
*i*Plants: THE WORLD'S PLANTS ONLINE

PRELIMINARY CONSERVATION ASSESSMENT & GEOREFERENCING: TOOLS AND PROCEDURES

APPENDIX 4

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

Author(s): Moat, J

Location: iPlants document library: www.iplants.intranets.com

Directory: Group documents / Outputs – Final Report /

Filename: preliminary conservation assements & georeferencing

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

Date:

Revision History of this document

Date	Who	What	Version
5 th Dec 2004	BA	Created	1.1
7 th Dec 2004	BA	Reviewed / revised for release as appendix with Final Report for Pilot Project	1.2
13 th Dec 2004	BA	Outstanding issues reviewed & updated by Justin Moat.	1.3
23 rd Mar 2005	BA	Incl. feedback NYBG	1.4

Revision History of major components

This document was created initially by merging and rearranging the contents of three previous documents. The revision history of these earlier documents is given below

1) Preliminary Conservation Assessments: tools

Date	Who	What	Version
5 th Nov 2004	JM	Created	1.0
8 th Nov 2004	KH	Minor changes and notes. Addition of how we actually get a value from the Metrics (JM)	1.1
19 th Nov 2004	JM	Comments and additions from Steve Bachmann – Kew (SPB).	1.2

2) Georeference Guidelines

Date	Who	What	Version
3 rd Nov 2004	JM	Created	1.0
4 th Nov 2004	KH	Changes and comments added	1.1
9 th Nov 2004	JM.	Comments from JW and Mark Jackson	
19 th Nov 2004	JM	Fields added for Georeferencing checks	1.2
26 th Nov 2004	TC & EA	Comments added from Trish Consiglio (TC) - Missouri and Emily Ashley (EA)	1.3

3) Georeferencing Automated Checks

Date	Who	What	Version
4 th Nov 2004	JM	Created	1.0
5 th Nov 2004	KH	Minor checks and changes	1.1
9 th Nov 2004	JM	Comment from Mark Jackson on <i>Syagrus</i> data	
19 th Nov 2004	JM	Added in results Lecythidaceae and Madagascar Endemics database. Also Minor changes added	
19 th Nov 2004	KH	Minor checks and TDWG endemism added for <i>Syagrus</i>	1.2

TABLE OF CONTENTS

1. Introduction	6
1.1. The iPlants project	6
1.2. Further Information:	6
1.3. Purpose of this document.....	6
1.4. Outstanding Issues	6
2. Preamble	8
2.1. Procedures for preliminary conservation assessments	8
2.2. Guidelines for georeferencing.....	8
2.3. Automated checks.....	8
3. Preliminary Conservation Assessments	10
3.1. Prerequisites.....	10
3.2. Preliminary Conservation Assessments Described.....	10
3.2.1. Extent of occurrence:	10
3.2.2. Area of Occupancy:.....	10
3.2.3. Number of Sub-populations.	11
3.3. Preliminary conservation values.....	11
3.4. More comprehensive Conservation Assessments	11
3.5. Re-assessment (from the web)	12
3.6. Procedure for preliminary conservation assessments.....	12
3.7. Additional products	12
3.8. Linkages.....	13
4. Guidelines for Georeferencing.....	14
4.1. Prerequisites.....	14
4.2. Effort taken to Georeference (what to georeference).....	15
4.3. What not to georeference?.....	15
4.4. Error calculations	16
4.5. Problem Georeferences	16
5. Database fields for use when georeferencing.....	17
5.1. Tables of fields	17
5.1.1. Georeference fields	17
5.1.2. Further Fields (Not GIS, but useful).....	20
5.2. Details of Fields.....	21
6. Automated checks of georeference records	25
6.1. Prerequisites.....	25
6.2. Checks made	25
6.3. What next	25
7. Results of checks run during the iPlants Prototype	26

7.1.	Syagrus database	26
7.2.	Madagascar Palms.....	26
7.3.	Madagascar Endemics	27
7.4.	Lecythidaceae.....	27
Appendix:	References	28
Appendix:	Glossary	29

1. Introduction

1.1. The iPlants 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.

1.2. Further Information:

For further information please contact:

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

1.3. Purpose of this document

This document aims to

- 1) Outline the procedures to allow preliminary conservation assessment.
- 2) Outline guidelines for georeferencing of specimen data
 - a. Outline handling of accuracy/errors in georeferencing
 - b. List fields needed for Georeferencing and Conservation assessments within the DIGIR schema.
 - c. Provide details of fields for use in georeferencing.
- 3) Outline the automated georeferencing checks that will be performed on georeferenced specimen data.
 - a. Present results from the prototype georeferencing.

1.4. Outstanding Issues

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

1. Procedures for preliminary conservation assessments
 - a. What level are we selling for the online conservation checks? We need to define the purpose of these. We need to further assess the user requirement, this will give us detail on how to supply the conservation ratings
 - b. Re-assessments can be done, but, we could get into problems if we store the users lat-long data, also would we be happy to have the re-assessments in the checklist (only if verified). I think only comments can be sent back and attached to the checklist.
 - c. How will the conservation rating/specimen data be stored in the Checklist?
 - d. Where/what will the Conservation assessment be done in!!! (i.e. GIS, web/Java, Databases)? – this may depend on the user requirement, i.e. if users wish to do their own assessments, then it would be better to go the web way!
 - e. We need to check to see if IUCN have a rating for this species, I assume this is done in the Checklist?

- f. Do we display both best case and worst case scenarios or a range, probably worst (ie more endangered)?

