Biodiversity Data Quality

An Overview

Eastern Bearded-dragon (Pogona barbata) - Toowoomba, Australia

Arthur D. Chapman

Australian Biodiversity Information Services

The data equation

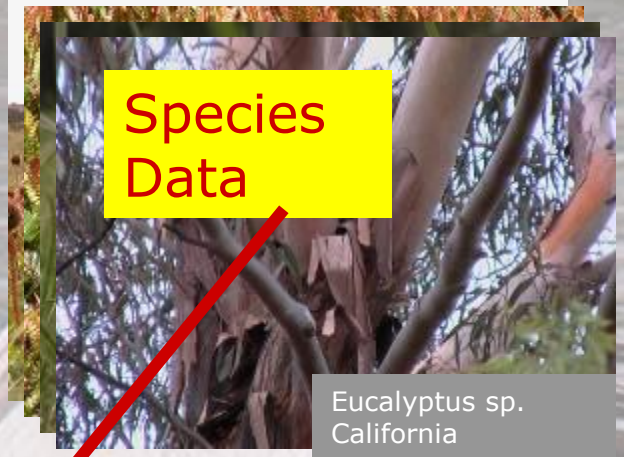
(Nix 1984)

Taking data to information

Species Data

Stick Insect
Campinas, Brazil

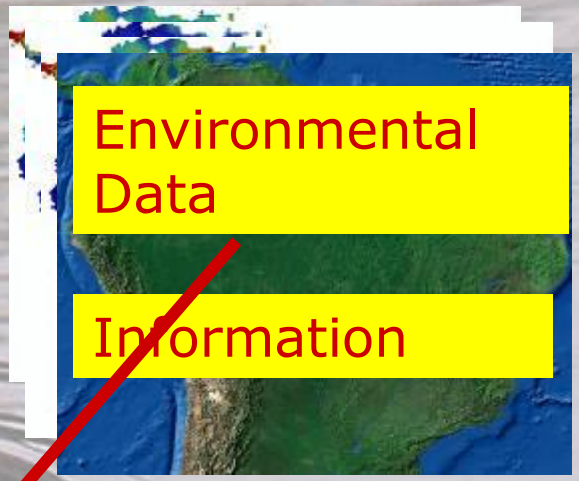
Species Data



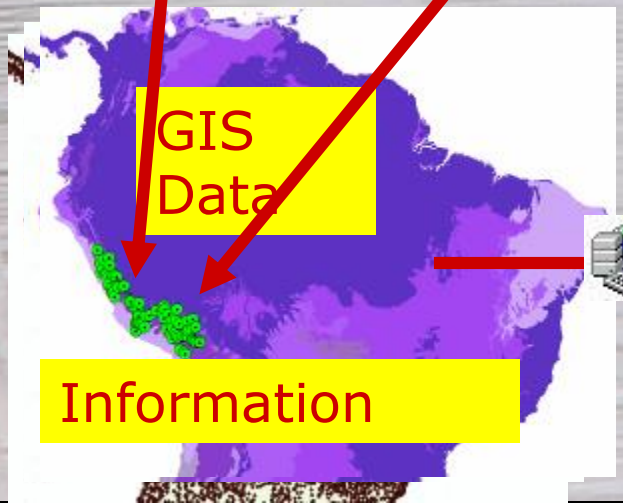
Eucalyptus sp.
California

Environmental Data

Information

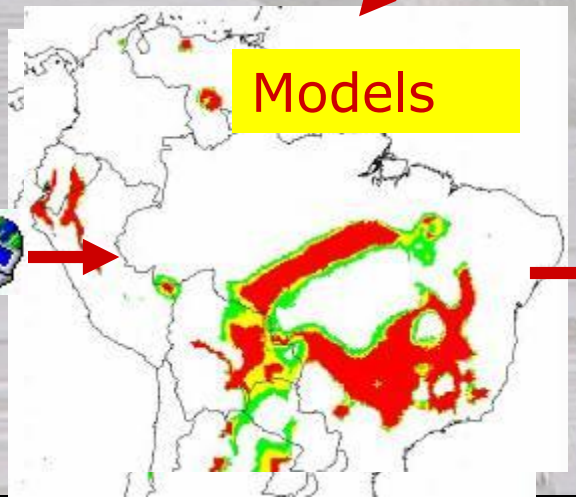


Decisions
Policy
Conservation
Management

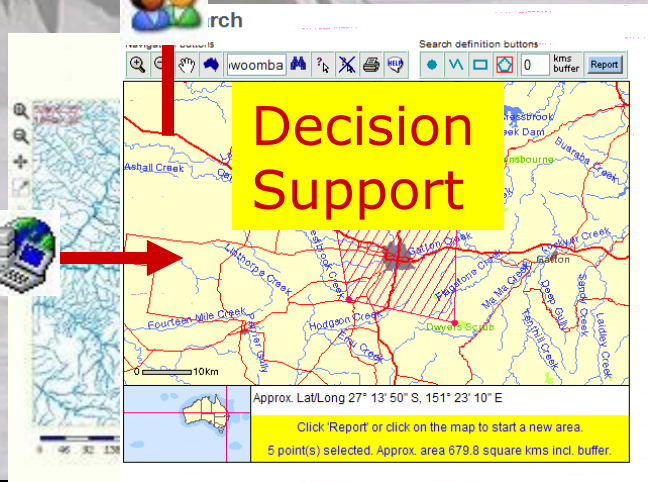


GIS Data

Information



Models



What do we mean by 'Data Quality'?

An essential or distinguishing characteristic necessary for [spatial] data to be fit for use.
SDTS 02/92

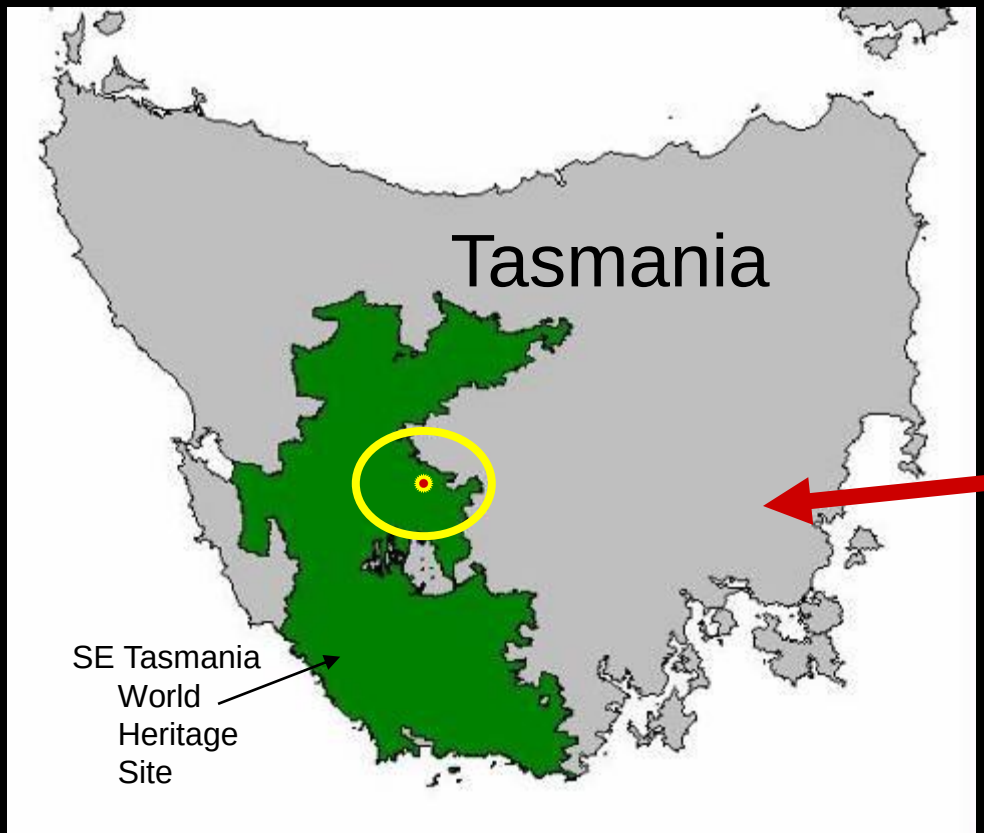
The general intent of describing the quality of a particular dataset or record is to describe the fitness of that dataset or record for a particular use that one may have in mind for the data. (Chrisman 1991)

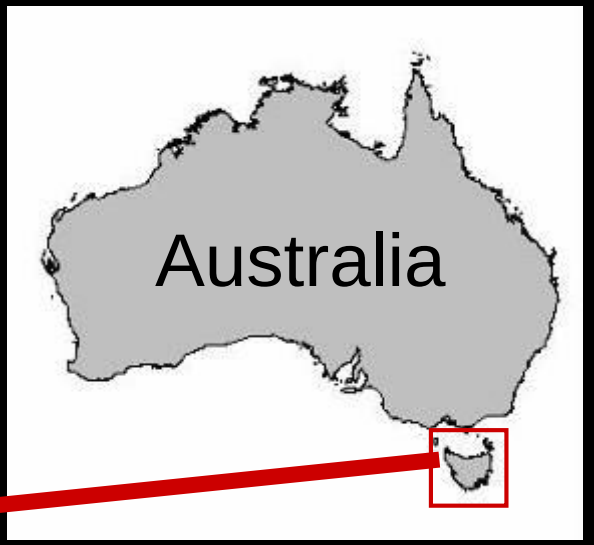
Data quality - fitness for use?

Fitness for use

Does species 'A' occur in Tasmania?

Does species 'A' occur in National Park 'y'





Data Quality

- Data Quality varies with the user
- Users don't require the same level of quality
- Having high quality data is often not as important to users as knowing what the quality is so they can decide whether to use it or not
- Users need to know the quality
- Comes down to documentation – i.e. the metadata of quality

Often not as important to improve the data quality as to assess its quality and to document that quality

Biodiversity data uses

Taxonomic Studies, Ecological Biogeography, Phylogenies

Biogeographic Studies, Species Modelling

Species Diversity and Population studies

Life Histories and Phenologies

Studies of Threatened and Migratory species

Climate Change Impacts

Ecology, Ecosystems, Evolution and Genetics

Environmental Regionalisations

Conservation Planning

Natural Resource Management

Using species data

Agriculture, Forestry, Fisheries and Mining

Health and Public Safety

Bioprospecting

Forensics

Border Control and Wildlife Trade

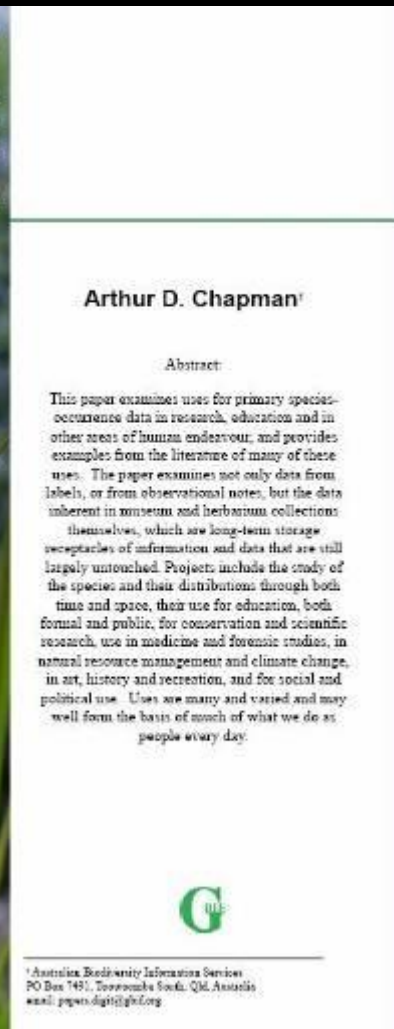
Education and Public Outreach

Ecotourism

Art and History, Science and Politics

Recreation

Human Infrastructure Planning



Quality versus Error

- Poor/low quality does not always equate to error
- Quality can't always be improved
- Data itself may be of high quality, but your view of it may not be.

Loss of Data Quality

- Errors
  - Mistakes, misinterpretations etc.
  - Data in wrong fields
  - Missing data
- Uncertainties
  - Locations/taxonomy
- Degrees of resolution
  - Rounding
    - During collecting (e.g. grids)
    - Added later (e.g. conversions)
  - Deliberate fuzzying
    - Sensitive taxa/locations
    - political, commercial, legal

Errors in data

In general, error must not be treated as a potentially embarrassing inconvenience, because error provides a critical component in judging fitness for use.

Chrisman, 1991

Although most data gathering disciplines treat error as an embarrassing issue to be expunged, the error inherent in (spatial) data deserves closer attention and public understanding.

