Royal Botanic Gardens

Kew

The Georeferencing Process

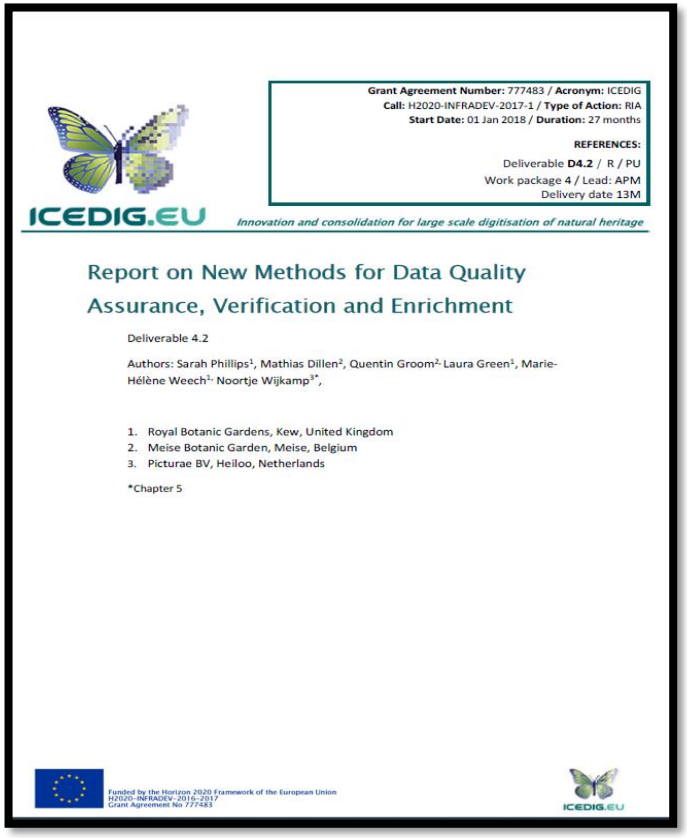
An evaluation of available georeferencing tools and protocols, advantages and shortcomings by Sarah Phillips and Jack Plummer

Contents

1. An overview of current software (“Innovation and Consolidation for Large Scale Digitisation of Natural Heritage”ICEDIG)
2. Georeferencing in practice: IUCN Red List assessments

An overview of current software (ICEDIG)

1. R biogeo (Robertson *et al.* 2016)
2. R GeoNames (Rowlingson 2016)
3. R dismo (Hijmans *et al.* 2017)
4. BioGeomancer (Guralnick *et al.* 2006)
5. Geo-referencing Calculator
6. The Edinburgh Geoparser
7. GEOLocate https://www.geo-locate.org/
8. SpeciesgeocodeR (Zizka & Antonelli 2015)
9. CoordinateCleaner (Zizka *et al.* 2019)



ICEDIG.EU

Grant Agreement Number: 777483 / **Acronym:** ICEDIG
Call: H2020-INFRADEV-2017-1 / **Type of Action:** RIA
Start Date: 01 Jan 2018 / **Duration:** 27 months

REFERENCES:
Deliverable **D4.2** / R / PU
Work package 4 / Lead: APM
Delivery date 13M

Innovation and consolidation for large scale digitisation of natural heritage

Report on New Methods for Data Quality Assurance, Verification and Enrichment

Deliverable 4.2

Authors: Sarah Phillips[1], Mathias Dillen[2], Quentin Groom[2], Laura Green[1], Marie-Hélène Weech[1], Noortje Wijkamp[3*],

1. Royal Botanic Gardens, Kew, United Kingdom
2. Meise Botanic Garden, Meise, Belgium
3. Picturae BV, Heiloo, Netherlands