2. Guidelines for georeferencing

- a. Need to formalize the cut off figures for Georeferencing effort.
- b. Each institution needs to agree on how to georeference and maybe some minor changes need to be made to their specimen databases.
- c. It would be preferable to split up Habitat/Ecology Into several fields, is this possible for each institution?
- d. Do we need to document the Gazetteer?

3. Automated Checks

2. Preamble

2.1. Procedures for preliminary conservation assessments

Under the Global Strategy for Plant Conservation, it was agreed that one of the primary targets is “A preliminary assessment of the conservation status of all known plant species, at national, regional and international levels”(Target (a)(ii)). This is to be achieved by 2010. The tools described here allow preliminary conservation assessments to be produced automatically for species that have been databased and georeferenced. Also described here is an outline of a more comprehensive assessment, which can be achieved, through the use of more comprehensive techniques.

2.2. Guidelines for georeferencing

Production of spatial data (location data, e.g. where a species occurs) is a fundamental part of databasing specimens. This data can range from place names in a specimen database, localities on labels, field notes, or the exact location of a plant using GPS (global positioning systems). To allow links to specimen databases whether this is just for dot maps or full-blown GIS analysis, the following standard fields will allow you to get the maximum from your data and ease the transfer of data to a GIS.

These are the fields that will describe the geographic (spatial) position of a collection. I have also included several fields that will help to identify where the locality is (georeferencing) and help in the checking of this position.

These fields follow closely the HISPID3 standard and MANIS, but incorporate other useful fields and the USBGN standards for locality types. A very detailed review of Georeferencing is in [Murphey et al 2004](#).

One of the main things we need to keep in mind is to keep a balance of throughput (how many) against quality (both number per species and georeferencing quality/accuracy).

2.3. Automated checks

Georeferenced specimen data will need to be checked to make sure that we do not have any incorrect localities, this process will be automated and results fed back to the specimen databases, for correction.

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

4. Procedures for preliminary conservation assessments

- a. What level are we selling for the online conservation checks? We need to define the purpose of these. We need to further assess the user requirement, this will give us detail on how to supply the conservation ratings
- b. Re-assessments can be done, but, we could get into problems if we store the users lat-long data, also would we be happy to have the re-assessments in the checklist (only if verified). I think only comments can be sent back and attached to the checklist.
- c. How will the conservation rating/specimen data be stored in the Checklist?
- d. Where/what will the Conservation assessment be done in (i.e. GIS, web/Java, Databases)? – this may depend on the user requirement, i.e. if users wish to do their own assessments, then it would be better to go the web way!

- e. We need to check to see if IUCN have a rating for this species, I assume this is done in the Checklist?
- f. Do we display both best case and worst case scenarios or a range, probably worst (i.e. more endangered)?

5. Guidelines for georeferencing

- a. Need to formalize the cut off figures for Georeferencing effort.
- b. Each institution needs to agree on how to georeference and maybe some minor changes need to be made to their specimen databases.
- c. It would be preferable to split up Habitat/Ecology Into several fields, is this possible for each institution?
- d. Do we need to document the Gazetteer?

6. Automated Checks

3. Preliminary Conservation Assessments

3.1. Prerequisites

- Specimen data has been georeferenced.
- Specimen data georeferencing has been cleaned.
- No duplicate specimen records.
- Only natural collections (no cultivated specimens).
- Checklist will store rating and specimen ID's.
- See Georeferencing Document and Georeferencing Checks document.

3.2. Preliminary Conservation Assessments Described

This initial process will be all automated. At present we have 3 algorithms that produce three measures; extent of occurrence, area of occupancy, number of sub-populations), we are in the process of adding others to this toolbox (i.e. extinction algorithms using collection date, protection and hopefully fragmentation). All of these are calculated from the specimens' locality and collection information (Habit, date, etc).

3.2.1. Extent of occurrence:

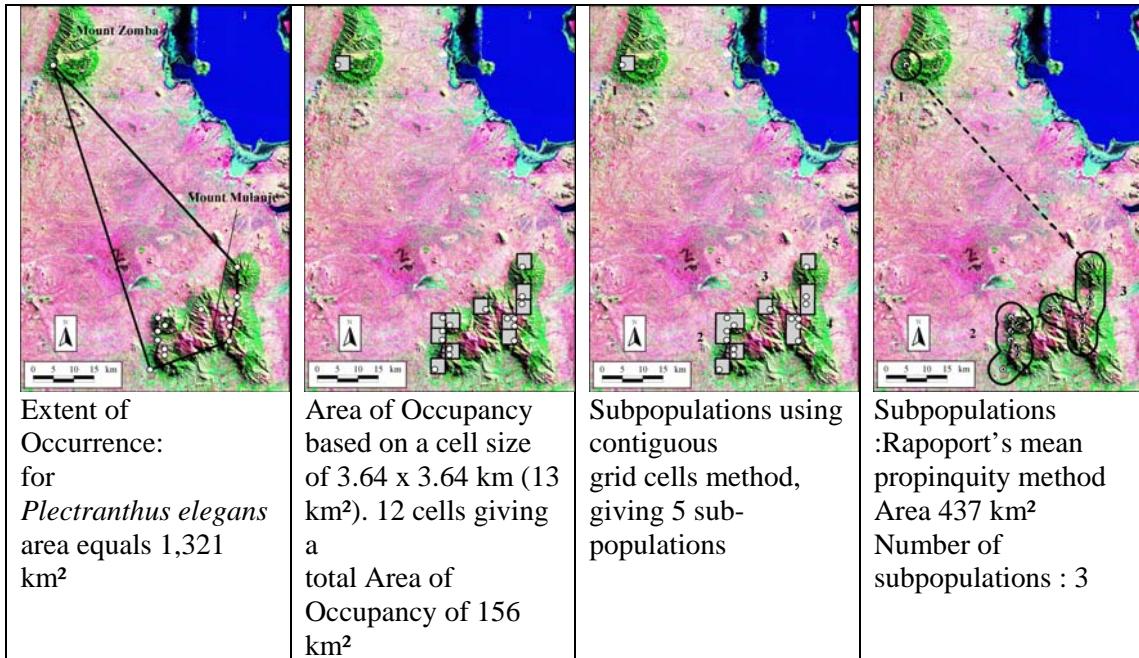
This is calculated using a convex hull; polygon of least tension enclosing all points. It is likely in the future that we will use an Alpha hull, which allows a more natural polygon/s. Both of these are well described in the literature (see Burgman & Fox 2003). See example below

