Note: This is Online Appendix 1 of Rebelo, A.G., Holmes, P.M., Spear, D., Klopper, R.R. & van Wilgen, N.J., 2025, 'Lessons learned from compiling a flora checklist for the Cape Peninsula, South Africa', *Koedoe* 67(1), a1856. https://doi.org/10.4102/koedoe.v67i1.1856

Supplementary Materials

Supplementary Material 1. Steps for processing species distributional data for regions in South Africa.

1. Define your study area, (e.g., a national park, nature reserve or magisterial district and its surrounding [representative] areas in South Africa or another country.
2. Obtain the complete taxonomic backbone from SANBI if your study area is within South Africa (South African National Plant Checklist, sources from BODATSA, for plants or SPECIFY for animals). Alternatively, use the GBIF taxonomic backbone or backbone relevant to your study area.
3. Identify data sources for distributional data
    a. Unpublished institutional databases
    b. Citizen Science sources
    c. Literature sources (e.g., publications, checklists, surveys, field guides)
    d. Other (e.g., Facebook, blogs, personal lists, unaccessioned reserve herbaria)
    e. Determine special data fields (e.g., date accuracy, altitude) needed from sources where they exist
4. Compile Metadata database for sources
    a. Data source
    b. Institution
    c. Number of records per type
    d. Type of records (e.g., count, estimate, specimen locality, checklist)
    e. Contact person
    f. Restrictions on data use
    g. Contracts if needed
    h. Turnaround times for fixing data
5. Obtain the data from the sources
    a. Decide how to determine scope (e.g., will the full list be processed with records outside of the scope clipped later or will the data be preclipped to a particular area (along with location accuracy of a site centroid or other relevant non-GIS based site descriptor)
6. Integrate the species lists
    a. If a Species Checklist already exists, check and update this against the latest taxonomic backbone
    b. Species/subspecies on the backbone (OK)
        i. Names match – no issue
        ii. Synonyms – Integrate old names to current name
    c. Problems (handling will depend on source, data type, and timeframe; document these in a field to track changes and progress or unresolvable)
        i. Species
            1. spelling or unassigned names or “c.f.”-species
            2. Species without applicable subspecies (or genera without species)
            3. Old species currently split

4. Probably incorrectly identified species
5. Data without a record of ID certainty
6. Data with specimens collected but still no names

ii. Localities
1. Locality outside of known distribution range, but possible
2. Likely wrong localities based on distribution
3. Likely wrong localities due to coding or name issues
4. Data without a locality error/resolution specified

iii. Dates
1. Invalid dates and nonsense
2. Unlikely dates

d. Add new records to the Species Checklist
i. Flag extralimital and alien species
ii. Document wrong species to be excluded from Species Checklist

e. Send feedback to data provider as well as national curator (here SANBI) for any new records not in backbone
i. Consider publication of new records

7. Send data back to sources for cleaning; await updated data
a. Or skip this step if a once-off desktop exercise

8. Receive cleaned data
a. Incorporate into Specimen Checklist as a snapshot
i. Database (data source)
ii. Source accession number
iii. BODATSA unique identifier (or other unique identifier as per backbone)
iv. Date
v. Latitude
vi. Longitude
vii. Locality name
viii. Locality resolution
ix. Collector
x. Other data redundant (e.g., Family: link from backbone more accurate)
xi. Other data unique (e.g., SANParks management block number)
xii. Other data shared fields (e.g., abundance using CREW codes)

b. Add administration fields to the Specimen Checklist
i. ID issues
ii. Location issues
iii. Data issues
iv. Other issues

c. Construct metadata for Specimen Checklist
i. Explicitly put “use by date”

ii. Include citation requirements
iii. Include use limitations and confidentiality agreements

9. Make data available to collaborators
10. Attribute Species Data
    a. Determine what fields are required
        i. Determine if data already exist and values for each field
        ii. If new, determine what values are required for each field
            1. Determine if it is possible to get data
                a. From literature
                b. From experts
                c. From workshops
                d. Drop field if data not available
            2. Ease of acquiring data
                a. Which fields are family and generic features
                b. Which fields are highly labile and need to be tackled species by species
    b. Determine which species the fields are needed
    c. Arrange staff-funding to obtain the data
    d. Negotiate for useful fields with extensive data to be stored on BODATSA (or other source database) for sharing across platforms and species checklists

Supplementary Figure S1. Schematic of taxonomic checking query steps for compiling a checklist.

