*i*Plants: The World's Plants Online

The Compilation Process

APPENDIX 2

Final Report of Pilot Project: 1 April to 30 Nov 2004

Version 1.6

Author(s): Jackson, M

Location: ***i*Plants** document library: **www.iplants.intranets.com**

Directory: Group docs / Management / Outputs / System Procedures

Filename: compilation procedures

Approval for Public Release if appropriate: Formally approved for release by Steering Committee

Date:

Revision History

Date	Who	What	Version
12/09/04	MJ	Created first draft	1.0
30/9/04	AP	Commented / + minor edits BA	1.1
19/10/04	MJ	Revised in light of comments from BA	1.2
11/11/04	MJ	Revised following feedback from Kew meeting October 2004	1.3
24/11/04	MJ / BA	Revised following further feedback	1.4
20/12/04	BA	Minor edits	
20/12/04	BA	Published for Moore	1.5
23/03/05	BA	Changes from NYBG included	1.6

TABLE OF CONTENTS

1. Introduction

1.1. The *i*Plants project

iPlants aims to produce an index of all the world’s plant species together with, where possible, an image and a preliminary conservation assessment. This index will be made available online.

1.1.1. Further Information:

For further information please contact:

The *i*Plants Initiative,
c/o Alan Paton,
Royal Botanic Gardens Kew,
Richmond, Surrey, UK
TW9 3AB
information@iplants.org

1.2. Purpose of this document

This document aims to
1) Set out the compilation process which is recommended for use in the iPlants project
2) Raise the issues that require resolution before the process can be finalised

1.3. Outstanding Issues

The following issues are raised in this document and have yet to be addressed.

1. A better name for the 'compilation' activity which forms part of the overall Compilation Process, OR agreement to rename 'Compilation Process' to 'Checklist Creation Process'.
2. What priority iPlants should assign to providing online expert review facilities.
3. Decisions on when a checklist might be made accessible to the end-user.
4. What the maintenance procedures should be.

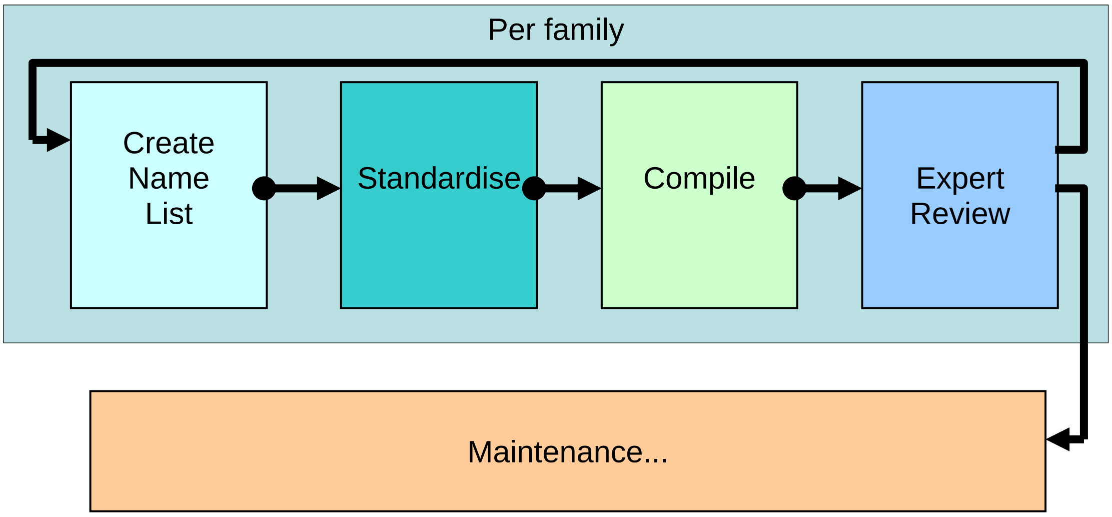
2. Introduction

2.1. Overview

The compilation process consists of the following sequential steps per family:

