

Lessons learnt from a herbarium specimen mass digitisation pilot

A Pilot between RBG Kew and Natural History Museum London

Sarah Phillips, Alan Paton, Sandy Knapp, Laura Green



Overview

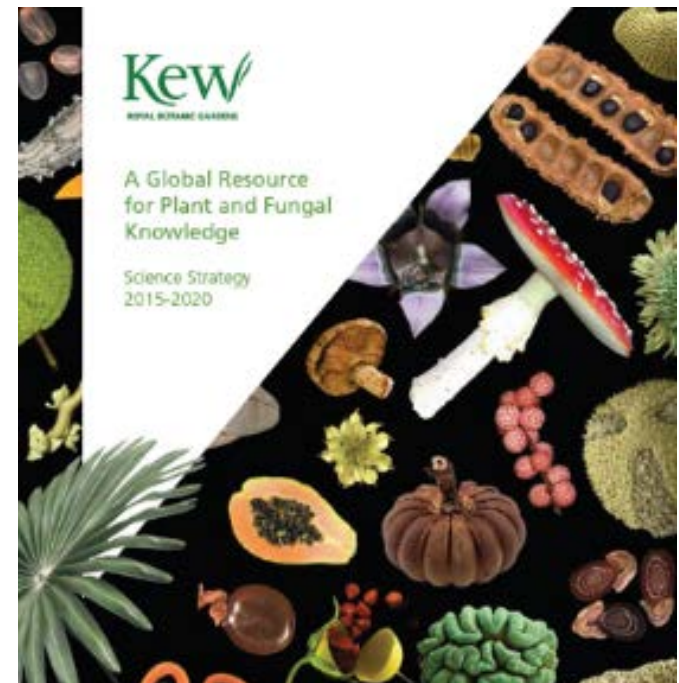
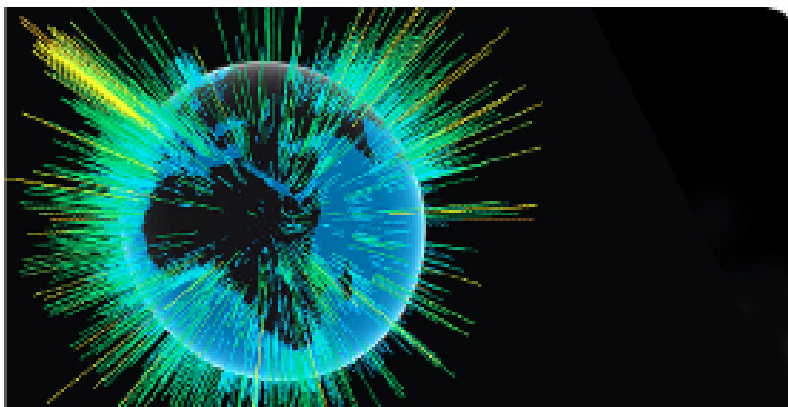
- Background to collaboration
- Mass digitisation pilot: What we did
- Lessons learned
- Next steps and challenges



- New science strategies emphasise, and have targets for, digitisation
- Establish proof of concept to stimulate new funding
- Large scale digitisation now possible



Science strategy 2013 - 2017



K/NHM Mass digitisation pilot

- Image all

Solanum, Dioscoreaceae,
Hypericum (NHM only)

67,500 specimens

420 ppi images

3000 - 4000 per day

- Transcribe specimen data - Team in Suriname
- Evaluate different workflows
- Derive robust costings