3.2.2. Area of Occupancy:

This is calculated, by counting the number of cells a species occupies. The problem is deciding the cell size, at present we use 1/10 of the longest EOO side which give good results. See example below

3.2.3. Number of Sub-populations.

Two methods are used, one uses contiguous grid cells method (using the cells from above). The other uses Rapoport's mean propinquity method. See examples below.



3.3. Preliminary conservation values.

The above techniques and algorithms return metrics for a taxon. Each of these metrics can be used for a conservation rating (using the values as quoted by IUCN). For each taxon we should return the range of conservation ratings with an ‘average’. As rating values are incremental, but not actual values, we will have to employ some fuzzy logic to give an “average” preliminary conservation assessment.

3.4. More comprehensive Conservation Assessments

For some taxa that are threatened, then a more detailed conservation assessment will be performed, this will involve an expert using the tools above and other additional tools. Due to the large datasets involved, this analysis would be performed in a desktop GIS (i.e. ARCGIS). Full conservation ratings for very endangered taxa will have to be passed on to a GIS application to proceed, as each case will be different and will need different data to assess it (i.e. satellite imagery, deforestation history, elevation, climate etc). Examples below of indicators to review:



- Fragmentation (from Satellite imagery)
- Encroachment (from Satellite imagery)
- Loss of habitat
- Change of habitat through time (many images needed)
- Additional information (i.e. uses, habit, dispersal ability, climate, deforestation history, elevation etc)

3.5. Re-assessment (from the web)

It is possible to allow others to re-assess the conservation rating of a species, by allowing the addition of other specimen data or even the removal of localities. The simplest method would be to allow the user to upload a list of Latitude and Longitudes, which would add to the present set, then the georeference check and the preliminary conservation assessment would be run again. The user would then be presented with a simple interactive web GIS, allow them to review the data and select specimens to run assessment on. Expert users would be allowed to send comments on that species to be presented with the main checklist. This would also allow conservation organizations to see the impact of management decisions on species i.e. if species are removed from one region how will this affect the conservation status of this species?

3.6. Procedure for preliminary conservation assessments.

Following one taxon (process to be automated):



1. Specimen data received, georeferenced
2. Convert Degrees, minutes, seconds to decimal degrees. Convert error to km.
3. Do georeference checks (see georeference checks document), return any records that are incorrect to georeferencer.
 - a. Do nearest neighbour analysis (to check for outliers).
4. Do preliminary assessment measurements
 - a. Remove duplicates (localities)
 - b. Calculate Extent of occurrence (Jarvis March will do)
 - c. Calculate grid cell size (1/10 longest axis of above)
 - d. Calculate Grid cell area
 - e. Calculate cell contiguity (by Merging and the count polygons)
 - f. Get results from Rapoport analysis
5. Return results (areas, etc) to checklist – store with time, date stamped
6. Return range of preliminary conservation values and an average.
7. Return used Specimen data ID's to checklist – store with time date stamped

Full conservation ratings for very endangered taxa will have to be passed on to a GIS (see 4.3 above).

3.7. Additional products

- Documentation for the automated procedure will be made available from the website to allow others to follow suit
- The algorithms used (and others as we develop them), will be made available from the website. Probably as scripts for ArcGIS.
- It is possible to build an application with all of the above in it and allow users to freely download it (we would have to pull ESRI for their support and permission).

3.8. Linkages

What is connected to the Procedures described?

- Specimen databasing very closely connected
- Georeferencing linked
- Gazetteer
- IUCN (do they have a rating)

4. Guidelines for Georeferencing

4.1. Prerequisites

- **Specimen data entered and grouped.** Specimens are databased at each institution and parsed to some central area for georeferencing.
- **Specimens with determination not matching the checklist,** should be removed. i.e. hybrids etc.
- **Secondary specimens are identified.** That is collection of cultivated collection or similar.
- **Duplicates.** Duplicates are marked flagged from each institution, so we can only georeference one, but we do still need to see all records (all duplicates), as they can have different information on them (i.e. the first duplicate in one institute will sometimes have more detail than the subsequent ones). Duplicates are probably best identified by using collector, collection number and date. Notes on duplications:
 - It should be possible in most cases to spot duplicates from last name and collection number, the cases where this is not possible are when collector start number again (ie each year), or if collector has the same names.
 - Some collectors use the same number for a locality (not collection), for georeferencing this is an advantage, as long as the georeference is propagated back to each record.
 - A few collectors number each duplicate with a different number, this will cause problems, but I think we will have to just put up with this.
 - It is assumed that any georeferencing will be propagated from one specimen to its duplicates.
- **Gazetteer.** As the project is running a Gazetteer will be produced for iPlants. Initially this will contain: institutional Gazetteers and probably the GNS names. As the georeferencing progresses these will be updated to the gazetteer. This should all help to speed the process of Georeferencing. Within the Gazetteer will be a tool to help with further georeferencing (offset, conversions, etc).