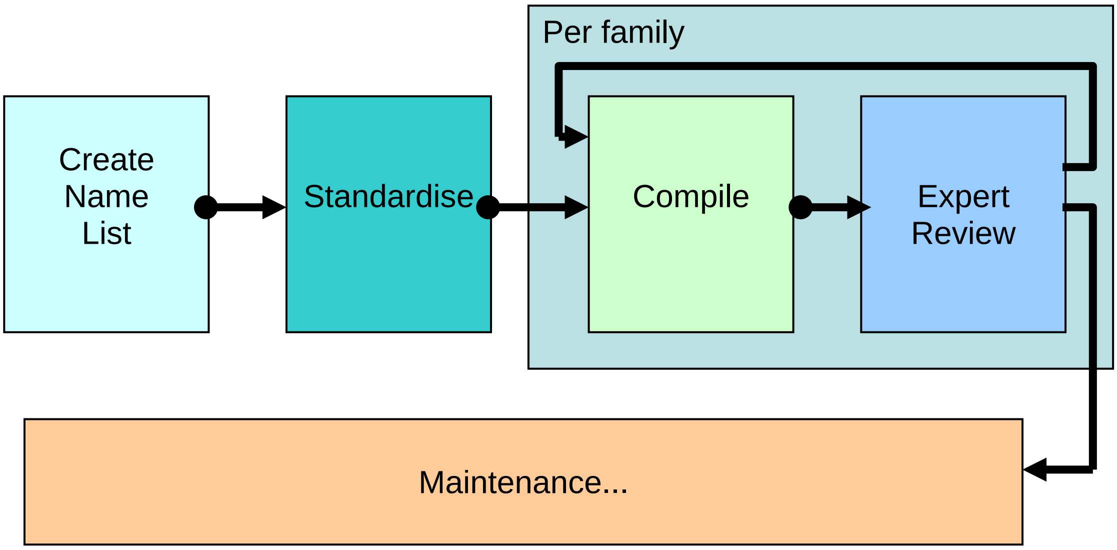
- Create Name List
- Standardise
- Compilation (assignment of taxonomic status and distribution)
- Expert Review
- Maintenance

Each of these steps must be carried out in the sequence shown, and is described below in more detail. Inputs to, and outputs from each step are listed at the top of each step. Issues, some requiring decisions, are indicated within the text. The following diagram summarises the major processes.



2.2. Alternative Approaches

An alternative approach is considered preferable in terms of improving the overall efficiency of the project, but with a probable impact on initial throughput of compiled datasets. This involves carrying out the name list creation and standardisation on *all names* en masse, which will avoid duplicating standardisation efforts betwene family datasets, and increase the potential gains from automating parts of the process. This approach will redefine the overall process as per the diagram below.



The first stage of this process would be to apply automated processes to split the place of publication. Across the board standardisation could then be achieved by the following steps:

a) Sort dataset by place of publication
b) Cut dataset into one part per person
c) Each person standardises publications in their part
d) Merge the parts together and check
e) Sort dataset by primary author
f) Cut dataset into one part per person
g) Each person standardises authors in their part
h) Merge together and check
i) Sort dataset by parenthetical author
j) Cut dataset into one part per person
k) Each person standardises authors in their part
l) Merge together and check

It may well be possible to automate i) through l) by searching for basionyms using the name as a text string, and where matches are found, inserting the already standardised author from the basionym.

Even if this solution is not adopted, and standardisation proceeds on a family by family basis, it may still make sense to carry out stages 3.1 through 3.3 en masse.

2.3. General processes

Throughout the process there is a need for the compilers to be able to

a) access the source data from which the current data originated. This is necessary in order to check for mistakes and to help decision-making.
b) view the entire current dataset sorted into alphabetic sequence, as a means of checking it.
c) generally have a flexible way of defining and viewing ad-hoc reports

2.4. Existing Lists

iPlants might consider importing an existing list, which may be just a new list of names, or have been compiled but require standardisation to iPlants standards, or be incomplete or unfinished. This is expected to be an infrequent occurrence. The most appropriate place and method of incorporating such lists into the iPlants process and system will be decided on a case-by-case basis. Attention will be given to acknowledging the contribution and feeding back updated data.

