Digitisation using Automated File Renaming and Image Processing

SYNTHESYS3 JRA

Task 1.1 - Automated data collection

Report Authors: Louise Allan[1], Ben Price[1], Laurence Livermore[1], and Vince Smith[1]

[1] The Natural History Museum, London

Last updated: 27 November 2017

Version: 1 (SYNTHESYS Report)

Table of Contents

Summary

Efficient mass digitisation requires pre-digitisation expert curation. By integrating temporary barcode labels during curation, which encode basic metadata, collections become available for downstream mass digitisation using automated file renaming, image processing, and bulk ingest into the collection management system. The automated file renaming process, developed during the Phthiraptera Slide Digitisation Project at the Natural History Museum (NHM), reduces the image post processing time and potential for human error when renaming files. This automated renaming process is accomplished by using two additional barcodes in each image that encode the Location and Taxon information. Temporary labels with these additional

metadata encoded barcodes were generated from the collection management system and inserted into the collection prior to digitisation.

Background

An inventory “specimen” record, for the purposes of collections management, requires at a minimum three aspects of metadata:

1. Unique identifier (UID) called the “Barcode” in the NHM’s collection management system
2. Taxon (eg. Species)
3. Location in the collection (eg. drawer / cabinet)

Two types of scripts can be used for the bulk ingest of images and metadata:

A. **Specimen record creation**
   This script takes individual images with metadata encoded in the filename and creates a specimen record with appropriate attachments to the relevant modules such as taxonomy and location.
   Example of format of encoded metadata: “UIDBarcode_Location_Taxon.jpg”

B. **Record attachment**
   This script takes individual images and attaches them to an existing record by matching the UID barcode value in the filename with an existing record in the collection management system.
   Example of the format of encoded metadata: “UIDBarcode_suffix.jpg”
   Note: suffix is used to ensure unique filename and can be changed depending on the nature of the image i.e. label only images, labels on reverse side of slide, high resolution specimen images etc.

The Phthiraptera Slide Digitisation used internal record numbers (IRNs) generated from the collections management system for the Taxon and Location information. The strength of this approach is that the use of IRNs ensures a 1:1 match in the collection management system; however, the weaknesses are that the manual renaming of files to numerical values is liable to human error, and numerical values cannot be easily verified before import as they are not easily human readable.

The previous Slide Digitisation Project, conducted by the NHM in 2015, developed a workaround for the manual renaming of files through the use of Inselect[1], where human readable values in a drop-down list were associated with the appropriate IRN values. Using Inselect specimens were tagged to their corresponding Location and Taxon information using

[1] https://github.com/NaturalHistoryMuseum/inselect/

these drop-down lists. The image files were then exported and renamed with the appropriate IRN metadata.

Limitations of this approach:

1. the specimens have to be manually tagged using the drop-down lists.
2. discrepancies between the drop-down list and the collection i.e. Location and/or Taxon missing from the list, thus halting the tagging process and increases the post processing time of specimen images.
3. manual tagging of specimen images can be prone to errors.
4. no attempt was made at verifying the numerical values as this functionality was not available and would have required a second image tagging and comparison step.

Automated File Renaming and Processing

In order to increase efficiency and accuracy of digitisation workflows automated processes are needed. During the previous Slide Digitisation Project (2015) automated processes were developed for (1) bulk transfer of image files from the imaging PC to the data managers, (2) bulk ingest of images and specimen record creation using metadata encoded in the file names, and (3) the clear-down of the original image files on the imaging PC once successfully ingested into the collections management system (Workflow 3, Appendix). This previous project, however, had a number of manual and semi-automated steps to process and rename the image files, which resulted in a substantial amount of time taken for post-processing before the image files were ready for transfer. The current Phthiraptera Slide Digitisation Project (2017) reused the previously developed file transfer and ingest scripts but also developed additional automated post-processing steps to increase the efficiency of specimen digitisation while reducing the potential for human error by using a Multiple Barcode Digitisation Workflow (Workflow 1, Appendix).