4.2. Effort taken to Georeference (what to georeference).

Each specimen will require a different amount of effort to georeference. There comes a time when it may not be productive to georeference a collection as the time required will not be outweighed by any gain. Below are some notes on minimising effort.

1. Georeference in batches and by region if possible (so you can have maps, gazetteer, websites, handy).
2. Sort specimens for georeference by locality name then district name etc, this should make life easier, giving you localities that are close to each other at one time.
3. Georeference all those that are easy to do (i.e. ones that can be looked up in a Gazetteer electronic or paper). 
4. If a species has plenty of georeferenced collections (say over 20) then leave any more geo-referencing. 
5. If a species has few collection (say 7-20) then it may be possible to do a rough geo-reference (i.e. locate the region, but will have a large accuracy value). Please add a note in the ‘Geo House Keeping field’ to say something like “Could be geo-referenced more accurately” and in the ‘Georeferencing notes’ field “Only geo-reference to ????”
6. If the species have very few collections (7 or below) then it is probably worth the effort. Effort increases from Gazetteer –to maps and Atlas – to archives to checking field note books.
7. It maybe worth leaving all the very difficult-to-geo-reference specimens to do as a batch (especially if you have to go to the archives for a field collector notebook). Mark these left as “Return to”  the ‘Geo House Keeping field’.

NB. We need to take a sensible stance on this; these are not hard and fast rules, just some guidelines. So for example if you have 20 specimens that are all the same locality then it would be better to database more specimens from other localities i.e. do not stop at 10 in this case! 

4.3. What not to georeference?

- Cultivated specimens
- Specimens identified only to genus
- Specimens with determination not matching the checklist, i.e. hybrids etc.



4.4. Error calculations

With each locality georeferenced there is an associated error, we need to be pragmatic about recording this error (i.e. we do not wish to spend more time recording error, than finding the actual locality).

The field ‘Accuracy of Location’, is used to record the Maximum Error Distance as defined by MANIS (see <http://elib.cs.berkeley.edu/manis/GeorefGuide.html>). The MANIS guidelines are very very comprehensive and to use them could potentially be very time consuming. I would suggest that this field is used for records that have to be looked up from sources other than gazetteers or Label data. For example if you need to lookup a georeference from a paper map then the accuracy can be calculated from the scale of the map (see field 14 below, for more details and on handling other errors). The online iPlants Gazetteer should be able to handle all these conversion and calculated errors. To start with for the majority of Georeferencing we will be using GNS (<http://gnpswww.nima.mil/geonames/GNS/index.jsp>) and other Gazetteers, with these we can calculate the bulk of errors automatically. The important thing is to have the source and reference fields filled (errors can then be calculated retrospectively). Also see the Manis Georeferencing Calculator (<http://elib.cs.berkeley.edu/manis/gc.html>) and manual (<http://elib.cs.berkeley.edu/manis/CoordCalcManual.html>).



4.5. Problem Georeferences

In the field descriptions below some of the problem georeferences are described, but ones that can not be dealt with within a field are:

- Places with the same name? Guess at the best one and record (probably in the one of the fields), that you are unsure in the ‘Geo house Keeping field’. If the locality is incorrect then hopefully this should be picked up with the georeferencing check done before the conservation assessment. This should not be such a problem if you can see the localities on a map (as the iPlants Gazetteer should be able to do).

5. Database fields for use when georeferencing

5.1. Tables of fields

5.1.1. Georeference fields

The table below gives the details of the fields, I have given the fields a priority from 1 to 3, where 1 is a core field, 2 is very useful for GIS work, 3 is not essential but can be useful. For more details see the section for DB users.

Priority Redlist	Priority Georeferencing	ID	Field Long Name	Data Type	Notes	Other Notes/ Possibilities
1	1	1	Latitude Degrees	byte <= 90	If unknown then fields to be blank/Null (i.e. not to default to zero)	Possible to lookup these entries from a internal Gazetteer. i.e. from place name 2 nd subdivision etc.
1	1	2	Latitude Minutes	byte < 60	If unknown then fields to be blank/Null (i.e. not to default to zero)	
1	1	3	Latitude Seconds	Single < 60	If unknown then fields to be blank/Null (i.e. not to default to zero)	
1	1	4	Latitude Direction	1 character Text E or W	Orientation! Direction must be filled in if any of above Degree, Minutes, Seconds (DMS) entries are inputted	
1	1	5	Longitude Degrees	byte <= 180	If unknown then fields to be blank/Null (i.e. not to default to zero)	
1	1	6	Longitude Minutes	byte < 60	If unknown then fields to be blank/Null (i.e. not to default to zero)	
1	1	7	Longitude Seconds	Single < 60	If unknown then fields to be blank/Null (i.e. not to default to zero)	
1	1	8	Longitude Direction	1 character Text S or N	Orientation! Direction must be filled in if any of DMS entries are inputted	
2	1	9	Source of Lat/Long - (text) lookup	text	Lookup: <ul style="list-style-type: none"> • Label • Map estimate • GPS (before 2000) • GPS (after 2000) • Gazetteer • Internal gazetteer • Literature • Others. - to be filled in i.e. from Tropicos DB etc. 	
2	2	10	Datum	Text	Datum used for Georeference <ul style="list-style-type: none"> • Usually WGS84 	Usually only used for GPS readings
3	1	11	Geo House Keeping	Yes/No	Default to No. A Flag field for uncertain localities. Lookup <ul style="list-style-type: none"> • Coordinates Added • Could not find locality Not Georeferenced • Return to • Unsure • Could be geo-referenced more accurately. 	