2.5. Management Information

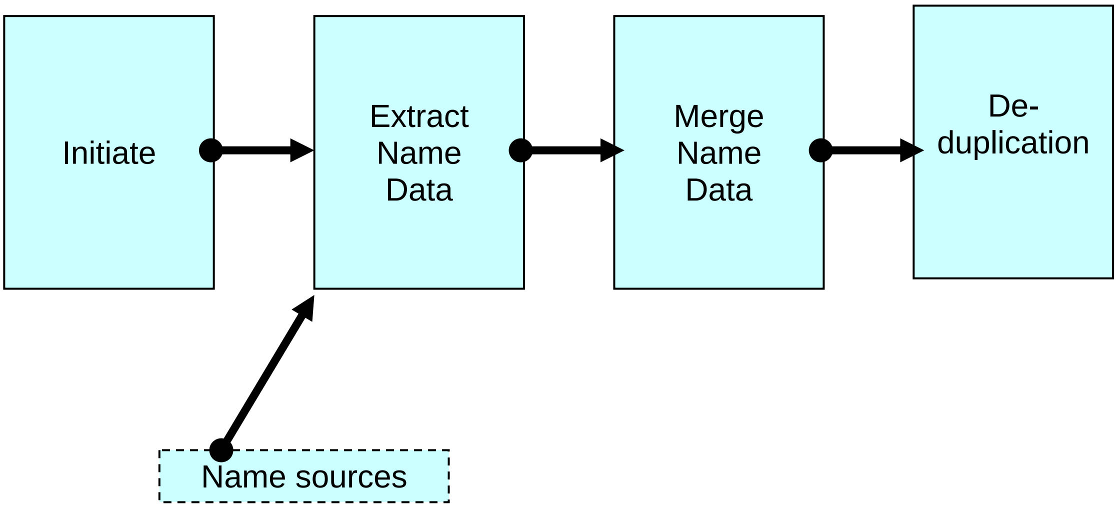
A 'master list' of all genera will be required to manage the compilation process. This list will be used to

a) identify a starting circumscription for each plant family
b) record which names have been extracted from the source databases and therefore which names remain to be dealt with
c) provide a list to check the availability of taxonomic revisions for use during compilation

This list can be created from Brummitt 1992 with additions from IPNI, Tropicos and NYVH.

3. Create Name List

This process concerns the creation of a single list of names from distributed sources. The major steps are diagrammed below.



3.1. Initiation

Input: List of families for which checklists exist and for which checklists are needed. Also list of all generic names showing whether data extracted yet or not.
Output: Family allocated to particular institution and generic circumscription of 'family' documented.
Who & how: Steering Committee, manual process.

Before the compilation process can begin, the Steering Committee will decide which families will be dealt with (compiled and conservation rated), by which lead institution and what the target dates should be.

Before the name data extract can take place, it is also necessary to define the circumscription of the family being dealt with to ensure that all relevant data is extracted. This circumscription will be defined by the Steering Committee (in consultation with family experts where necessary) in terms of genera to be included.

3.2. Extract name data

Input: Datasets from IPNI, Govaerts' working database, Tropicos and NYVH. Family circumscription.
Output: Set of name records per source in broadly standard format. Updated list of all genera to show data extraction.
Who & how: Data management staff in MO, NY and K. Manual process.

The generic definition passed down by the Steering Committee will then need to be translated into database-specific criteria for data extraction. This may be different from source to source due to the varying use of historical family names.

The raw material of compilation is the plant name data held in numerous databases, namely IPNI, the Govaerts' working database (based latterly on data extracts from IPNI), w3Tropicos and the NYVH. This stage of the compilation process is concerned with producing a single collated list of that name data. To do this, the data has to be extracted from the sources, merged, and deduplicated.

Using the taxon circumscription, the relevant data is extracted from the sources, and formatted in as standard a fashion as possible. The list of all genera is then updated to show which names have been extracted.

Issue: Will name data extraction always be initiated and controlled from one place, or by the lead institution per family?
Discussion: The data extraction itself depends on input from source database managers as to the search criteria that should be used. Although delays would impact on the work schedule, the frequency of name data extraction is likely to be low, which mitigates against the development of automated tools.

3.3. Merge data

Input: set of name records per source in broadly standard format.
Output: one set of name records containing all of the above.
Who & how: Data management staff in (lead institute/nominated institute). Manual process.

Data from all of these sources is merged into one list.

Issue: Will the data merge be carried out at one point rather than by the lead institution (see above)?

3.4. Deduplication

Input: set of name records.
Output: deduplicated set of name records.
Who & how: Data editing staff, automated process.