Chrisman, 1991

Errors in data

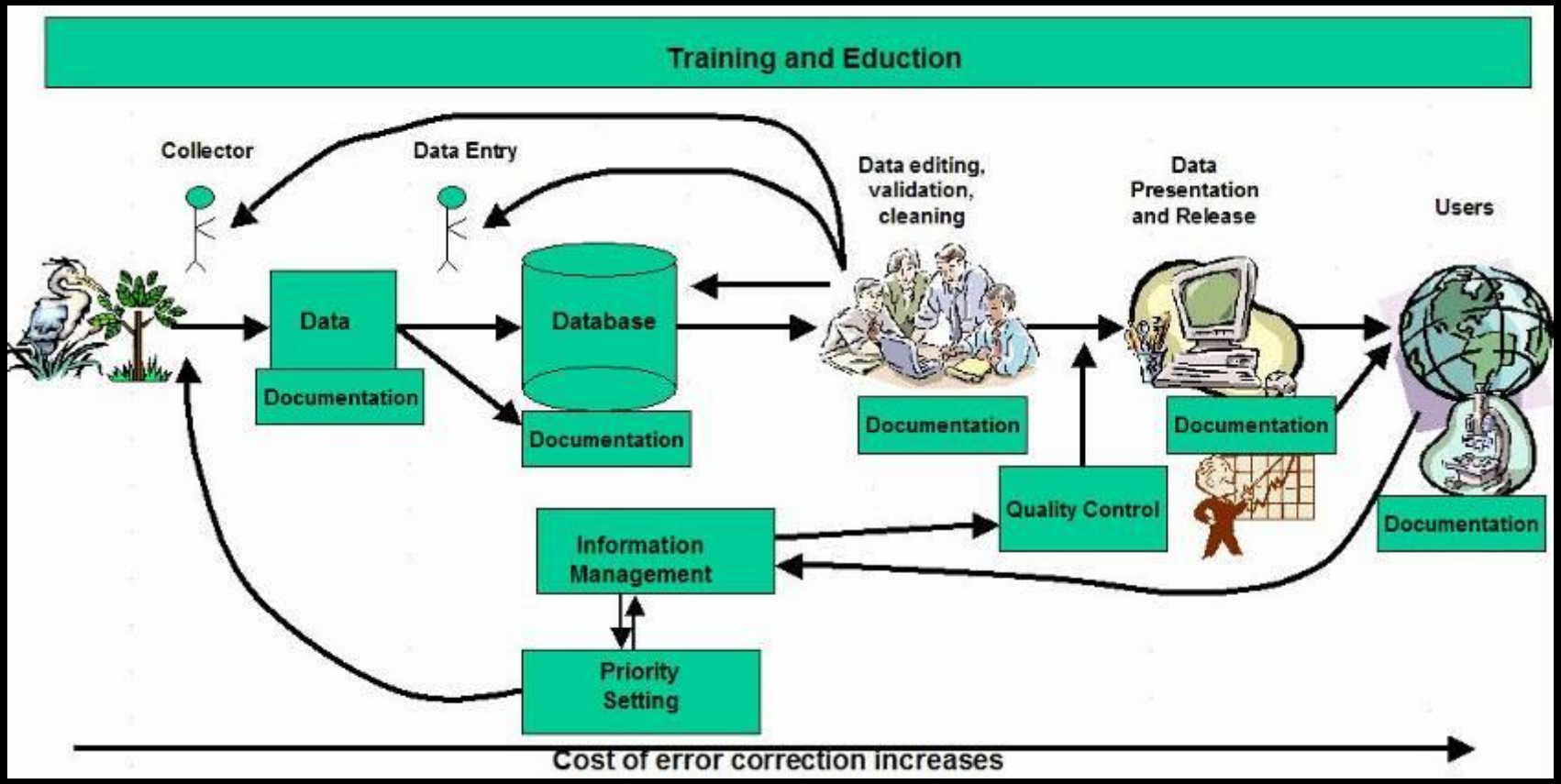
In general, uncertainty must not be treated as a potentially embarrassing inconvenience, because uncertainty provides a critical component in judging fitness for use.

Chapman, 2016

Although most data gathering disciplines treat uncertainty as an embarrassing issue to be expunged, the uncertainty inherent in (spatial) data deserves closer attention and public understanding.

Chapman, 2016

Data Quality Information Chain



Assign responsibility for the quality of data to those who create them. If this is not possible, assign responsibility as close to data creation as possible (Redman 2001)

Key Quality Fields

Two key areas of quality are:

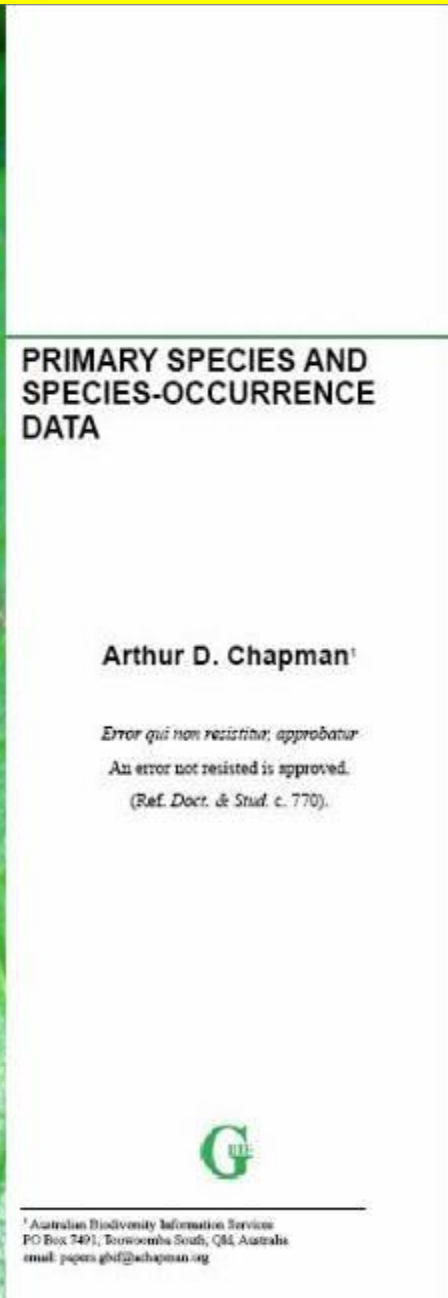
- Taxonomic names
- Georeferences (lat's and long's)

Methods for identifying error

Documented here ----------->

available via GBIF web site

http://www.gbif.org

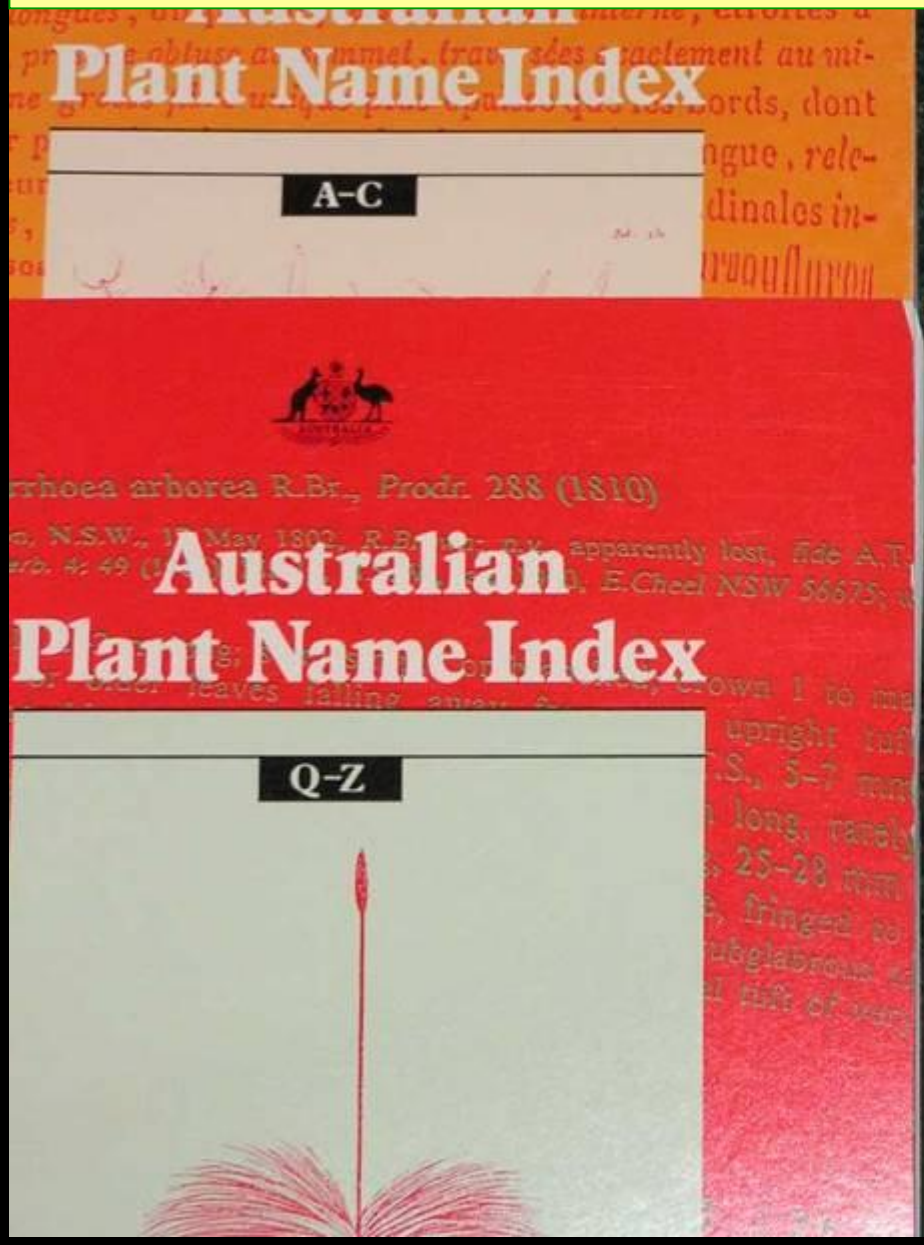


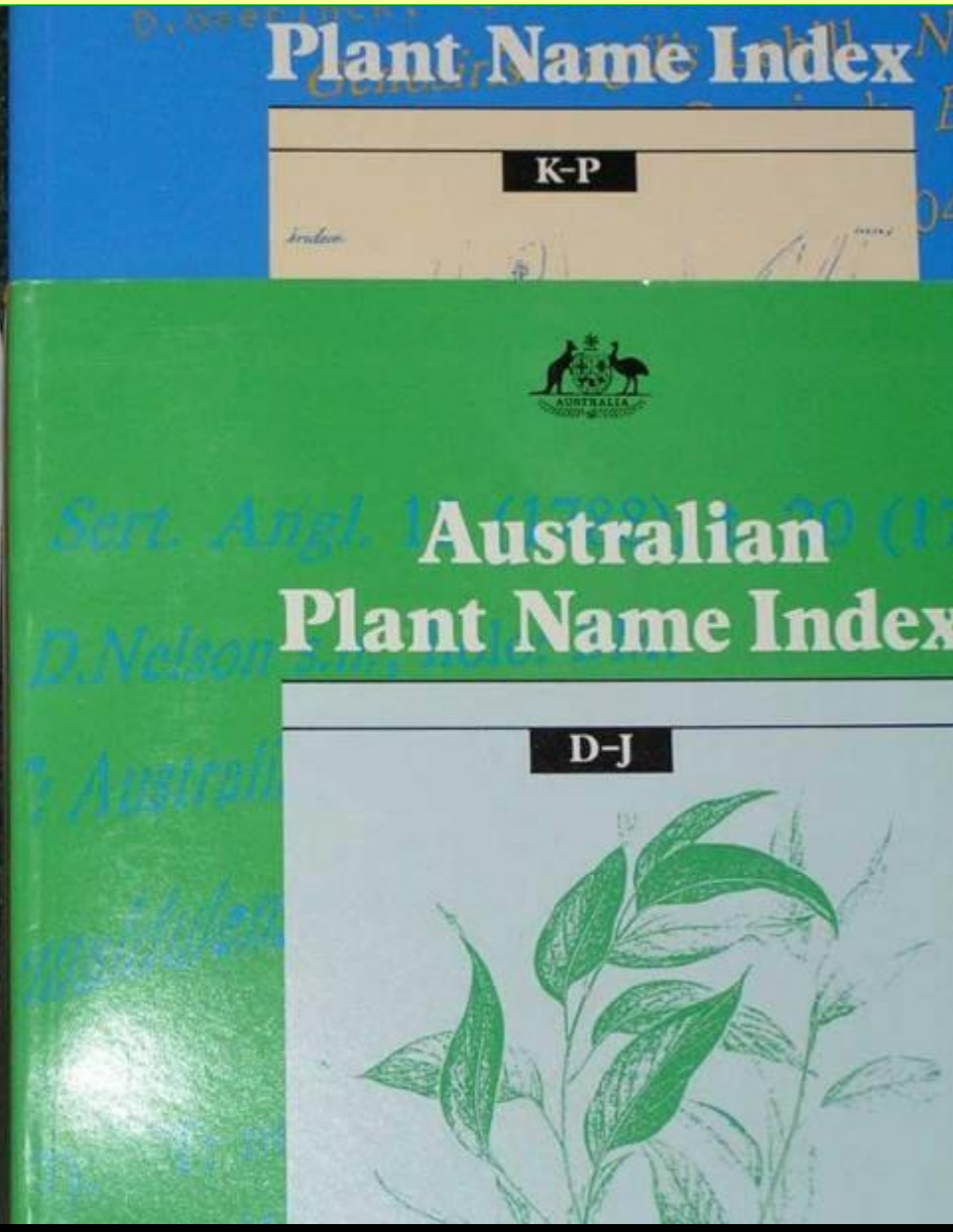
Data Cleaning

- Individual Museums
- Aggregators (such as ALA, GBIF, etc.)
- Citizen Science Projects
  - iSpot
  - iNaturalist
  - Expert assistance
  - Vetting (e.g. Birds Australia)
  - On-line collaboration (Flickr)
- Feedback from users