We use a Unique Taxon identifier (in the case of the SANBI SANPC, this field is called RecordGUID) to link with the taxonomic backbone to other databases. Each taxon name has its own unique identifier.

In our process, we first created a copy of the taxonomic backbone database where edits could be made without accidentally overwriting species names. This is not strictly necessary, and column additions can be made directly to the main backbone. This backbone checklist should include all taxonomic groups and species that are likely to be encountered in sources. We added the fields “Friendly name” and “Super-friendly name” to this table to enable easier matching across other sources and identify the appropriate Unique Taxon ID. For the case of the SANPC backbone, we also included the Spnumber, which is a unique identifier that matches to previous versions of the backbone and may be useful for matching to older datasets.

BACKBONE INFORMATION					ADDED FIELDS		EXTRA FIELDS	
Unique Taxon ID	**taxstat**	**Accepted Unique ID**	**Full Name**	**Spnumber**	**Friendly name**	**Super friendly Name**	**Red List status**	**Traits**
6e8e8189-de77-43de-a6a9-59c6ba95ad06	acc	6e8e8189-de77-43de-a6a9-59c6ba95ad06	*Schoenus auritus* (Nees) T.L.Elliott & Muasya	128142	*Schoenus auritus*	*Schoenus auritus*	LC	…
5ca8fbaf-fcbb-4ffe-a816-eb6992c5d6aa	syn	6e8e8189-de77-43de-a6a9-59c6ba95ad06	*Tetraria sylvatica* (Nees) C.B.Clarke	59921	*Tetraria sylvatica*	*Tetraria sylvatica*	LC	…
a2334e0c-dbd9-4cce-91b5-4f2f61d569c0	inc	a2334e0c-dbd9-4cce-91b5-4f2f61d569c0	*Erica glabella* Thunb.	43866	*Erica glabella*	*Erica glabella*	LC	…
ffe0068b-2c6c-4a6c-ae0d-8919ab0a22b0	acc	ffe0068b-2c6c-4a6c-ae0d-8919ab0a22b0	*Erica glabella* Thunb. subsp. *glabella*	43429	*Erica glabella* subsp. *glabella*	*Erica glabella glabella*	LC	…
6e8e8189-de77-43de-a6a9-59c6ba95ad06	acc	6e8e8189-de77-43de-a6a9-59c6ba95ad06	*Leucadendron conicum* (Lam.) I.Williams	20502	*Leucadendron conicum*	*Leucadendron conicum*	LC	…

Then, working with a copy of source data for processing, we added fields to describe updates to names based on the match to the taxonomic backbone. The “Original species” (table below) name (also called the verbatim name) is matched to the friendly name or superfriendly name created in the backbone table using queries or where this is not possible, manually. In the example, the data submitted by various data providers included the following taxa: *Tetraria sylvatica*, *Erica glabella* and *Leucadendron conicum.* These were matched to the backbone and the status of each name determined. A summary of matches to the data that includes the species original and current name, along

with curation notes and the Unique Taxon ID (here RecordGUID) of the matched name was then provided to the owner of the data source to enable them to make relevant updates to their own databases if so required.

SOURCE		ADDITIONAL FIELDS IN LINKED DISTRIBUTION DATA				
Source Acc number	**Original taxon**	**Latitude**	**Longitude**	**Accuracy**	**Collector**	**Date**
95	*Tetraria sylvatica*	-34.224	18.404	5m	Hugh Taylor	1966
19	*Erica glabella*	-34.207	18.374	5m	Hugh Taylor	1966
1123075148	*Leucadendron conicum*	-34.160	18.417	5m	Janeen Nichols & Claire McCartney	11/8/2013

ADDED FIELDS				
Match name	**Accepted name**	**Unique Taxon ID**	**Exclude**	**Curation notes**
Tetraria sylvatica (Nees) C.B.Clarke	*Schoenus auritus* (Nees) T.L.Elliott & Muasya	6e8e8189-de77-43de-a6a9-59c6ba95ad06	FALSE	Matched to synonym
Erica glabella Thunb.	*Erica glabella* Thunb. subsp. *glabella*	a2334e0c-dbd9-4cce-91b5-4f2f61d569c0	FALSE	Missing subspecies or variety; Appropriate subspecies determined based on distribution
Leucadendron conicum (Lam.) I.Williams	*Leucadendron conicum* (Lam.) I.Williams	6e8e8189-de77-43de-a6a9-59c6ba95ad06	TRUE	Check for exclusion: way out of range: probably typo for Ld coniferum.

As species are added during the processing of sources, the Masterlist of the checklist is updated, so that it includes all names used across sources and indicates their status according to the backbone, along with any relevant curation notes.

