarding the correct name that should be used for the taxa in question or taxa whose names cannot be officially verified.

Naturalized Asteraceae. We used the Australian Plant Name Index (APNI, http://www.anbg.gov.au/apni/) and the Australian Plant Census (APC, http://www.chah.gov.au/apc/about-APC.html) to determine currently accepted names for all taxa in our working list. The system of nomenclature adopted for APC is endorsed by the Council of Heads of Australasian Herbaria (CHAH), while APNI is maintained by the Australian National Botanic Gardens in collaboration with the Centre for Australian National Biodiversity Research and the Australian Biological Resources Study.

Step 6: Confirming occurrence in target region

Protocol. If a research goal is to include all taxa within a specific geographic region, then taxa in the working list are verified for their occurrence within that target region. This

step may also include the requirement that taxa are identified as native or exotic to the region. Official plant censuses and herbarium records curated and maintained by national herbaria or botanic gardens, among other sources of reliable information, can be inspected closely to provide such verification. Ground truthing in the field may be required if there is real uncertainty about the occurrence of taxa in the region.

Taxa are removed from the working list, placed in the CD and checked off against the *non-region* category if there are no verified records of them in the target region (Fig. 1). This can happen, for instance, when specimens collected well outside the region are kept in herbaria and then those records are incorrectly entered into distributional datasets for the region which are then used as source datasets in dataset assembly.

Taxa are removed from the working list and placed in the CD and checked off against the *island* category if they are not found in the mainland target region, but are found on nearby external islands (Fig. 1). It is desirable to keep such taxa separate from those in the *non-region* category, as it might be argued for some studies, for instance, that it is important to perform analyses with and without nearby island species. For example, taxa in the *island* category might be excluded if seeking to identify those taxa that have naturalized within a mainland study region. These taxa might be included, however, if the goal is to identify taxa that have penetrated broader national biosecurity and quarantine systems where the island is considered part of the nation.

Taxa that only occur in the target region because they have been cultivated there, and which do not occur naturally in the wild, are removed from the working list, placed in the CD and checked off against the *cultivated* category (Fig. 1).

If a study is focused specifically on taxa native to the region, then exotic taxa are excluded from the working list and placed in the CD and checked off against the *residence* category (Fig. 1). If a study is about exotic taxa, then native taxa are excluded and placed in the *residence* category. Alternatively, this category need not be included in the CD, but rather a separate column distinguishing native from exotic taxa can be included in the FD if comparisons between natives and exotics are desired in the study.

Naturalized Asteraceae. We used APNI and APC to determine non-region, island and cultivated taxa or native residency of taxa in Australia that would exclude them from the FD. If a name wasn't found in APNI, which provides a comprehensive record of every scientific plant name in taxonomic literature concerning Australia, this meant that the name had not been used in the scientific literature as referring to a taxon occurring within Australia. If a name was excluded from APC, this meant that the name was not considered by CHAH to be in Australia. We then scrutinized herbarium records in Australia's Virtual Herbarium (AVH, www.avh.chah.org.au) to seek further evidence of occurrence of species in Australia. The AVH resource is maintained by CHAH and provides on-line access to Commonwealth, State and Territory herbarium records. These records provide important information on the date and location of collection and if specimens were obtained overseas, from islands or cultivated plants, or from plants occurring in natural habitats.

Step 7: Verifying ecological status

Protocol. Dataset assembly often requires a final clean so that only taxon names with a particular ecological status or statuses, related to their distribution and abundance within the target region, are included. These might include, for example, datasets comprised of taxa classified as either naturalized, invasive, declining, or threatened. We have included this step in the taxonomic cleaning process because this a particular area where taxonomy and ecology overlap considerably and they should not be considered separately (Graham et al. 2004, Wheeler 2004, Halme et al. 2015).