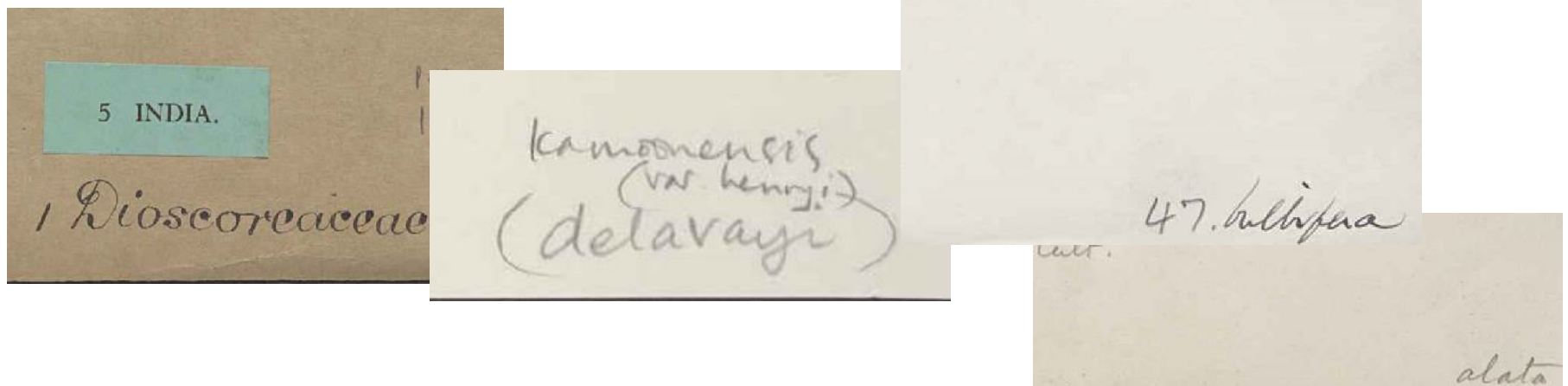


Differing workflows to test

Creation of stub record (barcoded in herbarium –
'filed as' name recorded)