3	1	12	Georeferencing notes	Text (255)	Any problems with geo-referencing. i.e. place with same name near etc. Part of above	
3	2	13	Reference for Source of Lat/Long	Text (255)	Only need if above (9) is not label. Self lookup	
2	2	14	Accuracy of location in km	Double	The precision of DMS.	N.B. HISPID3 is in Metres.
2	2	15	Locality Text	Text/Memo	Locality description (usually as written on label, field book etc)	
1	1	16	Country (ISO)/TDWG	Text	From iso/ and maybe TDWG list	Can lookup from each other. Only one of these needed
1	1	17	Country ISO	text	Can lookup from cou	
1	1	18	TDWG	text	Can lookup from iso	
2	2	19	1 st Subdivision of country	Text	Update - Self lookup	Possible to lookup from internal Gaz. USBGN!
2	2	20	2 nd Subdivision of country	Text	Update - Self lookup	
3	2	21	Locality Name	Text	Place name or nearest place name - possible self lookup	
3	3	22	Within Nearest or actual?	Yes/No	With above to identify if locality is at place or near? Defaults to no.	
3	2	23	Locality Place type.	Text	Lookup (most popular) <ul style="list-style-type: none"> • LCTY - Locality (Default) • PPL - Populated place • PPK - Park • STM River\stream • MT- Peak\Mountain The USBGN lookup	Other code to be added from full USBGN code when needed.
1	1	24	Secondary source	Text	Lookup. <ul style="list-style-type: none"> • No (default) • Cultivated • Non-native • Introduced • Naturalised 	Not actually a GIS field, but is very useful. Probably better put with the specimen details. In some ways we do not want any of these specimens (should they be geo-referenced?)
3	1	25	Geo-ref determined by	Text	The person and organization making the coordinate and error determination.	
3	1	26	Geo-ref Determined Date	Date Time	The date on which the georeference was made.	
3	1	27	Geo ref check error message	Text	Returns errors from the automated georeference checks	
3	1	28	Geo_ref_check	Text	3 entries: OK, Corrected, Error.	
1	1	29	Latitude Digital degrees	Double	Hidden - Updated on entry in locations	

1	1	30	Longitude Digital degrees	Double	Hidden - Updated on entry in locations	
1	1	31	Error diameter km	Double	Hidden - Updated on entry in locations. Fills in default values for accuracy if not entered.	

5.1.2. Further Fields (Not GIS, but useful)

I assume these will be handled else where. All of these can help with georeferencing, and preliminary conservation ratings.

Priority Redlist	Priority Georeferencing	ID	Field Long Name	Data Type	Notes	Other Notes/ Possibilities
1	1	32	Name	Text	Unique entry for species name (can be built from Genus species etc)	
3	1	33	Collector Name	Text	Used for georeferencing	
3	1	34	Collector Number	Text	Used for georeferencing	
1	1	35	Collection Date	Date	Used for Georeferencing and possible to use for Extinction algorithms	
2	3	36	Elevation /Ecology	Text	Would be better to have this against a standardised list (GLCC), but unlikely to be used often. describes the physical and environmental characteristics of the locality It is possible to split this up as in HISPID3,	It will be assumed that source of the elevation will be the same as the locality data (occasionally this will be different). I feel that HISPID3 makes this field too difficult to enter (i.e. having source and accuracy for the both min and max, which will only very occasionally be different. Again this field can be derived from GIS (If so this needs to be noted).
3	2	37	Elevation Min	Int	Metres If only one elevation quoted then this field entered.	
3	2	38	Elevation Max	Int	Metres	

5.2. Details of Fields

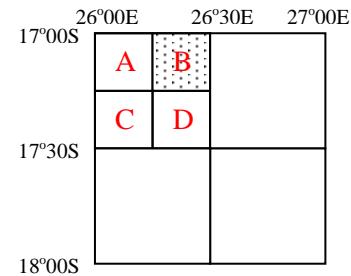
Latitude and Longitude. (1-8)

Purpose. Allows the collection locality to be georeferenced.

Details:

Problem cases: There are several usual types of georeferences that you may come across.

1. Locations described in degrees and digital minutes. Several GPS will default to degree and decimal minutes. Readings will look like $37^{\circ} 05.857'$ for example. The degrees and the minutes are correct, but to get the seconds you have to multiply the figures after the decimal by 60. So $37^{\circ} 05.857'$ is $37^{\circ} 05' 51.4''$. N.B. Sometimes occurs in TROPICOS DB. Online Gazetteer will have this conversion.
2. Locations described in decimal degrees. Again several GPS show this. The reading looks like 37.09761° for example. The degree is correct, but to get the minutes multiply all figures after the decimal minutes by 60 and round it down. Then take the remainder and multiply by 60 again to get the seconds. So
 $9761 \Rightarrow 0.9761 \times 60 = 5.8566$ (5 minutes), and then $0.8566 * 60 = 51.396$ (51.396 seconds) $\therefore 37^{\circ} 5' 51.396''$. Online Gazetteer will have this conversion.
3. Degrees of longitude over 180. Occasionally degrees of longitude are over 180 degrees, i.e. Hawaii can be described at 204° longitude East. To convert, just take the co-ordinate away from 360. i.e. $204 - 360 = 156^{\circ}$ and is West of 0° . Online Gazetteer will have this conversion.
4. Local co-ordinate systems. e.g references in meters, UK Ordnance Survey co-ordinates, etc. It is possible to convert these as long as the original co-ordinate system is known. If it is a one-off then it is probably easiest to check on a map, but for multiple entries contact Computing and GIS so that an automatic converter can be implemented. We will be implementing a converter for UTM on the online Gazetteer.
5. Quadrat co-ordinates. Occasionally used (Africa mainly), where the degree square is split into four (a,b,c,d), this can then be split into four again. Co-ordinates look like 1725Ab. Which has a centre of $17^{\circ}7'30''$ $25^{\circ}22'30''$. If possible convert to DMS using the centre point of the cell and record the source (9) as quadrat cell. Again if multiple entries contact Computing and GIS, so an automatic converter can be implemented. We will eventually provide a converter for this in the Online Gazetteer.

