

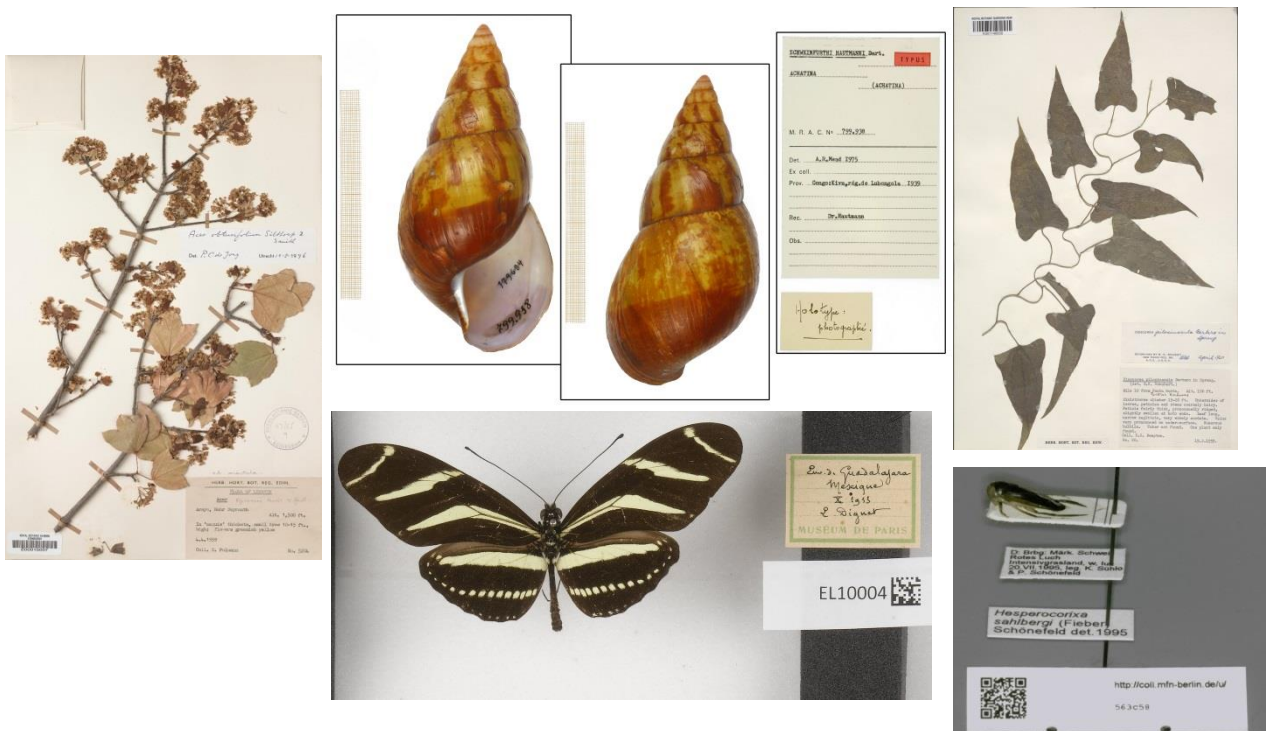


Project:	Synthesis of systematic resources
Project acronym:	SYNTHESYS3
Grant Agreement number:	312253
Workpackage:	WP4: Joint Research Activity (JRA)
Deliverable number:	Deliverable 4.2
Deliverable title:	Automating data capture from natural history specimens
Deliverable author(s):	Haston, E, Albenga, L, Chagnoux, S, Drinkwater, R, Durrant, J, Gilbert, E, Glöckler, F, Green, L, Harris, D, Holetschek, J, Hudson, L, Kahle, P, King, S, Kirchhoff, A, Kroupa, A, Kvacek, J, Le Bras, G, Livermore, L, Mühlenberger, G, Paul, D, Philips, S, Smirnova, L, Vacek, F
Date:	18 September 2015

AUTOMATING DATA CAPTURE FROM NATURAL HISTORY SPECIMENS

SYNTHESYS3 WORK PACKAGE 4 (JOINT RESEARCH ACTIVITY)

TASK 1.2 - AUTOMATIC METADATA CAPTURE



SUMMARY FROM THE DESCRIPTION OF WORK (17 JAN 2013)

Develop software that will automatically identify properties of an image. These data “facets” will be automatically captured without human intervention and provide categories of information that allow Users to easily search and browse virtual collections more effectively.

Specimen label data will be subjected to Optical Character Recognition (OCR) software to extract the text string and research methods to improve the accuracy of OCR use on handwritten labels. OCR-extracted text collected from handwritten labels will need to be subject to further processing and validation, such as via crowdsourcing methodologies (obj. 2).

D4.2) Optimal automated metadata capture: Report on optimal automated metadata capture for natural history collections [month 24]

EXECUTIVE SUMMARY

The need for the development and implementation of tools to speed up the process of data capture from natural history specimens is described in the Introduction. Through this project, the eight partners involved in this Task have collaborated to review, trial and develop tools for Optical Character Recognition (OCR), Natural Language Processing (NLP), Handwriting Text Recognition (HTR), template matching and pattern recognition.

The work was divided into four sections, each focussing on a different technology and process.

Section 1: Review of development of tools and workflows which incorporate automatic or semi-automatic metadata capture using Optical Character Recognition (OCR).

- Partners analysed data from existing trials of OCR software and, using the results as a guide, carried out additional trials using six OCR software programmes.
- OCR processing was carried out by three institutes. The images were provided by six partner institutes as well as a set from several US institutes supplied by iDigBio (Integrated Digitization of Biocollections). The images represented a range of material including plants, insects, molluscs and fossils.
- The results emphasise the usefulness of using OCR technology in the digitisation workflow, and discovered two options which provide the best results.
 - A server-based option (ABBYY Recognition Server v3)
 - A PC option (ABBYY FineReader v12 Professional)
 - Two online service options (Onlineocr.net and Newocr.com) were the best of the online services but did not perform as well as the ABBYY software.
- For some specimens the OCR output was up to 100% correct when compared to manual transcriptions.
- An example of a workflow incorporating OCR processing is presented
- The use of OCR output text and its integration into collections database software is an ongoing challenge. Some institutes are using the text successfully as a search tool to find specimens which have been databased with very minimal data or for pulling together batches of specimens for manual data entry.

Section 2: Review of development of Natural Language Processing (NLP) for parsing OCR text into Darwin core fields

- A short review of the current state of progress was carried out which discovered some of the key projects and individuals involved in this area
- Contact was made with Ed Gilbert of Arizona State University and Symbiota
- Arrangements are now being made to test three Portals which have incorporated the use of NLP in their workflow

Section 3: Review of Handwritten Text Recognition (HTR) and (semi) automatic specimen image classification, i.e. (semi) automatic tagging of specimen images from certain collectors or expeditions, using template matching software

- In Part 1 of this section, work was carried out to determine whether specimens could be automatically classified based on the classification of features holding data. In Part 2, software developed by tranScriptorium for historical handwritten documents was tested on natural history specimen labels
- A case study was based at the herbarium of the Botanic Garden and Botanical Museum Berlin-Dahlem (BGBM) as part of StanDAP-Herb, a joint project with the University of Applied Sciences, Hannover. 'Linienextraktor', the software used for this study, implements feature recognition algorithms that can be used on herbarium specimens herbarium specimens
 - This software was used to detect 'Herbarium botanicum Berolinense' specimens from a set of 465 randomly selected images (81 correctly found and 1 missed) and to then detect all specimens of Dr Albert Peter from a set of 916 specimens (906 correctly assigned).
 - The study produced a series of recommendations and guidelines for future work using this software, including the required resolution of the images
- Contact was made with a separate EU-funded FP7 project, tranScriptorium, who have developed software incorporating Handwritten Text Recognition (HTR) technology
- One of the tools developed within the tranScriptorium project, Transkribus, was installed locally by four partners for consideration. Two HTR training sets were uploaded, marked up and transcribed: 136 specimens collected by George Forrest (RBGE) and 200 specimens collected by Kerr (RBGK)
- The Transkribus team then processed these datasets to create HTR models for each collector and additional datasets were then uploaded, marked up and then processed using the appropriate HTR model
 - In total 750 specimens which had been databased with very minimal data were processed, this resulted in opening up some of the data in these collections which can now be searched
 - The results were promising, suggesting that further collaboration between tranScriptorium and natural history collections would be beneficial including exploring the use of crowdsourcing to help with the marking up process

Section 4: Review of automatic capture of character including colour, shape as well as exif data.