Automated file renaming was accomplished by using two additional barcodes in each image that encoded the Location IRN and the Taxon IRN (Workflow 1 and 2, Appendix). These IRN encoded barcodes, together with the UID Barcode, were used to rename the files using BardecodeFiler (purchased software that requires a licence for each imaging PC). The renamed image files are then rotated (if required due to camera's orientation) and cropped using XnConvert in order to remove the temporary IRN encoded barcode label from the final specimen image (Workflow 2, Appendix). The automated post-processing of images is streamlined further through the use of hot folders enabling the renaming and image processing to occur in the background during active digitisation.

In comparison to the 2015 Slide Digitisation Workflow the current image capture and processing rates are approximately 1.5 times faster (high performance rate: 2015 vs current project = 144 vs 228 slides per person per hour). Furthermore, by automating the renaming of image files human error can be kept to a minimum as the physical barcode labels are produced directly from the collections management system and then associated with each Location and Taxon before digitisation is started, ensuring all species and locations are accounted for in the collection management system before digitisation. This step essentially pushes the “tagging” of the specimens to the collection preparation step prior to image capture, and enables a second confirmation step when the slides are imaged, unlike the 2015 Slide Digitisation Project in which these steps were done during the post-processing stage.

Future Development

BardecodeFiler includes the ability to read multiple barcodes and regular expressions to reformat barcode values into appropriate file names, enabling modification of the Multiple Barcode Digitisation Workflow for other mass digitisation projects. The use of purchased software, such as BardecodeFiler, per imaging station must be taken into consideration when scaling up digitisation projects using this workflow. Unfortunately, as free barcode reading software is often prone to errors this is not feasible option when accuracy of barcode reading is imperative to the workflow.

The current workflow uses two pieces of software and hot folders to rename and process the images. As this process is open to error the development of a flexible barcode reading software would be beneficial. The solution would be to combine the cropping of XnConvert with the barcode reading of BardecodeFiler and the flexible templates and CSV export / file renaming of Inselect into a sustainable (open access) solution in order to implement the Multiple Barcode Digitisation Workflow on a larger scale.

Author Contributions

Louise Allan and Ben Price designed and developed the automated file renaming workflow using metadata encoded barcodes. Louise Allan developed the multiple barcode digitisation workflow. Laurence Livermore and Vince Smith conceived the idea for digitising the Phthiraptera (louse) collection, which was developed and supported by Louise Allan. Louise Allan compiled the report and all authors contributed to the revision of the report.

Acknowledgements

We would like to thank the following people for assistance during the Phthiraptera Slide Digitisation Project (2017): Olha Schedrina, Paul Brown, Steen Dupont, Vladimir Blagoderov, Sam Broom, Charlotte Barclay, Darrell Siebert, Chris Sleep and Paul Ward.

