Mass Digitization of a Scientific Biodiversity Collection

Olaf Slijkhuis

Picturae BV, Heiloo, The Netherlands
o.slijkhuis@picturae.nl

Abstract. In late 2012 the Dutch biodiversity center Naturalis published a tender for the mass digitization of herbarium sheets. Picturae proposed a solution with an industrial approach consisting of a conveyor belt construction with a fully automated processing workflow. From checking the technical parameters of the digital images to cropping and making all the derivatives. The digital images can be delivered to the network right after scanning. This approach makes it possible to produce high quality digital images without expertise of the operator.

Keywords: Biodiversity, Digitisation, Herbarium, Industrial, Workflow.

1 Introduction

In 2012 the Dutch biodiversity center Naturalis published a tender[1] for the digitization and data entry of herbarium sheets. It was clear from the start that if Picturae wanted to participate in the bidding process we had to come up with an innovative and efficient approach. To make the offer competitive the entire workflow was analyzed. This led to the idea of an industrialized approach, to standardize and automate the process where possible. Yet is has to be taken into consideration that the human element is still very important when digitizing cultural heritage collections. The material in herbarium collections is often brittle or stored in envelopes. This human intervention was incorporated at predefined moments in the workflow.

The new digitization system was built around a conveyor belt where the people care-fully handle the objects to position them for capturing and return them in the same order to the boxes. After careful consideration of the different stages of the workflow it was possible to automate everything else but the physical handling of the objects. The software to operate the digitization system, do the post-processing and the delivery of the digital files to the data-entry department and delivery to the client's servers was custom built for this project but has since involved into a multi-purpose tool which can be used for all our digitization equipment and post-processing workflows.

[1] Tender publication number: 2012/S217-357910

Fig. 1. Herbarium Digistreet.

2 Workflow Design

Picturae has gained a lot of experience from working for over fifteen years in the digitization of cultural heritage collections. Digitization systems are custom build to suit a variety of original material. One of the biggest challenges is to prevent errors which can only be solved by making a so called retake. A retake is making a new digital image of the original object because it has been rejected by our internal quality control or the client's quality control. This can be caused by a technical defect of the equipment or because of a cropping mistake. This retake process is cumbersome because it usually is not a standardized operation. To develop a workflow process which will prevent errors was one of the biggest challenges.

The digitization software was designed with automatic failsafe procedures to prevent technical defects of the digital images. The first step is to calibrate the hardware to make sure the system is performing according to agreed specifications. To calibrate the system you need parameters and the bandwidth in which these parameters can operate. These parameters are provided by the Dutch Metamorfoze guidelines and the US FADGI guidelines. Both guidelines are developed to test the performance of the digitization equipment. Technical targets[2] are used for analyzing color accuracy, sharpness, resolution, noise, tonal distribution, etc. and dedicated software is used to measure the targets and provide an objective result.

It is known that the specimen in a herbarium collection can have 3D characteristics. For this purpose we have developed a target to measure the depth of field (DoF). This is done by measuring slanted edge targets fixed in different heights. If the sharp-

[2] Digital ColorChecker SG by X-Rite.
QA-62 by Applied Image Inc.
Golden Thread by Image Science Associates.

ness results are within the bandwidth we know the system will perform within a depth of field range of 4 centimeters. Sufficient for the most extreme specimens mounted on a herbarium sheet.

Fig. 2. QA-62 Slanted edge targets used for DoF analysis.

These targets are chosen because they use either ISO standards and/or are widely accepted by the digital imaging community. From the start, over fifteen years ago, the use of technical targets have been a standard part of our workflow, but with the advance of software programs to objectively measure them, other opportunities presented themselves. A web based analyzing tool[3] was developed by Picturae to facilitate in an easy and efficient way of checking the system's performance. This technology was incorporated in the new workflow software for the herbarium sheet digitization system. Every day before the start of a production run targets are captured and analyzed to see if the system is still performing according to the parameters. The production can start after this first phase has been approved.

[3] http://delt.ae

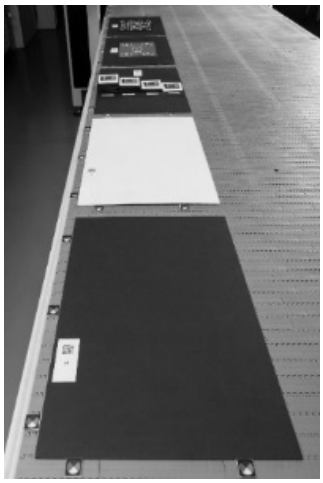

Fig. 3. Capturing the target sequence before production.

Another control measure is implemented by shooting a so called object level target in every image. This target is analyzed every time an image is shot. If one of the parameters is rejected the software will stop the system. The conveyor belt is transported back until the rejected item is in the right position to be digitized again. All images that were digitized afterwards are deleted and the process will start again with a new digital image. By building in this safety procedure we are sure the files which are going to be processed are technically sound and will not be rejected by the client's quality control for acceptance.