The merged dataset will contain many duplicates (entries for the same name). This process identifies the duplicates, creates a best record from them, and links them to it. A pre-requisite for this process is to nominate the precedence of the various name sources. A default priority order is given below, but this can be varied if the source data quality for the given family warrants it:

a) Govaert's working (unstandardised) database covering genera A-I together with Govaerts' standardised database
b) IPNI names originating from the Index Kewensis for genera J-Z
c) IPNI names originating from the Gray Cards for genera J-Z
d) IPNI names originating from the Australian Plant Names Index for genera J-Z
e) Tropicos names for all genera
f) NYVH names for all genera

NB: Govaerts' standardised database is used to check that duplicates are not entered, but as the names in are already standardised it is not incorporated with the unstandardised data (see 4.5 below).

Duplicates are defined as records where all of the epithets match. NB: IPNI consists of three merged datasets, and so duplicates may exist within it.

The process will work in the following fashion. Every time a new name is detected, a new iPlants record will be created. The content of the record will come from the first data source in the order of precedence, with the following exceptions:

a) place of publication data already parsed into separate strings can be preferred over unparsed data
b) missing data can be filled from subsequent datasources
c) standardised authors can be preferred over unstandardised ones

The above exceptions can be selected per family and per datasource.

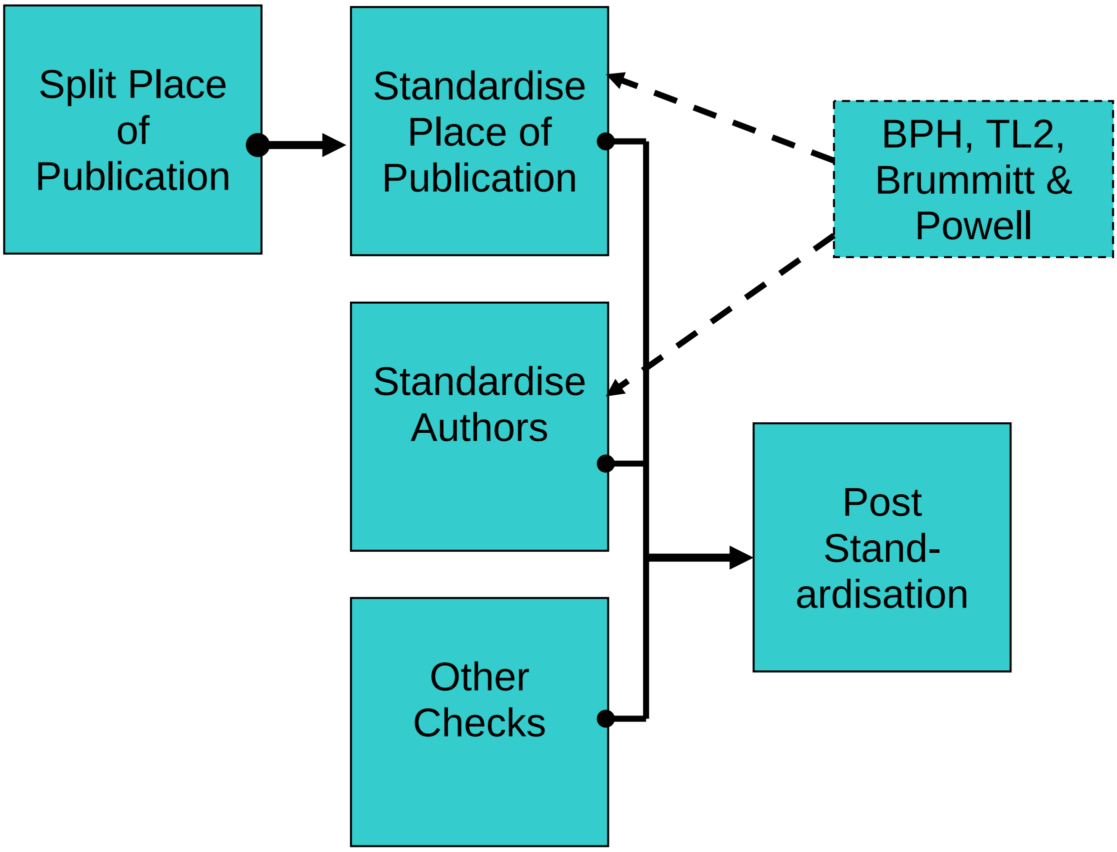
All of the source records are preserved and linked to the iPlants name record, so that

a) they can be examined at any later stage in the compilation process
b) once standardised, the data in the master record can be re-exported to the source databases enabling them to update their records (an activity considered to be outside of the scope of this project)