Breutelia affinis (Werribee Gorge, Australia)

Taxonomic and Nomenclature Data





Taxonomic Data

Consists of: (not all are always present):

- Name (scientific, common, hierarchy, rank)
- Nomenclatural status (synonym, accepted, typification)
- Reference (author, place and date of publication)
- Determination (by whom and when the record was identified)
- Type specimen citation
- Quality fields (accuracy of determination, qualifiers)

Determining Quality

- Not always easy
- Seldom carried out
- Use of Determinavit slips
- Qualifiers (aff., cf., s.str., s.lat., ?)
- Documentation?

Documenting Taxonomic Data Quality

- Several methods exist for documenting taxonomic verification - none are completely satisfactory
  - Herbarium Information Standards and Protocols for the Interchange of Data (HISPID)
  - Australian National Fish Collection (1993)
  - Several others restricted to one or two institutions
- Proposal – four level:
  - Who determined the specimen and when
  - What was used (type specimen, local flora, monograph, etc.)
  - Level of expertise of the determiner
  - What confidence did the determiner have in the determination.

Documenting Quality - 2

From: Herbarium Information Standards and Protocols for the Interchange of Data (HISPID)

0 The name of the record has not been checked by any authority

1 The name of the record determined by comparison with other named plants/animals

2 The name of the record determined by a taxonomist or by other competent persons using collections and/or library and/or documented living material

3 The name of the plant determined by taxonomist engaged in systematic revision of the group

4 The record is part of the type gathering

Documenting Quality - 3

From: Australian National Fish Collection (in use since 1993)

Level 1: ***Highly reliable identification***

Specimen identified by (**a**) an internationally recognised authority of the group, or (**b**) a specialist that is presently studying or has reviewed the group in the Australian region.

Level 2: ***Identification made with high degree of confidence at all levels***

Specimen identified by a trained identifier who had prior knowledge of the group in the Australian region or used available literature to identify the specimen.

Level 3: ***Identification made with high confidence to genus but less so to species***

Specimen identified by (**a**) a trained identifier who was confident of its generic placement but did not substantiate their species identification using the literature, or (**b**) a trained identifier who used the literature but still could not make a positive identification to species, or (**c**) an untrained identifier who used most of the available literature to make the identification.

Level 4: ***Identification made with limited confidence***

Specimen identified by (**a**) a trained identifier who was confident of its family placement but unsure of generic or species identifications (no literature used apart from illustrations), or (**b**) an untrained identifier who had/used limited literature to make the identification.

Level 5: ***Identification superficial***

Specimen identified by (**a**) a trained identifier who is uncertain of the family placement of the species (cataloguing identification only), (**b**) an untrained identifier using, at best, figures in a guide, or (**c**) where the status & expertise of the identifier is unknown.

Taxon Verification Status

Name of determinor:

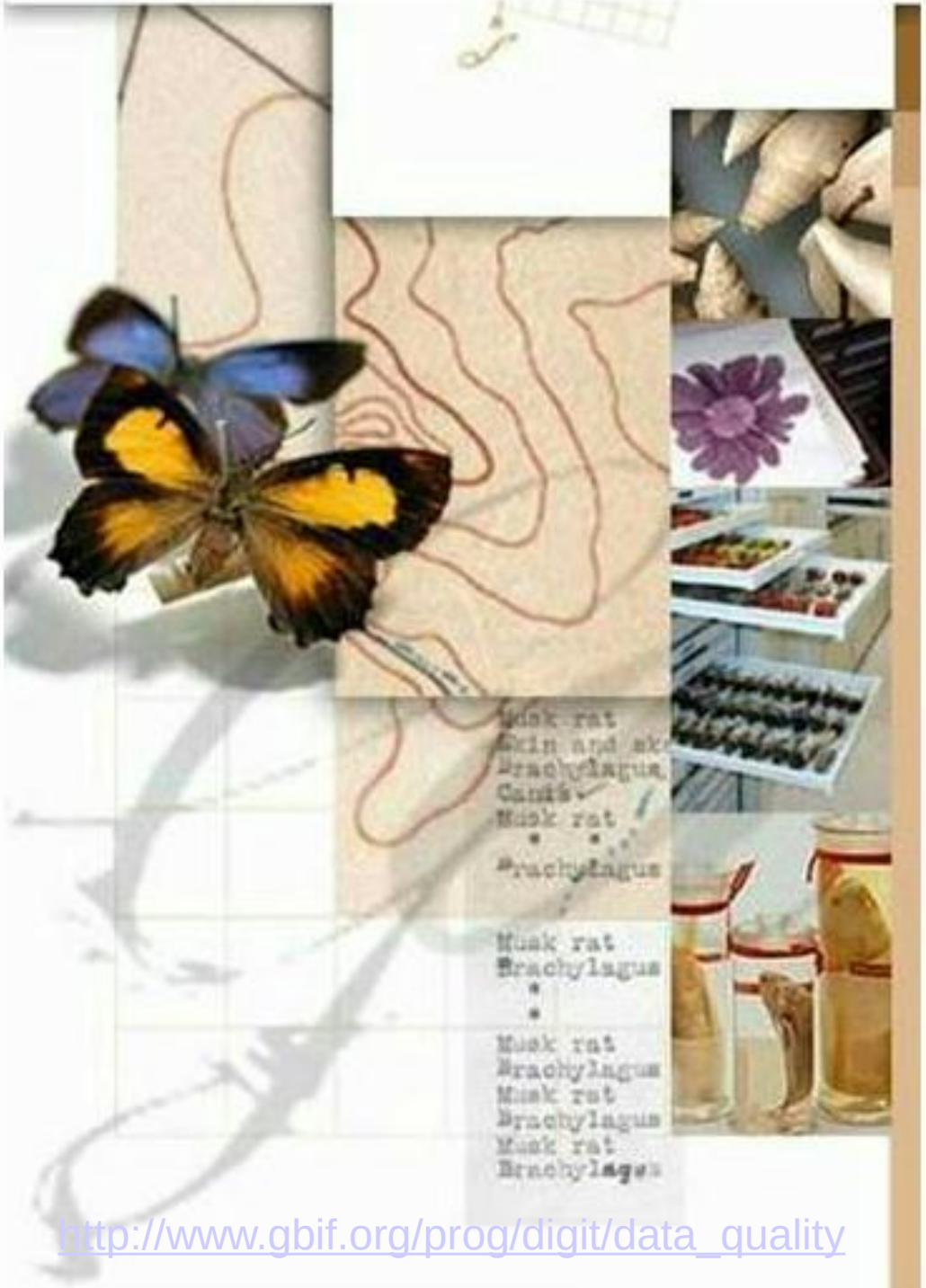
Date of determination:

Source of determination: (e.g. compared with holotype, used national flora)

- identified by **World expert** in the taxon with **high certainty**
- identified by **World expert** in the taxon with **reasonable certainty**
- identified by **World expert** in the taxon with **some doubt**
- identified by **regional expert** in the taxon with **high certainty**
- identified by **regional expert** in the taxon with **reasonable certainty**
- identified by **regional expert** in the taxon with **some doubt**
- identified by **non-expert** in the taxa taxon **high certainty**
- identified by **non-expert** in the taxa taxon **reasonable certainty**
- identified by **non-expert** in the taxa taxon **some doubt**
- identified by **the collector** with **high certainty**
- identified by **the collector** with **reasonable certainty**
- identified by **the collector** with some doubt.

From: Chapman (2005) Principles of Data Quality. GBIF

Spatial Data Cleaning
Species Occurrence Data



BioGeomancer

Guide to Best Practices for Georeferencing

http://www.gbif.org/prog/digit/data_quality

Chapman & Wieczorek (eds) (2006)

Georeferencing Guidelines

Georeferencing Collections and Recording Uncertainty

The document provides guidelines to World's Best Practice for georeferencing, including guidance on

- determining a georeference
- determining the spatial uncertainty
- recording the georeferences and uncertainties

Database Fields

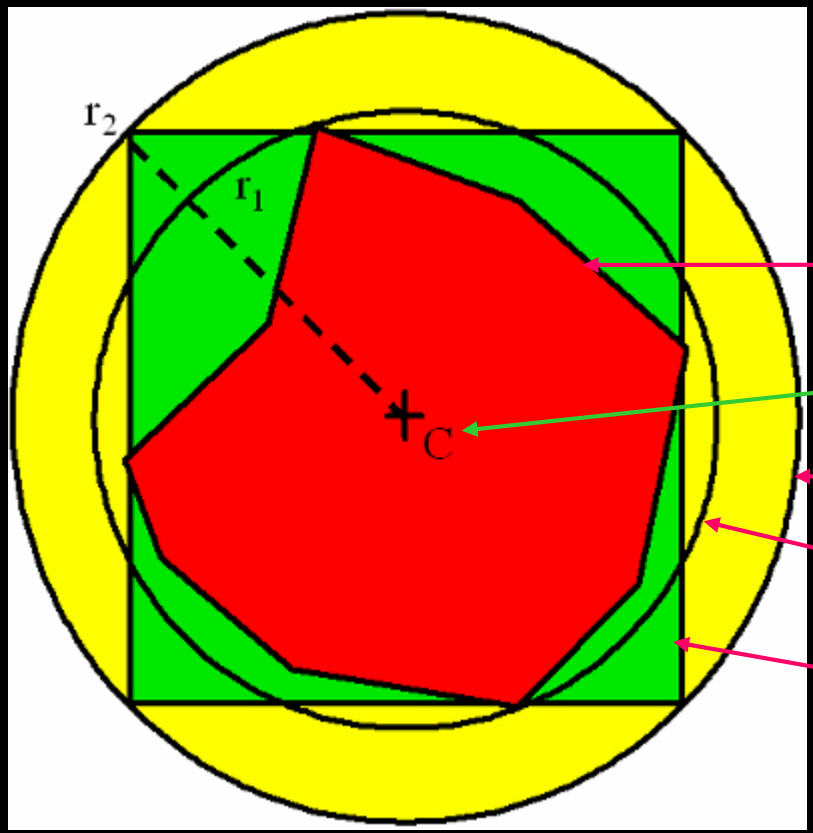
See: **Geospatial Element Definitions v1.4 (extension to Darwin Core)**

Decimal Latitude
Decimal Longitude
Geodetic Datum
Maximum Uncertainty Estimate
Maximum Uncertainty Unit
Verbatim Coordinates
Verbatim Coordinate System
Georeference Source (e.g. USGS Gosford Quad map 1:24000, 1973)
Verification Status (e.g.: "requires verification", "verified by collector")
Validation Status
Georeference Determined by
Georeference Determined date
Remarks

[Spatial Fit]

Spatial Fit

A measure of how well the geometric representation matches the original spatial representation.