v

No pre curation: barcoded at Picturae, names from
imaged covers



Transport



Training



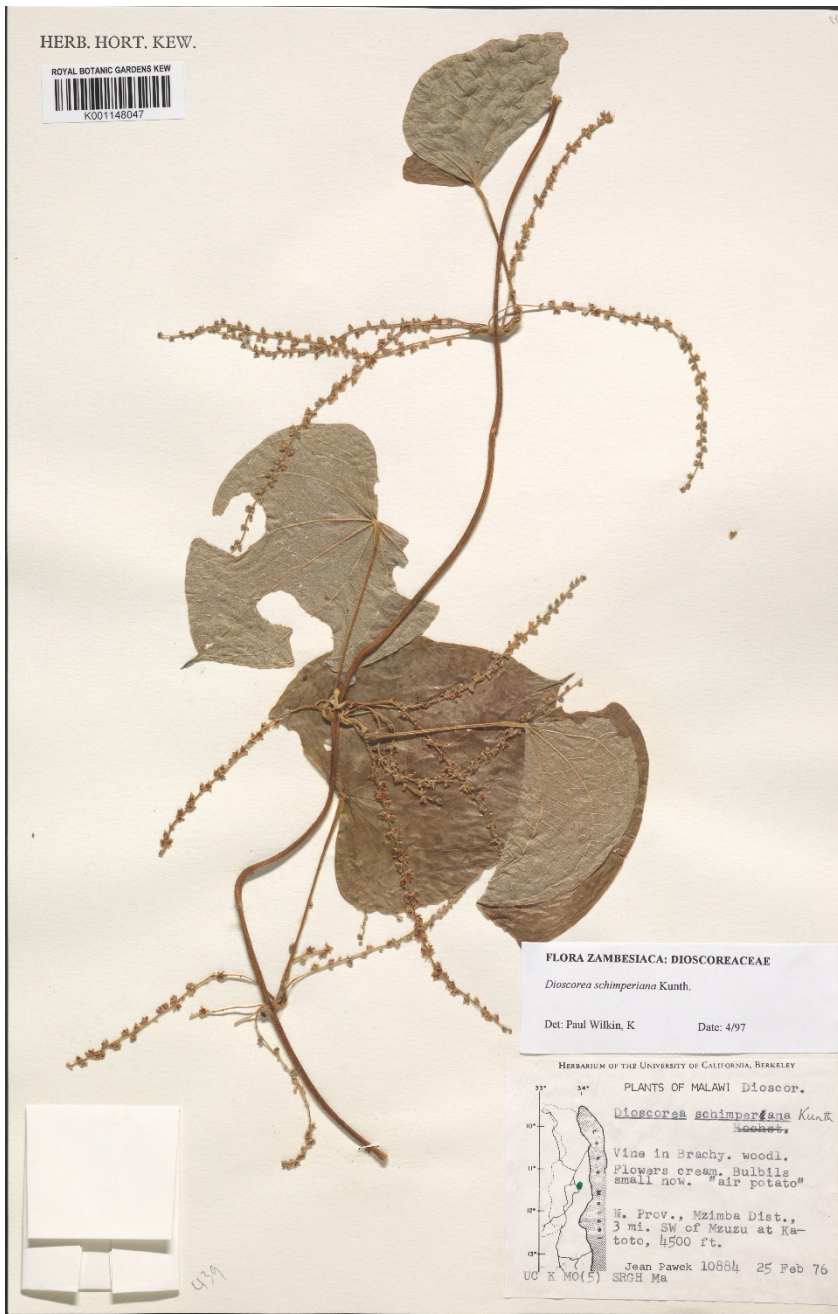
Picturae digistreet



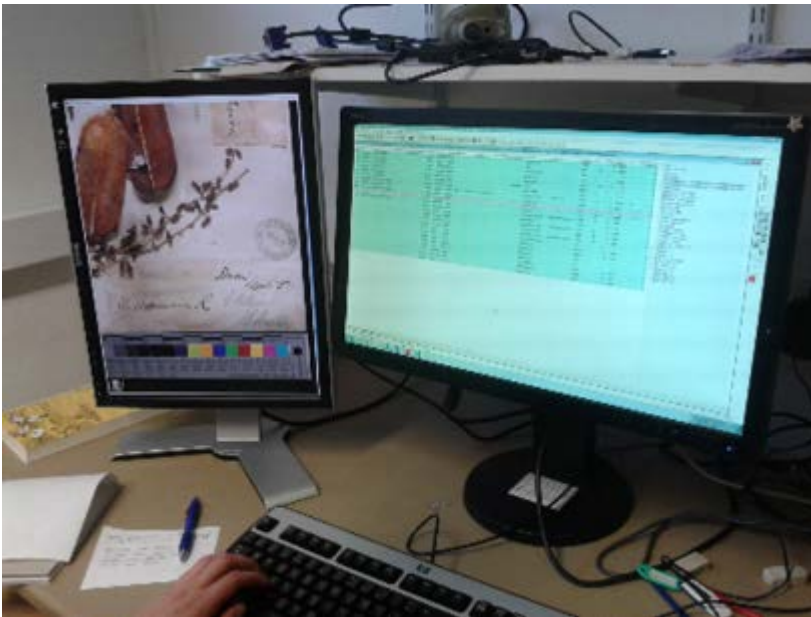
Sheet and cover order preserved



Image Quality Control



Data Quality Control



PICURAE Dataentry Herbaria

2017-01-23 10:27:02

Herbarium Bogoriense (BO) - Herbarium Bogoriense (BO)

| NO | NO | NO | NO | NO | NO | NO |
|----|----|----|----|----|----|----|
| 1 | BO | BO | BO | BO | BO | BO |
| 2 | BO | BO | BO | BO | BO | BO |
| 3 | BO | BO | BO | BO | BO | BO |
| 4 | BO | BO | BO | BO | BO | BO |
| 5 | BO | BO | BO | BO | BO | BO |
| 6 | BO | BO | BO | BO | BO | BO |
| 7 | BO | BO | BO | BO | BO | BO |
| 8 | BO | BO | BO | BO | BO | BO |
| 9 | BO | BO | BO | BO | BO | BO |
| 10 | BO | BO | BO | BO | BO | BO |