THIS PAGE MARKS THE END OF THE FORMAL REPORT

SUBSEQUENT PAGES CONTAIN APPENDIX INFORMATION

Appendix

Workflow 1 - Multiple Barcode Digitisation

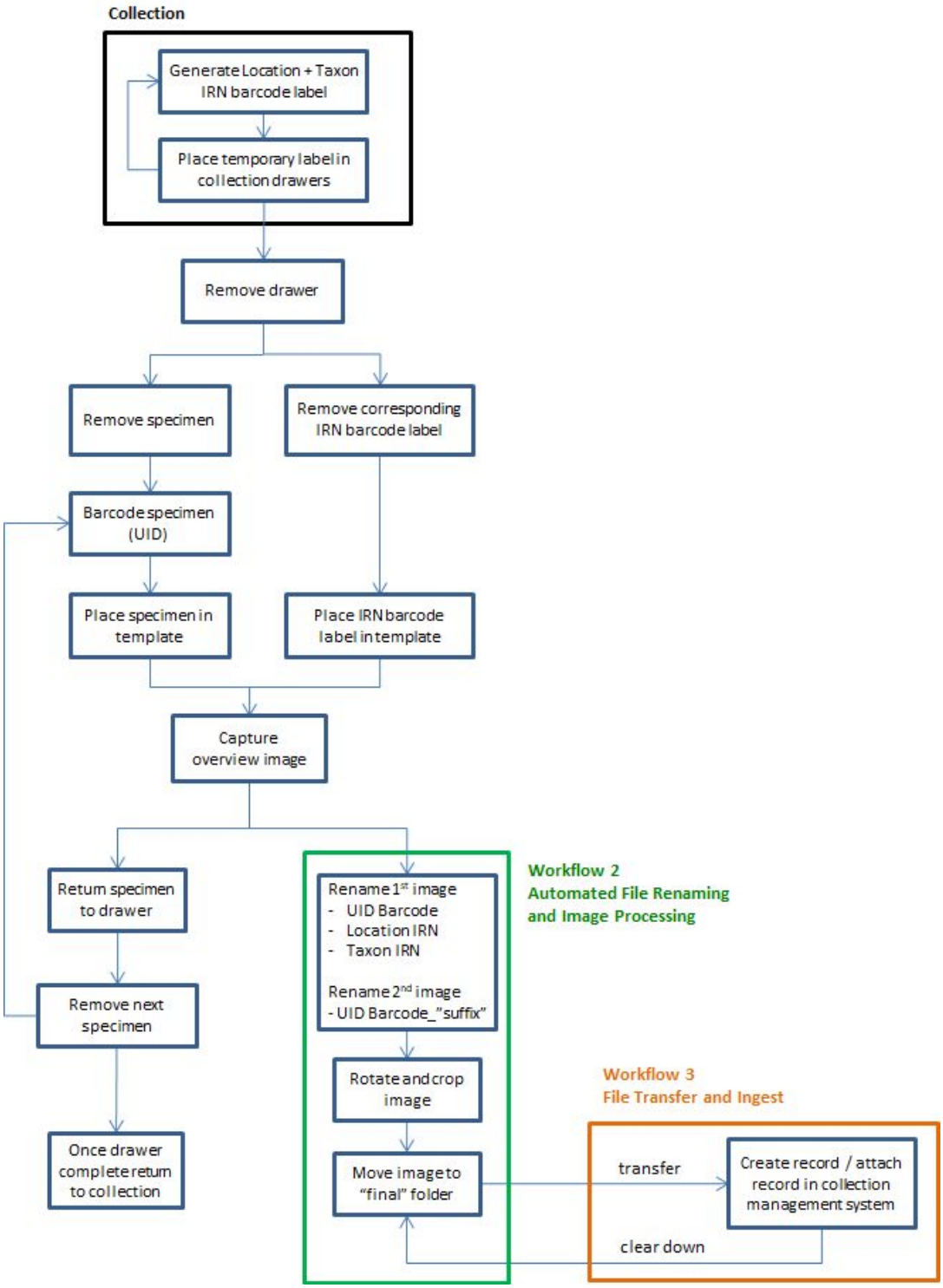

Figure 1. Workflow 1 - high throughput digitisation using multiple barcodes encoding metadata to enable automated file renaming and bulk ingest into a collection management system.

A) IRN barcode incorporation

1) Export metadata such as Location and Taxon information, for a particular collection from the collections management system (Figure 2a and b).
   (Two options, vertical and horizontal, should be implemented to standardise image import while allowing flexibility with various imaging techniques).
2) Print labels and insert them into the collection prior to digitisation (Figure 2c and d).