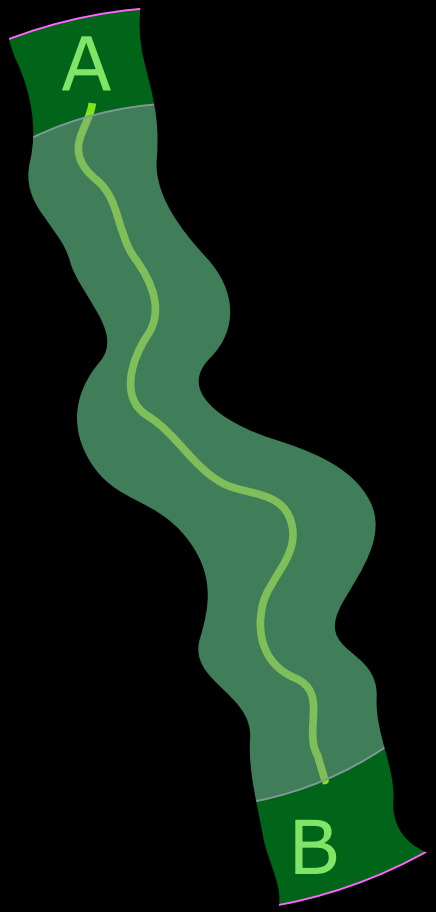


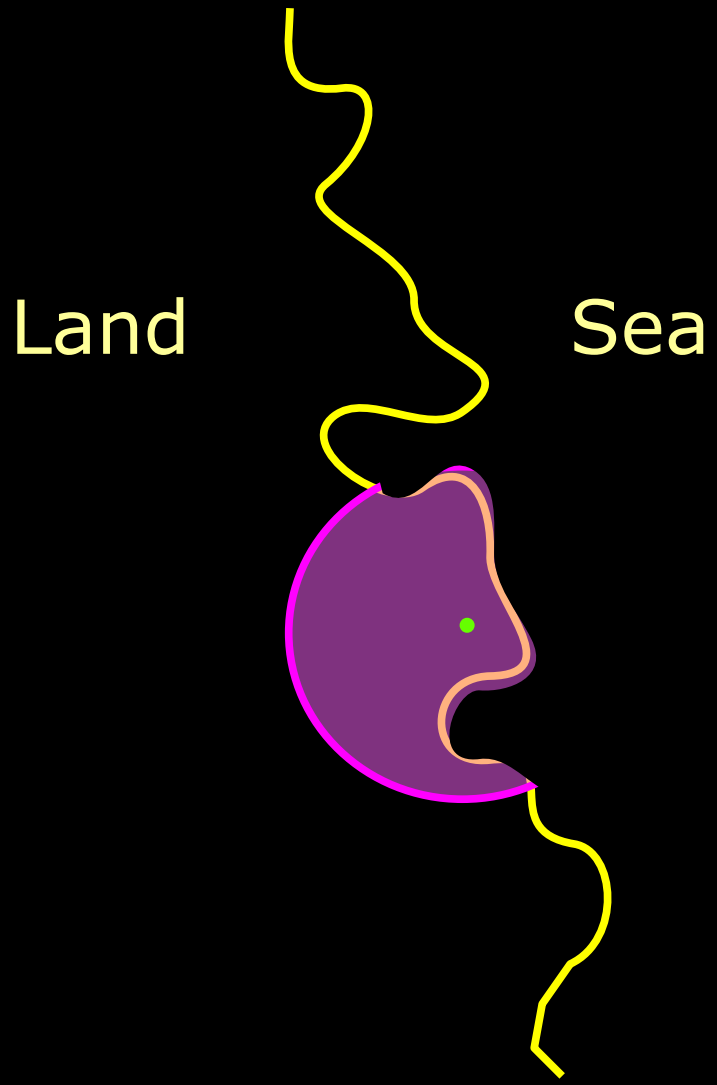
For an area where the original spatial representation of a locality is the **red polygon** with area 'A'. The spatial fit is:

- 1.0
- 0
- $(Pi*r_2^2)/A$
- $(Pi*r_1^2)/A$
- $(2*r_2^2)/A$

From J. Wieczorek, in Chapman and Wieczorek (eds) (2006)

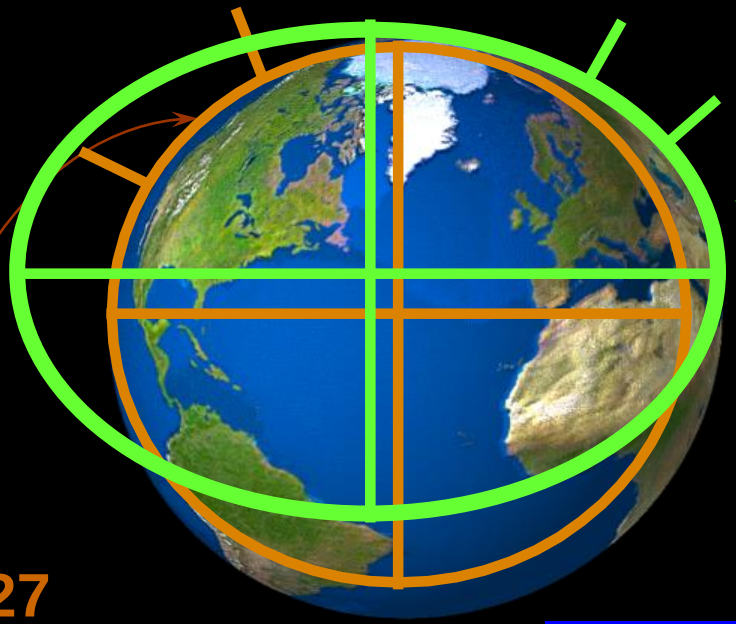
Spatial Fit and Uncertainty





Traditional Horizontal Datums

Different datums can mean a difference in location of from a few cms to 3.552 km.

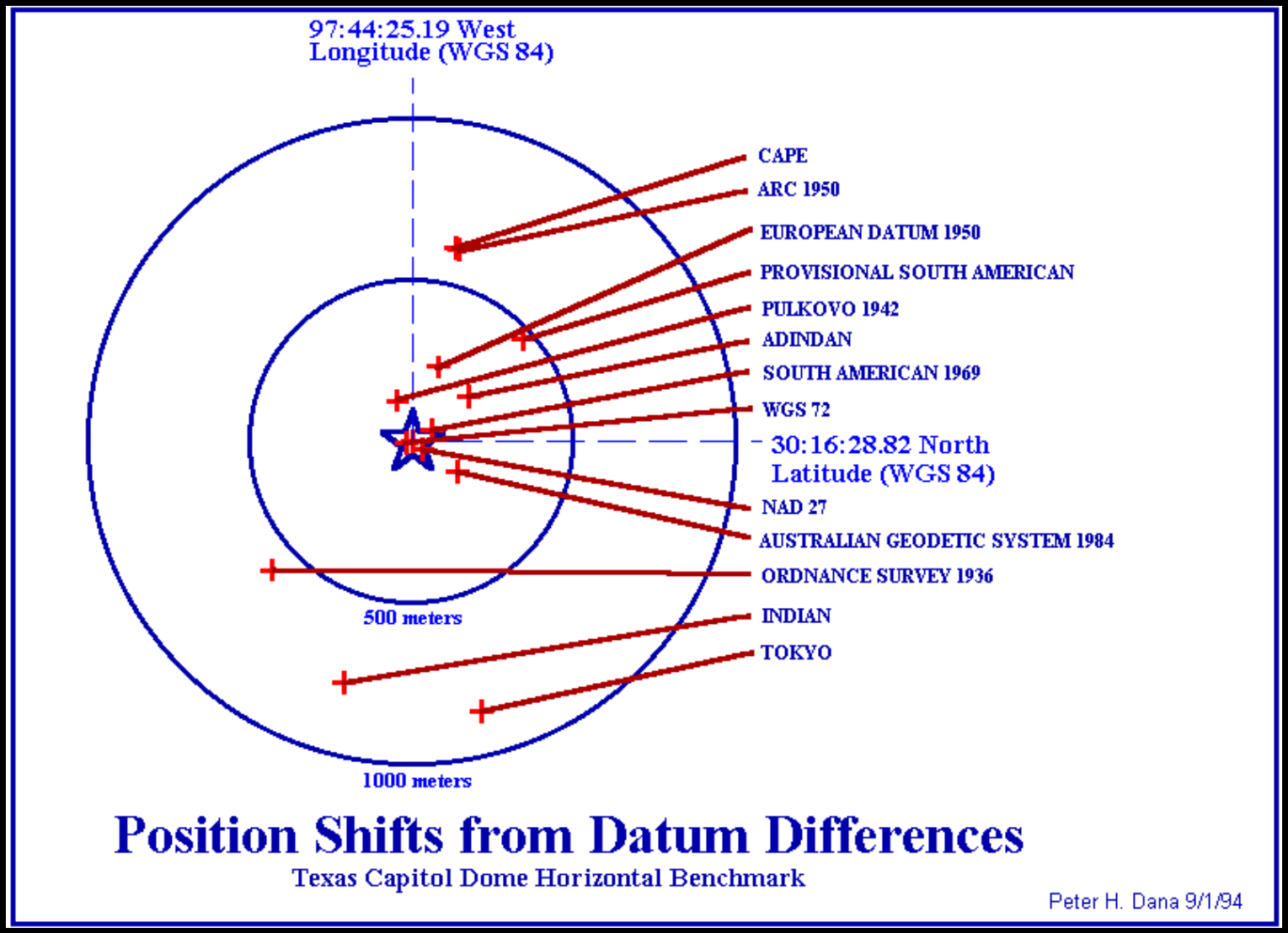


NAD 27
(Clarke Ellipsoid)

ED 50
(International Ellipsoid)

From US Navy (n.dat.)

Horizontal Datum Shifts



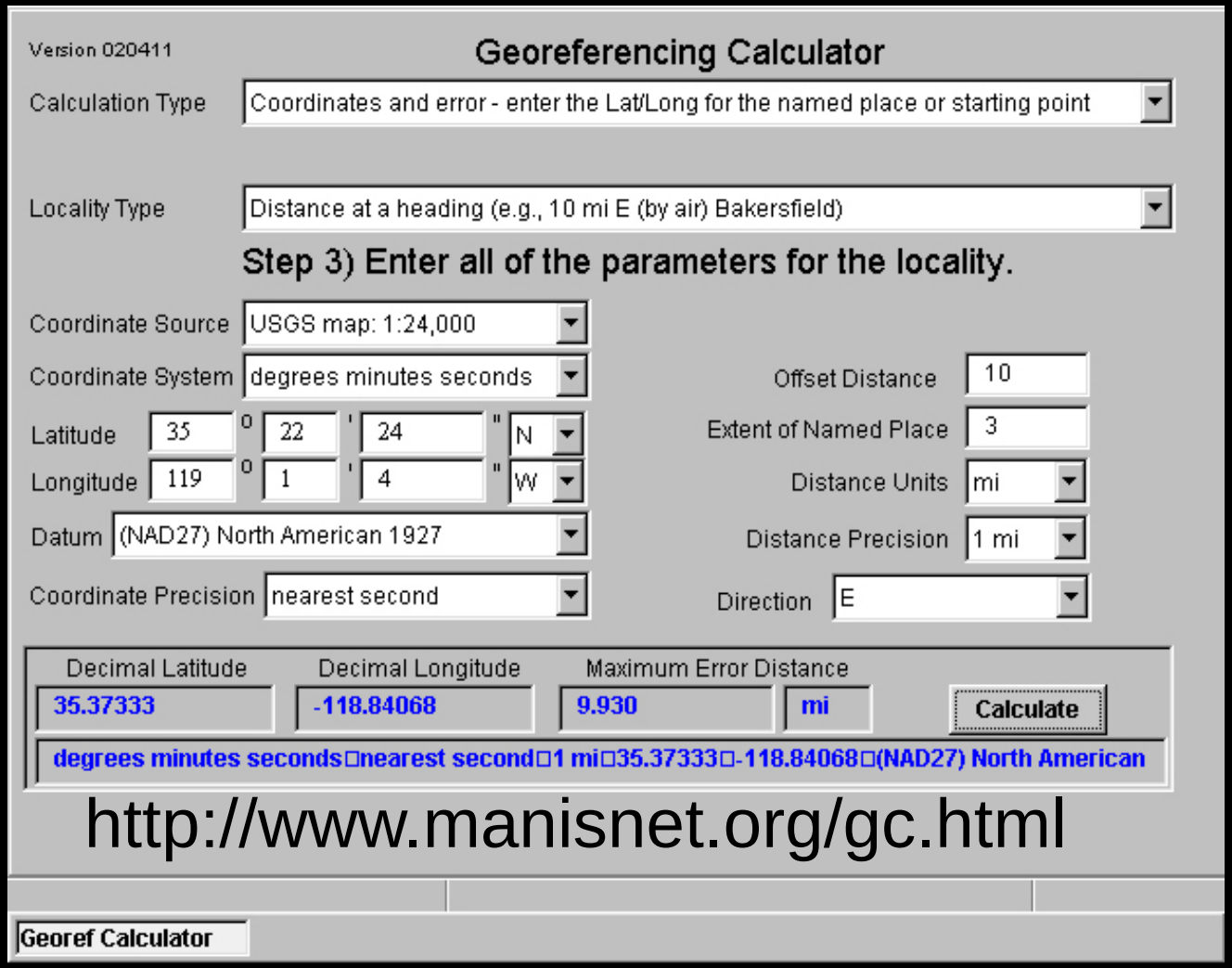
Vertical Datums

Like horizontal measurements, elevation only has meaning when referenced to some start point.