Issue: Authors have not been standardised at this stage and so cannot be used to detect duplicates, with the unfortunate side-effect that most (some homonyms are perpetrated by the same author) homonyms cannot be detected and would be removed.
Discussion: Three options are available which could overcome most of these situations. Firstly, a test could be made for the year of publication and page number. If these differed, then homonymy is likely and the records could be preserved. A problem with this is that these data fields will often not have been parsed out at this stage. Secondly, if one data source was thought likely to contain most of the homonyms (e.g. the IK part of IPNI, which generally has the most names), then it could be given primacy in this comparison process and duplicates within it could be preserved. This is the main reason for the default priority order shown above. Thirdly, even if the source priority order was to be different, a special overiding exception could be made whereby if a name is detected more than once in Index Kewensis thne these records could be preferred.

4. Standardise

This process concerns the standardisation of place of publication and plant name authors. The standardisation of these two data items is not interdependent and can take place simultaneously.


Split Place of Publication
Standardise Place of Publication
BPH, TL2, Brummitt & Powell
Standardise Authors
Post Stand-ardisation
Other Checks

4.1. Split Place of Publication

Input: deduplicated set of name records
Output: name records with parsed and formatted place of publication
Who & how: Data editing staff, semi-automated process.

In order to standardise the place of publication, it is first split into its constituent parts: publication author, place of publication, volume and page number, and publication date. The 'volume and page number' field contains a combination of the following items: edition, series, volume, part and page number, each with their own formatting rules.

Where this data was already split in the source databases, this should have been preserved. It should also be possible to automate some of this parsing process, so that batch processes can be run on the data. The degree to which this will be possible requires further investigation. However, certain principles are clear

a) the unparsed data would be preserved, and the parsed data inserted into separate fields. This will allow iPlants staff to check that the routines have operated correctly.
b) some data will not parse correctly. iPlants staff will have to manually process these records.

All of the records will need to be checked. For instance, pre-parsed data may still not have been formatted exactly according to iPlants standards. Sorting the data by place of publication, volume and page, and publication date will help spot outliers and obvious formatting errors.

4.2. Standardise Place of Publication

Input: name records with place of publication split into separate fields
Output: name records with standardised place of publication
Consulted: BPH/S, BPH, Abbreviations database, Index Herbariorum, TL/2, Kew Library Catalogue, Natural History Museum Library Catalogue, Original data
Who & how: Data editing staff, manual process.

During this process, the place of publication string is standardised. The place of publication can be a journal, serial flora or book. Standard terms are selected primarily from BPH/S and BPH (journals), the Abbreviations database (Serial Flora) and TL/2 (books). Each time the publication date given in these sources should be checked, and added if not cited. The original source data may need to be checked.

Once the splitting and standardisation is complete, the data is checked by sorting the data by place of publication, volume and page, and publication date. This will help spot outliers and obvious formatting errors.

Issue: Freely accessible online versions of BPH/S and BPH, the Abbreviations database and TL2 would be useful resources.
Discussion: TL/2 is now available online, but only as a subscription service and there appear not to be any web services that iPlants software could link in to. In addition, the 'standard' abbreviations are sometimes not unique. BPH exists in digital form but is not widely available or online. It would probably benefit iPlants data entry if these resources were available as linked authority files.

Issue: where publications are not found in BPH and/or TL/2 and are otherwise unknown a choice needs to be made. Either the text used will be left as is, or a new standard term will need to be coined.
Discussion: In order to facilitate subsequent data entry it makes sense to coin new terms, and make them available for use within iPlants. Either way, the text not found and any new term coined will need to be recorded so that they can be fed back to BPH and TL/2 at an appropriate juncture.

4.3. Standardise Authors

Input: deduplicated set of name records
Output: name records with standardised authors
Consulted: TL/2, Authors of Plant Names (Brummitt & Powell 1992), Original data
Who & how: Data editing staff, manual process.