Showing records 1 - 10 of 10 records


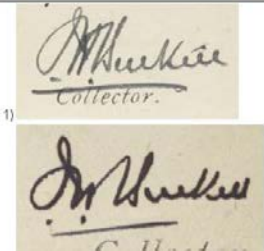
| NO | NO | NO | NO | NO | NO | NO | NO | NO | NO |
|----|----|----|----|----|----|----|----|----|----|
| 1 | BO | BO | BO | BO | BO | BO | BO | BO | BO |
| 2 | BO | BO | BO | BO | BO | BO | BO | BO | BO |
| 3 | BO | BO | BO | BO | BO | BO | BO | BO | BO |
| 4 | BO | BO | BO | BO | BO | BO | BO | BO | BO |
| 5 | BO | BO | BO | BO | BO | BO | BO | BO | BO |
| 6 | BO | BO | BO | BO | BO | BO | BO | BO | BO |
| 7 | BO | BO | BO | BO | BO | BO | BO | BO | BO |
| 8 | BO | BO | BO | BO | BO | BO | BO | BO | BO |
| 9 | BO | BO | BO | BO | BO | BO | BO | BO | BO |
| 10 | BO | BO | BO | BO | BO | BO | BO | BO | BO |

Showing records 1 - 10 of 10 records

manis

Transcription QC

- Weekly conference call to discuss issues
- Feedback written in individual Brahms records
- Shared Google Doc for question on labels and examples
- Increase level of QC at start of project
- Important to have a clear tested transcription protocol

| | | | | |
|----|---------|---|---|--|
| 86 | 16 June |  | Date has been entered 8/10/1895 | Unfortunately Rob Combs is a collector that uses the american format for dates. Therefore this is actually 10th August 1895. [Andrew Budden 16 June 2015] |
| 85 | 16 June |  | Collector has been entered as 1) Burkill, J.H. 2) Burkill, H.M. 3) Burkill, I.H. | This collector is 3) Burkill, I.H. It is in the lookup as opposed to 1). The specimens transcribed as being collected by 2) had a date range of 1916-1920, which makes it highly unlikely given that he was born in 1914 (listed in the lookup too). [Andrew Budden 16 June 2015] |

Error Values Table

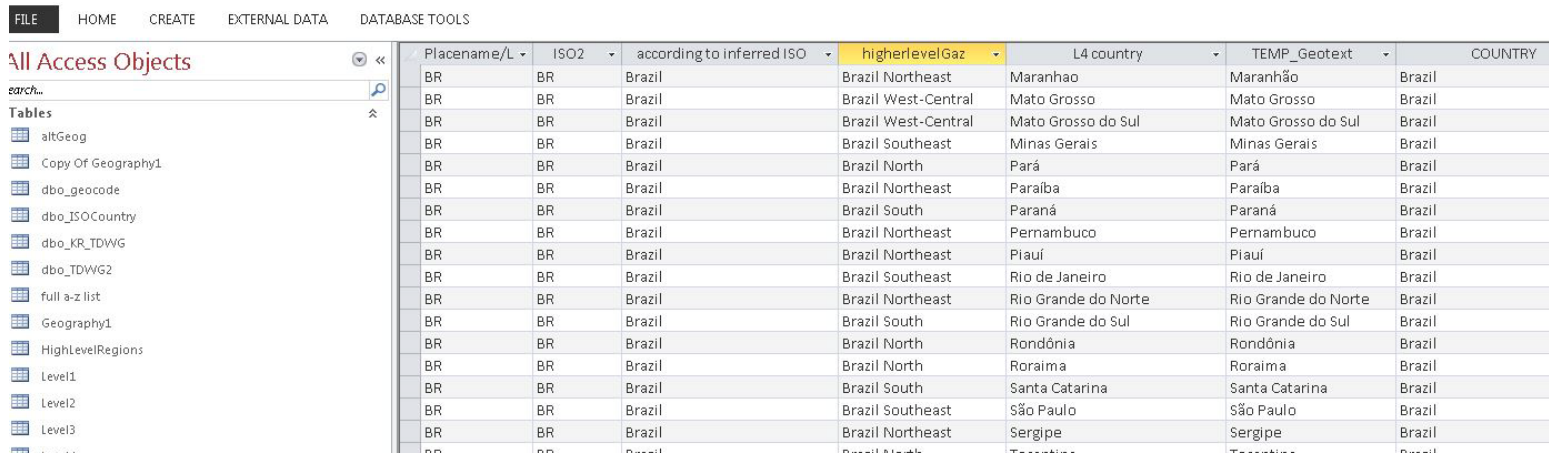
| Field | Transcription error | Identification error | |
|-----------|---------------------|-------------------------|------------------|
| | | Incorrect data / column | Data not entered |
| Collector | 0.5 | 0.5 | 0.5 |
| Addcoll | 0.5 | 0.5 | 0.5 |
| Number | 0.5 | 0.3 | 0.5 |
| ColIDD1/2 | 0.3 | 0.3 | 0.3 |
| ColIMM1/2 | 0.3 | 0.3 | 0.3 |
| ColIYY1/2 | 0.5 | 0.5 | 0.5 |
| Datetext | 0.3 | 0.3 | 0.3 |