From US Navy (n.dat.)

Mean sea level is the most common vertical datum.

MaNiS Georeferencing Calculator



Methods for Validating Georeferences

- Internal Database Checks
  - Logical inconsistencies within the database
  - Checking one field against another
    - Text location vs geocode or District/State
- External Database Checks
  - Checking one database against another
    - Gazetteers
    - DEM
    - Collectors
- Outliers in Geographic Space - GIS
- Outliers in Environmental Space - Models
- Statistical outliers

Error

Error is inescapable and it should be recognised as a fundamental dimension of data.

Chrisman 1991

Bolax gummifera, Argentina

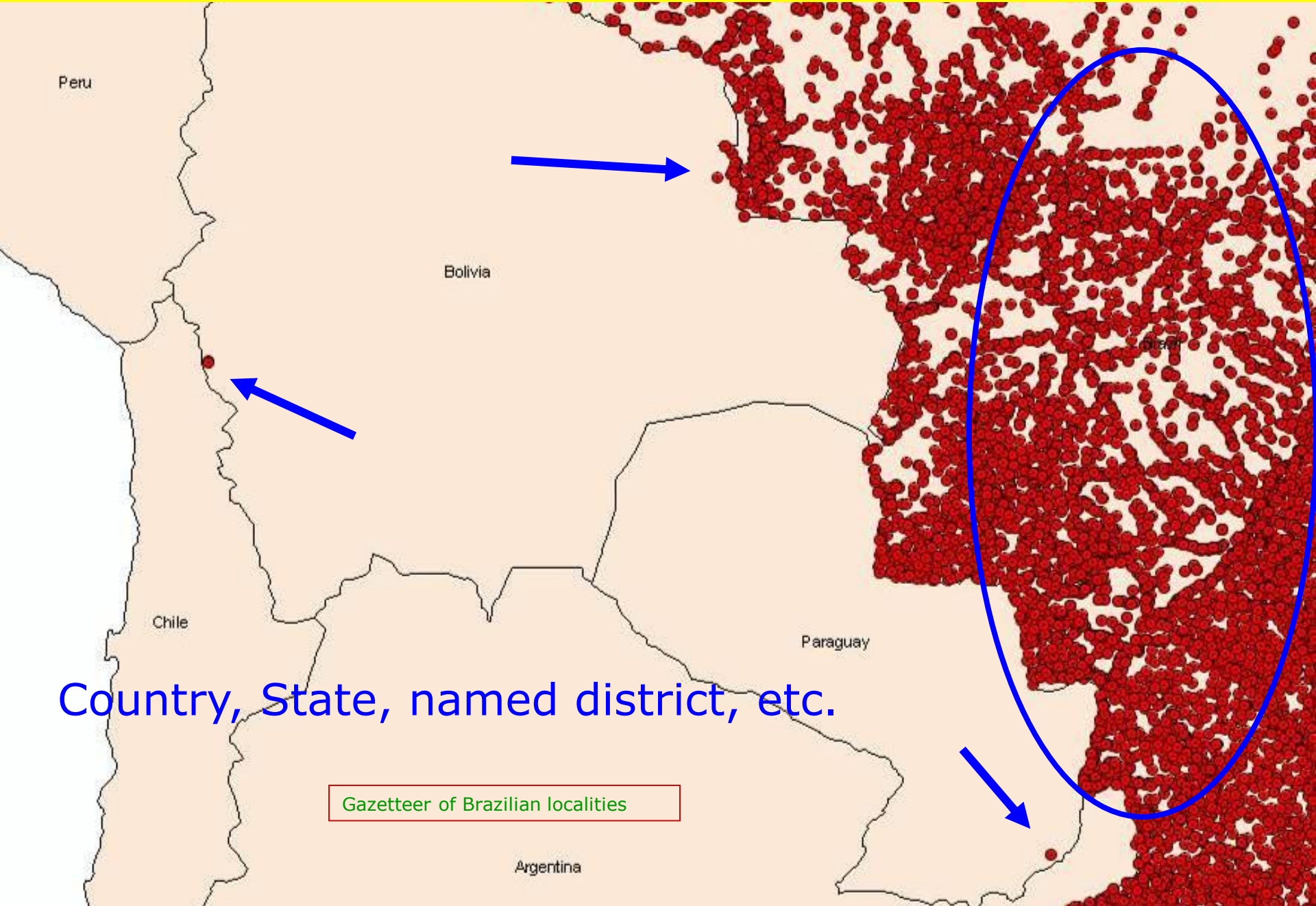
Uncertainty

Uncertainty is inescapable and it should be recognised as a fundamental dimension of data.

Chapman, 2016

Bolax gummifera, Argentina

Geographic outliers - GIS



Country, State, named district, etc.

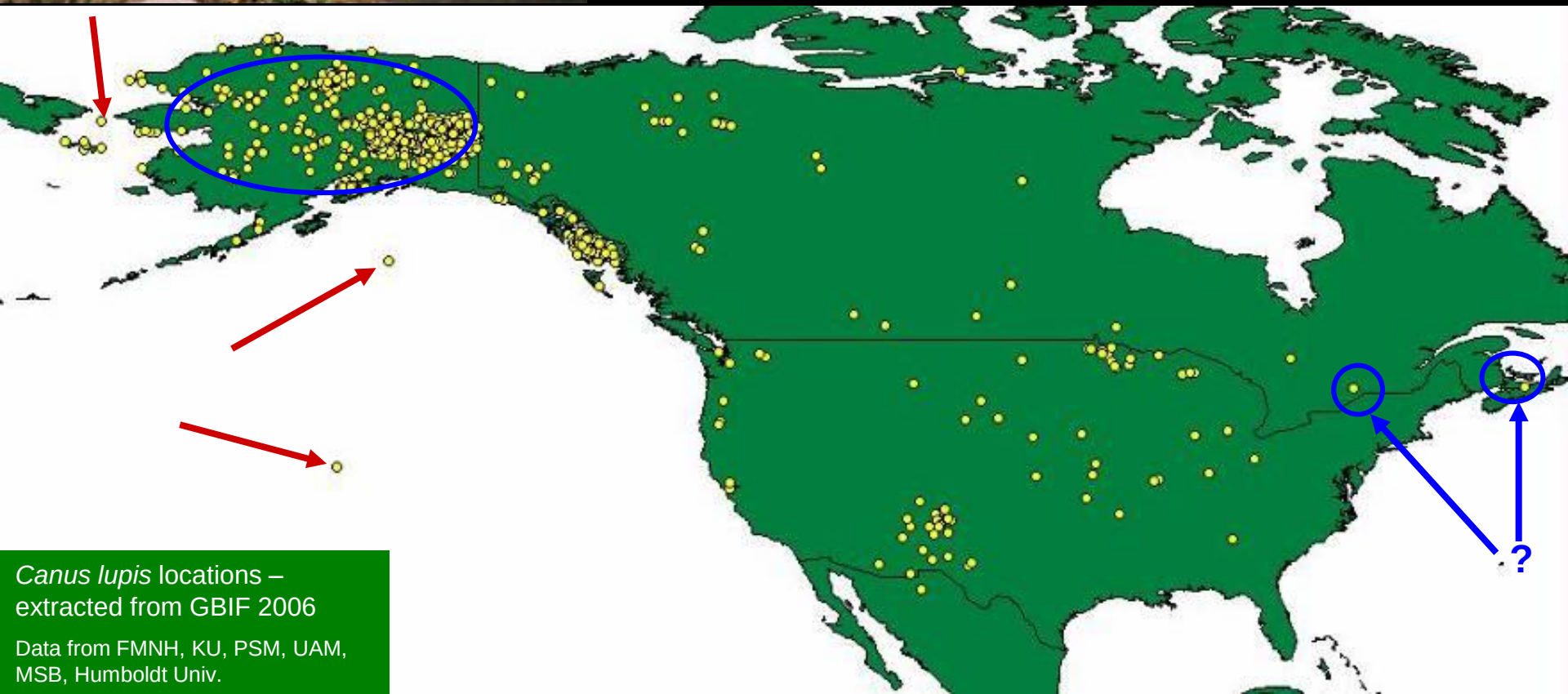
Gazetteer of Brazilian localities

How do we find the suspect records?

Some errors are easy to find!

But!

What does this say about the others?



Canus lupis locations – extracted from GBIF 2006

Data from FMNH, KU, PSM, UAM, MSB, Humboldt Univ.

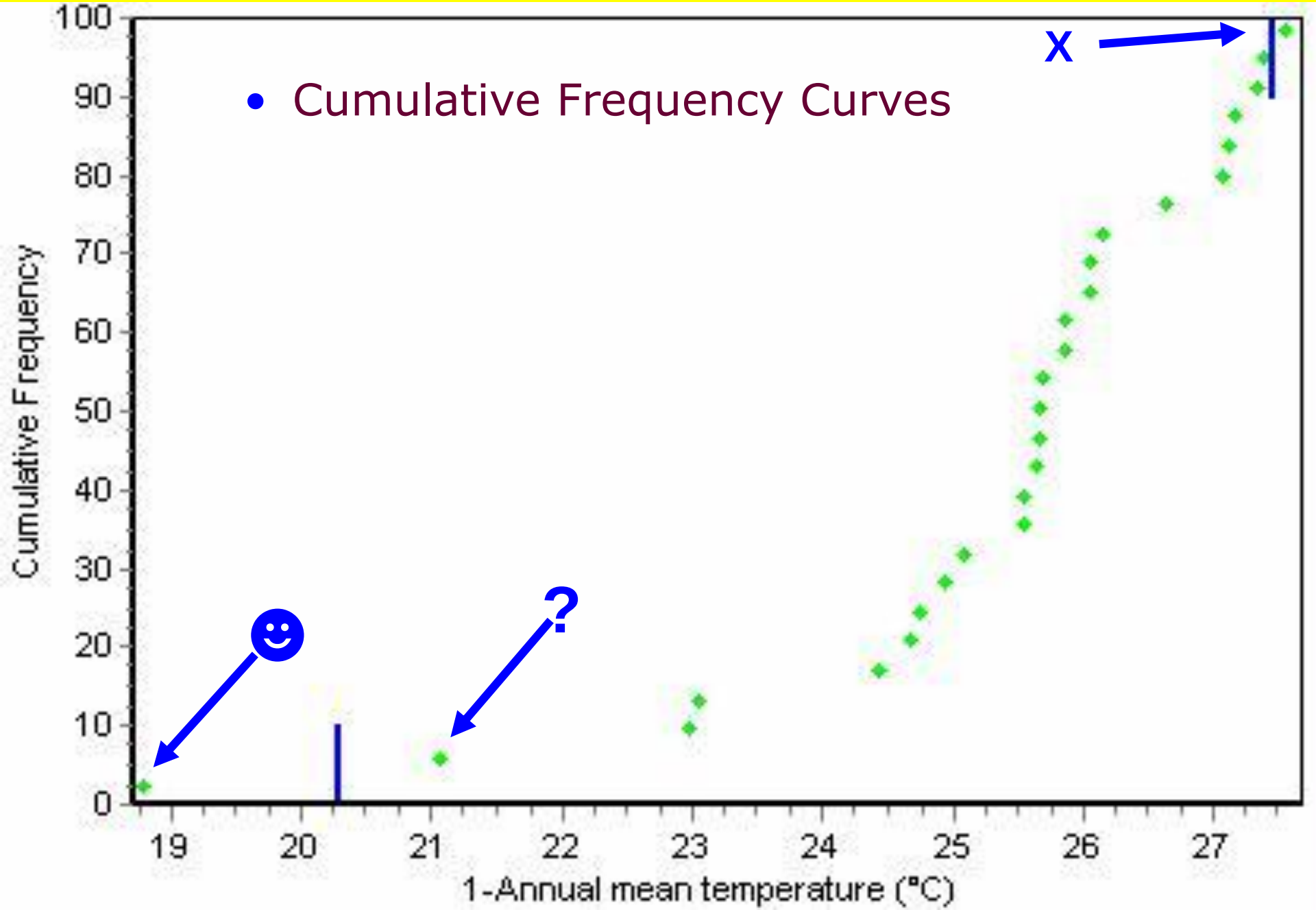
Geographic Outliers - GIS

Collectors – location vs date

▲ *Karel Domin* (1882–1953), profesor botaniky Karlovy univerzity, světově proslulý odborník a cestovatel, autor monografií o Brdech a Kokořínsku.