- Within this section, work focussed on analysing specimen images to capture non-text specimen data.
- A series of open source prototypes were developed by NHM to do the following:
 - segment specimens from their backgrounds and segment regions of interest (eg, particular body parts)
 - detect morphological features to be used for classification (eg, markings that indicate gender)
 - calculate of physical dimensions from images (eg, wing length)
 - colour analysis to be used for classification (eg, wing colours)
 - heat maps for regions of interest
- The code for these tools is available in a GitHub repository:
https://github.com/NaturalHistoryMuseum/insect_analysis
- In conjunction with Work Package NA2, trials are being carried out by RBGK and RBGE using the colour analysis algorithm to identify any correlation between leaf colour and quality of DNA and to determine whether the tool can be used to aid material selection for sequencing

CONTENTS

Summary from the Description of Work (17 Jan 2013)	1
Executive Summary	2
Introduction	8
Section 1: Review of development of tools and workflows which incorporate automatic or semi-automatic metadata capture using OCR.....	10
Introduction.....	10
Trial 1: Comparing a range of OCR software tools	10
Materials and Methods	10
Results.....	13
Discussion	15
Trial 2: Comparing OCR tools being used in herbaria	16
Materials and Methods	16
Results.....	19
Discussion	23
Trial 3: Multiple OCR trials of diverse specimens.....	24
Materials and Methods	24
Results.....	25
Workflows: Incorporating OCR into digitisation workflows.....	28
Discussion	30
Section 2: Review of development of NLP for parsing ocr text into Darwin core fields.....	31
Introduction.....	31
Review	31
Section 3: Review of (semi) automatic specimen image classification, i.e. (semi) automatic tagging of specimen images from certain collectors or expeditions, using template matching software.....	33
Part 1: Semi-automated Classification of Herbarium Specimens by means of Template Matching Algorithms	33
Part 2: Review and trials of Handwritten Text Recognition (HTR)	34
Introduction	34
Materials and Methods	35
Results.....	39
Discussion	45
Section 4: Review of automatic capture of character including colour, shape as well as exif data.	46

Part 1: Computer vision for specimen classification	46
Summary	47
Tools Used	48
Software Prototypes	48
Specimen segmentation	48
Method	48
Morphological feature detection	51
Calculating physical dimensions	54
Colour analysis	55
Heat maps for regions of interest.....	56
Dissemination.....	56
Links	56
References.....	56
Part 2: Correlation of leaf colour and DNA quality	57
Introduction	57
Materials and Methods	57
Results.....	57
References	58
Software and Projects	58
Appendix 1A: Settings for ABBYY Recognition Server v3 at RBGE	60
Appendix 1B: Trial 2 - Summary of OCR output for one specimen from each institute.....	63
Appendix 1C: Settings for ABBYY FineReader v12 Professional at RBGK	77
Appendix 1D: File preparation at RBGK	78
Appendix 1E: Scores for each specimen from each institute by word	79
Appendix 1F: OCR Software Results from RBGK testing of different formatting options	101
Appendix 2: Screenshots of portals using.....	105
Appendix 3: Protocol for using Transkribus for natural history collections	107
Introduction.....	107
Step 1: Register and download software.....	107
Step 2: Log in	108
Step 3: Upload documents to your private collection.....	108
Step 4: Segment your document into text blocks and baselines	109
Step 5: Manually transcribe a training dataset of 100 pages.	111

Step 6: Training the HTR model.....	113
Step 7: Running the HTR model.....	113
Appendix 4: Protocols for sampling and extracting DNA from herbarium specimens at RBGE	115
DNA Extraction Methodology: using the QIAGEN automated QIAxtractor	116

INTRODUCTION

There are an estimated 1.5 to 3 billion specimens held in natural history collections around the world (Smith & Blagoderov, 2012; Arino, 2010; Duckworth et al., 1993). A relatively small number of these specimens have been electronically catalogued with data accessible online. Making specimen data accessible is a priority to enable their inclusion in critical research to discover, document and conserve the world's biodiversity.

The aims of this task were to discover and develop tools to make specimen data capture more efficient. These tools will enable us to bring the goal of opening data for millions of specimens achievable.

The specimens in natural history collections can be considered as an aggregation of: a) a physical item; b) the label data attached to the specimen; c) the curatorial data which often does not appear on the label; and d) supplementary data held in other repositories (Haston et al., 2012). Each of these elements has an impact on how the specimen can be used for research.

The physical item provides the evidential basis for a large part of the data. It can be used to verify the taxonomic identity of the specimen, as well as enabling researchers to verify published trait data and record new trait data. Imaging the specimen will enable some of this work to be carried out remotely. The level of research which can be undertaken will depend of the kind of specimen (important traits for some taxonomic groups are more visually accessible than others) and the kind of image (resolution, 2D vs 3D, internal scanning). Examining and measuring specimens can be a slow process, often taking up a large part of a research project. By imaging the specimen, it may be possible to use automated or semi-automated tools to speed up this process of data capture. The analysis of images in relation to data capture has been explored here.

The label data attached to the specimen consist of the original collection information as well as later annotations such as identifications, destructive sampling use and nomenclatural data. The information may be handwritten, typed or printed. The amount of data present on specimens is highly variable, with many early specimens containing very little collection information, although they may have a higher number of important annotations. Imaging the labels can open access to these data very rapidly but will push the time required to capture the data in electronic format to the user rather than the provider. This will usually result in duplication of effort if more than one researcher wants to use the data. Optical Character Recognition (OCR) and Handwritten Text Recognition (HTR) software have the capacity to help speed up the process of label data capture by the provider, and their use has been tested and reviewed here.

The importance of curatorial data, which includes information such as filing name and filing geographical region, can be overlooked in planning digitisation programmes. These data are frequently not visible in the image of the specimen or labels, but they may provide critically important information relating to the identity and the collecting locality which are required for both research and curation of the specimens. These data are a key part of the digitisation process but,

whilst tools are being developed to speed up this part of data capture, this area was not included in the work of this project.

Supplementary data held in other repositories include collectors' journals, published research based on the specimens, molecular data, trait data, correspondence and botanical illustrations of the specimen. The written forms of supplementary data may be handwritten, typed or printed. Supplementary data are often not easily discoverable and opening access to the specimens will help discover and enable the links to be made between the data and the physical specimen. OCR software has been used for automating text capture from books and journals for some time, but is not yet extensively used for non-published literature. HTR software is being developed for the use of handwritten text including botanical text by an EU funded project (tranScriptorium) and has the potential to be more widely used for journals and correspondence.

The aim of the digitisation process is to enable people to discover and use the digital object as well as any associated data. For this to happen, digitisation cannot be about simply taking a photograph of the specimen. The digital images need to be held in a management system, curated, made freely available online and linked to the associated data to aid discoverability. The data need to be in a standardised structure, securely preserved, curated and made available in both human and machine readable format.

Here we report on the tools that are available to automate the electronic data capture from labels and some formats of supplementary data. We also report on some of the obstacles to implementation that currently exist for natural history institutes. Details of tools and projects are given in the References.

The report is divided into four sections corresponding to different tools and processes to capture label data, supplementary data and data from the image of the physical object:

Section 1: Review development of tools and workflows which incorporate automatic or semi-automatic metadata capture using OCR.

Section 2: Review of development of NLP for parsing ocr text into Darwin core fields

Section 3: Review of HTR and (semi) automatic specimen image classification, i.e. (semi) automatic tagging of specimen images from certain collectors or expeditions, using template matching software

Section 4: Review of automatic capture of character including colour, shape as well as exif data.

SECTION 1: REVIEW OF DEVELOPMENT OF TOOLS AND WORKFLOWS WHICH INCORPORATE AUTOMATIC OR SEMI-AUTOMATIC METADATA CAPTURE USING OCR.

INTRODUCTION

The development of optical character recognition (OCR) software has focussed to date on the recognition of typewritten or printed text. Herbarium collections started in the 16th century, the oldest surviving herbarium dating back to 1532. Although printing for publications was used when even the earliest surviving specimens were being collected, the nature of specimen labels has not facilitated the use of printing for their production. Most specimens were annotated with handwritten text for hundreds of years. As the use of printing became more available, some institutes and individual collectors encouraged the use of preprinted labels which would be completed by hand. The first working typewriter was built in 1808, but it was not until the 1880s that the typewriter started to achieve more widespread use in offices. Their use for specimen labels for herbarium specimens started from the early 20th century. Handwriting labels persisted from this time, although becoming less common. Printed labels started becoming the standard after the introduction of computer printers in the 1970s.

This project has concentrated on the use of OCR for typewritten and printed text. Whilst the main benefit of OCR will therefore be for specimens collected after 1900, it can still prove to be extremely useful for earlier, pre-printed labels which often contain the name of the collector, the country and a year. Additional investigations are being carried out on the use of automated handwriting recognition tools and are included later in this report.

Tests were carried out to determine the accuracy, effectiveness and usefulness of OCR software for natural history specimens.

TRIAL 1: COMPARING A RANGE OF OCR SOFTWARE TOOLS

MATERIALS AND METHODS

In order to investigate the wider range of OCR software options available, one set of five specimen images were processed through 19 different OCR tools (Table 1). The original, unformatted specimen images were processed. In addition, each specimen image was cropped, retaining the collection label only, and these cropped images were then also processed. The processing of specimens in this trial was carried out in 2012. The results were not fully analysed at the time and the data have therefore been brought into the SYNTHESYS3 JRA project for analysis and to help inform the additional trials being undertaken.