| | | | | | |
|--|------|-------|-------|--------|--------|
| Amount of records/ Batch size | ≤150 | ≤ 280 | ≤ 500 | ≤ 1200 | ≤ 3100 |
| Sample size | 20 | 32 | 50 | 80 | 125 |
| Accepted if amount of errors | < 1 | < 1.5 | < 2 | < 3 | < 4 |

- e.g. For a batch of 440 records, 50 records are sampled. If error count more than 2 batch is rejected

Data Cleaning

- Concentrated on high priority fields e.g. Country
- Cleaned in batches using queries in MS Access
e.g. Missing country information added by looking in locality text field for matching entries in a geography table
- 16.8% of Dioscoreaceae records had missing country information – reduced to 8.6%



The screenshot shows the Microsoft Access interface with a table view. The table has the following columns: Placename/L, ISO2, according to inferred ISO, higherlevelGaz, L4 country, TEMP_Geotext, and COUNTRY. The data rows list various Brazilian states and their corresponding country codes.

| Placename/L | ISO2 | according to inferred ISO | higherlevelGaz | L4 country | TEMP_Geotext | COUNTRY |
|-------------|------|---------------------------|---------------------|---------------------|---------------------|---------|
| BR | BR | Brazil | Brazil Northeast | Maranhao | Maranhão | Brazil |
| BR | BR | Brazil | Brazil West-Central | Mato Grosso | Mato Grosso | Brazil |
| BR | BR | Brazil | Brazil West-Central | Mato Grosso do Sul | Mato Grosso do Sul | Brazil |
| BR | BR | Brazil | Brazil Southeast | Minas Gerais | Minas Gerais | Brazil |
| BR | BR | Brazil | Brazil North | Pará | Pará | Brazil |
| BR | BR | Brazil | Brazil Northeast | Paraíba | Paraíba | Brazil |
| BR | BR | Brazil | Brazil South | Paraná | Paraná | Brazil |
| BR | BR | Brazil | Brazil Northeast | Pernambuco | Pernambuco | Brazil |
| BR | BR | Brazil | Brazil Northeast | Piauí | Piauí | Brazil |
| BR | BR | Brazil | Brazil Southeast | Rio de Janeiro | Rio de Janeiro | Brazil |
| BR | BR | Brazil | Brazil Northeast | Rio Grande do Norte | Rio Grande do Norte | Brazil |
| BR | BR | Brazil | Brazil South | Rio Grande do Sul | Rio Grande do Sul | Brazil |
| BR | BR | Brazil | Brazil North | Rondônia | Rondônia | Brazil |
| BR | BR | Brazil | Brazil North | Roraima | Roraima | Brazil |
| BR | BR | Brazil | Brazil South | Santa Catarina | Santa Catarina | Brazil |
| BR | BR | Brazil | Brazil Southeast | São Paulo | São Paulo | Brazil |
| BR | BR | Brazil | Brazil Northeast | Sergipe | Sergipe | Brazil |