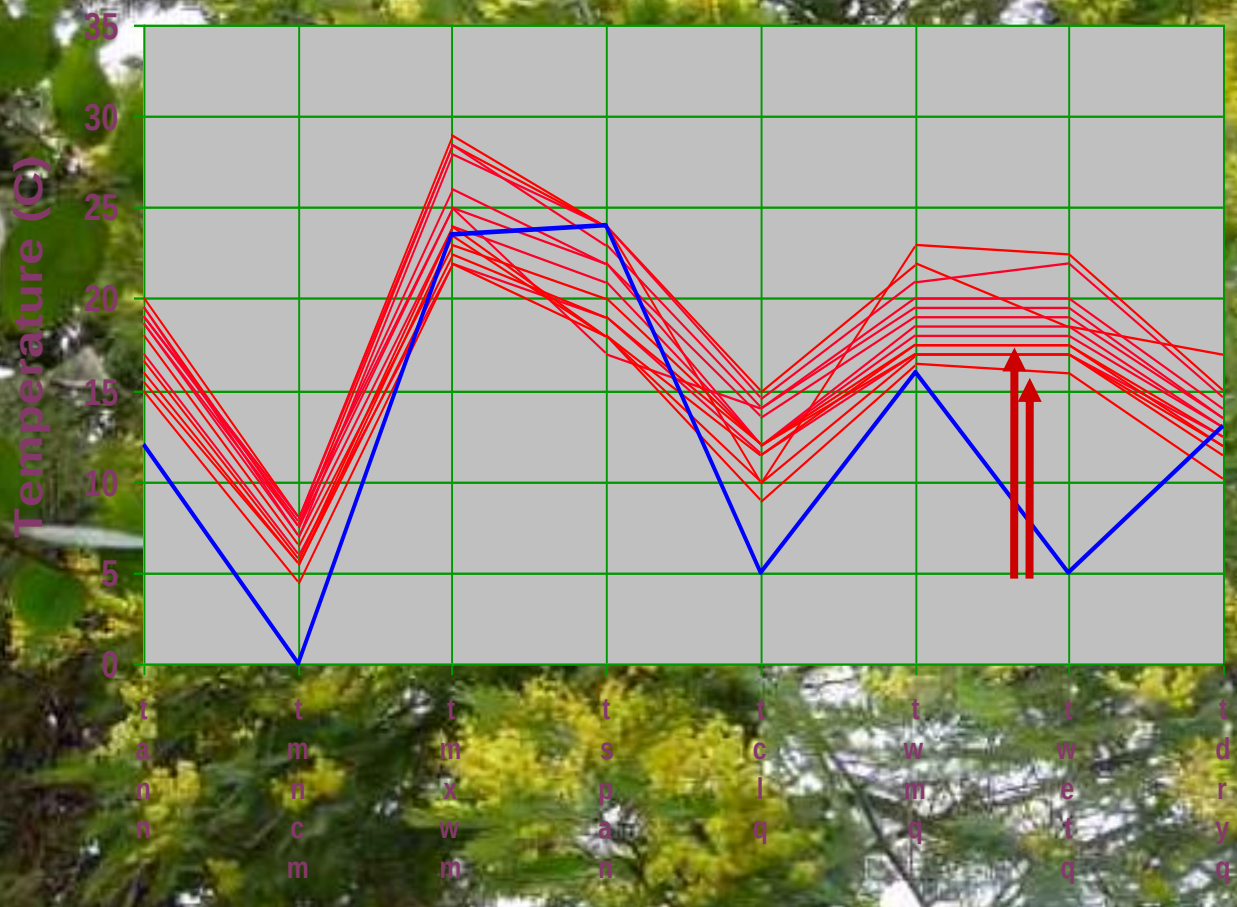
March
Townsville
1 Apr
Charters Towers
21-25 Apr
Lolworth
16-20 Apr
Mt Remarkable
2-6 Apr
11-15 Apr
Pentland
6-11 Apr
26 Apr
Hughenden
27-28 Apr
Mt White

Environmental Outliers



- Cumulative Frequency Curves

Using Climate to Identify Outliers



$$x < \bar{x}$$

if

$$y_{(i)} = \left(x_{(i+1)} - x_{(i)}\right)\left(\bar{x} - x_{(i)}\right)$$

else

$$y_{(i)} = \left(x_{(i+1)} - x_{(i)}\right)\left(x_{(i+1)} - \bar{x}\right)$$

then

$$C = \frac{y_{(i)}}{\sqrt{\frac{\sum_{i=1}^{n}\left(y_{(i)} - \bar{y}\right)^2}{n-1}}}$$

Reverse Jack-knife

Acacia orites - 19 records -
9 Temperature parameters

NB. Because the value of 'C' relates to it's nearest point, successive values may be very small, so we ensure that if 'x[i]' is an outlier, then all points beyond are outliers too (even if they are clustered)

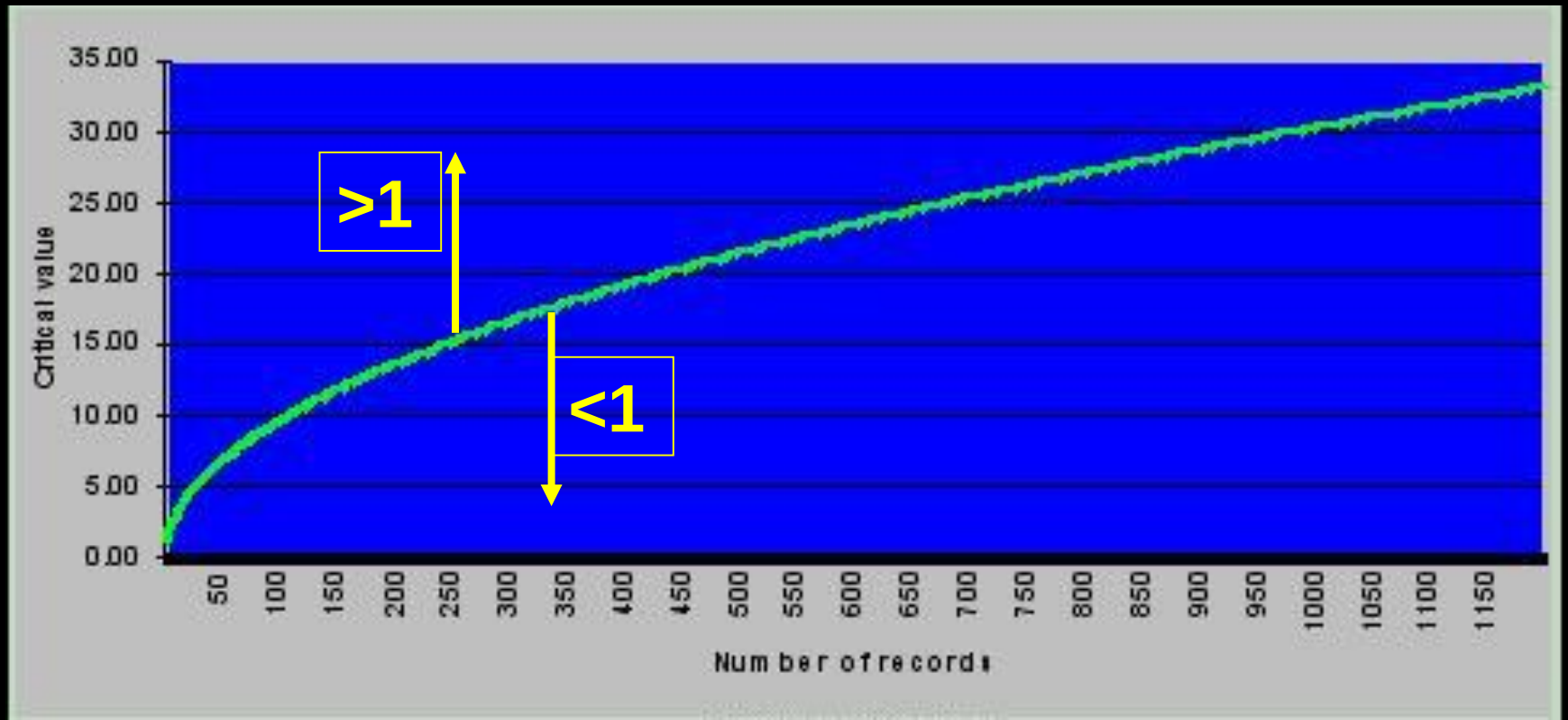
Concept of *"Outlierness"*

$$T=((0.95(\sqrt{n})+0.2)\ X\ (Range/50))$$

where 'n' is the number of records

"Outlierness" is the degree to which a record is an outlier

Outlierness = c[i]/ T



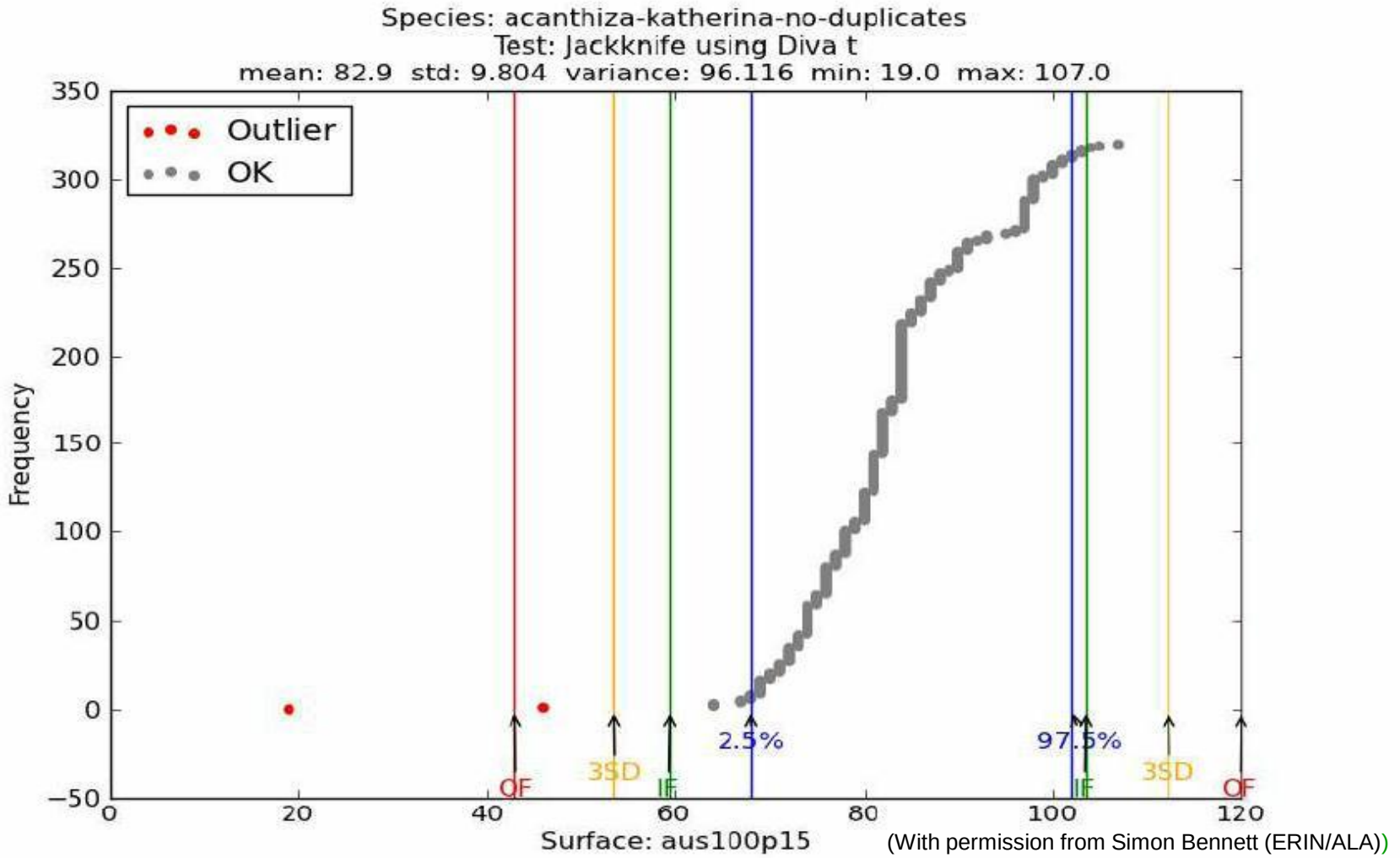
Acanthiza katherina

(With permission from Simon Bennett (ERIN/ALA))

- Typically an aggregated set of species occurrence data contains a small proportion of erroneous records.
- The map below shows records of the Mountain Thornbill *Acanthiza katherina* held by the Atlas of Living Australia.
- As this species is confined to upland rainforest of north-eastern Queensland the three locations in central Queensland, Victoria and on the South Australian-Northern Territory border are most unlikely to be natural occurrences and should be considered to be suspect records.
- These notes look at some methods that have been used, or could be used to flag outlier records requiring closer scrutiny.