*Chapter 5

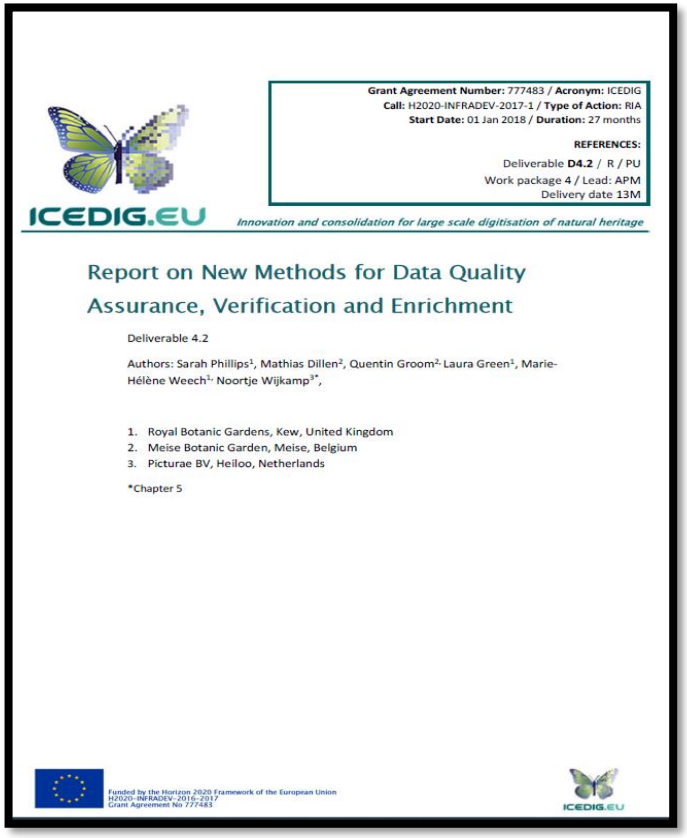
Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

ICEDIG.EU

An overview of current software (ICEDIG)

1. R biogeo (Robertson *et al.* 2016)
2. R GeoNames (Rowlingson 2016)
3. R dismo (Hijmans *et al.* 2017)
4. BioGeomancer (Guralnick *et al.* 2006)
5. Geo-referencing Calculator
6. The Edinburgh Geoparser
7. GEOLocate
8. Spec… Antonelli 2015)
9. Coor… *t al.* 2019)

Available to install and tested



ICEDIG.EU

Grant Agreement Number: 777483 / **Acronym:** ICEDIG
Call: H2020-INFRADEV-2017-1 / **Type of Action:** RIA
Start Date: 01 Jan 2018 / **Duration:** 27 months

REFERENCES:
Deliverable **D4.2** / R / PU
Work package 4 / Lead: APM
Delivery date 13M

Innovation and consolidation for large scale digitisation of natural heritage

Report on New Methods for Data Quality Assurance, Verification and Enrichment

Deliverable 4.2

Authors: Sarah Phillips[1], Mathias Dillen[2], Quentin Groom[2], Laura Green[1], Marie-Hélène Weech[1], Noortje Wijkamp[3*],

1. Royal Botanic Gardens, Kew, United Kingdom
2. Meise Botanic Garden, Meise, Belgium
3. Picturae BV, Heiloo, Netherlands

*Chapter 5

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

ICEDIG.EU

An overview of current software (ICEDIG)

Royal Botanic Gardens Kew

R dismo (Hijmans *et al.* 2017)

- Developed for species distribution modelling.
- function – geocode – was used to provide coordinates based on a locality description. Uses Google geocoding webservice. This required a relatively precise location unsuitable for vague locality descriptions.