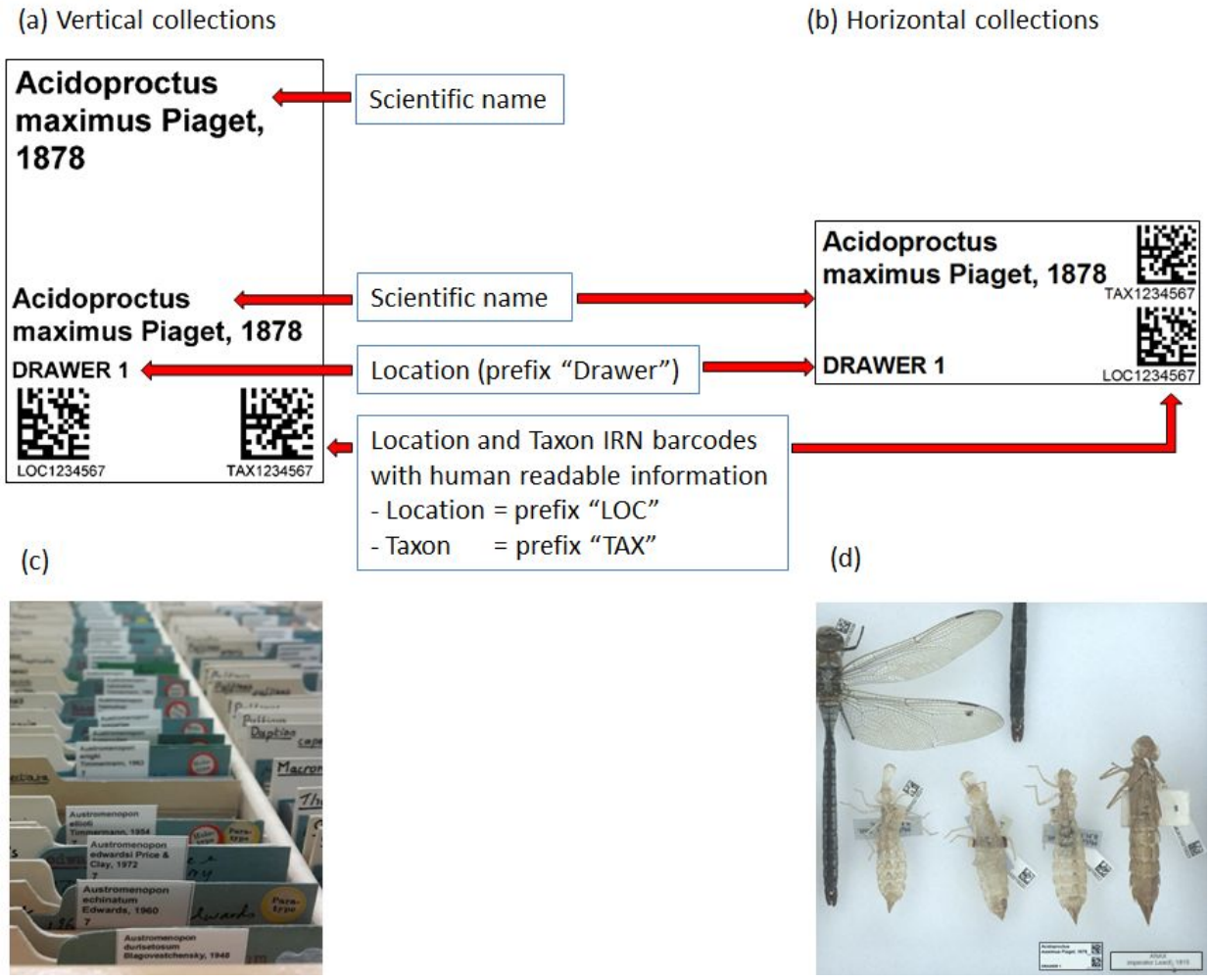

Figure 2. Example of a vertical (**a**) and a horizontal (**b**) temporary barcode label encoding the Location and Taxon IRNs; and the temporary labels inserted into a vertical (slide) collection (**c**) and horizontal (pinned specimen) collection (**d**) prior to digitisation.
Note: Taxon name repeated in vertical labels: lower name included in image, top name for viewing in vertical slide collection.

B) Imaging

3) Remove the specimen from the collection and place a unique identifier (UID) Barcode on the specimen (if not already present).
4) Place the specimen with its UID Barcode in the imaging template.
5) Remove the specimen's corresponding Location and Taxon IRN label from the collection and place in the imaging template.
6) Image the specimen (Figure 3), saving the image file to the hot folder "input".
7) Place the specimen back in the collection.