The definition of ecological status in the source datasets must be clear and should preferably comply for the most part with published and widely adopted descriptions. In the field of invasion ecology, for instance, there are widely adopted schemes for consistent terminology (e.g. Richardson et al. 2000, Blackburn et al. 2011). Only if these definitions are similar should source datasets be put through the process simultaneously. If two or more source datasets differ substantially in their classification schemes, and these differences cannot be resolved, it is advisable to treat the datasets independently and put them through the process separately to produce two separate FDs. For example, species invasiveness might be determined as level of impact in one source and as rate of spread and geographic range size in another, and it is important that these two definitions of invasiveness are not considered the same. If a taxon name does not have the appropriate ecological status, it is excluded from the working list, placed in the CD and checked off against a *status* category in the companion dataset (Fig. 1). If more than one ecological status is assessed, then separate columns are included in the CD representing each status. As an alternative, this category need not be included in the CD, but rather a separate column distinguishing the status of each taxon can be included in the FD if comparisons between or among statuses are the focus of the study (e.g. the study seeks to compare rare and common taxa in the dataset).

Naturalized Asteraceae. The naturalized status of each taxon in Australia was reviewed by carefully examining source datasets in conjunction with APC, APNI and AVH. In particular, the APC states clearly if taxa are doubtfully naturalized, and we excluded those taxa from the FD.

Step 8: Producing the final and companion datasets

Protocol. The working list at this stage of the process becomes the FD of taxa linked to the CD. The FD has now been cleaned and is the primary, up-to-date inventory of species that can be used with confidence and transparency in dataset studies. In both the FD and CD, it is important to ensure that the language and terminology used in the comments columns are consistent, to ensure ease of use when cross-walking the datasets.

Naturalized Asteraceae. The FD is presented in Suppl. material 1 and the CD is presented in Suppl. material 2.

Results

Summary patterns in the FD and CD

The FD of naturalized Asteraceae in Australia contained 257 taxa. Four of these taxa (1.6%) were new, recorded as naturalized in Australia since the publication of the source datasets. There were 278 taxa in the CD. A total of 173 taxa (67.3% of the FD) were clones across the FD and CD with the same currently accepted name in both source datasets. There were 54 taxa (21.0%) in the FD that were either found only in GD (23 taxa, 8.9%) or only in RD (31 taxa, 12.1%) under their currently accepted name. Thus, a total of 227 taxa (88.3%) in the FD were unchanged from the source datasets. A total of 26 updated names (10.1%) not found in GD or RD were included in the FD.

A walk-through of the taxonomic cleaning process

The source datasets GD and RD were selected (step 1, Fig. 1) with the working list containing 537 taxon names after their merger (step 2, Fig. 1). Management of clones led to the removal of 209 duplicate taxon names (e.g. *Ambrosia artemisiifolia*) leaving 328 distinct taxon names in the working list (step 3, Fig. 1). We added 4 new taxa (e.g. *Pentzia globosa*) resulting in 332 taxa in the working list (step 4, Fig. 1). A total of 88 taxon names were excluded as they were either problematic (e.g. *Chrysocoma comaaurea*); they were [genus + species] names that were replaced with valid [genus + species + infra-species] names (e.g. *Chrysanthemoides monilifera* in RD was excluded and *Chrysanthemoides monilifera* ssp. *monilifera* and *Chrysanthemoides monilifera* ssp. *rotundata* in GD were included); and/or they were old synonyms that required updating with currently accepted names (step 5, Fig. 1, e.g. four taxon names in GD, *Xanthium cavanillesii*, *Xanthium italicum*, *Xanthium occidentale*, *Xanthium orientale* were excluded and the currently accepted name *Xanthium strumarium* in RD was included). In some cases during this step, the currently accepted names or [genus + species + infra-species] names appeared in one or both of GD and RD. For example, *Cineraria lyrata* in RD was updated to *Cineraria lyratiformis* which appeared in both GD and RD (Suppl. material 2). In other cases, the old synonyms were replaced with a total of 26 updated names that were not in the source datasets (e.g. *Oligocarpus calendulaceus* was included and its synonym *Osteospermum calendulaceum* in GD and RD was excluded).