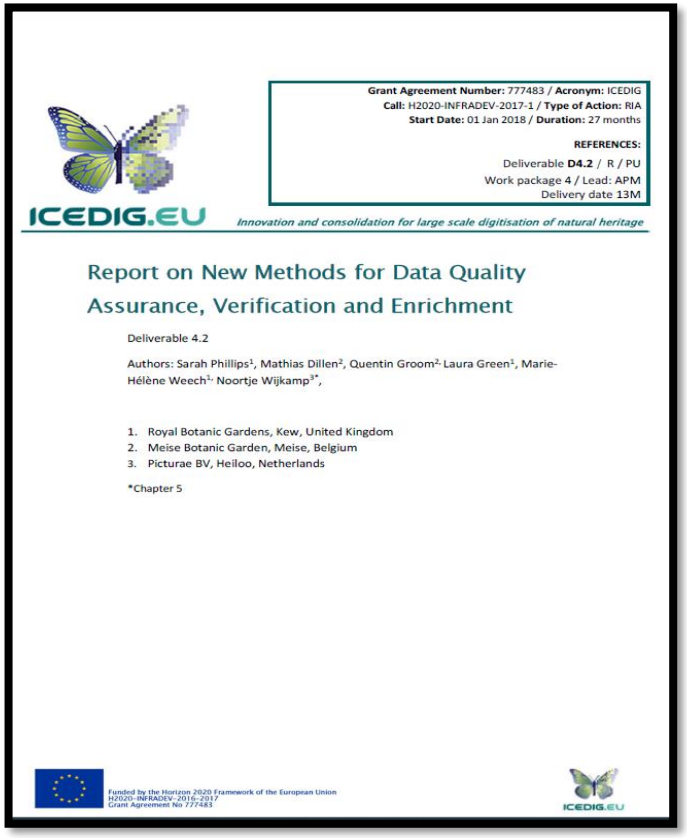
GEOLocate

- A web-based platform to georeference from a text string. The program will split the text string into country, county, locality, etc.
- It is possible to type, cut and paste a single locality string, or upload a CSV file and batch process it.
- Returns an output of latitude and longitude in decimal degrees, with an accuracy value in meters.
- Outputs a number of points that need to be reviewed to choose the one that is the most likely
- Accuracy values varied greatly.

An overview of current software (ICEDIG)

1. R biogeo (Robertson *et al.* 2016)
2. R GeoNames (Rowlingson 2016)
3. R dismo (Hijmans *et al.* 2017)
4. BioGeomancer (Guralnick *et al.* 2006)
5. Geo-referencing Calculator
6. The Edinburgh Geoparser
7. GEO
8. Spe Antonelli 2015)
9. Coc *et al.* 2019)

Unavailable or could not be installed



ICEDIG.EU

Grant Agreement Number: 777483 / **Acronym:** ICEDIG
Call: H2020-INFRADEV-2017-1 / **Type of Action:** RIA
Start Date: 01 Jan 2018 / **Duration:** 27 months

REFERENCES:
Deliverable **D4.2** / R / PU
Work package 4 / Lead: APM
Delivery date 13M

Innovation and consolidation for large scale digitisation of natural heritage

Report on New Methods for Data Quality Assurance, Verification and Enrichment

Deliverable 4.2

Authors: Sarah Phillips[1], Mathias Dillen[2], Quentin Groom[2], Laura Green[1], Marie-Hélène Weech[1], Noortje Wijkamp[3*],

1. Royal Botanic Gardens, Kew, United Kingdom
2. Meise Botanic Garden, Meise, Belgium
3. Picturae BV, Heiloo, Netherlands

*Chapter 5

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

ICEDIG.EU

An overview of current software (ICEDIG)

Royal Botanic Gardens Kew

- R biogeo (Robertson *et al.* 2016)
- Developed for detecting and correcting errors and for assessing data quality
- Finds coordinates for localities that have no coordinates
- Functions are also available for converting coordinates that are in various text formats into degrees, minutes and seconds and then into decimal degrees.

- R GeoNames (Rowlingson 2016)
- A geographical database that can be used to georeference from a bank of > 8 million place names
- Able to use functions to input north, south, east and west text values to find places a certain distance in a given direction from the locality.

An overview of current software (ICEDIG)

- BioGeomancer (Guralnick *et al.* 2006)
- After records containing locality information are uploaded to the website, one or more methods for natural language processing parse parts of a locality description into data fields. Named places are then looked up in a Gazetteer
- estimates uncertainty associated with records' coordinates
- Validation tools

- The Edinburgh Geoparser
- A system that is able to automatically recognise place names, within a text file or string, which can disambiguate them with respect to a gazetteer.
- Can be used with several gazetteers – Unlock and GeoNames
- Only use with Mac or Linux. Tried with a Mac but needed coding skills.