Occurrences of the Mountain Thornbill Acanthiza katherina.
Source: Atlas of Living Australia.

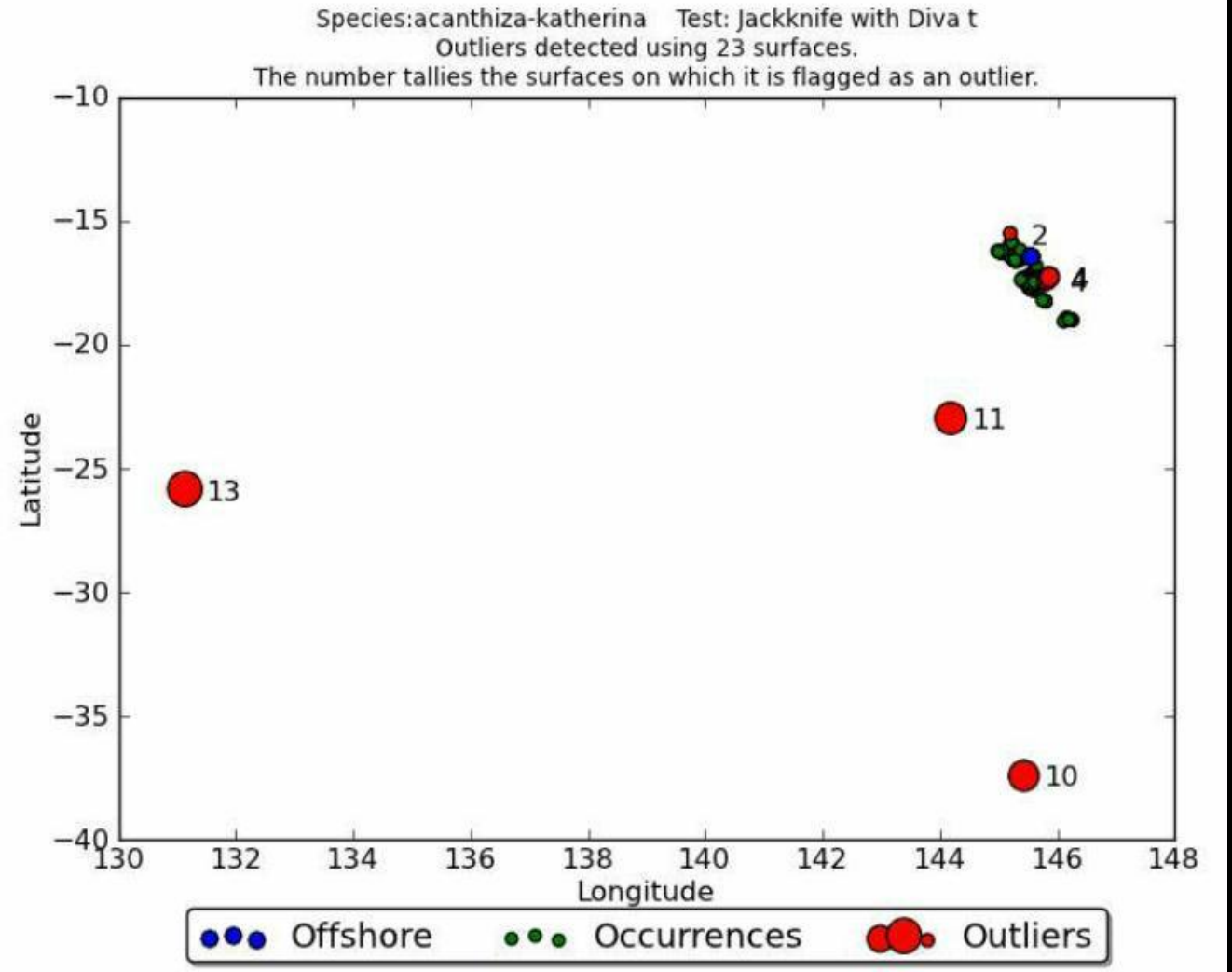
Reverse Jack-knife



(With permission from Simon Bennett (ERIN/ALA))

An example of the a cumulative frequency curve used by this and other methods, e.g. Bioclim. The figure plots occurrences against a climate variable (p15 Precipitation Seasonality (Coefficient of Variation). The Outliers detected using the **Reverse Jackknife** method are shown in **red circles**. Also indicated are the 95 inter-percentile range (2.5% and 97.5%), as well as values three standard deviations from the mean (untransformed) (3SD), and inner (IF) and outer (OF) fences from Tukey boxplot method.

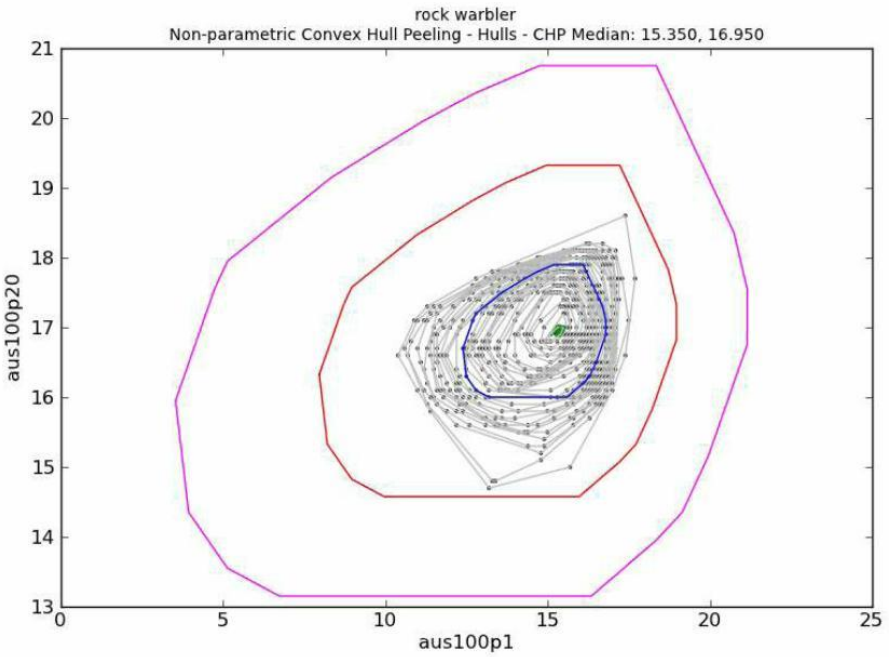
Reverse Jack-knife

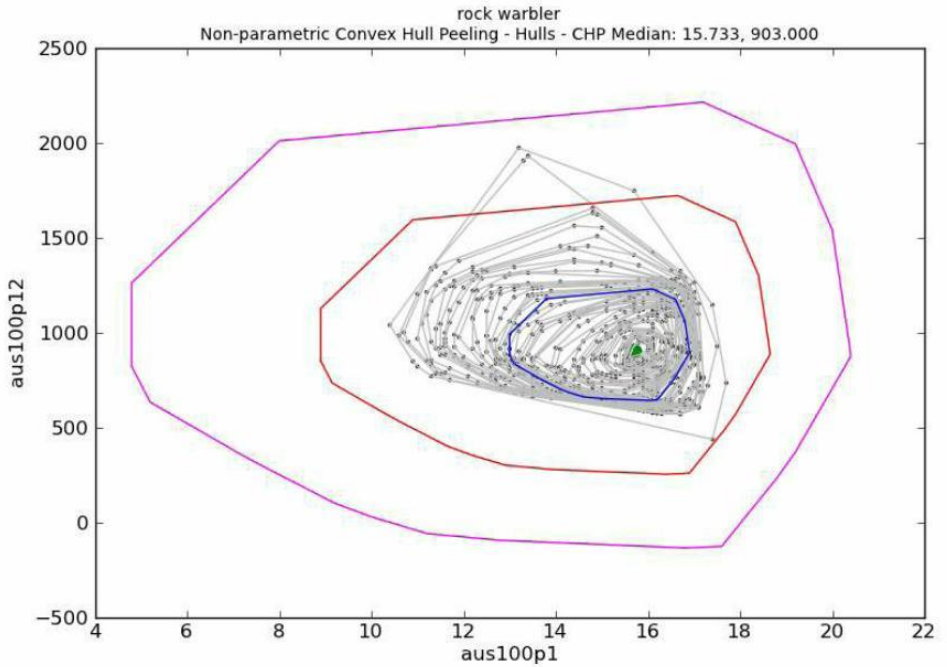


(With permission from Simon Bennett (ERIN/ALA))

This plot shows the overall results of applying the jackknife test to 23 surfaces: (latitude, longitude, elevation and 20 climate surfaces) to occurrences of the *Acanthiza katherina*. **Detected outlier values are shown in red** with the number of surfaces for which it was an outlier indicated. Occurrences falling in the sea are shown in blue. The plot indicates three obviously erroneous to the south west of the core range, along with number of local scale potential outliers requiring investigation. These may be altitudinal outliers.

Convex Hull



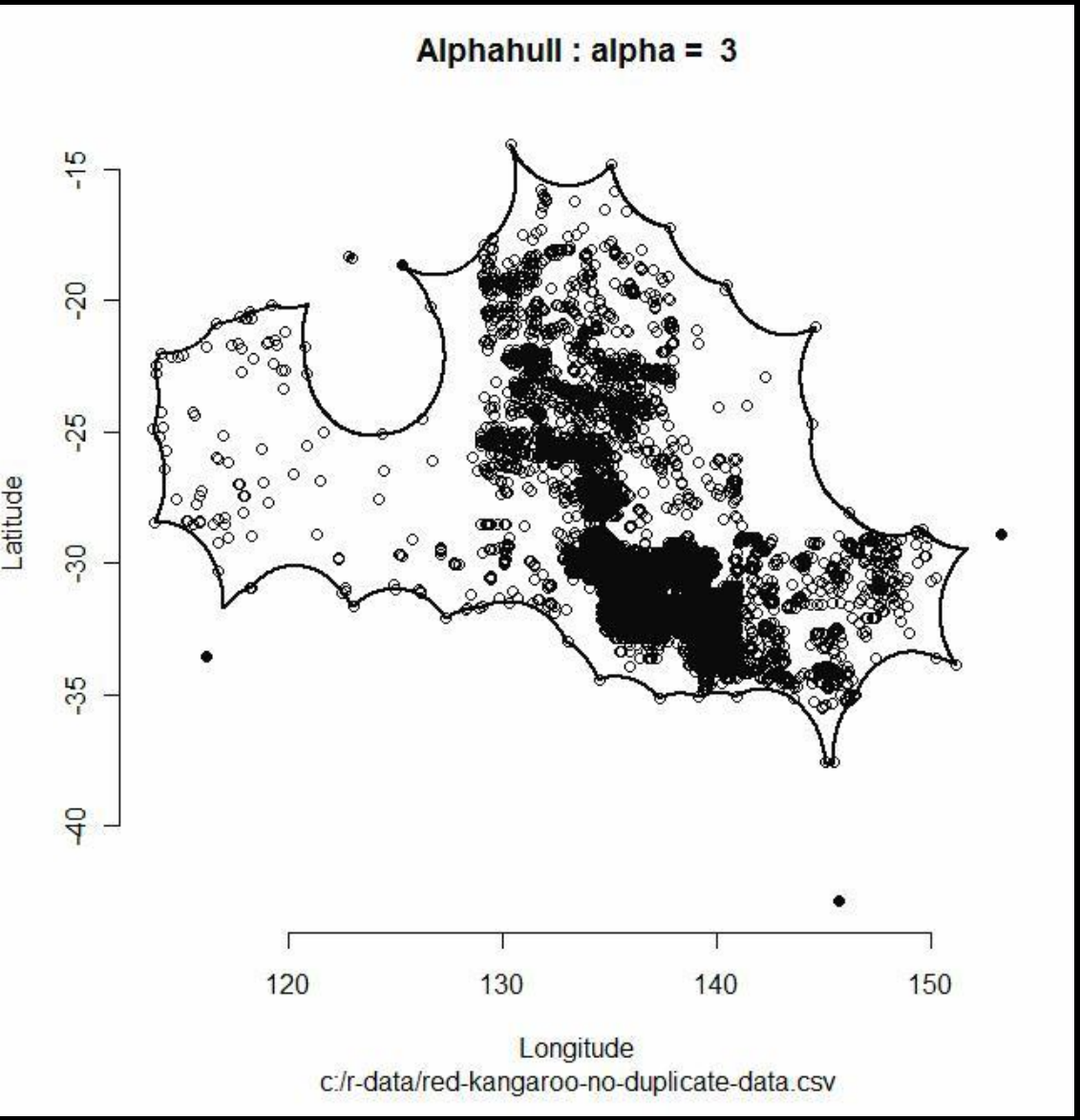


(With permission from Simon Bennett (ERIN/ALA))

Example of Convex Hull using Lee Belbin's algorithm. The two plots use different environmental surfaces. Interquartile range is blue, inner fence is red and outer fence magenta.
NB No outliers are indicated on the left hand plot – several outliers detected on the right hand plot.