At the end of step 5, there were 270 taxa in the working list. Five taxa were found not to be present in Australia (e.g. *Gazania serrata*) and their removal left 265 taxa in the working list (step 6, Fig. 1). Eight taxa were identified as doubtfully naturalized (step 7, Fig. 1, e.g. *Cichorium endivia*) and their removal left 257 taxa in the working list which became the FD (step 8, Fig. 1). Among these eight taxa, *Brachylaena discolor* was excluded both because its two herbarium records were collected overseas and because it is considered doubtfully naturalized, while *Picris hieracioides* was excluded

because it does occur on mainland Australia, its presence on an external island (Norfolk Island) is questionable due to misidentification and because it is also considered doubtfully naturalized.

Discussion

Several outcomes of our dataset assembly of naturalized Asteraceae in Australia demonstrate how critical it is to implement taxonomic cleaning. Although our study only dealt with a few hundred taxa, the outcomes of the study have direct implications for even bigger data studies involving thousands of taxa. First, the cleaned dataset contained 257 taxa. Had the cleaning protocol not been implemented, and a dataset constructed simply by merging the two source datasets (with just the straightforward removal of duplicate names), the assembled dataset would have contained 328 taxa. This equates to a considerable and unacceptable overestimate of taxon richness of naturalized Asteraceae in Australia by 71 taxa (27.6%). Such a high level of taxonomic inaccuracy is especially unsuitable for comparative plant studies that require accurate representations of phylogenetic relationships (Gotelli 2004). Second, any taxonomic cleaning process must account not just for nomenclatural issues (step 5), it must also include careful scrutiny of the occurrence (step 6) and ecological status (step 7) of each taxon. Had we not manually inspected the actual distributional records of each taxon, the assembled dataset would have contained 270 taxa, an overestimate of taxon richness by 13 taxa. Third, where there is any reasonable gap in time between dataset assembly and the construction of the source datasets, the literature must be scoured for evidence of new taxa that need to be added to the dataset (step 4). While in our case, this involved searching for and finding four recently naturalized taxa in Australia, in other cases, this might include newly described taxa within study clades.

Implementation of our cleaning protocol has also demonstrated that it is unlikely that a reliance on automated processes for cleaning will be all that is required to completely clean and prepare datasets. Indeed, previous work has described data cleaning and taxonomic scrutiny of Big Data as 'intelligent processes' (Chavan 2007), requiring the involvement of skilled individuals with taxonomic expertise to be fully effective. Kooyman et al. (2012) pointed out that even after automated taxonomic cleaning, each taxon in an assembled dataset must be individually inspected to ensure all taxonomic inaccuracies have been dealt with. While there have been recent efforts to automate the most time-consuming process (step 5) of cleaning (Cayuela et al. 2012, Boyle et al. 2013, Pennell et al. 2016), coordination of effort across a global or even more regional scales to provide combined automation of step 5 with steps 4 (new taxa), 6 (confirming occurrence) and 7 (verifying status) in particular will be much harder to achieve in the foreseeable future, and these steps will for some time require human vetting and expertise. This is especially so for step 6, as distributional records need to be inspected. With these records, there is often much detail and little consistency in how comments and notes are provided, making efforts to establish an automated process

rather difficult. In addition, it is critical to understand that automated approaches are still reliant on the sources used for nomenclatural cleaning being regularly maintained and updated to reflect current taxonomic knowledge. Unfortunately, the approach to ensuring currently valid names are used has generally been haphazard in broader curatorial practice (Costello and Wieczorek 2013), but there is much scope for it to become more systematic as these datasets grow and receive more attention based on their value in the age of Big Data (Zermoglio et al. 2016).