OCR Software	url	OS	Online	Free	Reads barcode
ABBYY FineReader	http://www.abbyy.com/finereader/	Windows, Mac OSX	Y, also Desktop	N	Y
ABBYY Recognition Server Version 2	http://www.abbyy.com/recognition-server/	Windows server		N	Y
ABBYY Recognition Server Version 3	http://www.abbyy.com/recognition-server/	Windows server	N	N	Y
Cuneiform	http://www.filesriver.com/app/107/openocr	Windows, MacOS X, Linux	N	Y	
Custom OCR	http://www.customocr.com/	n/a	Y	Y	
Free OCR	http://www.free-ocr.com/	Windows	N	Y	
Free online OCR	http://www.free-online-ocr.com/	n/a	Y	Y	
GImageReader	http://dottech.org/21372/gimagereader-open-source-google-powered-ocr-optical-character-recognition-program-that-actually-works/	Linux, Windows	N	Y	
I2OCR	http://www.i2ocr.com/	n/a	Y	Y	
OCReXTRACT	http://www.cvisiontech.com/ocr/best-ocr/best-ocr-extract.html	n/a	Y	Y	
OCRonline.com	http://www.ocronline.com/				
Omni Page Professional 18	http://www.nuance.co.uk/for-business/by-product/omnipage/standard/index.htm	Windows	N	N	N
Presto!OCR Pro 4.0	http://us.newsoft.com.tw/company/news_style.php?NT_Id=1&N_Id=313	Windows			N
Pumanet	http://pumanet.codeplex.com/	Windows	N	Y	
Salix	http://daryllafferty.com/salix/	Windows (difficult to find other info) Linux?	N	Y	
Scanitto	https://www.scanitto.com/	Windows	N	N	
Simple OCR	http://www.simpleocr.com/	Windows	N	Y	
TopOCR	http://www.topocr.com/	Windows	N	Y	
TypeReader	http://www.expervision.com/ocr-software	Windows	Y	N	N?
WeOCR	http://weocr.ocrgid.org/	n/a	Y	Y	

Table 1a: OCR Software options which were tested. The information relating to the software was gathered in 2012 and may have changed since then.

OCR Software	Input formats	Output formats	Notes
ABBYY FineReader	BMP, PCX/DCX, JPEG, JPEG2000, JBIG2, PNG, GIF, TIFF, PDF, DjVu, WDP	DOC, DOCX, XLS, XLSX, PPTX, RTF, PDF, HTM, CSV, TXT, ODT, DjVu, EPUB, FB2	Input & Output formats may be slightly different for MacOSX.
ABBYY Recognition Server Version 2	BMP, PCX/DCX, JPEG, JPEG2000, JBIG2, PNG, GIF, TIFF, PDF, DjVu, WDP	DOC, DOCX, XLS, XLSX, RTF, XML, PDF, HTML, CSV, TXT, TIFF, JPG, J2K	
ABBYY Recognition Server Version 3	BMP, PCX/DCX, JPEG, JPEG2000, JBIG2, PNG, GIF, TIFF, PDF, DjVu, WDP	DOC, DOCX, XLS, XLSX, RTF, XML, PDF, HTM, CSV, TXT, TIFF, JPG, J2K	
Cuneiform	JPG (difficult to find other info)	TXT (formatted, table and unformatted), HTML, FED, RTF, DBF	
Custom OCR	JPG, PNG, TIFF	Copy and paste from online result	Based on Tesseract OCR engine
Free OCR	PDF, JPG (unable to find more information)	TXT (unable to find more information)	
Free online OCR	PDF, GIF, BMP, JPEG, TIFF, PNG	DOC, PDF, TXT, RTF	
GImageReader	JPEG, GIF, PNG, TIFF, PDF	Copy from programme, download as TXT	
I2OCR	TIF, JPEG, PNG, BMP, GIF, PBM, PGM, PPM	Copy and paste into desired format, download as DOC	Max size 10MB
OCReXtrACT	PDF, TIFF, BMP, PBM/PGM/PPM	Copy and paste from online result Download TXT	5MB file size limit Based on Tesseract OCR engine
OCRonline.com			
Omni Page Professional 18	DOC, XLS, PPTX, RTF, WPD, TIF, JPG, BMP, PCX, GIF, PDF, MAX	DOC, XML, DOCX, XLS, XLSX, PPTX, TXT, CSV, PDF, XPS	
Presto!OCR Pro 4.0	BMP, PCX, DCX, JPEG, TIFF, PNG, couldn't find out about text formats	RTF, TXT, DOC, CSV, XLS, DBF, PD, HTML	
Pumanet	BMP, GIF, EXIG, JPG, PNG, TIFF	TXT, RTF, HTML	Cuneiform wrapper
Salix	JPG (difficult to find other info)	Copied from software Possible to Parse to DwC	Based on Tesseract OCR engine? Or Abbyy
Scanitto	BMP, JPG, TIFF, JP2, PNG	TXT, RTF	
Simple OCR	TIFF, TWAIN standards	TXT, RTF	
TopOCR	GIF, JPEG, TIFF, BMP	PDF, TXT, RTF, HTML	
TypeReader	JPG (unable to find other information)	.DOC, .XLS (although I couldn't get it to save in these outputs)	
WeOCR	BMP, JPEG, PBM/PGM/PPM	TXT (Unicode UTF-8, Latin 9 ISO-8859-15)	Based on Tesseract OCR engine

Table 1b: OCR Software options which were tested. The information relating to the software was gathered in 2012 and may have changed since then.

Transcription & OCR processing

Each specimen was manually transcribed and the number of lines of text counted. The OCR processing was carried out using the OCR software versions which were available in 2012. There are updated versions for most of the software tested and performance may have changed. In general, all the OCR processing was carried out on a jpg version of the full specimen. For some software the size of the jpg had to be reduced and in some cases, the OCR would only run if the label was selected manually.

Marking up and scoring OCR output by line

The OCR text output was compared to the original transcription and marked up (Figure 1). For each line, three primary levels of accuracy were used, 100% correct, partially correct and no correct text. Within the partially correct group, the correct and incorrect text was marked up as well as any text which had not been included in the original transcription (eg, text on the ruler and colour chart). The OCR text output was then scored by line: 1 for 100% correct, ½ for partially correct and 0 for no correct text.

RESULTS

The results of the OCR processing for the various software options tested varied across the specimens and no one OCR software tool performed better than any other for every specimen. For the full specimens, the two options which scored the highest for more than one specimen were ABBYY Recognition Server v.3 and OnlineOCR.net (both highest for two specimens) (Table 2).

OCR Software	E00037202 full (/14)	E000150 07 full (/14)	E00262827 full (/10)	E00262858 full (/10)	E00448970 full (/12)
ABBYY Recognition Server Version 2	11	1	5	1	1
ABBYY Recognition Server Version 3	13	6.5	8	6	11.5
Google	1	4	1	3	0
Free-ocr.com	10.5	9	4	4	9
OnlineOCR.net	11	10	6	9	10
Puma	5	5	2	2.5	9
Cuneiform	2.5	3	1.5	0.5	9
Simple OCR	7	0.5	3.5	0.5	5
Free online OCR	1	2.5	0.5	Not able to process	2
I2OCR	8.5	7	2	1.5	8
Newocr.com	8	10	3	2.5	4.5
WeOCR	9.5	5.5	failed	1	Not able to process
OCReXTRACT	11.5	8.5	3	2	8.5
Custom OCR	9.5	5.5	2	0	6
Salix	8	4.5	2	2	4.5
GImageReader	12	8.5	1.5	2.5	9.5
Scanitto	10	10	1.5	3	7

Free OCR	6	7	4	2	5.5
TopOCR	9.5	4	3.5	3	9
TypeReader	11.5	11	1	3	10
ABBYY FineReader	14	9.5	7.5	7.5	10
Omni Page Professional 18	12	1.5	6	7	9.5
Presto!OCR Pro 4.0	11	7	4	5.5	8.5

Table 2: OCR text output scores by line for full specimen images.

For the cropped specimens, there were only two software options which scored the highest for any of the specimens. ABBYY Recognition Server v.3 (highest for four specimens), and ABBYY FineReader (highest for one or three specimens) (Table 3).

OCR Software	E00037202 cropped (/14)	E00015007 cropped (/14)	E00262827 cropped (/10)	E00262858 cropped (/10)	E00448970 cropped (/12)
ABBYY Recognition Server Version 2					
ABBYY Recognition Server Version 3	12	12	7	10	11.5
Google	8	5	1.5	3.5	8
Free-ocr.com	9	0	0	0	0
OnlineOCR.net	10	10	5.5	7	8.5
Puma	11.5	6.5	2	1	7.5
Cuneiform	11.5	6.5	2	1	7
Simple OCR	8.5	3.5	0	2	0
Free online OCR	1	3	0.5	0.5	9
I2OCR	9.5	0	0	0	8
Newocr.com	9.5	6.5	5	4.5	9
WeOCR	7.5	1.5	1.5	2	7
OCRextrACT	11.5	0	0	Server error	8
Custom OCR	9.5	0.5	0	0	9
Salix	7.5	0.5	3.5	4	6.5
GImageReader	9	9	2	5	7.5
Scanitto	11	0	0.5	4.5	10
Free OCR	9.5	1	0	0	8.5
TopOCR	11.5	5	5	4.5	9
TypeReader	11	0	0	0	9
ABBYY FineReader	14	9	9	7.5	11.5
Omni Page Professional 18	11.5	11.5	7	6	7.5
Presto!OCR Pro 4.0	11.5	11.5	6	5.5	10

Table 3: OCR text output scores by line for cropped specimen images.