Source of Latitude and Longitude (9)

Purpose: Identifies where the Latitude and Longitude came from: label, GPS etc. This allows default errors to be applied and also allows the original source to be identified for any checks, corrections etc.

Details: If entry is not in the default list just type it in e.g.: *From other database, Inferred from GPS, Quadrat cell.*

Datum (10)

Purpose: To record any datum used for the Lat Long

Details: Only really useful for GPS readings or for very small scale mapping (towns), where the datum will be important. Most likely to use WGS84 (if not it would be better to convert to this if poss)


Geo House Keeping (11)

Purpose: To identify problem Geo-references. This is a drop down field, but the user can add other entries

Details: Use this field if you are unsure about the DMS typed in or if you wish to return to this record at a later date. Occasionally this may be used when a collection locality name occurs in several places in one country. e.g.: There are around 40 places called Ham in the UK, the most likely one has been chosen, but it could still be one of the others, so need more research or viewing within the GIS. Also label data may just give one name that is at several levels e.g. Rio do Contas and Morro do Chapeu are both municipio and cities. Also use this field if you can not find a locality as this stops anyone else from wasting effort to look for it again (unless they really want to??!)

Georeferencing notes (12)

Purpose: Notes for any problem georeferences

Details: Use with above. Use to show any corrections or assumption that you may have made. ie “maybe spell incorrectly Londen – London” or “probably in Devon not Cornwall”

Reference for source of lat/long: (13)

Purpose: Not needed for label data, adds details of the actual source of the lat/long.

Details: the reference source (e.g., the specific map, gazetteer, or software) used to determine the coordinates. Such information should provide enough detail so that anyone can locate the actual reference that was used (e.g., name, edition or version, year). e.g. *Madagascar map series 1:500,000 FTM 1926, reprinted, 1986*. Or for GPS: the make, model, etc. Also most of the entries will be the same, so you only have to add a few at the beginning of the database, then just use lookup after that. This field is not essential, but can be very useful.

Accuracy of Location (14)

Purpose: Gives some idea as to how accurate the lat/long is. Default values will be used from Source of Latitude and Longitude (9) if nothing is entered.

Details: Enter the Diameter (in km) of the circle of error. Not needed if georeference is from label data. If it is from a paper map and you are sure of the location, just use the following table (This is assuming that the map can not be read to better than 1 mm, if the map is more inaccurate then adjust appropriately).

1:5,000,000 1 mm = 5 km

1:1,500,000 1mm = 1.5 km

1:1,000,000 1mm = 1 km

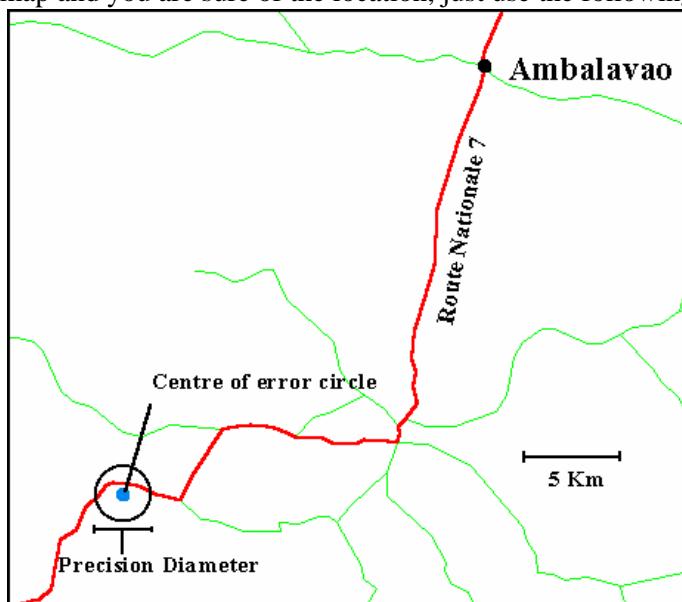
1:500,000 1mm = 0.5 km

1:250,000 1 mm = 0.25 km,

etc.

Other examples: a park name or mountain range, the diameter of the circle should encompass the whole park.

If looking up an uncertain co-ordinate on a map:



Precision	Label description
4 km	Fianarantsoa: 35 km South of Ambalavao on Route Nationale 7

Locality Text (15)

Details: Text as it is on the label, notebook etc. I think it should be as on label and any translations and inferences should be in square brackets [].

Country/ISO/TDWG (16 17 18)

Details: Only one of these needs to be entered. Not sure what to do about old country names, etc? Suggests using square brackets [] for these country/area names.

The three fields (19-21) below are not core, but are very useful for georeferencing the collections. Basically, if the data is easily at hand or known then is should be entered.

1st Subdivision of the country (19)

Purpose: Not core, but very useful for gazetteer work and for error checking.

Details: Enter the primary division each locality is within. First divisions may include: State, Province, County.

**2nd Subdivision of the country (20)**

Purpose: Again not core but very useful for gazetteer work and for error checking.

Details: If available enter the secondary division the locality is within. e.g. District, Municipio.

Locality Name (21)

Purpose: Again not core but very useful for gazetteer work and for error checking. Becomes very useful if the latitude and longitude are not known.

Details: Enter the nearest place name if available.

Within? (22)

Purpose: A flag value that is used as an Identify if the actual location is at/within the named locality. Only really needed for gazetteer work.