The number of clones in the FD, taxa found in both GD and RD under their currently accepted names, was moderately high (67%). This is probably unsurprising given the meticulous nature with which the source datasets were constructed. Nevertheless, the differences between the two source datasets point to issues that need to be considered when merging datasets. For instance, the 21% of taxa in the FD that were either found only in GD or only in RD under their currently accepted name demonstrate that using more than one source dataset when possible is likely to lead to a higher number of relevant taxa in the FD and that disparate source datasets are likely to differ in their taxonomic content (e.g. Hulme and Weser 2011). At this stage, it is unclear why our two source datasets each contained taxa that the other did not. It is also interesting to note that in nearly ten years since the publication of the latest dataset (RD), 26 updated taxon names needed to be inserted into the FD with the removal of 88 other names for issues related to synonymy, infra-species epithets and problematic circumstances. These numbers are not insignificant and indicate that even in a short period of time, taxonomy is incredibly dynamic.

A key strength of the protocol presented in this paper is that it presents a simple step-by-step approach for taxonomic cleaning that can easily be adopted by non-specialists who are assembling a plant ecological dataset, perhaps for the first time. In addition, it systematically coordinates steps in a way that especially targets the construction of plant ecological datasets, particularly because it includes ecological aspects (i.e. occurrence, status) and the need to search the most up-to-date sources for taxa new to study regions (if a target area approach is used) or taxonomic clades (if a broader comparative study is involved). Further detailed descriptions of taxonomic cleaning can be obtained by consulting sources such as Chapman (2005) and Mathew et al. (2014). The production of both a final dataset and a companion dataset via our protocol make it very clear that we believe it important to be transparent about not only which taxa are included in a study, but also about those taxa not included and the reasons for their exclusion. The recent retraction of a published paper from the journal Biology Letters (Hanna and Cardillo 2014) on the grounds that the ecological dataset contained substantial errors lends weight to the argument of transparency in dataset presentation. Analysis of a revised dataset has produced considerably different outcomes compared with the original study, and will lead to a new publication in a different journal (Retraction Watch at http://retractionwatch.com/2016/12/27/error-laden-database-kills-paper-extinction-patterns/).

This is the first botanical study that details the types and amounts of taxonomically-related errors that arise when source datasets are merged to assemble an ecological

dataset. A small number of studies, however, have begun to empirically address the issue of taxonomic reliability in the sorts of large datasets available for use in large dataset studies in animal ecology. Zermoglio et al. (2016), for example, analysed 1000 scientific names taken at random from VertNet, an aggregator of vertebrate biodiversity data from natural history collections. They found that less than 47% of names were currently valid. Our cleaning protocol removed 27% (88 out of 328 taxon names at step 2) based on similar nomenclatural issues. Although this percentage is not as high as that reported in Zermoglio et al. (2016), it still represents the highest number of taxa requiring attention (excluding the removal of duplicate names). The high prevalence of synonymy is not surprising as this type of issue is the most difficult and time consuming to solve (Zermoglio et al. 2016). In this context, consultation with specialist taxonomists is highly desired (Gotelli 2004). However, such expertise has become less available in recent times (Wheeler 2014). In the long term, this problem will best be solved by increasing the taxonomic expertise of ecologists building and using datasets containing large numbers of species. Thus, our study provides further evidence to support calls for continued investment in plant systematics and the representation of taxonomy and botany in university curricula (Wheeler et al. 2012, Pyšek et al. 2013, Bebber et al. 2014, Wheeler 2014, Deng 2015).