The primary and parenthetical authors, then the replaced synonym author, are first standardised by replacing the names given with standard abbreviated terms from Authors of Plant Names (Brummitt & Powell 1992). If in doubt, consult TL/2 or go back to the orginal source data. If the name isn't to be found, create a new abbreviated term following the Brummitt and Powell approach.

Where authors are not found in Authors of Plant Names a new term is coined. The new term and the original text will be preserved for feeding back to the Authors of Plant Names at an appropriate juncture.

Once the standardisation is complete, the data can be checked by sorting the data by publication author. This will help spot outliers and obvious formatting errors.

Issue: The Authors of Plant Names is available online. It may be possible to cost-effectively integrate it into the iPlants data entry system and provide a more convenient method for staff to access values from it.

4.4. Other Checks

Input: deduplicated set of name records
Output: checked name records
Consulted: Brummitt 1992
Who & how: Data editing staff, semi-automated process.

Several other checks and edits need to be made, including spelling of epithets, presence of hybrid signs, and stray nomenclatural remarks. Authors and dates are checked, and book authors standardised. Duplicate names and older invalid names are deleted. These checks require staff to view the records individually, and have been carried out to date at the same time as the splitting of place of publication. However, if the latter can be automated to any degree, then it makes sense to separate these processes.

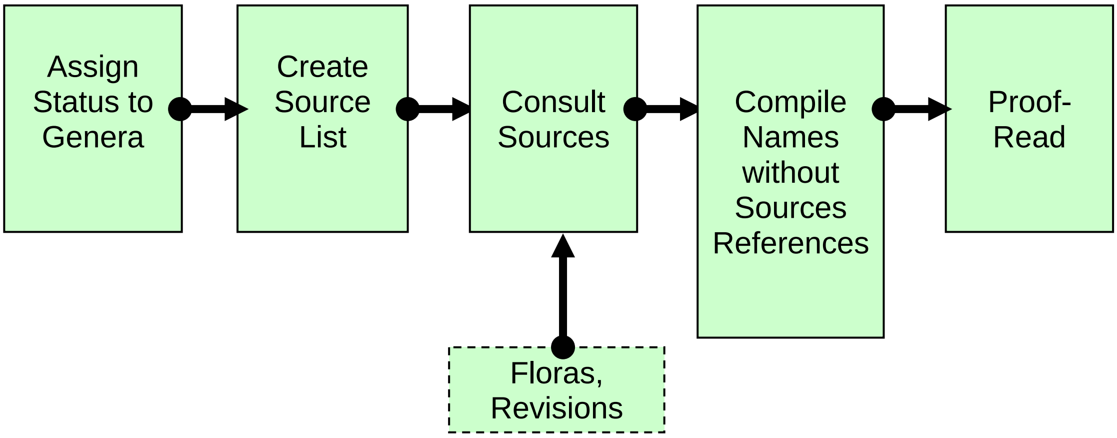
4.5. Post-standardisation

Input: list of standardised name records, Govaerts' standardised database
Output: comprehensive list of standardised name records
Who & how: Data editing staff, manual process.

The final data source to be added into the list of names is Govaerts' standardised database for genera A through I.

5. Compilation

This process is where taxonomic status and linkages are applied, and distribution recorded.



5.1. Assign Status to Genera

Input: List of names
Output: List of names with some taxonomic statuses assigned
Consulted: Brummitt 1992
Who & how: Data management staff. Automated process.

Genus records are assigned taxonomic statuses at this stage, based on those in Brummitt 1992 (potentially moderated by expert opinion), and if the name is a synonym, the accepted name that it links to is recorded. All species names of synonymous genera are also be assigned a default synonym status at this stage.

5.2. Create Source List

Output: List of sources to be consulted
Consulted: Frodin 2001, Govaerts' general list, Family circumscription, Mabberley 1997, Kew Record
Who & how: Compilers in consultation with taxonomic experts. Manual process.

The starting point for compilation is a list of sources to be referenced, split into floras and revisions.

Floras that cover the family distribution should be included. To identify these, consult the Frodin 2001 and Govaerts' general list.