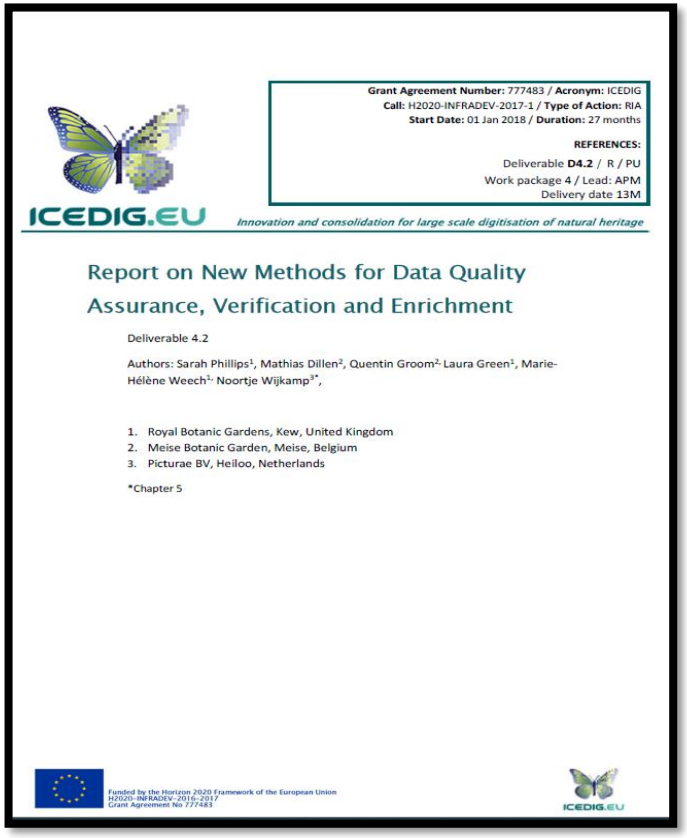
An overview of current software (ICEDIG)

Royal Botanic Gardens Kew

1. R bi… …6)
2. R Ge… …6)
3. R dis…

Tools for complementary functions

4. BioGeomancer (Guralnick *et al.* 2006)
5. Geo-referencing Calculator
6. The Edinburgh Geoparser
7. GEOLocate
8. SpeciesgeocodeR (Zizka & Antonelli 2015)
9. CoordinateCleaner (Zizka *et al.* 2019)



ICEDIG.EU

Grant Agreement Number: 777483 / **Acronym:** ICEDIG
Call: H2020-INFRADEV-2017-1 / **Type of Action:** RIA
Start Date: 01 Jan 2018 / **Duration:** 27 months

REFERENCES:
Deliverable **D4.2** / R / PU
Work package 4 / Lead: APM
Delivery date 13M

Innovation and consolidation for large scale digitisation of natural heritage

Report on New Methods for Data Quality Assurance, Verification and Enrichment

Deliverable 4.2

Authors: Sarah Phillips[1], Mathias Dillen[2], Quentin Groom[2], Laura Green[1], Marie-Hélène Weech[1], Noortje Wijkamp[3*],

1. Royal Botanic Gardens, Kew, United Kingdom
2. Meise Botanic Garden, Meise, Belgium
3. Picturae BV, Heiloo, Netherlands

*Chapter 5

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

ICEDIG.EU

An overview of current software (ICEDIG)

Royal Botanic Gardens Kew

- Georeferencing Calculator

- Calculates all the factors that contribute to the uncertainty in a georeference

- SpeciesgeocodeR (Zizka & Antonelli 2015)

- An R package for automatically cleaning, processing and analysing species occurrence data
- The *GeoClean* function offers an automated flagging of potentially problematic records. The function includes basic tests for coordinate validity

- CoordinateCleaner (Zizka *et al.* 2019)

- Tool for speeding up the identification of problematic records and common problems in a data set for further verification

An overview of current software (ICEDIG)

- Can be difficult to find the tools or know how to get hold of them
- Sustainability issues with some tools developed under projects
- Users often need to be comfortable with use of Github, R or API's
- Some institutions/projects trying to build own pipelines with some automation steps (Luomus)
- Full automation currently not possible but there are useful tools out there that we are not using to their full potential

ICEDIG.EU

Funded by the Horizon 2020 Framework Programme of the European Union

Other methods to speed up Georeferencing

Royal Botanic Gardens Kew

- Georeference by collector– Georeference a collecting trip especially useful if you have other information you can use e.g. field notebooks
- Georeference by Locality – sort by locality and georeference all specimens from that site at once.
- Collaborative georeferencing Tools e.g. GeoLocate– detects duplicates. Allocates different records to different users.
- Collection Management Systems missing important fields e.g. Notes for how georeference was determined.
- Currently lots of duplication of effort between and within institutions.