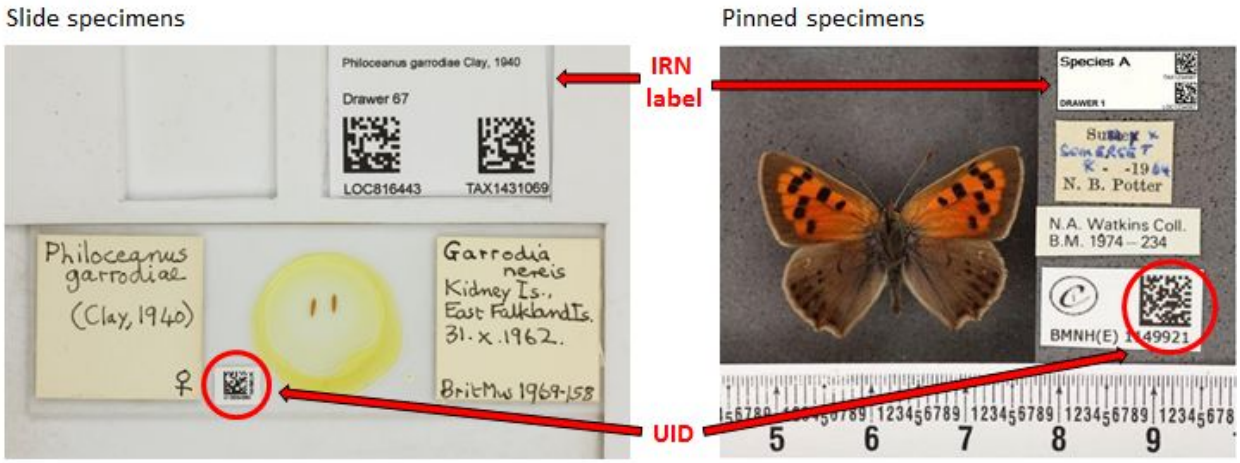

Figure 3. Example of imaging template with specimen (UID Barcode), and the corresponding Location and Taxon IRN label.
Note: original image is captured upside down and rotated using XnConvert.