Conclusions

Big data can be used effectively in a targeted way in ecological studies to address major scientific and societal problems (Hampton et al. 2013). However, the value of any analysis of large ecological datasets depends on the quality of the underlying taxonomic data (Valdecasas and Camacho 2003). The challenge is that taxonomic cleaning during dataset assembly is an incredibly difficult task. This difficulty, compounded by the global decline of taxonomic expertise, leads to situations where ecological datasets are often used without much attention being given to the quality of the underlying taxonomic data (Maldonado et al. 2015). This is concerning because a lack of appropriate taxonomic consideration can have serious impacts on the robustness of outcomes from large dataset studies (Jansen and Dengler 2010, Duarte et al. 2014, Zermoglio et al. 2016). The protocol we have presented here is helpful because it brings together an integrated management plan that combines usually disparate elements of dataset assembly which are not always considered together in a systematic way for plant ecological datasets. Our study has clearly shown that ignoring the critical process of taxonomic cleaning can lead to serious dataset problems that will likely lead to incorrect ecological conclusions.

Acknowledgements

BM, LM and MP thank the members of the Murray Ecology Lab at the University of Technology Sydney for helpful discussions, and Joyce Byers for comments on a

draft of the manuscript. PP was supported by project no. 14-36079G Centre of Excellence PLADIAS (Czech Science Foundation), long-term research development project RVO 67985939 and Praemium Academiae award (The Czech Academy of Sciences).

References

Bebber DP, Wood JRI, Barker C, Scotland RW (2014) Author inflation masks global capacity for species discovery in flowering plants. New Phytologist 201: 700–706. https://doi.org/10.1111/nph.12522

Blackburn TM, Pyšek P, Bacher S, Carlton JT, Duncan RP, Jarošík V, Wilson JRU, Richardson DM (2011) A proposed unified framework for biological invasions. Trends in Ecology and Evolution 26: 333–339. https://doi.org/10.1016/j.tree.2011.03.023

Boyle B, Hopkins N, Lu Z, Garay JAR, Mozzherin D, Rees T, Matasci N, Narro ML, Piel WH, Mckay SJ, Lowry S, Freeland C, Peet RK, Enquist BJ (2013) The taxonomic name resolution service: an online tool for automated standardization of plant names. BMC Bioinformatics 14: 16. https://doi.org/10.1186/1471-2105-14-16

Cayuela L, Granzow-de la Cerda Í, Albuquerque FS, Golicher DJ (2012) TAXONSTAND: An R package for species names standardization in vegetation datasets. Methods in Ecology and Evolution 3: 1078–1083. https://doi.org/10.1111/j.2041-210X.2012.00232.x

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

Chapman AD (2005) Principles and Methods of Data Cleaning: Primary Species and Species-Occurrence Data. Version 1. Report for the Global Biodiversity Information Facility, Copenhagen.

Chavan V (2007) Design and Implementation of a Biodiversity Information Management System: Electronic Catalogue of Known Indian Fauna – A Case Study. PhD Thesis, National Chemical Laboratory, Pune.

Chytrý M, Pyšek P, Tichý L, Knollová I, Danihelka J (2005) Invasions by alien plants in the Czech Republic: a quantitative assessment across habitats. Preslia 77: 339–354.

Chytrý M, Wild J, Pyšek P, Tichý L, Danihelka J, Knollová I (2009) Maps of the level of invasion of the Czech Republic by alien plants. Preslia 81: 187–207.

Costello MJ, Wieczorek J (2013) Best practice for biodiversity data management and publication. Biological Conservation 173: 68–73. https://doi.org/10.1016/j.biocon.2013.10.018

Deng B (2015) Plant collections left in the cold by cuts. Nature 523: 16. https://doi.org/10.1038/523016a

Dodd AJ, Burgman MA, McCarthy MA, Ainsworth N (2015) The changing patterns of plant naturalization in Australia. Diversity and Distributions 21: 1038–1050. https://doi.org/10.1111/ddi.12351

Duarte M, Guerrero PC, Carvallo G, Bustamante RO (2014) Conservation network design for endemic cacti under taxonomic uncertainty. Biological Conservation 176: 236–242. https://doi.org/10.1016/j.biocon.2014.05.028

Gotelli NJ (2004) A taxonomic wish-list for community ecology. Philosophical Transactions of the Royal Society of London B 359: 585–597. https://doi.org/10.1098/rstb.2003.1443

Graham CH, Ferrier S, Huettman F, Moritz C, Peterson AT (2004) New developments in museum-based informatics and applications in biodiversity analysis. Trends in Ecology and Evolution 19: 497–503. https://doi.org/10.1016/j.tree.2004.07.006

Groves RH (2002) Robert Brown and the naturalised flora of Australia. Cunninghamia **7**: 623–629.