Royal Botanic Gardens Kew

2. Application of georeferenced data: IUCN Red List assessments

Application of georeferenced data

- Underpinning IUCN Red List assessments is the creation of a distribution map
- “Approximately half of the species on the IUCN Red List were listed on the basis of ***only*** **geographic range criteria**” (Gaston and Fuller 2009)
- “Additions of specimen records and taxonomic remodelling had relatively little impact in driving changes in conservation category compared with corrections of misidentifications and **enhanced georeferencing**” (Nic Lughadha *et al*. 2019)

Accurate georeferencing is essential for the creation of informed assessments

Application of georeferenced data

Specimen data collated (from various sources e.g. literature, GBIF download, herbaria records, bespoke collector database etc.)

Locality information entered to Google Earth

If not found, online gazetteer (e.g. GeoNames) consulted

If not found, a variety of further sources used to identify locality e.g. physical gazetteer, online search, expedition notes, maps

Coordinates and error diameter assigned based on all available info

Point file with verified coordinates used to create distribution map

Typical workflow when georeferencing for Red List Assessments

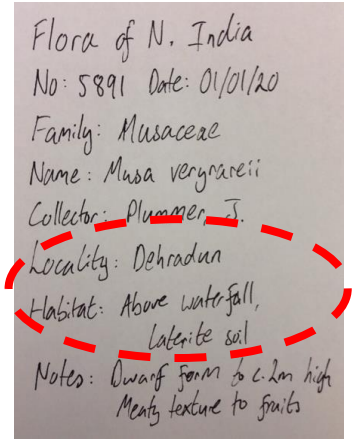
Record-by-record manual georeferencing

In our process, on average, georeferencing/data collation takes at least as much time as writing the assessment itself

Application of georeferenced data

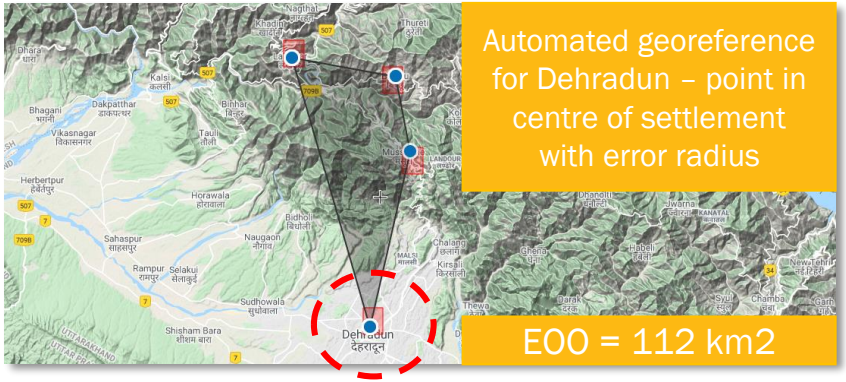
Made-up species Example 1:

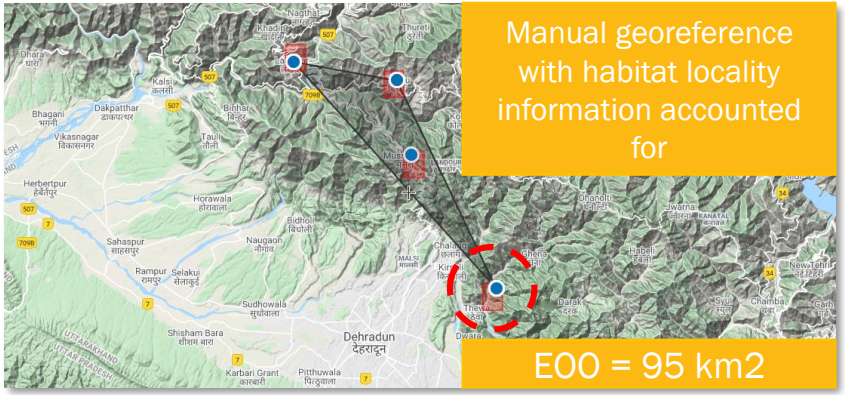
Musa veryrareii



- Known from 4 records
- Severely fragmented
- Threatened by over-collection across range
- If EOO < 100 km2, Critically Endangered;
  If EOO > 100 km2, Endangered
- Automated georeferencing could lead to inaccuracy ultimately affecting the assessment outcome