MASTERLIST						
Unique Taxon ID	**Full Name**	**Taxon status**	**Exclude**	**Area status**	**Curation notes**	**Source Acc number**
6e8e8189-de77-43de-a6a9-59c6ba95ad06	*Schoenus auritus* (Nees) T.L.Elliott & Muasya	acc	FALSE	Indigenous	Matched to synonym	95
ffe0068b-2c6c-4a6c-ae0d-8919ab0a22b0	*Erica glabella* Thunb. subsp. *glabella*	acc	FALSE	Indigenous	Missing subspecies or variety; Appropriate subspecies determined based on distribution	19
6e8e8189-de77-43de-a6a9-59c6ba95ad06	*Leucadendron conicum* (Lam.) I.Williams	acc	TRUE	Not present	Exclude: way out of range: probably typo for Ld coniferum.	1123075148

Supplementary Material 2. Detailed steps used in the processing of the Master List

The process outlined below is best conducted within a database (e.g., MS Access). Fields and values are displayed below in alternate fonts (these are based on the fields in the South African National Plant Checklist – SANPC, formerly BODATSA, and will differ slightly if you use another backbone source).

Steps:

1. Set up a table for the species **Master List** (Checklist) as follows:
    a. The important fields (columns):
        i. Old name on list (if any) – if you have a preliminary list, use this
        ii. BODATSA RecordGUID or other unique identifier (to link to the BODATSA / SANPC backbone or another backbone data source)
        iii. New_Name (to store the new name for checking – this will be a duplicate of the name in the SANPC or another backbone, but it will be useful for rapid responses to queries, and checking procedural errors)
        iv. Notes: to record any issues
        v. Species_Status: the residence status of the species in the area: Invasive | Naturalized | Extralimital | Alien | Indigenous | Near_Endemic | Endemic | Exclude | Unknown [note "Exclude" is for taxa mistakenly recorded in the area (e.g., taxonomic change, misidentification or locality error), or where all records are planted (e.g., in recorded at Kirstenbosch National Botanical Garden)]
        vi. Exclusion: flag for names that will not be used.

2. Get the SANPC or other backbone. Use the latest version for download, updated annually, and to it:
    a. Create a field: Friendly_Name as trim(genus + species + subspecies + variety + forma) - [with one space between each, and by adding the rank denoting terms for infraspecific names as required (e.g., subsp., var., or forma)] – this is essential as it links the names to datasets that do not include authors [SANBI is considering adding a FullName_NoAuthors field to the next annual release (2026 onwards) to facilitate this process].
    b. Create a field for iNaturalist: Super_Friendly_Name as trim(genus + species + subspecies + variety + forma) - [with one space between each] – this is essential as iNaturalist uses trinomials without ranks.
3. For each dataset, create a table with each column as a separate field, and then:
    a. Check fields.
    b. Add field CheckList_Notes to the checklist.
    c. Add a field CheckList_Friendly_Name if a suitable field does not exist (some datasets have genus, species and subspecies as separate fields), and construct it as trim(genus + species + rank1 + rank1_name + rank 2 + Rank2_name) - [with one space between each]

= no authors, and with standardized ranks [Rank1/2= “subsp.” or “var.” or “forma”]. If the rank is not provided, construct a field CHECKLIST_SUPER_FRIENDLY_NAME as in Step 2.

4. For each dataset, add to the Master List (or first time, to check the preliminary Master List) and assess the status of these names according to the backbone. In the SANPC, there is a column TAXSTAT that indicates whether a name is accepted or not:
    a. Check matched accepted names – SANPC: TAX = ACC - accept these, fill in fields RECORDGUID and NEW_NAME.
    b. Check matched inclusive species names - SANPC: TAX = INC - accept these names, fill in fields RECORDGUID and NEW_NAME
        i. In the Master List Notes: add a WARNING: this species has infraspecific taxa
        ii. In the checklist LIST_NOTES field: add a WARNING: this species has infraspecific taxa
    c. Check matched synonyms - SANPC: TAX = SYN – accept the current name, fill in fields. RECORDGUID and NEW_NAME with the current name
        i. In the Master List Notes add: this species was originally added from a synonym
        ii. In the checklist LIST_NOTES field: add a NOTE: this name is now a synonym for *xyz*.
    d. Check multiple matched synonyms or a synonym already flagged as matched
        i. In the Master List Notes: add a WARNING: this species has multiple synonyms
            1. These should be checked and the outcome reported in the LIST_NOTES if relevant
        ii. In the checklist LIST_NOTES field: add a WARNING: this species has subspecies