Main Successes

- NHM and Kew worked effectively together
- Communication generally good, but details of scheduling and implementation would benefit from more face to face meetings
- Good specimen care during digitisation
- High imaging throughput
- Collaboration allowed more testing of variables and workflows



Challenges Identified

- Logistics for imaging off-site resource heavy
- Picturae struggled with multi-specimen sheets
- A more effective protocol regarding guidelines for quality standards needed
- Time needed for uploading images and data into Institutional Systems



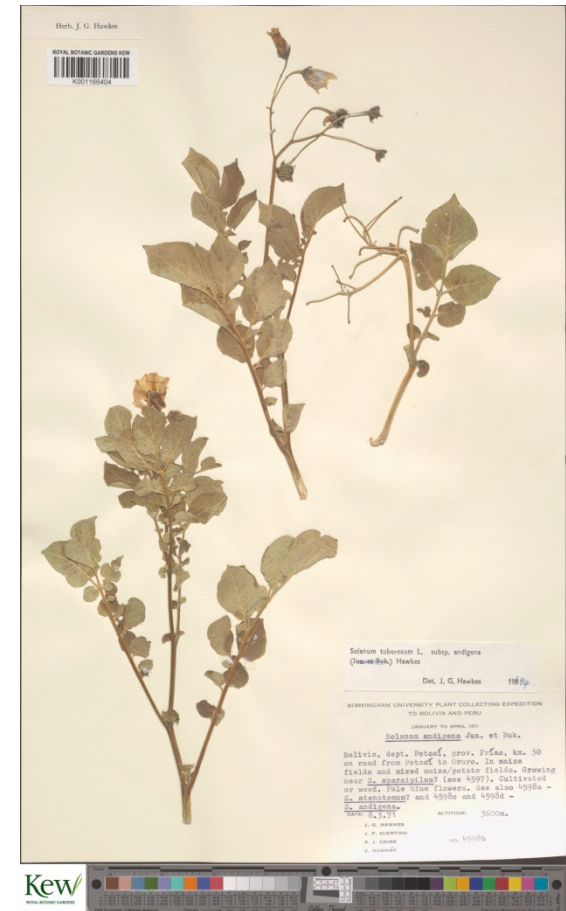
Costs

- Costs of stub record creation in house c 50%
outsourcing of imaging and capture of filed as names
- Total Cost per Sheet -
£2 or €2.5



Post Mass Digitisation Challenges

- Digitisation of new accessions
Kew C.30,000 a year
- Keeping the digital records up to date and aligned with physical collection
- Need to define resources needed for these activities
- Improve collection management system for more streamlined updating of specimen records



THANK YOU

Alan Paton, Laura Green, Andrew Budden, Sandy Knapp, Steve Cafferty, Jonathon Gregson, Jacek Wajer, Lawrence Brooks, Ben Atkinson, Charlotte Couch, Marie-Helene Weech, Anna Haigh, Joanne Osborne, Keith Lyons, Elizabeth Woodgyer, Xander Van der Burgt, Lee Davies, Lauren Phelan, Frances Crawford, Nina Davies, Sarah Blyth, Ken Bailey, Jeroen Bloothoofd, Noortje Wijkamp, Chrissie Hendricks, everyone at Picturae and Alembo.