Workflow 2 - Automated File Renaming and Image Processing

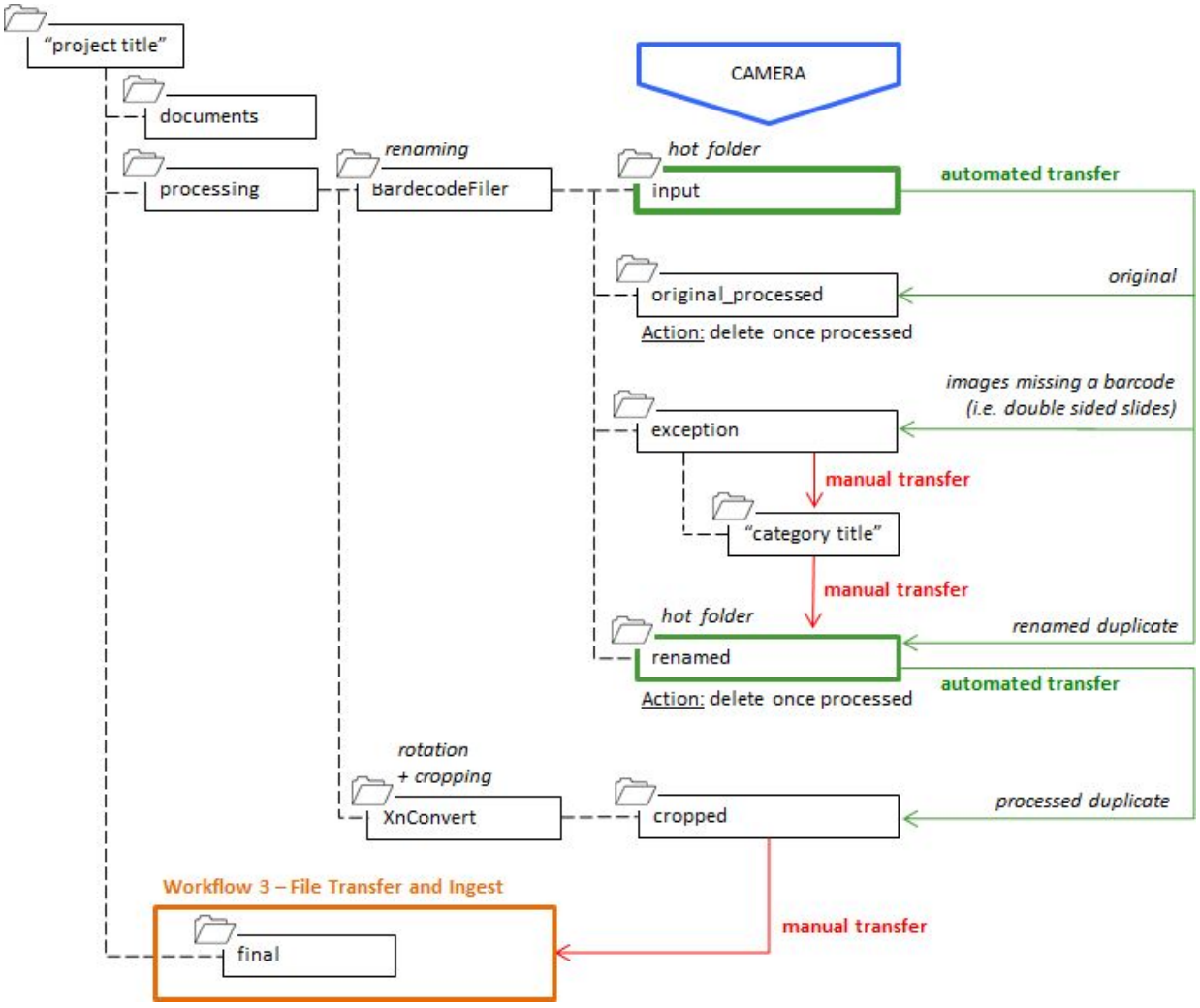

Figure 4. Workflow 2 - automated file renaming and image processing developed to run alongside the Multiple Barcode Digitisation Workflow for low resolution specimen imaging.
Note: Folder structure: dashed lines; Workflow: green lines automated steps; red lines manual steps.

C) File renaming using BardecodeFiler

8) BardecodeFiler watches the hot folder "input" for image files.
9) The image file is copied and the three barcodes (UID Barcode, Location IRN and Taxon IRN) in the image are read in a specific order according to the following rules:
   i) Value 1 greedy match: "010" followed by digits
   ii) Value 2 greedy match: "LOC" followed by digits
   iii) Value 3 greedy match: "TAX" followed by digits
10) The image is renamed "value1_value2_value3", where "LOC" and "TAX" are trimmed from value2 and value3, and will appear as follows: "UIDBarcode_LocationIRN_TaxonIRN.jpg"

11) The renamed image is then automatically saved to a second hot folder “renamed”, while the original file is moved from the folder “input” to “original_processed”.
12) If a barcode is missing in the image then the image file will be saved to the folder “exceptions” and will be renamed to “UIDBarcode_reverse”, where it uses the UID barcode from the previous image. This step was developed to deal with specimens where additional image(s) are needed but the UID Barcode will not be present in those image(s) i.e. slides with label information on the reverse side to that of the coverslip.

D) Image rotation and cropping using XnConvert

13) XnConvert watches the hot folder “renamed”.
14) The renamed image file is copied, then rotate 180° and cropped to specified coordinates to remove the temporary Location and Taxon IRN label from the final image (Figure 5).
15) The cropped image is then automatically saved to the folder “cropped”.
16) At the end of each day the renamed and processed image files are manually transferred from “cropped” to a “final” folder, ready for file transfer and ingest into the collection management system.
17) Image files in folders “original_processed” and “renamed” are manually deleted daily.

Figure 5. Example of final specimen image, rotated and cropped using XnConvert, ready for file transfer and ingest into the collection management system.

Workflow 3 - File Transfer and Ingest in the Collection Management System

E) File transfer

18) The images in the “final” folder are manually copied daily into a network share folder for retrieval by the data managers.

F) Ingestion of images and metadata

19) Metadata encoded in the filename of the image is used as follows:
    i) specimen record creation - for images with no existing record the unique identifier (UID) Barcode, Location and Taxon information are used to create an inventory record to which the image is attached.
    ii) record attachment - for images with an existing specimen record the UID Barcode is used to attached the image file to the corresponding specimen record.

G) Clear-down of image files

20) Once image files are successfully ingested into the collection management system the original images on the imaging PC are automatically deleted from the “final” folder.