Add notes to checklist LIST_NOTES field: record could not be used <or> this concept accepted: needs checking

    e. Check unmatched names
        i. Manually find matches in the SANPC or other backbone used. These will be –
            1. Spelling errors: fix spelling and update RECORDGUID and NEW_NAME. Document in Notes
                a. Document in checklist LIST_NOTES field: provided corrected name
            2. Rank errors: fix the rank (subsp., var., forma) and update RECORDGUID and NEW_NAME. Document in LIST_NOTES
                a. Document in checklist LIST_NOTES field: provide corrected rank
            3. Flag all names that have no obvious matches, or where the matches are dubious or unlikely. Document in LIST_NOTES that they are not useable
                a. Flag these for EXCLUSION (these can be deleted later if desired, but preferably not if flagged “Exclude” in SPECIES_STATUS)
                b. Document in checklist LIST_NOTES field: <as appropriate>
            4. Note all problems or errors in the SANPC or other backbone used
                a. On completion, send a list of issues to the Plant Checklist Coordinators at SANBI or curators of other backbones used.

5. Tidy up the Master List:
    a. Remove duplicates
    b. Delete unwanted entries from checklists
    c. Document excluded and erroneous species.
6. Check for updated synonyms to wrong species.
    a. These are usually duplicate names that can accidentally be linked to the wrong species. They are hard to detect, unless careful documentation notes of updates are kept in the Master List: Usually picked up by strange distributions, extralimital species, and prior knowledge when double checking the lists.

The steps for processing distribution data are the same, except that names not on the Master List need to be further checked and then added to the list.

Supplementary Material 3. Issues encountered in processing data.

The taxonomic issues encountered during processing data, summarized by taxon (i.e., not by number of specimens / observations, but bundled as taxa) and presented in decreasing order of frequency.

Status (total = 3176 taxa)	Explanation	Solution	# taxa	%
Current	Records matching the taxonomic backbone	No action required	2489	78.4
BEWARE: has infraspecific taxa, including subspecies	Species name is current, but subspecies has not been specified	If treating it at species level is sufficient, then no further action required. Where subspecies and varieties exist, literature was consulted to assign it to an infraspecific name, where possible.	362	11.4
WARNING DUPLICATE[1]	The name could refer to more than one possible species as a result of taxon splits. In some cases, the incorrect name might have been assigned – these require checking if it is possible to do so. Older lists will have more of this problem.	If only one of the taxa occurs on the peninsula, this one was assigned. If more than one taxon was present on the peninsula, the original source needed to be consulted to assign the correct taxon.	115	3.6
Synonym – updated	This name was a synonym and has been corrected	Updated to current name as per BODATSA	67	2.1
ERROR: cannot find match in BODATSA [to be discarded]	This is an unknown name and cannot be used. These are often species put in the wrong genus, temporary fieldwork names, or misreadings. In rare cases, they may be unrecorded species from elsewhere.	Usually these are unresolvable and simple errors are routinely corrected.	28	0.8
ERROR: spelling wrong	Name recognized, but spelling incorrect.	Name has been updated to match correct spelling.	12	0.3
Synonym – updated - duplicates	The name was a synonym, and was updated, but the updating has more than one possible taxon.	Where this occurred, names were manually checked and the correct taxon	9	
WARNING TRIPLICATE[1]	Three or more species have this name. In some cases the incorrect name might have been assigned. Older lists will have more of this problem.	If only one of the taxa occurs on the peninsula, this one was assigned. If more than one taxon was present on the peninsula, the original source needed to be consulted to assign the correct taxon.	8	

Status (total = 3176 taxa)	**Explanation**	**Solution**	**# taxa**	**%**
Temp mismatch	A temporary matching error, made while integrating the lists (e.g., assigning wrong rank to unranked lists – var. instead of subsp., etc.). For example, iNaturalist can only handle trinomial names and not quadrinomial names.	We used BODATSA to assign the correct taxonomic rank.	5	
ERROR: wrong rank	The entry had the wrong rank requiring fixing (var. instead of subsp., etc.).	We used BODATSA to assign the correct taxonomic rank	4	
NOTE: BODATSA treatment rejected: taxa maintained as separate	This is for cases where we disagree with the BODATSA species (e.g., we treat *Erica gilva* as separate from *E. mammosa*).	No action required. A note to this effect might be needed.	4	
Synonym – updated – triplicates	The name has several synonyms, but the updating has more than one possible taxon.	Where this occurred, names were manually checked and the correct taxon	2	
Described since SANPC list version accessed	New taxon since the backbone was imported. Probably already in current backbone.	Checked and added to backbone if not already there. This will have to be correctly linked in future following a subsequent release of the SANPC that contains this name.	1	
ERROR: wrong rank and spelling wrong	Double error - refer to above descriptions and explanations of the two error types	Made required updates as described above.	1	
ERROR: wrong subspecies	The listed subspecies does not occur on the Peninsula	Updated to the subspecies known to occur locally.	1	