Groves RH, Hosking JR, Batianoff GN, Cooke DA, Cowie ID, Johnson RW, Keighery GJ, Lepschi BJ, Mitchell AA, Moerkerk M, Randall RP, Rozefelds AC, Walsh NG, Waterhouse BM (2003) Weed Categories for Natural and Agricultural Ecosystem Management. Bureau of Rural Sciences, Canberra, Australia. https://www.researchgate.net/publication/235980851_Weed_categories_for_natural_and_agricultural_ecosystem_management

Groves RH, Boden R, Lonsdale WM (2005) Jumping the Garden Fence: Invasive Garden Plants in Australia and Their Environmental and Agricultural Impacts. WWF-Australia, Ultimo, NSW.

Halme P, Kuusela S, Juslén A (2015) Why taxonomists and ecologists are not, but should be, carpooling? Biodiversity Conservation 24: 1831–1836. https://doi.org/10.1007/s10531-015-0899-3

Hamilton MA, Murray BR, Cadotte MW, Hose GC, Baker AC, Harris CJ, Licari D (2005) Life-history correlates of plant invasiveness at regional and continental scales. Ecology Letters 8: 1066–1074. https://doi.org/10.1111/j.1461-0248.2005.00809.x

Hampton SE, Strasser CA, Tewksbury JJ, Gram WK, Budden AE, Batcheller AL, Duke CS, Porter JH (2013) Big data and the future of ecology. Frontiers in Ecology and the Environment 11: 156–162. https://doi.org/10.1890/120103

Hanna E, Cardillo M (2014) Predation selectively culls medium-sized species from island mammal faunas. Biology Letters 10: 20131066. https://doi.org/10.1098/rsbl.2013.1066

Hosking JR, Conn BJ, Lepschi BJ, Barker CH (2007) Plant species first recognised as naturalised or naturalising for New South Wales in 2002 and 2003, with additional comments on species recognised as naturalised in 2000-2001. Cunninghamia 10: 139–166.

Hosking JR, Conn BJ, Lepschi BJ, Barker CH (2011) Plant species first recognised as naturalised or naturalising for New South Wales in 2004 and 2005. Cunninghamia 12: 85–114. https://doi.org/10.1111/j.1472-4642.2011.00800.x

Hulme PE, Weser C (2011) Mixed messages from multiple information sources on invasive species: a case of too much of a good thing? Diversity and Distributions 17: 1152–1160. https://doi.org/10.1111/j.1654-1103.2010.01209.x

Jansen F, Dengler J (2010) Plant names in vegetation datasets – a neglected source of bias. Journal of Vegetation Science 21: 1179–1186.

Kooyman R, Rossetto M, Laffan S (2012) Using Australian Virtual Herbarium data to find all the woody rain forest plants in Australia. Cunninghamia 12: 177–180. https://doi.org/10.7751/cunninghamia.2012.12.014

Křivánek M, Pyšek P (2006) Predicting invasions by woody species in a temperate zone: a test of three risk assessment schemes in the Czech Republic (Central Europe). Diversity & Distributions 12: 319–327. https://doi.org/10.1111/j.1366-9516.2006.00249.x

Maldonado C, Molina CI, Zizka A, Persson C, Taylor CM, Albán J, Chilquillo E, Rønsted N, Antonelli A (2015) Estimating species diversity and distribution in the era of Big Data: to what extent can we trust public datasets? Global Ecology and Biogeography 24: 973–984.