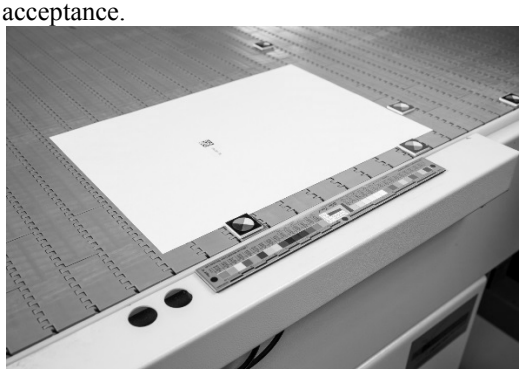

Fig. 4. Golden Thread Object level target 0.5x

Other errors by which images can be rejected in the quality control phase are caused by for instance file naming errors and post processing errors. To prevent file naming errors every scanned item has to have a unique identifier. For the digitization of the herbarium sheets we were allowed to stick data matrix labels on the sheets giving each sheet a unique identifier. The data matrix labels were also applied to the box and the coversheets to keep track of the contents and order of a box.

Fig. 5. Applying a data matrix label to the herbarium sheet.

These data matrix labels in turn can be detected and read by the workflow software. If there is no data matrix detected in the image it has no identifier and therefore cannot be saved. Again the conveyor belt will move backwards to position the item correctly in order to apply a data matrix label. After the item has been captured and approved (it has a data matrix and all the technical parameters are in accordance with the specifications) the file undergoes post processing. A color profile is applied, the item is cropped and all the necessary data is saved in the header of the file. If there is no mistake in the post processing stage the file is immediately ready for delivery to the client.

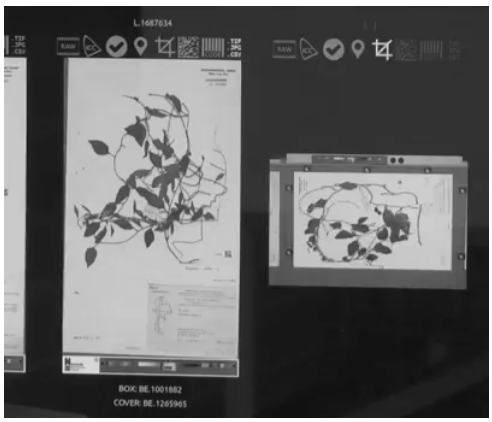

Fig. 6. Automated post-processing workflow.

A mock-up of the physical workflow was used to determine how many people were actually needed to do all the, still necessary, manual steps. It turned out that 4 people could operate the machine and produce an output of approximately 8000 images per 8 hour working day.

3 Data Entry

The technical metadata is inserted in the header of the files at the digitization stage as part of the workflow software. Another part of the project is to transcribe the information which is written on the original labels to make the herbarium sheets searchable. Since a large part of the labels are handwritten using OCR is out of the question. A dedicated team of data entry specialists were chosen to transcribe the labels manually. This is a still ongoing part of the project because of the volume of data that needs to be transcribed (over 4 million labels). The results are checked by staff of Picturae before handing over the definitive results to the Naturalis staff for final acceptance. Another possibility of transcribing handwritten text is using crowdsourcing technology. For the last two years Picturae has been working on a very successful crowdsourcing platform called VeleHanden[4] (translated as ManyHands). The digital images of another project of Naturalis were used as a pilot for describing biodiversity collections. The project involved the transcription of labels on 100,000 microscopic plates. This crowdsourcing project uses a double entry system with a control function. In the end 500 volunteers finished the description and checking of 300,000 items in nine months. All this metadata is automatically ingested in BRAHMS[5] (a widely used database in the botanical sciences). After this ingest the metadata is already linked to the digital files through the data matrix code and can be used immediately.

4 Conclusion

Analyzing the digitization workflow and breaking it down into distinguishable stages provides the opportunity to innovate the process on several levels. It also leads to a more industrial approach of digitizing cultural heritage objects. Yet an industrial approach doesn't mean a decrease in quality or the further deterioration of the original object instead it leads to an improvement in quality and more attention to the handling of fragile items. The installation of three similar systems made it possible to complete the digitization of over 4 million sheets in a record time of eight months. The completion of the data-entry is still continuing until mid-2015 and the online publication of the images with the accompanying metadata is soon to follow afterwards. Although the digitization process was very efficient the data-entry in comparison takes a very long time to complete. There is still room for innovation in this part of the project. For instance OCR of handwritten text is an efficient way of data-entry. This technology is researched at the University of Groningen in the Netherlands and the results are promising but not yet applicable on such a large and diverse scale.

[4] http://velehanden.nl

[5] http://herbaria.plants.ox.ac.uk/bol/brahms/Software