The Family circumscription should be consulted to get a list of genera in the family. Revisions of any of these genera should be included. A starting list can be identified by referring Mabberley 1997 and the Kew Record online database. However, most of the sources will be the places of publication given in the list of names (as new name lists are frequently connected with revisions), and other sources listed in the bibliographies to these publications.

The consolidated list should be reviewed and amended where necessary with appropriate taxonomic experts.

5.3. Consult Sources

Input: comprehensive list of names with some synonymy (at generic level)
Output: comprehensive list of names, with incomplete synonymy and distribution
Who & how: Compilers, manual process.

The sources are now consulted in sequence. The preferred sequence may vary from family to family, but there is some evidence that consulting floras before revisions has marginal efficiency benefits because it enables the bulk of the data to be captured early on.

Each source is recorded in the system as it is used. Based on the source, and a combination of a set of rules backed up by the compiler's judgement, taxonomic statuses are applied to names, and distribution data to taxa/accepted names. A variety of remarks are recorded depending on the taxonomic status assigned, and other considerations.

The following categories of names are added:

i) autonyms for taxa with accepted infraspecies – geography has to be corrected, and all heterotypic synonyms should now refer to the autonym and not the species
ii) new records provided they have a place of publication, or if they are accepted or a basionym (with a question mark in the publishing author field)

and the following categories of names are deleted:

i) duplicates
ii) invalid records (except perhaps those in current use, tautonyms and autonyms)
iii) misapplied names (sensu auct. or pp.)
iv) autonyms if no accepted infraspecies exist

Each record is checked for e.g. spelling of names, internal consistency, and formatting.

5.4. Compile Names without Source References

Input: comprehensive list of names, with incomplete synonymy and distribution
Output: comprehensive list of taxa and synonyms with distribution
Who & how: Compilers, manual process.

After the sources have been worked through, there may be some names which have not been dealt with. To deal with these, the entire list is sorted alphabetically and checked. Based on a set of business rules, these are assigned taxonomic statuses and distributions. Names not sourced in the literature are referred back to the originating name list and the compiler.

5.5. Proof-Read

Input: comprehensive list of taxa and synonyms with distribution
Output: comprehensive list of taxa and synonyms with distribution – checked and corrected
Who & how: Compilers, manual process.

The checklist is now printed out in taxonomic format, with synonymy indicated. A variety of checks are carried out on spelling and internal consistency.

6. Expert Review

This process is where a compiled list is reviewed by taxonomic expert(s) and feedback incorporated.

Input: comprehensive list of taxa and synonyms with distribution
Output: comprehensive list of taxa and synonyms with distribution – reviewed and corrected
Consulted: taxonomic experts
Who & how: Taxonomic experts and compiler, manual.

This process is likely to consist of referring the list to taxonomic experts for review, and actioning review comments. One or more experts may be asked to review a family, or part of family.

Initially this process will involve sending printed or emailed lists and receiving them back annotated by the expert. iPlants Compilers will then make the necessary changes. Generally compilers will not query experts taxonomic judgements, but they may challenge or override factual errors or inconsistencies. It may be necessary for some dialogue to ensue, but this would normally be focussed on individual queries and shortlived.

iPlants would like to develop online reviewing facilities which may extend the feedback facilities anticipated for the maintenance phase, and these would eventually become the preferred method for conducting reviews.

Lists may be released without expert review. In some cases there may be no experts, and in others it may not be possible for whatever reason for experts to respond within a reasonable timescale. Experience has shown that the product of compilation, provided it is conscientiously created and carefully checked, will be largely accepted by experts. Therefore, iPlants intends to release lists (clearly labelled as provisional) where necessary.

Issue: In recompense for conscientious review taxonomic experts will expect appropriate acknowledgement, as this is an important aspect of how their careers are currently judged. 'Reviewer' may not imply enough input, so some sort of co-authorship status may have to be considered.

7. Maintenance

As soon as a dataset is made accessible on the iPlants website it may be considered under maintenance (i.e. maintenance activity will begin well before the completion of the project and online website). Changes to the data will be prompted by the following events:

i) new publications (or old ones being drawn to the attention of the iPlants editors)
ii) feedback
iii) periodic reviews

Editorial responsibility would lie with a team of editors. One option is for maintenance to lead to versioned releases of the taxonomy per family, as opposed to dynamic updating of individual records, and past releases would be preserved. This would allow

a) stable views which change infrequently and may be securely referenced by user community
b) more effective scheduling of editorial and expert review workload

However, given that changes affect an estimated less than 1% of the data each year, a dynamic database also has its attractions. Many changes are not 'controversial' but simply factual - new names and distribution data could be added as soon as they are detected, factual errors corrected, and all feedback made visible as soon as received and vetted. This will enable contributors to see their contributions, and for end-users to make decisions based on the latest information.