BODATSA / SANPC ISSUES

These are issues relating to the taxonomic backbone and South African National Plant Checklist produced from this backbone and are presented in order of the frequency in which they were encountered.

Status (total = 3176 taxa)	Explanation	Solution	# taxa	%
* = gen-spec/spnumber issue[2]	Refer to footnote 2	New numbers were manually assigned. This matter has been resolved in the post-2022 releases of the SANPC.	(60)	1.9
!! SANPC error: synonym links to a synonym	A synonym must link to a current name.	These issues were manually corrected to the current name as per BODATSA. This matter has been resolved in the post-2022 releases of the SANPC.	23	
!! SANPC error: technical – nominotypical issues	All the subspecific taxa have been elevated to species and the nominotypical name has been removed from the yearly SANPC release as it is now an unused autonym. Consequently, the subspecific taxa cannot be matched to the backbone.	These taxa have been manually assigned to species. The SANPC editors are looking into this problem.	13	
!! SANPC error: missing spelling variant (which was in previous version)*	A database error where a spelling was corrected resulting in a mismatch. The correct procedure in the backbone should be to regard this as a synonym and not delete it. [Orthographic variants due to spelling errors are not retained in the SANPC as separate records. Notes regarding orthographic variants may be available in the Nomenclatural Notes field of the yearly release.]	Taxon names were manually assigned according to the correct spelling.	7	
!! BODATSA error: missing species (alien)	Species not in BODATSA and should be added.	These species were added and BODATSA informed of the presence of the taxon.	6	
!! BODATSA error: missing species	Species not in BODATSA and should be added.	These species were added and BODATSA informed of the presence of the taxon.	5	
!! ERROR: cannot find match in BODATSA (but was in previous version)*	The name cannot be found in BODATSA for an unknown reason. It was present in the previous version.	BODATSA were notified to investigate why name no longer appears. It might have been a misapplication.	2	

Status (total = 3176 taxa)	Explanation	Solution	# taxa	%
!! BODATSA error: species has vars., but no species rank name	Database error where the species name has been omitted or wrongly categorized.	BODATSA were notified to investigate why taxon name was omitted.	1	
!! BODATSA error: species listed as "inc" but no subspecific taxa exist	Database error where the species name has been wrongly categorized.	BODATSA were notified to investigate the issue.	1	
!! BODATSA error: species has subspecies, but no species rank name	Database error where the species name has been omitted or wrongly categorized.	BODATSA were notified to investigate why taxon name was omitted.	1	
!! BODATSA error: spelling error:	There is a spelling error in BODATSA.	BODATSA were notified to investigate the issue.	1	

Notes

1. The duplicates and triplicates are cases where names have to be checked to ensure that the appropriate name is used. We will assume that there are no issues, but the reality is that if the list is old, there may be a proportion of names that have changed in concept since the list was compiled, and thus may be of the wrong species [i.e., (1) a change in species concept, where species have been split, or (2) – less likely: where there have been name swaps – which are easily identifiable where lists have authors, but not where lists do not include author names]. As a rule this won't usually apply to up-to-date herbarium lists, and iNaturalist that updates its taxonomy continuously, but may apply to the smaller herbaria, other lists, and especially to older literature listings.
2. * = gen-spec/spnumber issue: BODATSA no longer uses the PRECIS Gen-Spec reference numbers and following migration to BRAHMS v.8, the BRAHMS v.7 SpNumbers have been replaced with the RecordGUID as the unique identifier for name records. As a result, data links in older dataset that used these numbers have been lost. Since the 2023 yearly release of the SANPC, the old PRECIS GenNo and SpNo, and the BRAHMS7 SpNumber is included for relevant records to enable linking older datasets using these unique identifiers, and so that the new RecordGUID for a name can be obtained.

Additionally, it must be noted that the starting list that was used is a relatively recent and clean list (prepared for the city in 2017). Older lists will generate far more issues. There are 28 names that are irreconcilable (unfindable or ambiguous) and will have to be discarded.