When the results were averaged across all five specimens (Table 4), the top OCR tool for whole specimens was OnlineOCR.net (77%), followed by ABBYY Recognition Server v.3 (75%) and ABBYY FineReader (73%). For cropped images the top OCR tool was ABBYY Recognition Server v.3 (88%), followed by ABBYY FineReader (75 or 77%) and Presto! OCR Pro 4.0.

Software	Whole specimen	Label only
Abbyy 2	32%	
Abbyy 3	75%	88%
Google	15%	43%
Free-ocr.com	60%	15%
OnlineOCR.net	77%	68%
Puma.net	39%	48%
Cuneiform	28%	47%
Simple OCR	28%	23%
Free Online OCR	10%	23%
I2OCR	45%	29%
Newocr.com	47%	55%
WeOCR	27%	33%
OCReXTRACT	56%	33%
Custom OCR	38%	32%
SALIX	35%	37%
GImage Reader	57%	54%
Scanitto	53%	43%
FreeOCR	41%	32%
TopOCR	48%	58%
TypeReader	61%	33%
Abbyy FineReader	73%	77%
OmniPage Professional 18	60%	73%
Presto! OCR Pro 4.0	61%	74%

Table 4. A summary of the results by OCR software, with the average across all specimens, scored by line. All the results over 50% are highlighted.

DISCUSSION

A range of OCR software tools were compared in Trial 1, based on the processing of five RBGE specimens, both full specimen images and also images consisting of the label which had been cropped. The results indicated that three OCR software options were generally performing better than the others for herbarium specimens. These were ABBYY Recognition Server v.3, ABBYY FineReader and OnlineOCR.net.

TRIAL 2: COMPARING OCR TOOLS BEING USED IN HERBARIA

MATERIALS AND METHODS

Three sets of specimen images were gathered from the herbaria of the Royal Botanic Garden Edinburgh (RBGE), the Muséum national d'Histoire naturelle (MNHN) and New York Botanical Garden (NYBG). Each institute selected five specimens and supplied the images in the original format as well as in the format that they currently use for the OCR processing. RBGE do not do any formatting of the images prior to processing. NYBG automatically crop the images to the lower half of the specimen only and remove colour. MNHN reduce the size of their images (Table 5).

Institute	Filename	Filesize	Image format	Label description
NYBG	v-212-01295628_f	535 KB	jpg. Lower half of specimen only	Printed, with map above text. Handwritten det.
NYBG	v-160-01341337_f	692 KB	jpg. Lower half of specimen only	Printed or typed, with map above text. Typed det.
NYBG	01499735_f	540 KB	jpg. Lower half of specimen only	Label with typed and handwritten information. Det slip with typed and handwritten information.
NYBG	01496256_f	509 KB	jpg. Lower half of specimen only	Label with typed and handwritten information. Det slip typed. Handwritten det on specimen.
NYBG	01493584_f	517 KB	jpg. Lower half of specimen only	Printed, with map above text.
NYBG	v-212-01295628_o	619 KB	jpg. Lower half of specimen only	Printed, with map above text. Handwritten det.
NYBG	v-160-01341337_o	744 KB	jpg. Lower half of specimen only	Printed or typed, with map above text. Typed det.
NYBG	01499735_o	689 KB	jpg. Lower half of specimen only	Label with typed and handwritten information. Det slip with typed and handwritten information.
NYBG	01496256_o	563 KB	jpg. Lower half of specimen only	Label with typed and handwritten information. Det slip typed. Handwritten det on specimen.
NYBG	01493584_o	746 KB	jpg. Lower half of specimen only	Printed, with map above text.
MNHN	PC0559998_f	534 KB	jpg	One sheet with four specimens each with its own label which are printed and stamped. Additional stamps on each specimen. Handwritten number of each specimen.
MNHN	P01596658_f	7 MB	jpg	Label printed and handwritten. Handwritten number.
MNHN	P01583601_f	2 MB	jpg	Label printed and

				handwritten. Two det slips with both printing and handwriting.
MNHN	P01583356_f	5 MB	jpg	Printed label with some typing.
MNHN	P01523160_f	4 MB	jpg	Printed label. One det slip printed and one with a mix of printing and handwriting. Handwritten note.
MNHN	PC0559998_o	48 MB	jpg	One sheet with four specimens each with its own label which are printed and stamped. Additional stamps on each specimen. Handwritten number of each specimen.
MNHN	P01596658_o	50 MB	jpg	Label printed and handwritten. Handwritten number.
MNHN	P01583601_o	50 MB	jpg	Label printed and handwritten. Two det slips with both printing and handwriting.
MNHN	P01583356_o	49 MB	jpg	Printed label with some typing.
MNHN	P01523160_o	47 MB	jpg	Printed label. One det slip printed and one with a mix of printing and handwriting. Handwritten note.
RBGE	E00015007	141 MB	tif	Printed and typed label.
RBGE	E00037202	141 MB	tif	Printed and handwritten label.
RBGE	E00262858	142 MB	tif	Printed and handwritten label.
RBGE	E00262827	143 MB	tif	Printed and handwritten label.
RBGE	E00448970	141 MB	tif	Printed label.

Table 5: A summary of the specimen images from each institute.

Transcription & OCR Processing

Each specimen was manually transcribed and the number of lines of text counted. The image files were then processed by each of the institutes using the OCR software currently in place at that institute: ABBYY Recognition Server (RBGE), ABBYY FineReader (NYBG) and Tesseract (MNHN).

The settings for ABBYY Recognition Server are included in Appendix 1A.

Marking up and scoring OCR output by line

The OCR text output was compared to the original transcription and marked up following the procedure for scoring by line in Trial 1.

ACTUAL /15
Det Utrecht 19 Herb. Hort. Bot. Reg. Edin. Flora of Lebanon Acer Araya, Nahr Beyrouth Alt. 1,500ft In 'maquis' thickets, small tree 10-15ft. high; Flowers greenish yellow 4.4.1959 Coll. O. Polunin No. 5204 Royal Botanic Garden Edinburgh E00015007 E00015007
NY ORIGINAL 13.5/15
hIUZS >lli g°sl E o o IN ROYAL BOTANIC GARDEN EDINBURGH E00015007 £ Det. F?Co*-* Je'-j S'UMs^ 3. Utrecht / « -/- 19 9^ fL.HA OF LEBANON A?,?£ ğienSS 'f if Araya, Nahr Beyrouth Alt. 1,500 ft. In 'maquis* thickets, small tree 10-15 ft., high; flowers greenish yellow 4.4.1959 Coll. O. Polunin No. 5204 HERB. HORT. BOT. REG. EDIN. csv-xäj copyright reserved E00015007

Figure 1. An example of the marking up of the OCR output text for a specimen from RBGE. The Actual is the original transcription with the number lines. The NY Original is the output from the ABBYY FineReader with the score based on number of lines correct or partially correct. Green was used to indicate correct OCR reading of the text. For partially correct lines, orange was used to indicate the words which were correct and red was used for the incorrect text. Blue text was used for additional words which had not been captured by the manual transcription, eg copyright reserved from the ruler.

Marking up and scoring OCR output by character

The OCR output text was then marked up and scored by character. The number of characters in the original transcription was counted. The total number of characters, the number of correct characters and the number of incorrect characters in the OCR output text was also counted. From these numbers, the following percentages were calculated:

1. The percentage of the correct characters (correct / actual characters)
2. The percentage of OCR correct characters (correct / output characters)
3. The percentage of OCR incorrect characters (incorrect / output characters)

The percentage of the correct characters effectively measures the quality of the OCR output, disregarding any output caused by non-label interference (plant material, rulers and colour charts, etc). The percentage of OCR correct characters gives an indication of the amount of output caused by non-label interference, particularly when compared to the percentage of correct characters value.

RESULTS

Scores by OCR text output line

The scores by line of OCR output text were collated (Table 6). A summary for each specimen was produced, including the specimen image and the output text from each institute. An example from each institute is presented in Appendix 1B.

Barcode	Processed by	Original	Original as %	Formatted	Formatted as %
P01523160	RBGE	11/14	0.79	12/14	0.86
P01523160	MNHN	10.5/14	0.75	/14	-
P01523160	NYBG	11/14	0.79	/14	-
P01583356	RBGE	18.5/23	0.80	19/23	0.83
P01583356	MNHN	17.5/23	0.76	/23	-
P01583356	NYBG	17/23	0.74	/23	-
P01583601	RBGE	8/13	0.62	10/13	0.77
P01583601	MNHN	5/13	0.38	/13	-
P01583601	NYBG	5.5/13	0.42	/13	-
P01596658	RBGE	10/11	0.91	10/11	0.91
P01596658	MNHN	8/11	0.73	/11	-
P01596658	NYBG	8/11	0.73	/11	-
PC0559998	RBGE	28/34	0.82	16.5/34	0.49
PC0559998	MNHN	19.5/34	0.57	/34	-
PC0559998	NYBG	18/34	0.53	/34	-
E00448970	RBGE	10.5/12	0.88	n/a	-
E00448970	MNHN	5/12	0.42	/12	-
E00448970	NYBG	10/12	0.83	/12	-
E00015007	RBGE	12/15	0.80	n/a	-
E00015007	MNHN	1/15	0.07	/15	-
E00015007	NYBG	13.5/15	0.90	/15	-
E00037202	RBGE	14.5/16	0.91	n/a	-
E00037202	MNHN	/16	-	/16	-
E00037202	NYBG	/16	-	/16	-
E00262827	RBGE	9/12	0.75	n/a	-

E00262827	MNHN	0/12	0	/12	-
E00262827	NYBG	9/12	0.75	/12	-
E00262858	RBGE	8.5/11	0.77	n/a	-
E00262858	MNHN	0/11	0	/11	-
E00262858	NYBG	1/11	0.09	/11	-
01493584	RBGE	14/18	0.78	16/18	0.89
01493584	MNHN	14.5/18	0.81	/18	-
01493584	NYBG	14.5/18	0.81	16.5/18	0.92
01499735	RBGE	4/11	0.36	7.5/11	0.68
01499735	MNHN	7/11	0.64	/11	-
01499735	NYBG	7/11	0.64	8.5/11	0.77
01496256	RBGE	11.5/15	0.77	12.5/15	0.83
01496256	MNHN	11/15	0.73	/15	-
01496256	NYBG	12/15	0.80	12/15	0.80
01341337	RBGE	9/15	0.60	10.5/15	0.70
01341337	MNHN	9.5/15	0.63	/15	-
01341337	NYBG	9/15	0.60	11/15	0.73
01295628	RBGE	12/17	0.71	13/17	0.76
01295628	MNHN	13/17	0.76	/17	-
01295628	NYBG	13/17	0.76	12.5/17	0.74

Table 6. The scores by line for each specimen image.