Issue: The policy on versioning needs to be finalised.

7.1. New publications

Input: new publications
Output: new entries to website
Consulted: IPNI, Tropicos, NYVH and Kew Record. Taxonomic experts.
Who & how: Editorial team. Manual process.

This consists of going through all of the incoming new publications which are currently scanned for IPNI, Tropicos and NYVH, and changing the database accordingly. This has not yet been defined well enough to document the expected procedures. Kew Record is a necessary reference here too as it may be the best way of picking up changes in synonomy rather than changes in nomenclature.

Issue: iPlants will need to determine the best workflow for this so as to streamline the procedure, and dovetail with existing efforts to maintain IPNI, Tropicos, NYVH and Kew Record.

7.2. Feedback

Input: feedback
Output: new entries to website,
Consulted: Taxonomic experts.
Who & how: Editorial team. Manual process.

iPlants will have procedures for receiving and responding to feedback, which may originate from the website, or from other sources. These have not yet been defined well enough to document the expected procedures.

Issue: A paper has been written on feedback (and also expert review) – see Group Documents / Website / ReviewFeedback.doc. Decisions will need to be taken on the issues raised.

7.3. Family review

Input: new publications
Output: new family version
Who & how: Compiler and taxonomic experts.

A review of a given family with the intention of checking its overall consistency may be initiated at any time. This may happen when there have been numerous minor changes and a new overview is thought advisable, when expert opinion becomes newly available, or if sufficient evidence (e.g. feedback, expert opinion) has accumulated to suggest a review is advisable, or significant new revisions are published.

Appendix: Glossary

Abbreviations database	A database of abbreviated terms compiled by Rafael Govaerts.
APNI	Component of IPNI; Australian Plant Name Index, compiled by the Australian National Herbarium, Canberra.
BPH and BPH/S	Botanico Periodicum Huntianum [Supplementum] – a source of standard abbreviated terms for botanical journals and periodicals publishjed by the Hunt Institute for Botanical Documentation 1968 and 1991.Print.
Brummitt 1992	Vascular Plant Families and Genera (Brummitt 1992). Online database and print.
Brummitt & Powell 1992	Authors of Plant Names (Brummitt & Powell 1992). Online databse and print.
Frodin 2001	Guide to Standard Floras of the World (Frodin 2001). Print.
Govaerts' general list	A short list of taxonomic works which should be consulted during compilation.
Govaerts' standardised database	A database of standardised plant names for genera A-I compiled by Rafael Govaerts at Kew.
Govaerts' working (unstandardised) database	A database of non-standardised plant names for genera A-I compiled by Rafael Govaerts at Kew from IPNI.
Gray Cards	Component of IPNI; List of New World Plant Names compiled by the Harvard University Herbaria.
Index Herbariorum	List of herbaria and botanic gardens with standard identifier codes. Online at The New York Botanical Garden's Virtual Herbarium.
Index Kewensis (IK)	Component of IPNI; global list of plant names compiled by Royal Botanic Gardens, Kew.
IPNI	International Plant Names Index. An internet accessible listing of all published plant names with their authors and place of publication. Additional nomenclatural information such as basionym, date of publication and type collections are supplied for some names where available.
K	Royal Botanic Gardens, Kew.
Kew Record	The Kew Record of Taxonomic Literature. Book and online database.
Mabberley 1997	The Plant Book (Mabberley 1997). Print.
MO	Missouri Botanical Gardens.
NY	The New York Botanical Garden
NYVH	Virtual Herbarium Database at NYBG. Online.
TL/2	Taxonomic Literature vol.2 and supplements (Stafleu and Cowan 1976 et seq.). A source for infromation about, and standard abbreviated terms for, taxonomic publications. Print and online.
Tropicos	Online Botanical Database of the Missouri Botanical Garden.