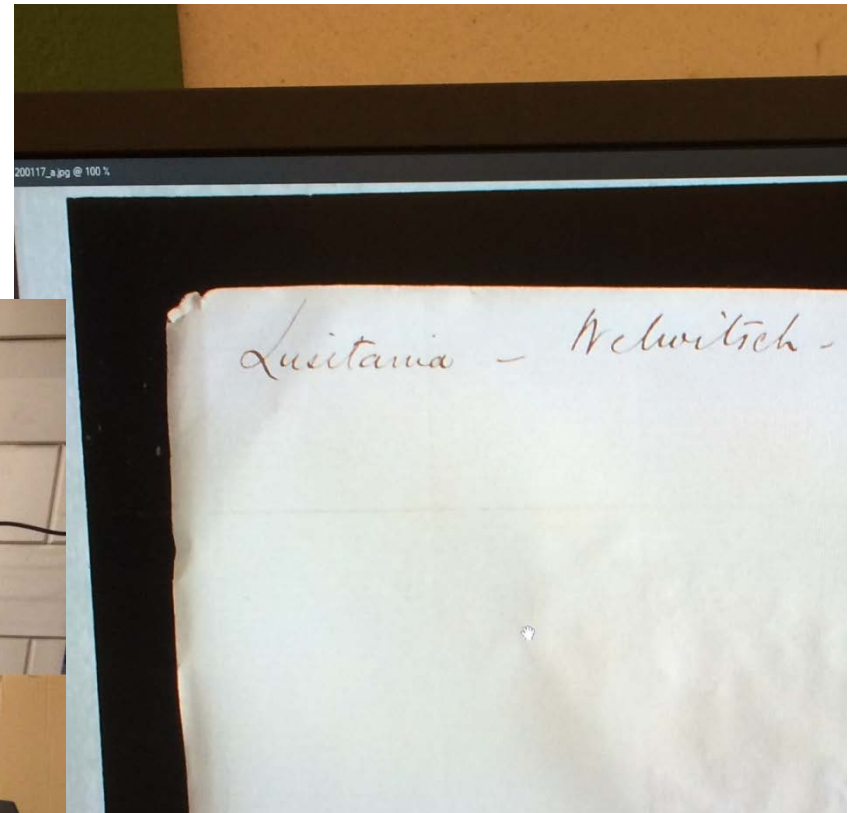


| Field | Transcription error | Identification error | |
|-----------|---------------------|-------------------------|------------------|
| | | Incorrect data / column | Data not entered |
| Collector | 0.5 | 0.5 | 0.5 |
| Addcoll | 0.5 | 0.5 | 0.5 |
| Number | 0.5 | 0.3 | 0.5 |
| CollDD1/2 | 0.3 | 0.3 | 0.3 |
| CollMM1/2 | 0.3 | 0.3 | 0.3 |
| CollYY1/2 | 0.5 | 0.5 | 0.5 |
| Datetext | 0.3 | 0.3 | 0.3 |
| Country | 0.5 | 0.5 | 0.5 |
| Locality* | 0.1/0.3 | 0.3 | 0.1/0.3 |
| Flags | 0.3 | 0.3 | 0.3 |
| Llunit | 0.3 | 0.3 | 0.3 |
| Lat/Long | 0.3 | 0.3 | 0.3 |
| NS/EW | 0.3 | 0.3 | 0.3 |

| Field | Unjustified use of interpretation marks |
|--------------|--|
| All fields | 0.3 |

| | | | | | |
|--|------------|------------|------------|-------------|-------------|
| Amount of records/ Batch size | ≤ 150 | ≤ 280 | ≤ 500 | ≤ 1200 | ≤ 3100 |
| Sample size | 20 | 32 | 50 | 80 | 125 |
| Accepted if amount of errors | < 1 | < 1.5 | < 2 | < 3 | < 4 |

Verso Imaging



Detailed description of specimen condition (noting any damage in particular): Specimen mounted on sheet with glue; Sheet clean and flat; capsule with paperclip containing a large number of flowers; two labels – one collection label in bottom right corner with a smaller yellow label mounted above; one collection tag attached to the specimen and glued to the sheet. A couple of leaves slightly damaged (possibly pre-mounting)

This section to include extra page(s) with annotated photograph(s)



Condition Reporting

1. At Kew/NHM
2. On arrival
3. After digitisation
4. Back at Kew/NHM