Preliminary evaluation: Lee Belbin's algorithm was original implemented with massive data sets using stellar data. It appears to work well in indicating outliers where data are continuous with a mounded distribution on both axes. As the method expands the shape of the interquartile hull, it does not appear suited to data with a constrained value range, such as occurs with the Rainfall in Driest Quarter surface, which has an exponential distribution, with most of Australia having a near zero value.

Alpha Hull



The alpha-hull method is based on Delaunay triangulation and works by joining lines between all species occurrence points to form triangles, then measuring the length of each edge and excluding those triangles that are more than a multiple (alpha) of the average edge length.

While IUCN (2009) advocated an alpha value of two as a good starting point, Burgman and Fox found that a value of three was consistently the most robust to sampling artefacts introduced by differing sampling intensities, spatial accuracies and spatial uniformities.

(With permission from Simon Bennett (ERIN/ALA))

Principles of data quality

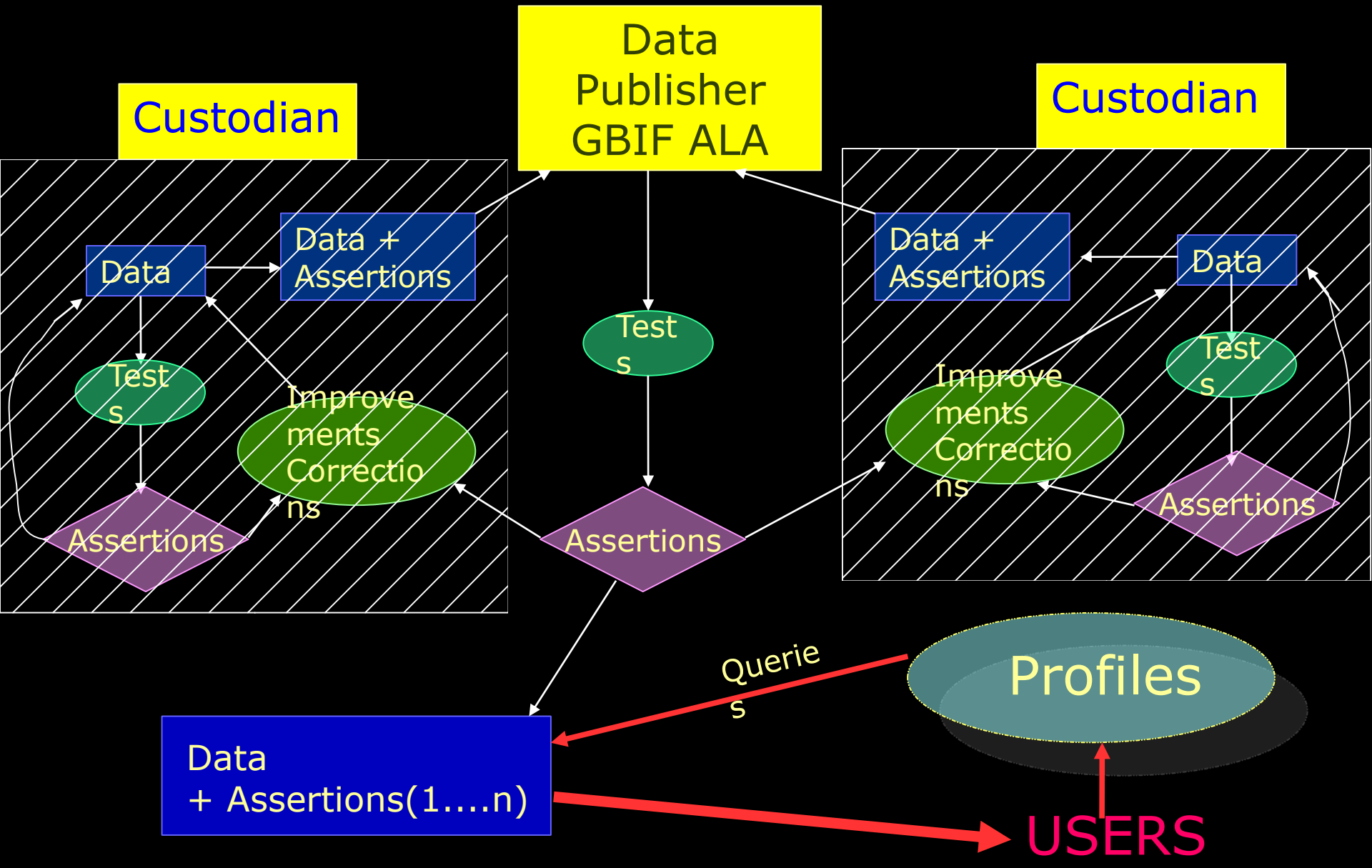
It is important for organizations to have

- a vision with respect to having good quality data;
- a policy to implement that vision; and
- a strategy for implementation.

Experience has shown that treating data as a long-term asset and managing it within a coordinated framework produces considerable savings and ongoing value.

(NLWRA 2003).

Data and Assertions

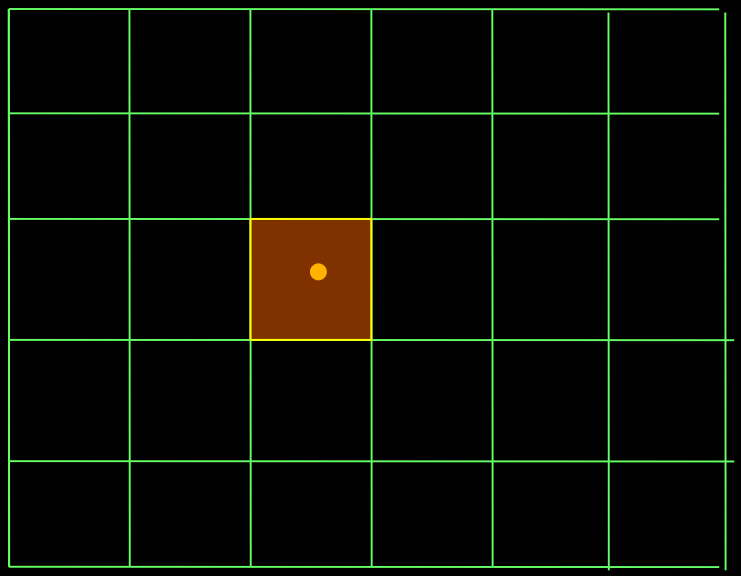


Responsibility of Users

- Users need to take responsibility for their use of the data quality information
  - Few users are extracting and using geocode uncertainty
  - Most users don't understand how to use Uncertainty
  - Users need to provide feedback on quality
  - We should have more papers discussing uncertainty and its use in analysis

Uncertainty and Modelling

Using georeferences for niche modelling

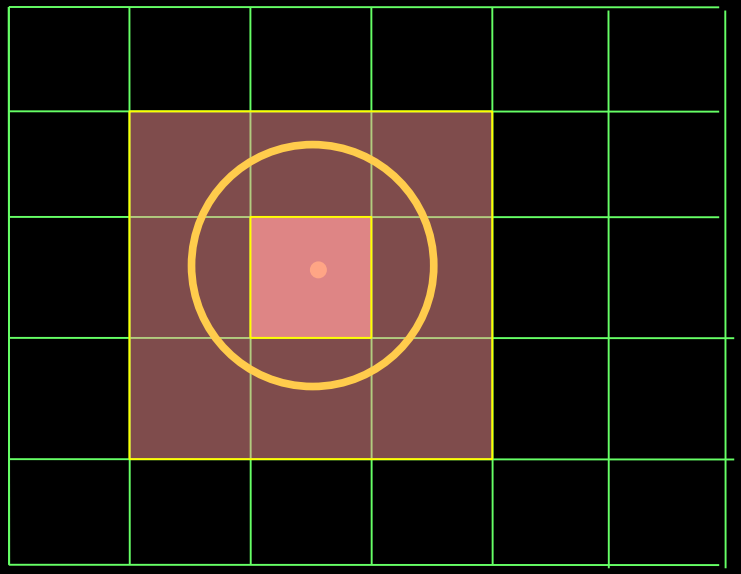


If we assume that a latitude and longitude refers to a point - you take that grid square – determine the climatic/environmental parameters and use that in your model to find other grids with the same climate or environments

This is the usual case with niche modelling as it exists today.

Uncertainty and Modelling

However georeferences aren't a just point but represent an area including its uncertainty



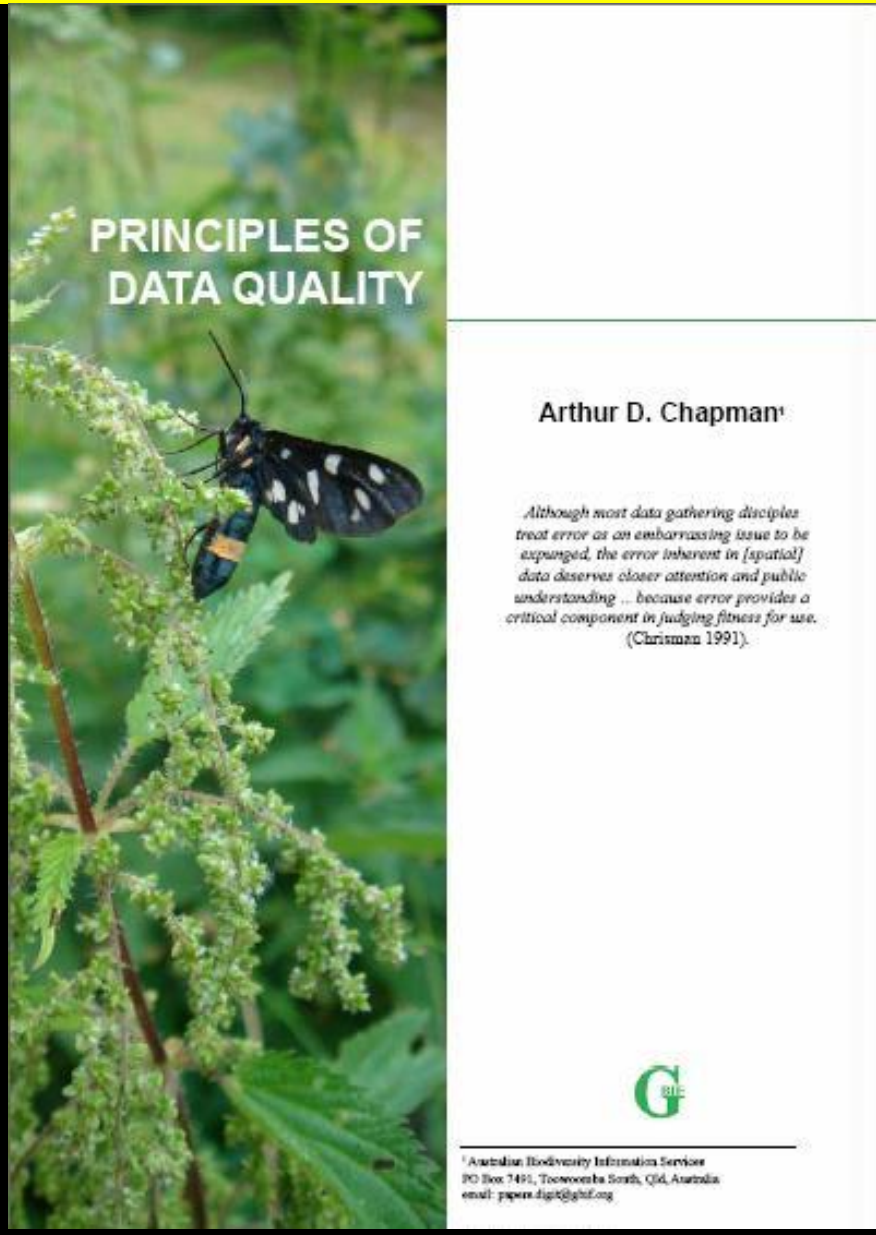
But taking in uncertainty, we see the point is no longer a point and perhaps we should be taking 9 grid squares – determining the environments for all those and apply those to the model

Then the question arises should we weight the grids in some way as one could assume that there is a higher likelihood of occurrence in the one grid than in the other 8?

Further reading

For further information see:

Chapman, A.D. (2005a).
Principles of Data Quality.
Report for the Global Biodiversity Information Facility. 61 pp.



http://www.gbif.org/prog/digit/data_quality/DataQuality.pdf

Thank You/Obrigado

Camponotus suffusus (Golden Flumed Sugar Ant) – Werribee Gorge, Australia