For processing the original specimens, ABBYY Recognition Server scored highest or highest equal in 9/15 specimens, ABBYY FineReader scored highest or highest equal in 7/15 specimens and Tesseract scored highest or highest equal in 4/15 specimens (Table 7). When these figures are broken down by institute specimens, it becomes clear that there is not a single OCR software option performing strongly across all specimens.

OCR Software	MNHN Specimens (5)	RBGE Specimens (5)	NYBG Specimens (5)	Total Specimens (15)
ABBYY Recognition Server	5	4	0	9
ABBYY FineReader	1	2	4	7
Tesseract	0	0	4	4

Table 7. The number of specimens for which the OCR software scored the highest or highest equal.

Scores by OCR text output character

The results of the OCR text output scored by character was collated (Tables 8-11).

BARCODE	Actual characters	Total Output characters	Correct characters	Incorrect Characters	% correct characters	% of OCR correct	% of OCR incorrect
01499735 (E)	199	174	148	29	74.3%	85.1%	16.6%
01499735 (NY)	199	240	143	96	71.8%	59.5%	40%
01499735 (P)							
01295628 (E)	500	527	486	40	97.2%	92.2%	7.5%
01295628 (NY)	500	478	443	35	88.6%	92.6%	7.3%
01295628 (P)							
01341337 (E)	428	403	329	76	76.8%	81.6%	18.8%
01341337 (NY)	428	402	324	76	75.7%	80.5%	18.9%
01341337 (P)							
01493584 (E)	515	542	499	43	96.8%	92.0%	7.9%
01493584 (NY)	515	496	447	48	86.7%	90%	9.6%
01493584 (P)							
01496256 (E)	251	330	246	83	98.0%	74.5%	25.1%
01496256 (NY)	251	351	217	131	86.4%	61.8%	37.3%
01496256 (P)							

Table 8. Comparison of NYBG, P and E processing (Original) (NY Specimens)

BARCODE	Actual characters	Total Output characters	Correct characters	Incorrect Characters	% correct characters	% of OCR correct	% of OCR incorrect
01499735 (E)	199	240	144	96	72.3%	60%	40.0%
01499735 (NY)	199	248	153	95	76.8%	61.6%	38.3%
01295628 (E)	500	480	469	11	93.8%	97.7%	2.2%
01295628 (NY)	500	467	442	25	88.4%	94.6%	5.3%
01341337 (E)	428	432	343	88	80.1%	79.3%	20.3%
01341337 (NY)	428	443	349	88	81.5%	78.7%	19.8%
01493584 (E)	515	526	499	27	96.8%	94.8%	5.1%
01493584 (NY)	515	543	486	57	94.3%	89.5%	10.4%
01496256 (E)	251	330	217	111	86.4%	65.7%	33.6%
01496256 (NY)	251	299	221	78	88.0%	73.9%	26.0%

Table 9. Comparison of NYBG and E processing (Formatted) (NY Specimens)

BARCODE	Actual characters	Total Output characters	Correct characters	Incorrect Characters	% correct characters	% of OCR correct	% of OCR incorrect
E00015007 (E)	218	445	197	248	90.3%	44.2%	55.7%
E00015007(NY)	218	303	211 (excl. blue text)	75	96.7%	69.6%	35.5%
E00015007(P)	218	11,488	13	11,475	5.9%	0.1%	99.9%
E00448970 (E)	260	353	242	111	93.1%	68.5%	31.4%
E00448970(NY)	260	228	215	12	82.6%	94.2%	5.2%
E00448970(P)	260	909	163	745	62.7%	17.9%	82.0%
E00262858 (E)	162	331	134	197	82.7%	40.5%	59.5%
E00262858(NY)	162	151	9	142	5.5%	5.9%	94.0%
E00262858(P)	162	13,351	0	13,351	0%	0%	100%
E00262827 (E)	260	327	215	112	82.7%	65.7%	34.2%
E00262827(NY)	260	272	241	30	92.6%	88.6%	11%
E00262827 (P)	260	14,710	0	14,710	0%	0%	100%
E00037202 (E)	235	404	205	199	87.2%	50.7%	49.2%
E00037202(NY)	235						
E00037202 (P)	235						

Table 10. Comparison of NYBG, P and E Processing (E specimens)

BARCODE	Actual characters	Total Output characters	Correct characters	Incorrect Characters	% correct characters	% of OCR correct	% of OCR incorrect
P01523160 (E)	218	233	186	47	85.3%	79.8%	20.1%
P01523160(NY)	218	225	189	35	86.6%	84.0%	15.5%
P01523160 (P)	218	225	186	39	85.3%	82.7%	17.3%
P01583356 (E)	757	753	680	68	89.8%	90.3%	9.0%
P01583356(NY)	757	728	692	36	91.4%	95.0%	4.9%
P01583356 (P)	757	728	674	54	89.0%	92.6%	7.4%
P01583601 (E)	115	270	79	190	68.6%	29.2%	70.3%
P01583601(NY)	115	74	62	11	53.9%	83.7%	14.8%
P01583601 (P)	115	75	54	20	47.0%	72%	26.6%
P01596658 (E)	201	214	178	36	88.5%	83.1%	16.8%
P01596658(NY)	201	184	163	20	81.0%	88.5%	10.8%
P01596658 (P)	201	184	163	21	81.1%	88.6%	11.4%
PC0559998 (E)	744	793	554	240	74.4%	69.8%	30.2%
PC0559998(NY)	744	794	453	341	60.8%	53.2%	42.9%
PC0559998 (P)	744	794	459	335	61.7%	57.8%	42.2%

Table 11. Comparison of E, P and NYBG Processing (P specimens)

When individual characters were used as a measure of accuracy the following results were observed (Table 12). For NYBG specimens which had not been formatted, the ABBYY Recognition Server scored highest for every specimen. When the specimen images were formatted prior to processing, then the ABBYY FineReader scored highest for three specimens. For the RBGE specimens, ABBYY Recognition Server scored highest for three specimens and ABBYY FineReader scoring highest for the remaining two specimens. For the MNHN specimens, ABBYY Recognition Server scored highest for three specimens, and ABBYY FineReader scored highest for the remaining two specimens.

OCR Software	MNHN Specimens (5)	RBGE Specimens (5)	NYBG Specimens (5)	NYBG Specimens formatted (5)	Total Specimens (20)
ABBYY Recognition Server	3	3	5	2	13
ABBYY FineReader	2	2	0	3	7
Tesseract	0	0	0	0	0

Table 12. The number of specimens for which the OCR software scored the highest or highest equal.

DISCUSSION

The results of Trial 2 in which preliminary testing of three OCR software options currently being used by three institutes, ABBYY Recognition Server, ABBYY FineReader and Tesseract was carried out, suggested that although there was not one software option consistently outperforming the others, there was clear support for using one of the ABBYY software options over Tesseract.

TRIAL 3: MULTIPLE OCR TRIALS OF DIVERSE SPECIMENS

MATERIALS AND METHODS

Specimens were selected from six partner institutes (MfN, MNHN, MRAC, NMP, RBGE, RBGK) as well as a set from several US institutes supplied by iDigBio. The images represented a range of material including plants, insects, molluscs and fossils.

OCR processing was carried out by three institutes, using different OCR options. RBGE processed images using ABBYY Recognition Server v3. RBGK processed images using ABBYY FineReader v12 (Professional). MfN processed images using four different online OCR services: Onlineocr.net, Newocr.com, Ocrgeek.com, Ocrconvert.com.

The settings used for ABBYY Recognition Server v3 are provided in Appendix 1A. The settings used for ABBYY FineReader v12 are provided in 1C.

RBGE currently run ABBYY Recognition Server v3 on a Windows server. Several workflows have been created for different departments and purposes in the institute. For this SYNTHESYS trial the Herbarium OCR Workflow, used for ad hoc ocr processing, was used rather than our main processing workflow. The settings, however, are the same in both workflows.