Details: Yes/No defaults to No. Only tick if the lat/long is the actual named locality.

Locality type: (23)

Purpose: Useful for gazetteer work

Details: Identifies the type of location. Enter the code. The most common entries are on the drop down list. Use lookup code from USBGN for more uncommon entries. Defaults to LCTY - locality.

Secondary Source (24)

Purpose: Not really a location field but is very useful.

Details: Defaults to No. Only really needs an entry if the collection is known to be cultivated, non-native, introduced or naturalised.

Geo-ref determined by (25)

Purpose: House Keeping, so reference can be traced back to owner

Details: User name and institute

Geo-ref Determined Date (26)

Purpose: House Keeping, Date/Time stamp the geo-reference

Details: Date and time.

Geo ref check error message (27)

Purpose: Returned by the georeference check

Details: Error message from check, i.e. not on land, Outlier 12 SD from rest of population

Geo ref check (28)

Purpose: Returned by the georeference check

Details: Either OK – no problems, Error - georeference check has identified a problem or corrected - User has corrected one of the problems. If a locality has been identify as error, but is correct (ie is out to sea, but the best localities is described as a bay, or a very small island), then the user can enter OK here (this will stop this being checked).

6. Automated checks of georeference records

6.1. Prerequisites

- Central or centrally accessible database of georeferenced specimens (See database integration document).
- See Georeferencing Guidelines document.

6.2. Checks made

Level	Check	Comment send back	Comments/Action
1	Not of this world (over 90 or 180 degrees)	Incorrect georeference over X degrees.	Error (check for typos or inversions).
2	Not on Land (i.e. out to sea)	Specimen is not on a land mass, x distance to coast	Distance to Coast (many be correct i.e. collection locality a bay).
3	Not in the country on the locality data	Incorrect country, X distance to country.	Distance to country boundary (need to check for historic boundary change, also localities close to boundary will probably correct)
4	In a region not on the checklist	Addition region to checklist.	Back to check list (flag specimen). Specimen maybe wrong or Checklist may need updating
5	Over 3+? Standard Deviations from the rest of the population	Point is X distance Y SD from it nearest neighbor.	Distance and SD flag specimen. Using closest neighbor, further analysis is need to determine cut off SD.

If an error is found at level X then checks below that are not needed for that specimen.

6.3. What next

The above check will be returned to the georeferencers, they will need to check all localities with a comment, if a locality is corrected or correct, then the checked georeference field needs to be checked!

It is possible that many of these checks will be part of the georeferencing process, so could be trapped immediately:

Level	Check	Comment
1	Not of this world (over 90 or 180 degrees)	Easy to check whilst georeferencing – could restrict database to only accept these values
2	Not on Land (i.e. out to sea)	Possible to check, but would need Spatial abilities of the DB
3	Not in the country on the locality data	Possible to check, but would need Spatial abilities of the DB
4	In a region not on the checklist	Possible to check, but would need Spatial abilities of the DB
5	Over 3+? Standard derivations from the rest of the population	Can only be checked when much specimen data is collected for that species.

7. Results of checks run during the iPlants Prototype

7.1. Syagrus database

N.B. Mobot data Decimal degrees came over with minutes divided by 100 i.e. 10 degrees 30 seconds came over as 10.005, this was corrected immediately.

Measure	Number	% of total
Total number of specimens	720	
Total Georeference	652	90.56%
Of above duplicated*	129	17.92%
Non endemics (i.e. over one TDWG level 3 region) **	490	68.06%
Cultivated	6	0.83%
Genus only or cf, af.	25	3.47%
Level 1 check	0	0.00%
Level 2 check***	13	1.81%
Level 3 check	5	0.69%
Level 4 check (not checked)	0	0.00%
Level 5 check	1	0.14%
Specimens used for Conservation analysis****	613	85.14%

* Quick estimate from collector and number

** These were included in analysis

** Many of these corrected (11 out of 13), due mainly to typos or Lat long inversion or N,S,E,W incorrect.

*** Any obvious error where corrected (i.e. lat long mixed or s/n incorrect), also level 5 were left in analysis as I can not be sure they are correct or incorrect. Duplicates were included in analysis (they are to some degree ignored by the GIS analysis).

7.2. Madagascar Palms

N.B. This database has had some prior cleaning! 5 hours of standardising names, duplicate checking etc. KH

Measure	Number	% of total
Total number of specimens	1410	
Total Georeference	1367	96.95%
Of above duplicated	327	23.19%
Non endemics (i.e. over one TDWG level 3 region), not used	41	2.91%
Cultivated	3	0.21%
Genus only or cf	4	0.28%
Level 1 check	0	0.00%
Level 2 check	4	0.28%
Level 3 check	3	0.21%
Level 4 check (not checked)	0	0.00%
Level 5 check*	2	0.14%
Specimens used for Conservation analysis	983	69.72%

* One specimen is probably cultivated *Dypsis lutescens* (Locke 66), the other *Bismarckia nobilis* has one collection very far south (J.Dransfield says that there is a problem with the taxonomy of this species).

7.3. Madagascar Endemics

Measure	Number	% of total
Total number of specimens	6606	
Total Georeference	6031	91.3%
Number of duplicates*	3219	48.7%
Non endemics (i.e. over one TDWG level 3 region), not used**		
Cultivated		
Genus only or cf	372	5.6%
Level 1 check	0	
Level 2 check	3	0.05%
Level 3 check	7	0.1%
Level 4 check (not checked)		
Level 5 check	3	0.05%
Specimens used for Conservation analysis***	6026	91.2%

* Estimate from collector, number and year. This seems very high, but most specimens seem to have been represented twice in this database!!!!