Mathew C, Güntsch A, Obst M, Vicario S, Haines R, Williams A, de Jong Y, Goble C (2014) A semiautomated workflow for biodiversity data retrieval, cleaning, and quality control. Biodiversity Data Journal 2: e4221. https://doi.org/10.3897/BDJ.2.e4221

Mihulka S, Pyšek P, Pyšek A (2003) *Oenothera coronifera*, a new alien species for the Czech flora, and *Oenothera stricta*, recorded again after two centuries. Preslia 75: 263–270.

Murray BR, Hose GC (2005) Life-history and ecological correlates of decline and extinction in the endemic Australian frog fauna. Austral Ecology 30: 564–571. https://doi.org/10.1111/j.1442-9993.2005.01471.x

Parsons RF (2012) The deliberate introduction to Australia of the shrub genus *Pentzia* (Asteraceae) and its subsequent persistence and spread. Cunninghamia 12: 239–246. https://doi.org/10.7751/cunninghamia.2012.12.019

Parsons WT, Cuthbertson EG (2001) Noxious Weeds of Australia. Second edition. CSIRO Publishing, Collingwood, Victoria, Australia.

Pennell MW, FitzJohn RG, Cornwell WK (2016) A simple approach for maximizing the overlap of phylogenetic and comparative data. Methods in Ecology and Evolution 7: 751–758. https://doi.org/10.1111/2041-210X.12517

Phillips ML, Murray BR, Pyšek P, Pergl J, Jarošík V, Chytrý M, Kühn I (2010) Plant species of the Central European flora as aliens in Australia. Preslia 82: 465–482.

Pyšek P, Sádlo J, Mandák B (2002) Catalogue of alien plants of the Czech Republic. Preslia 74: 97–186.

Pyšek P, Jarošík V, Kuèera T (2003) Inclusion of native and alien species in temperate nature reserves: an historical study from Central Europe. Conservation Biology 17: 1414–1424. https://doi.org/10.1046/j.1523-1739.2003.02248.x

Pyšek P, Richardson DM, Rejmánek M, Webster G, Williamson M, Kirschner J (2004) Alien plants in checklists and floras: towards better communication between taxonomists and ecologists. Taxon 53: 131–143. https://doi.org/10.2307/4135498

Pyšek P, Danihelka J, Sádlo J, Chrtek J Jr, Chytrý M, Jarošík V, Kaplan Z, Krahulec F, Moravcová L, Pergl J, Štajerová K, Tichý L (2012a) Catalogue of alien plants of the Czech Republic (2nd edition): checklist update, taxonomic diversity and invasion patterns. Preslia 84: 155–255.

Pyšek P, Chytrý M, Pergl J, Sádlo J, Wild J (2012b) Plant invasions in the Czech Republic: current state, introduction dynamics, invasive species and invaded habitats. Preslia 84: 576–630.

Pyšek P, Hulme PE, Meyerson LA, Smith GF, Boatwright JS, Crouch NR, Figueiredo E, Foxcroft LC, Jarošík V, Richardson DM, Suda J, Wilson JRU (2013) Hitting the right target: taxonomic challenges for, and of, plant invasions. AoB PLANTS 5: plt042. https://doi.org/10.1093/aobpla/plt042

Radford IJ, Cousens RD (2000) Invasiveness and comparative life-history traits of exotic and indigenous *Senecio* species in Australia. Oecologia 125: 531–542. https://doi.org/10.1007/s004420000474

Randall RP (2002) A Global Compendium of Weeds. RG and FJ Richardson, Melbourne.