RBGK trialled ABBYY FineReader v12 (Professional) using Windows 2013. Different settings and file preparations were tested in order to assess the form of processing that yields the best OCR result. A subset of images from RBGK were selected and were formatted in several different ways either prior to being read or as part of the FineReader process.

An initial basic scoring system was used based loosely on those used within the earlier trials that allowed for outcomes of different approaches to be compared. One point was awarded for each line of correct text recognised and the percentage of correct lines out of total lines was calculated. This was done for a subset of 5 RBGK specimens for each of the different image formats (the different formatting approaches are listed in Appendix 1D and the results of the scoring for each barcode and also the average are in Appendix 1F).

MfN trialled Onlineocr.net, Newocr.com, Ocrgeek.com, Ocrconvert.com. In the settings, the German language was selected for the MfN specimens, the French language was selected for the MNHN specimens and the Czech language was selected for the NMP specimens. The OCR output was then scored based on characters (Table 13).

The OCR output was then scored using the manually transcribed label as the control, using the word count function in Microsoft Office to do this. Given that a primary use of the OCR output is for filtering images based on keyword searching both for research and for creating batches for further

data entry, word accuracy is arguably the most important measure. Only information that was printed was considered in this word count, handwritten text was either not transcribed, or ignored (grey text). Accents in the original text were ignored, as were any that may have been 'read' by the OCR. It was also decided to ignore artefacts around a word (e.g. /, *, |, etc.), as well as single trailing letters (I,J were quite often seen). The correct words were highlighted, and the total of correct words again calculated using the word count. The number of 'words' 'read' by the OCR was also calculated using word count – this could include words that were on the specimen, but not considered part of the specimen information (e.g. colour target or ruler), as well as nonsense caused by handwriting or specimen.

The score was calculated using:

$$\text{Correct/Actual} \times 100$$

This gave a percentage for the correctness of the OCR result.

RESULTS

The results found that the two ABBYY OCR solutions consistently gave better results than the online options. In some cases, up to 100% of the label text was correctly transcribed (Tables 13-20).

Of the online services, Onlineocr.net and Newocr.net gave better results than the other two services. Full results by specimen are provided in Appendix 1E.

MFN SPECIMENS

Total output	Total output characters	Correct characters	Incorrect characters	Missing/ excessive characters	% correct characters	% of ocr correct
Onlineocr.net	95,55	51,64	43,73	39,57/ 20,5	42,89	44,75
Newocr.net	77,91	38,09	39,82	34,18	37,70	48,08
Ocrgeek.com	16,45	11,55	5,00	95,36	11,66	31,19
Ocrcoverg.com	29,55	21,91	17,55	81,55	9,41	11,85

Table 13. Summary of output scored by character

Service	Total word count	Total Output word count	Total Correct words	% of ocr correct (Correct/Actual)
ABBYY Recognition Server v3	220	105.0	130.2	59.2
ABBYY FineReader v12 Suggested protocol	220	n/a	16	7.3
Onlineocr.net	220	n/a	54	24.5
Newocr.com	220	169.0	21	9.5
Ocrgeek.com	220	88.0	8	3.6
Ocrconvert.com	220	138.0	30.9	14.0

Table 14. Summary of output scored by word

MNHN SPECIMENS

Service	Total word count	Total Output word count	Total Correct words	% of ocr correct (Correct/Actual)
ABBYY Recognition Server v3	376	453	266	70.7
ABBYY FineReader v12 Suggested protocol	376	372	233	62.0
Onlineocr.net TIFF	376	204	66	17.6
Onlineocr.net JPEG	376	269	87	23.1
Newocr.com	376	713	104	27.7
Ocrgeek.com TIFF	376	304	21	5.6
Ocrgeek.com JPEG	376	261	29	7.7
Ocrconvert.com TIFF	376	295	3	0.8
Ocrconvert.com JPEG	376	279	3	0.8

Table 15. Summary of output scored by word

MRAC SPECIMENS

Service	Total word count	Total Output word count	Total Correct words	% of ocr correct (Correct/Actual)
ABBYY Recognition Server v3	265	325	175	66.0
ABBYY FineReader v12 Suggested protocol	265	319	149	56.2
Onlineocr.net	265	n/a	n/a	n/a
Newocr.com	265	n/a	n/a	n/a
Ocrgeek.com	265	n/a	n/a	n/a
Ocrconvert.com	265	n/a	n/a	n/a

Table 16. Summary of output scored by word

NMP SPECIMENS

Summary of output scored by word:

Service	Total word count	Total Output word count	Total Correct words	% of ocr correct (Correct/Actual)
ABBYY Recognition Server v3	365	324	159	43.6
ABBYY FineReader v12 Suggested protocol	365	300	173	47.4
Onlineocr.net	365	233	134	36.7
Newocr.com	365	279	84	23.0
Ocrgeek.com	365	269	81	22.2
Ocrconvert.com	365	530	108	29.6

Table 17. Summary of output scored by word

IDIGBio SPECIMENS

Service	Total word count	Total Output word count	Total Correct words	% of ocr correct (Correct/Actual)
ABBYY Recognition Server v3	461	778	386	83.7
ABBYY FineReader v12 Suggested protocol	461	639	372	80.7
Onlineocr.net	461	n/a	n/a	n/a
Newocr.com	461	n/a	n/a	n/a
Ocrgeek.com	461	n/a	n/a	n/a
Ocrconvert.com	461	n/a	n/a	n/a

Table 18. Summary of output scored by word

RBGE SPECIMENS

Service	Total word count	Total Output word count	Total Correct words	% of ocr correct (Correct/Actual)
ABBYY Recognition Server v3	684	1268	613	89.6
ABBYY FineReader v12 Suggested protocol	684	902	535	78.2
Onlineocr.net	684	n/a	n/a	n/a
Newocr.com	684	n/a	n/a	n/a
Ocrgeek.com	684	n/a	n/a	n/a
Ocrconvert.com	684	n/a	n/a	n/a

Table 19. Summary of output scored by word

RBGK SPECIMENS

Service	Total word count	Total Output word count	Total Correct words	% of ocr correct (Correct/Actual)
ABBYY Recognition Server v3	944	2256	846	89.6
ABBYY FineReader v12 Suggested protocol	944	1000	835	88.5
Onlineocr.net	944	n/a	n/a	n/a
Newocr.com	944	n/a	n/a	n/a
Ocrgeek.com	944	n/a	n/a	n/a
Ocrconvert.com	944	n/a	n/a	n/a

Table 20. Summary of output scored by word

WORKFLOWS: INCORPORATING OCR INTO DIGITISATION WORKFLOWS

The development of digitisation workflows has been taking place in natural history collections around the world. The Andrew W. Mellon funded Global Plants Project was instrumental in bringing institutes together from around the world to develop standard formats and protocols for digitisation. It was recognised at an early stage that there would not be a single solution which would fit every institute so flexibility was also seen as an important factor in a large project. Each institute will have their own priorities and constraints. There were, however, some key principles which were recognised and which became standards for the project. These included standards for image quality, the use of colour charts and rulers in the images, and the data and metadata format.

More recently, the development of digitisation workflows has concentrated on scaling up the process to enable the digitisation of millions of specimens in a realistic timeframe. Revolutionary processes including outsourcing the imaging to be carried out in warehouses using a system of conveyor belts and cameras were introduced by Muséum national d'Histoire naturelle (MNHN). Similar systems were also being developed by Digitarium in Finland and at Naturalis. These large-scale digitisation projects are resulting in millions of digitised collection objects in Europe which have been catalogued with minimal data attached. There is a need now to find ways to efficiently transcribe the label data and make those data available electronically. The successful application of OCR technology to natural history collections as described above, needs to be integrated within existing and developing digitisation workflows.

In developing digitisation workflows which incorporate OCR technology we have looked at the whole pipeline of image and data management within the digitisation framework. The Royal Botanic Garden Edinburgh (RBGE) has put in place an integrated workflow in which the OCR output text has been used to speed up the process of transcribing over 100,000 specimen labels (Figure 2). MNHN and BGBM have also included OCR processing in their digitisation workflow.