For species with a restricted distribution, manual georeferencing will remain necessary

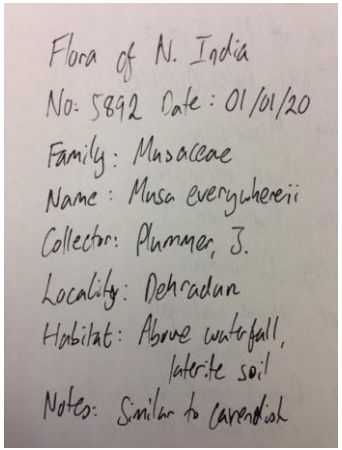


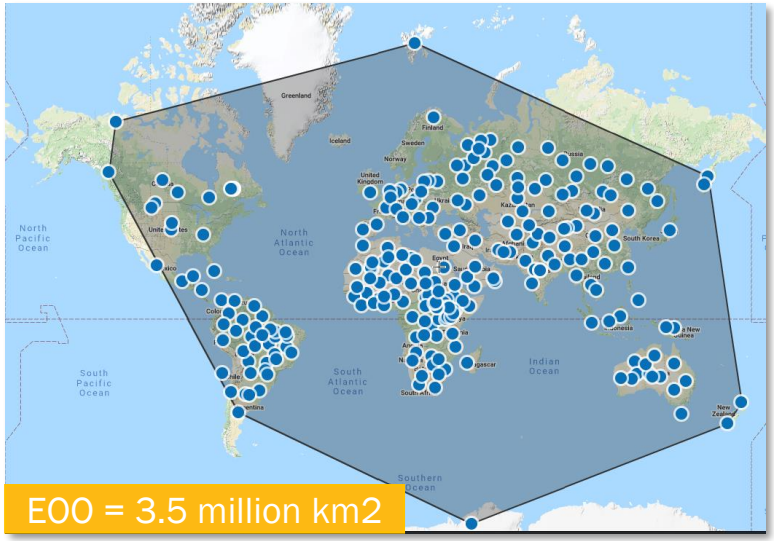


Application of georeferenced data

Made-up species Example 2:

