Incorporating OCR into a digitisation and curation workflow

Elspeth Haston, Robyn Drinkwater & Robert Cubey
Royal Botanic Garden Edinburgh

- Living collections, Herbarium and Library & Archives
- Nearly 3 million herbarium specimens
- $\frac{3}{4}$ million specimens databased
- $\frac{1}{4}$ million specimens imaged and online
Digitisation of herbaria: “gold standards” set by the Global Plants Initiative (600dpi + full data)

Shift to large scale digitisation with resulting changes in process necessary
Standard Workflow

1. Minimal curation

2. Assign unique identifier (eg attach barcode)

3. Initial minimal data capture

4. Image specimen (camera or scanner)

5. Additional data entry
Expanded Workflow

1. Minimal curation
2. Assign unique identifier (e.g., attach barcode)
3. Initial minimal data capture
4. Image specimen (camera or scanner)
   4a. Assess specimen condition
   4b. OCR processing
5. Additional data entry
5a. Additional curation
**Expanded Workflow**

1. Minimal curation
2. Assign unique identifier (e.g., attach barcode)
3. Initial minimal data capture
4. Image specimen (camera or scanner)
   4a. Assess specimen condition
   4b. OCR processing
5. Additional data entry
5a. Additional curation
4a. Assessing specimen condition

- Condition assessment now integrated into digitisation process
- Poster on display
Expanded Workflow

1. Minimal curation

2. Assign unique identifier (e.g., attach barcode)

3. Initial minimal data capture

4. Image specimen (camera or scanner)
   4a. Assess specimen condition
   4b. OCR processing

5. Additional data entry

5a. Additional curation
4b. OCR processing

- Optical Character Recognition (OCR) is now integrated into the digitisation process
- OCR output is currently captured in a single, unparsed field in the database
- ABBYY Recognition Server version 3
4b. OCR processing

- Barcode checking
  - Against image filename

- Sorting and filtering specimens
  - Significantly increases efficiency
  - Data have been added to over 100,000 specimens in presorted batches
  - Drinkwater et al., 2014

- Finding types
  - Terms including typus, sp. nov., etc.
  - Link to Global Plants Initiative project

- Sorting label type by uncertainty %
  - Potential to sort typed vs handwritten
4b. OCR processing in the future

OCR output could be used for:

• Text mining to pull out labels which appear to have no recognisable text and which are potentially handwritten
  – These specimens could go into a different workflow incorporating handwriting recognition

• Text mining to sort specimens automatically into batches
  – Addition of data becomes project based

• Comparing with OCR output from other institutes to find duplicates

• Parsing into database fields
Expanded Workflow

1. Minimal curation
2. Assign unique identifier (eg attach barcode)
3. Initial minimal data capture
4. Image specimen (camera or scanner)
   4a. Assess specimen condition
   4b. OCR processing
5. Additional data entry
5a. Additional curation
5a. Additional curation

• Correcting specimens which have been filed under the wrong geographical region
  – May be caused by same name in two countries, eg Tripoli
  – May be caused by changing borders
  – May be caused by lack of concentration

• Emphasises link between digitisation and curation
Collaboration

• iDigBio
  – Augmenting OCR Working Group

• Synthesys 3
  – Joint Research Activity (JRA) includes review and development of OCR within digitisation

• CETAF Digitisation Working Group

• Kew & Smithsonian
  – Finding duplicate specimens
  – Herbarium complementarity

• DINA

• OpenUp! & Europeana
Acknowledgements

• Deb Paul and the iDigBio aOCR Working Group
• Martin Pullan and Roger Hyam, RBGE
• Simon Chagnoux, Paris & Stephen Gottschalk, NYBG
• Nicky Nicolson, RBGKew & Rusty Russell, SI
• Andrew W Mellon Foundation
• Scottish Government
Dedicated to the digitisers at RBGE

Jaime Aguilar, Ruth Atkinson, David Bell, Boni Nieto Blazquez, David Braidwood, Cliodhna Ni Bhroin, Marie Briggs, Rebecca Camfield, Catherine Cooke, Anna Dennis, Robyn Drinkwater, Alan Elliott, Anna Dennis, Muhammad Ghazali, Lorna Glancy, Zoe Goodwin, Alan Gray, Lucy Head, Paulina Hechenleitner, Dorota Jaworska, Noreen Jennison, Neville Kilkenny, Cathy King, Lorna MacKinnon, Louise Olley, David Purvis, Markus Ruhsam, Nicky Sharp, Natalia de la Torre, Helen Townsend/Rupp, Julia Weintritt