Randall RP (2007) The Introduced Flora of Australia and Its Weed Status. CRC for Australian Weed Management, Department of Agriculture and Food, Western Australia, University of South Australia, Adelaide. http://www.iffa.org.au/introduced-flora-australia-and-its-weed-status

Richardson DM, Pyšek P, Rejmánek M, Barbour MG, Panetta FD, West CJ (2000) Naturalization and invasion of alien plants: concepts and definitions. Diversity & Distributions 6: 93–107. https://doi.org/10.1046/j.1472-4642.2000.00083.x

Robertson MP, Visser V, Hui C (2016) Biogeo: an R package for assessing and improving data quality of occurrence record datasets. Ecography 39: 394–401. https://doi.org/10.1111/ecog.02118

Soberón J, Peterson AT (2004) Biodiversity informatics: managing and applying primary biodiversity data. Philosophical Transactions of the Royal Society of London B 359: 689–698. https://doi.org/10.1098/rstb.2003.1439

Valdecasas AG, Camacho AI (2003) Conservation to the rescue of taxonomy. Biodiversity and Conservation 12: 1113–1117. https://doi.org/10.1023/A:1023082606162

van Kleunen M, Dawson W, Essl F, Pergl J, Winter M, Weber E, Kreft H, Weigelt P, Kartesz J, Nishino M, Antonova LA, Barcelona JF, Cabezas FJ, Cárdenas D, Cárdenas-Toro J, Castaño N, Chacón E, Chatelain C, Ebel AL, Figueiredo E, Fuentes N, Groom QJ, Henderson L, Inderjit, Kupriyanov A, Masciadri S, Meerman J, Morozova O, Moser D, Nickrent DL, Patzelt A, Pelser PB, Baptiste MP, Poopath M, Schulze M, Seebens H, Shu W, Thomas J, Velayos M, Wieringa JJ, Pyšek P (2015) Global exchange and accumulation of non-native plants. Nature 525: 100–103. https://doi.org/10.1038/nature14910

Wheeler QD (2004) Taxonomic triage and the poverty of phylogeny. Philosophical Transactions of the Royal Society of London B 359: 571–583. https://doi.org/10.1098/rstb.2003.1452

Wheeler QD, Knapp S, Stevenson DW, Stevenson J, Blum SD, Boom BM, Borisy GG, Buizer JL, de Carvalho MR, Cibrian A, Donoghue MJ, Doyle V, Gerson EM, Graham CH, Graves P, Graves SJ, Guralnick RP, Hamilton AL, Hanken J, Law W, Lipscomb DL, Lovejoy TE, Miller H, Miller JS, Naeem S, Novacek MJ, Page LM, Platnick NI, Porter-Morgan H, Raven PH, Solis MA, Valdecasas AG, Van Der Leeuw S, Vasco A, Vermeulen N, Vogel J, Walls RL, Wilson EO, Woolley JB (2012) Mapping the biosphere: exploring species to understand the origin, organization and sustainability of biodiversity. Systematics and Biodiversity 10: 1–20. https://doi.org/10.1080/14772000.2012.665095

Wheeler QD (2014) Are reports of the death of taxonomy an exaggeration? New Phytologist 201: 370–371. https://doi.org/10.1111/nph.12612

Zermoglio PF, Guralnick RP, Wieczorek JR (2016) A standardized reference data set for vertebrate taxon name resolution. PLoS ONE 11: e0146894. https://doi.org/10.1371/journal.pone.0146894

Supplementary material 1

Final dataset

Authors: Brad R. Murray, Leigh J. Martin, Megan L. Phillips, Petr Pyšek

Data type: species data

Explanation note: The final dataset (FD) of naturalized species and infra-species of Asteraceae in Australia. GD and RD refer to the Groves et al. (2003) and Randall (2007) source datasets respectively. Comments link the names of taxa to the companion dataset (CD, Suppl. material 2).

Supplementary material 2

Companion dataset

Authors: Brad R. Murray, Leigh J. Martin, Megan L. Phillips, Petr Pyšek

Data type: species data

Explanation note: The companion dataset (CD) for naturalized Asteraceae in Australia. GD and RD refer to the Groves et al. (2003) and Randall (2007) source datasets respectively. A 'Y' indicates that the taxon name is checked off against the header category. Comments link the names of taxa to the final dataset (FD, Suppl. material 1).