Musa everywhereii

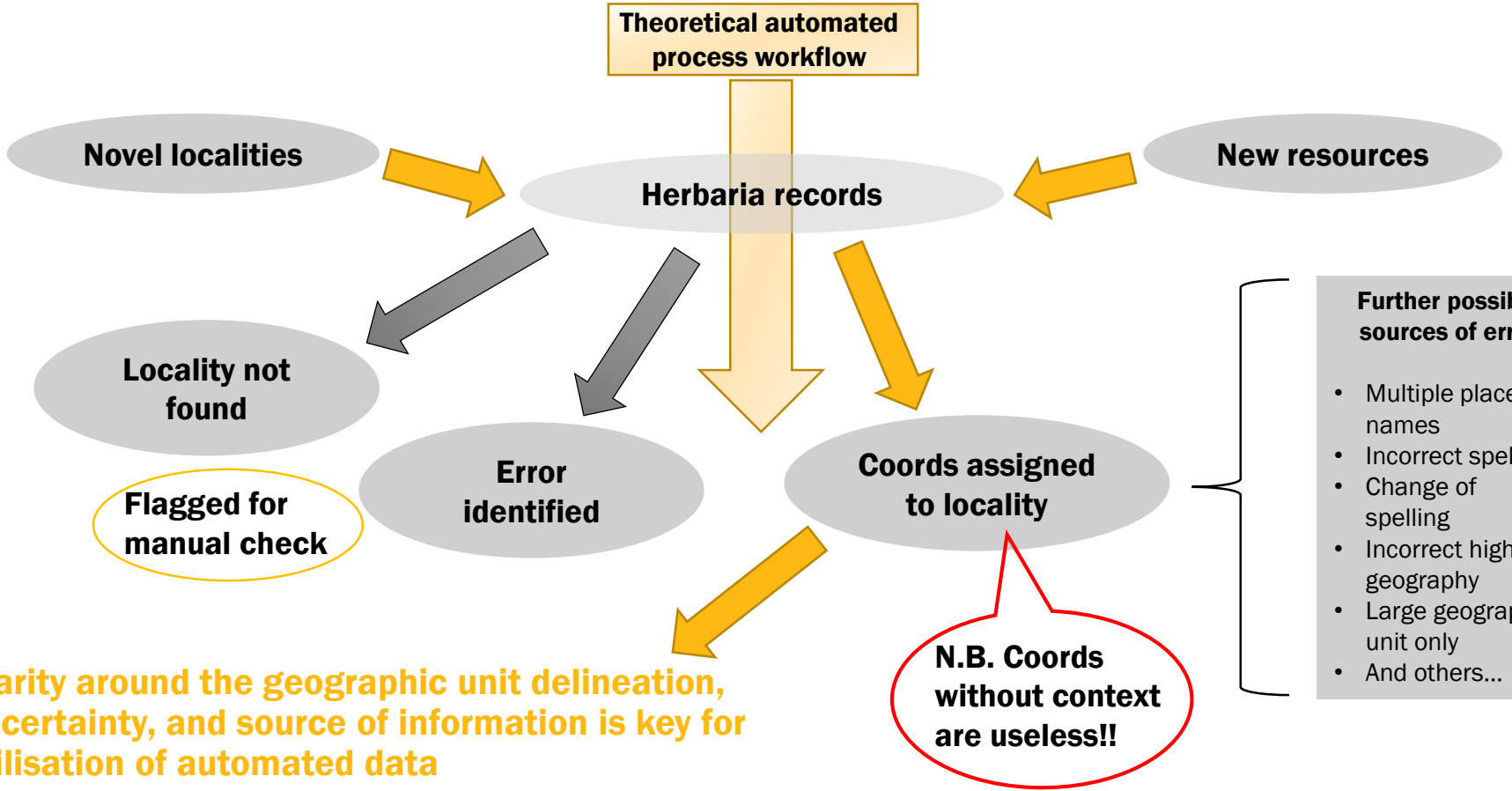




- Extremely widespread
- Abundant where found
- No major threats apparent
- High confidence that species is Least Concern
- Automated georeference for Dehradun – no problem
- Hundreds of records to manually check – being able to filter dataset by high confidence records would be very beneficial

For species with a widespread distribution, automation of georeferencing could hugely improve efficiency and permit greater allocation of resource to threatened species

Application of georeferenced data



Clarity around the geographic unit delineation, uncertainty, and source of information is key for utilisation of automated data

Application of georeferenced data

Royal Botanic Gardens Kew

1. Georeferenced collection data provides an invaluable resource for further research

2. However, the degree to which automated georeferencing and cleaning are appropriate/sufficient will depend on the end use of the data

3. As such, clear documentation of geographic unit delineation, uncertainty and source of information is key

4. In some cases, manual georeferencing will still be necessary

REFERENCES

Bachman, S., Moat, J., Hill, A.W., de la Torre, J. and Scott, B. 2011. Supporting Red List threat assessments with GeoCAT: geospatial conservation assessment tool. In: V. Smith and L. Penev (eds) e-Infrastructures for data publishing in biodiversity science. *Zookeys* 150: 117-126

Gaston, K. J., & Fuller, R. A. (2009). The sizes of species' geographic ranges. Journal of Applied Ecology, 46, 1– 9.

Nic Lughadha E.M., Graziele Staggemeier V., Nogales da Costa Vasconcelos T., Walker B.E., Canteiro C., Lucas E.J. 2019. Harnessing the potential of integrated systematics for conservation of taxonomically complex, megadiverse plant groups. *Conservation Biology* **33**: 510– 521

Phillips Sarah, Dillen Mathias, Groom Quentin, Green Laura, Weech Marie-Hélène, & Wijkamp Noortje. (2019). Report on New Methods for Data Quality Assurance, Verification and Enrichment. Zenodo https://doi.org/10.5281/zenodo.3364509

ACKNOWLEDGEMENTS

Maya Master and Nicholas Wells