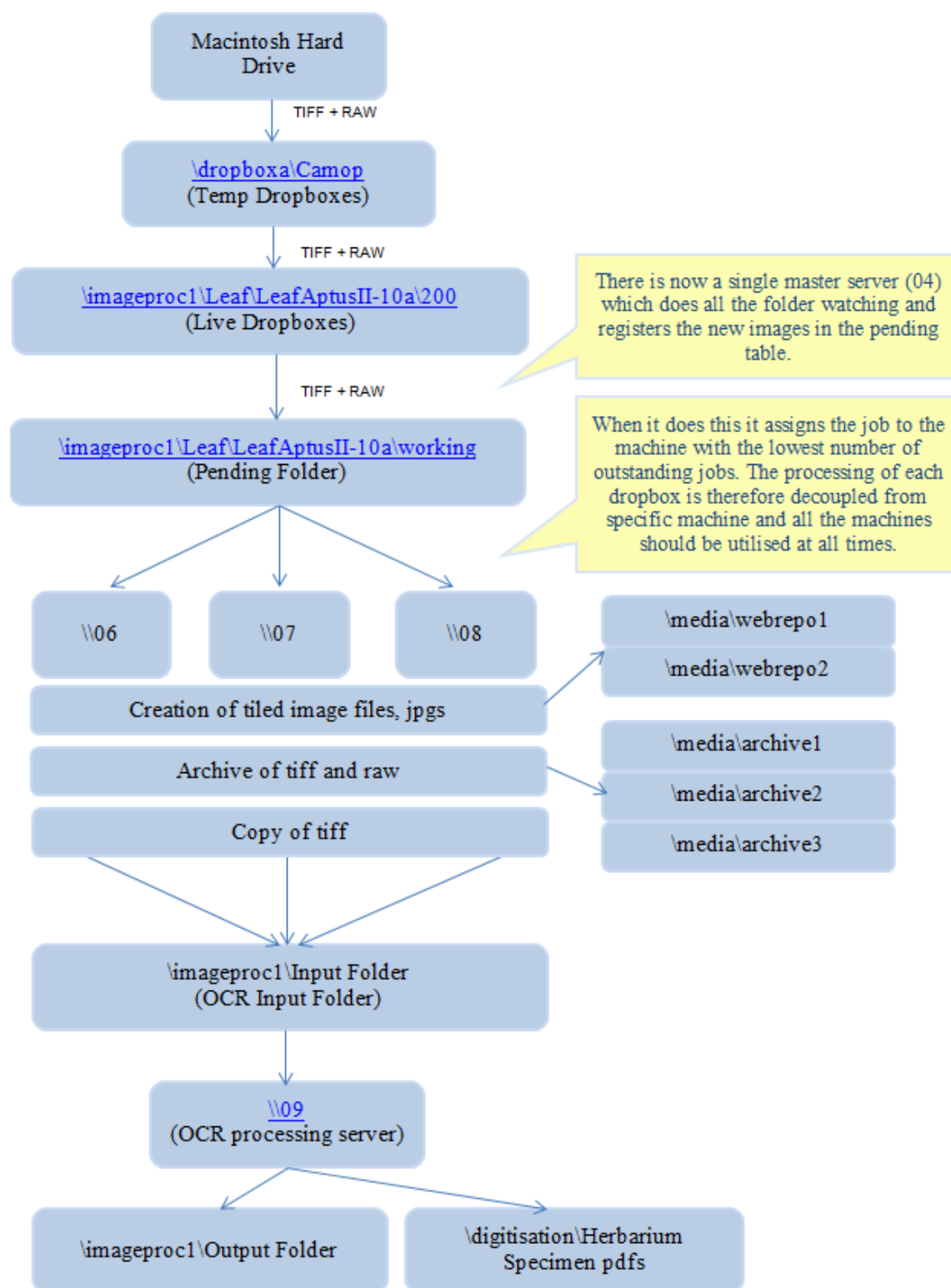


Figure 2. Example of digitisation workflow which incorporates OCR at RBGE

DISCUSSION

The aim of these trials were to determine the accuracy, effectiveness and usefulness of OCR software for natural history specimens. The trials found that ABBYY software gave the best results in most cases. However, no single software option produced the best results consistently.

The results of the trials show that the use of OCR software in automatically transcribing specimen labels can give excellent results of up to 100% correct transcription of label text, and provide guidance for institutes looking to incorporate OCR software into their digitisation workflow. The best options allow the institute to choose the option which suits their ICT and workflow system: a server option (ABBYY Recognition Server v3), a pc option (ABBYY FineReader).

Of the online options, OnlineOCR.net and Newocr.com came out higher although the online services did not perform as well and the testing was restricted to specimens from three institutes.

Integrating OCR workflows within institutional digitisation workflows will be key to the effective use of the technology. The reduction of any decision-making and human actions within the workflow will allow the process to work faster. This includes processing all images as they are digitised unless an automated selection can be carried out.

Some of the most problematic areas in the curation of collection data involve the interoperability of databases and data transfer. The options for the use of OCR software to capture label data are:

1. to retain the OCR text output in its raw state and save in a database to enable queries to be linked to the collections data
2. to run a parsing algorithm on the OCR text output and retain the data in a separate system
3. to run a parsing algorithm on the OCR text output and ingest the data into the collections database

SECTION 2: REVIEW OF DEVELOPMENT OF NLP FOR PARSING OCR TEXT INTO DARWIN CORE FIELDS.

INTRODUCTION

In some ways, the parsing of OCR text into Darwin core fields has been seen as the holy grail of digitisation. It would potentially allow the full automation of the post image capture process. However, it has proved to be extremely hard to achieve. A short review was carried out to identify the current state of progress in this area.

REVIEW

In 2013, a Hackathon was held by the Augment OCR Working Group of iDigBio (Integrated Digitized Biocollections). The challenge was “One of the most significant areas of interest for improving the utilization of OCR output is parsing. *Digitization* and *data curation* and *dissemination* of biodiversity museum collections specimen data can be sped up if the output from OCR can be parsed faster and more accurately and packaged into semantically meaningful units for insertion into a database.”

They had some success and several groups have been continuing work on the issue. Through this project we made contact with one of the major teams who have incorporated the parsing of OCR text into software. Symbiota was developed by a collaboration between the University of Wisconsin and Arizona State University, as a platform for creating voucher-based biodiversity information communities, allowing the communities to build virtual collection portals.

Three of these portals are currently using systems which includes the parsing of OCR output text: the Lichen Portal (<http://lichenportal.org/portal/>), the Bryophyte Portal (<http://bryophyteportal.org/portal/>), and the SERNEC (Southeast Regional Network of Expertise and Collections) Portal (<http://sernecportal.org/portal/index.php>). Arrangements are now being made to test these portals with specimen images from partners in the SYNTHESYS3 project. (Examples of screenshots of these portals are provided in Appendix 2.

Additional information of projects and publications relating to parsing of OCR output text into structure Darwin Core format with links to websites, posters and presentations is given here.

iDigBio: AugmentOCR Working Group

Hackathon 2013 and wiki page

(https://www.idigbio.org/wiki/index.php/2013_AOCR_Hackathon_Wiki)

A list of related projects can be found here:

https://www.idigbio.org/wiki/index.php/Participant_Related_Projects#DarwinCore_Parser

Beyond the Box competition

5th Level of Achievement (OCR Data Parsing and Natural Language Processing)

<https://beyondthebox.aibs.org/level-of-achievement.html>

LBCC (Lichens, Bryophytes and Climate Change)

<http://lbcc1.acis.ufl.edu/portals>

North American Bryophyte and Lichen TCN

Integrated OCR and NLP capabilities into their processing workflows and their Symbiota web portals
Darwin Score, Ben Brumfield

(<https://github.com/idigbio-citsci-hackathon/darwin-score/blob/master/README.md>)

SilverBiology

Business working with Cornell and University of Florida

<http://www.helpingscience.org/service/darwincoreprocessing/>

<http://www.helpingscience.org/service/darwincoreprocessing/examples/example3.html>

BiSciCol (Biological Sciences Collections)

Patrick Heidorn, University of Arizona

<http://grantome.com/grant/NSF/DBI-0956271>

Tracker: develop methods to facilitate and evaluate the creation of structured database records in extended Darwin Core from images of specimen labels from museums using records created from Optical Character Recognition

participated in aOCR Hackathon

developed a set of programmes which produce ordered xml from unordered csv

<https://github.com/BryanHeidorn/LABELX>

algorithms for scoring <https://github.com/idigbio-aocr/scoring>

CalBug, University of California

Looking for programmers to create a 'smart' parsing program

www.nature.berkeley.edu/~oboyski67/CalBug/CalBug.ppt

Salix, Semi-Automatic Label Information eXtraction System

<http://daryllafferty.com/salix/>

Daryl Lafferty, Arizona State University

Apiary Project, University of North Texas

www.apiaryproject.org

Anglin, R., Best, J., Figueiredo, R., Gilbert, E., Gnanasambandam, N., Gottschalk, S., Haston, E., Heidorn, P. B., Lafferty, D., Lang, P., Nelson, G., Paul, D., Ulate, W., Watson, K., & Zhang, Q. (2013). *Improving the Character of Optical Character Recognition (OCR): iDigBio Augmenting OCR Working Group Seeks Collaborators and Strategies to Improve OCR Output and Parsing of OCR Output for Faster, More Efficient, Cheaper Natural History Collections Specimen Label Digitization*. iConference 2013 Proceedings (pp.957-964).doi:10.9776/13493

(<https://www.ideals.illinois.edu/bitstream/handle/2142/42089/493.pdf?sequence=2>)

Moen, William E.; Huang, Jane Q.; McCotter, Melody; Best, Jason H. & Neill, Amanda K. *An Application Profile Using Darwin Core Rendered in the New Dublin Core Application Profile Framework*. UNT Digital Library.<http://digital.library.unt.edu/ark:/67531/metadc81371/>. Accessed September 15, 2015.

Heidorn, PB & Wei, Q. *Automatic metadata extraction from museum specimen labels*. Proc. Int'l Conf. on Dublin Core and Metadata Applications, 2008.

<http://dcpapers.dublincore.org/pubs/article/viewFile/919/915>

SECTION 3: REVIEW OF (SEMI) AUTOMATIC SPECIMEN IMAGE CLASSIFICATION, I.E. (SEMI) AUTOMATIC TAGGING OF SPECIMEN IMAGES FROM CERTAIN COLLECTORS OR EXPEDITIONS, USING TEMPLATE MATCHING SOFTWARE

PART 1: SEMI-AUTOMATED CLASSIFICATION OF HERBARIUM SPECIMENS BY MEANS OF TEMPLATE MATCHING ALGORITHMS

This report has been produced as a separate document and is inserted here.