** Not used as we are looking at only Madagascar endemics.

*** Number is high as any obvious error where corrected (i.e. lat long mixed or s/n incorrect), also level 5 were left in analysis as I can not be sure they are correct. Duplicates were included in analysis (they are some degree ignored by the GIS analysis)

7.4. Lecythidaceae

Measure	Number	% of total
Total number of specimens	849	
Total Georeference	813	95.7%
Of above duplicated*	166	13.7%
Non endemics (i.e. over one TDWG level 3 region), not used		
Cultivated		
Genus only or cf		
Level 1 check **	41	4.8%
Level 2 check	2	0.2%
Level 3 check	4	0.5%
Level 4 check (not checked)		
Level 5 check***	10	1.2%
Specimens used for Conservation analysis****	797	93.9%

* Estimate from collector, number and year.

** Most of these were due to spurious entries or entries with degree only or incorrect orientation (S/N or E/W). Many corrected quickly

*** 2 of these were incorrect country as well, so removed from analysis.

**** Any obvious error where corrected (i.e. lat long mixed or s/n incorrect), also many from level 5 were left in analysis as I can not be sure they are correct or incorrect. Duplicates were included in analysis (they are to some degree ignored by the GIS analysis).

Appendix: References

Guralnick, Robert P. , Robert Glaubitz, David Neufeld and J. Allen Ryan 2004
Georeferencing of museum collections: A review of problems and automated tools, and the methodology developed by the Mountain and Plains Spatio-Temporal Database-Informatics Initiative (Mapstedi) PhyloInformatics 3: 1-29 see:
<http://phyloinformatics.org/abstract.php?id=3>

HISPID Version 3. Conn, B. J. (ed.) (1998). HISPID - Herbarium Information Standards and Protocols for Interchange of Data (National Herbarium of New South Wales: Sydney). -
<http://www.rbgsyd.gov.au/HISCOM/>

IUCN Guidelines for using the IUCN Red List Categories and Criteria, 2004:
<http://www.iucn.org/themes/ssc/redlists/RedListGuidelines.pdf>

IUCN Red List Categories and Criteria (Version 3.1) 2001
<http://www.iucn.org/themes/ssc/redlists/RLcats2001booklet.html>

MANIS Georeferencing guidelines <http://elib.cs.berkeley.edu/manis/GeorefGuide.html>

Manis Error calculator <http://elib.cs.berkeley.edu/manis/CoordCalcManual.html>

Murphrey, Paul C. Robert P. Guralnick, Robert Glaubitz, David Neufeld and J. Allen Ryan 2004 Georeferencing of museum collections: A review of problems and automated tools, and the methodology developed by the Mountain and Plains Spatio-Temporal Database-Informatics Initiative (Mapstedi) PhyloInformatics 3: 1-29 see:
<http://phyloinformatics.org/abstract.php?id=3>

RBG KEW, Gazetteer links page, GIS unit- <http://www.kew.org/gis/links/gaz.html>

USBGN - United States Board on Geographic Names -
<http://www.nima.mil/gns/html/BGN.html>

Willis, F., J. Moat, Paton, A. 2003. "Defining a role for herbarium data in Red List assessments: a case study of Plectranthus from eastern and southern tropical." Biodiversity and Conservation 12(7): 1537-1552

Appendix: Glossary

Compilation System	The software, people and procedures used to compile the iPlants online list of the plants of the world
Darwin Core	Darwin Core data structure (an agreed set of data elements for exchanging Natural History collections data)
DiGIR	Distributed Generic Information Retrieval project which has implemented an XML-based API to access specimen data based on the Darwin Core
DIVERSITAS	An international initiative aiming to promote integrative biodiversity science, linking biological, ecological and social disciplines in an effort to produce socially relevant new knowledge.
GBIF	Global Biodiversity Information Framework. Making the world's biodiversity data freely and universally available. GBIF works cooperatively with and in support of several other international organizations concerned with biodiversity.
GenBank	Online database of sequence data at the US National Center for Biotechnology Information
GSPC	The Global Strategy for Plant Conservation. Convention on Biological Diversity adopted the Global Strategy for Plant Conservation (decision VI/9), including 16 outcome-oriented global targets for 2010.
GTI	The Global Taxonomic Initiative. The GTI was established by the Conference of the Parties to the Convention on Biological Diversity to address the lack of taxonomic information and expertise available in many parts of the world, and thereby to improve decision-making in conservation, sustainable use and equitable sharing of the benefits derived from genetic resources.
IOPI	International Organization for Plant Information. Manages a series of cooperative international projects that aim to create and link databases of plant taxonomic information.
iPlants	The <i>iPlants initiative</i>
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.
IT IS	Integrated Taxonomic Information System. Designed to supply authoritative taxonomic information on plants, animals, fungi, and microbes of North America and the world.
IUCN	International Union for the Conservation of Nature
K	See Kew
Kew	The Royal Botanic Gardens, Kew, London, UK
LUCID	Knowledge management tool for diagnosing biological organisms

MANIS	Mammal Networked Information System A network of distributed databases of mammal specimen data.
MBG	The Missouri Botanical Garden, St. Louis, MO, USA
MO	See MBG
NatureServe	A US non government agency networking science to conservation
NY	See NYBG
NYBG	The New York Botanical Garden, New York, USA
NYVH	The New York Botanical Garden's Virtual Herbarium
RBG Kew	See Kew
Sp2000	The Species 2000 initiative Has the objective of enumerating all known species of plants, animals, fungi and microbes on Earth as the baseline dataset for studies of global biodiversity.
Tropicos	Online Botanical Database of the Missouri Botanical Garden
UNEP	United Nations Environment Programme.
WCMC	World Conservation Monitoring Centre (Cambridge)