Semi-automated Classification of Herbarium Specimens by means of Template Matching Algorithms

Report for the SYNTHESYS III project

WP 4, Task 1.2.3

Jörg Holetschek
Botanic Museum & Botanical Garden Berlin-Dahlem
Königin-Luise-Str. 6-8
14195 Berlin-Dahlem
j.holetschek@bgbm.org

Inhalt

Motivation	3
Digital Specimens	3
Specimen Metadata	4
Use case at BGBM	6
Aim of studies	8
Feature Detection in the Herbar-Digital Project	8
Template Matching feature of Linienextraktor	9
Approach	10
Findings	12
Number of templates	12
Image resolution	12
Image format	13
Full/partial templates	14
Contrast, Image manipulations.....	14
Results for Peter Reisen	15
Conclusions	16
Guidelines for “good” templates	17
Outlook.....	18
Appendix: Terminology	19

Motivation

Digital Specimens

Specimen collections house huge amounts of preserved organisms gathered by collectors throughout the world over the past 300 years. This stock is the result of innumerable working years of botanists and zoologists past and present. The specimens provide rich and verifiable documentation of the planet's flora and fauna throughout the centuries covered by these specimens. They are a valuable source of information and material for today's biological research, for example for getting a better understanding of certain organisms by new analytical methods and for finding evidence for past and ongoing changes (and losses) of our biodiversity.

Typical examples for such collections are herbaria. In the traditional way, to make use of the material in the herbaria, researchers had to travel in order to get physical access to the specimens, or the specimens had to be sent on loan. The disadvantages of this are obvious: apart from being time- and cost-intensive, the specimens could be damaged during transport, even when handled with care. To overcome this, the efforts for digitising the collections have grown in the past. Instead of being sent via mail, loan requests are acted on by creating a high-resolution image of the specimen, which is then made available on a web server. This does not completely substitute personal travel and manual inspection of herbarium sheets, but it greatly reduces the number of specimens that have to be sent and the time required for getting access to the desired specimen. In addition, some of the collections are digitised systematically. Some of these herbaria house millions of sheets, so this process might take years, even decades. The ultimate aim is to fully digitise the herbaria, making them easily accessible over the web.

As the number of digitised specimens grows, the need for efficient search mechanisms becomes more important. Cataloguing digitised specimens is done by associating them with metadata, information usually taken from the labels attached to the specimens. Typical pieces of metadata are the catalogue number and a potentially existing barcode or field number; the taxon name of the specimen (species in most cases, or a higher taxon if identified only to a higher level); gathering agent, gathering time and gathering location (country, state, province, town, detailed description of locality, geographic coordinates); type status information, if appropriate; and potential annotations added by past researchers. This information will help users to find the specimens they are interested in, for example specimens of a certain species or taxonomic group, specimens gathered by a certain collector or during a specific expedition, or specimens gathered in a certain geographic region during a specific period of time.

The need of capturing as much of these metadata items as possible is intensified by another trend: initiatives such as SYNTHESYS and the Global Biodiversity Information Facility (GBIF) are setting up networks for connecting biodiversity data from distributed sources. This enables access to a multitude of collections, potentiating the amount of possibly useful data in which users have to find their pieces of interest. Currently (January 2015), GBIF offers access to 7.7 million images of specimens from 164 datasets (collections). Browsing these images is not feasible, even for a single one of these collections. Without metadata, finding the right specimens is like finding a needle in a haystack.

The collation of data from various sources can potentially enable findings that cannot be gained by examining individual data sources. Given an ample amount of underlying data, a temporal

analysis of occurrence data can provide insights into biodiversity changes for certain organisms in a given geographic region. Combined with climate models, potential changes of biodiversity in the future can be predicted, or at least probabilities of certain scenarios (ecological niche modelling). This is important for the evaluation of the threats by invasive species or the potential spread of disease vectors. Capturing geographic and temporal metadata as accurate as possible enables the use of specimen information for this kind of analyses.

Traditionally, metadata are captured manually in the course of the digitisation process. Before or after the image is taken, the agent will type in label or ledger information into a database that is used to organise the specimen images. When connected to biodiversity networks such as BioCASE (Biological Collection Access Service) or GBIF, these data will be published along with the digital image.

As imaging technology advances and the time required for taking the picture diminishes, metadata capture becomes a bottleneck. Recently, specimen digitisation has moved to a new stage by setting up streamlined, conveyor-based digitisation workflows, even with several lines in parallel, allowing for digitisation on an industrial level. With these setups digitising up to 10,000 specimens a day, full metadata capture becomes the crucial bottleneck that cannot be resolved with traditional approaches.

In the past, within the SYNTHESYS II project, the potential for metadata capture from label or ledger images by involving volunteers, so-called crowdsourcing, has been investigated. In contrast, work package 4, object 1, task 1.2 of the SYNTHESYS III project focuses on *automated* metadata capture from specimen images. This report documents the findings of the studies undertaken at the BGBM on this subject.

Specimen Metadata

A typical example of a collection specimen is a herbarium sheet as the one shown in fig. 1. Besides showing the actual specimen (the organism, 1) it may hold additional plant material (seeds, fruits: 2) in a paper pocket. Attached to the specimen sheet, written or stamped directly on the sheet, you may find the following types of information:

- The herbarium label (3) identifies the specimen either on species or a higher level. Apart from the identified taxon it usually records the identifier person and the identification date, the collector and the gathering date, and information on the gathering site – country, state or province, town or named area and a description of the locality. So the herbarium label holds very important pieces of metadata. It is mandatory for all sheets, but the amount of information can differ. It may be handwritten or typewritten.
- Additional identification labels (4; can be one or more) record subsequent identification events. They may confirm the first identification result or contradict; together with the identification result it lists the identifier and the identification date. Sometimes, they can cite a publication that refers to this specimen.
- If digitized, the sheet usually contains a scale (5), a colour scale (6) and a digitization stamp (7). Some specimens are digitised twice, so more than one digitisation stamp can be found. This happens when an organism gets re-identified with another result or a specimen is of such high importance that it needs to be re-digitised with a higher resolution (typical for type specimens).

- A barcode label (8), if the specimen has been assigned a barcode.
- A type designation (9) marks the specimen as a type. An additional label (10) may specify the type in more detail.
- A stamp identifies the herbarium (11) and potential previous herbaria owning the specimen (12).
- Loan numbers (13) can be used to tell when and how long the specimen was given on loan.



Figure 1: Typical herbarium specimen sheet with different objects attached

Accession numbers (not shown in fig. 1) are an additional metadata item often assigned to specimens and stamped or written on the sheet.

Apart from the herbarium label, all these pieces of information are optional. They may be missing, or they may exist several times. They may be handwritten or typed. There is no defined position on the sheet for them; moreover they can be sideways, upside down or even partly covered by parts of the plant.

Keeping these things in mind, the task of acquiring metadata from specimen labels can be subdivided into three subtasks, namely

- (1) Identification of labels or parts of the specimen sheet that hold metadata: the herbarium label, identification and annotation label(s), type markers and type specifications, notes on previous herbaria, accession and loan number(s), barcodes.
- (2) Classification of features holding metadata. This helps with the actual data capture in the next step, when OCR (optical character recognition) is used or when labels are assigned to specialists. In both cases knowing the type of information the feature is depicting is helpful; OCR can be supported by specialised dictionaries, and different people can be assigned to task like deciphering geographical information or taxonomic names. Moreover, different strategies can be applied for the different types of features, or it can be decided that only some of them get captured.
- (3) Capture of the metadata.

The first subtask will be addressed by task 1.1 of SYNTHESYS III, work package 4 (automatic segmentation of digital images). Focus of the studies at the BGBM was on task 1.2, namely the automated classification of specimens.

Use case at BGBM

The herbarium of the Botanic Garden and Botanical Museum Berlin-Dahlem (herbarium acronym: B) is the largest in Germany and holds a collection of more than 3.5 million preserved specimens. All plant groups – flowering plants, ferns, mosses, liverworts, and algae, as well as fungi and lichens – are represented in the collections which are worldwide in scope. Associated with the general herbarium are special collections of dried fruits and seeds, wood samples, and specimens preserved in alcohol. The collections of the herbarium are growing constantly through field research conducted by staff, and through gifts, acquisitions, and exchanges of specimens from other herbaria. Currently (January 2015), about 140,000 of the specimens are digitised.

Up to now there has been no sufficient funding for digitising the complete stock. Priorities on digitising have been on specimens that have been requested for loan - in order to avoid shipping the specimens - and type specimens. Especially the latter are often historic, meaning their labels, instead of being typed, are usually handwritten in an old-fashioned and sometimes very personal way. Current OCR software has proved to be unable to read this type of handwriting; error rates exceed an acceptable level. With current technologies, manual transcription is the only way of capturing metadata from these types of specimen labels.

This process of transcription often requires special expertise on the collector's handwriting and their habits of noting and abbreviating taxa, geographic names and habitats. This involves knowledge on the collector's field of research, expeditions and expedition routes as well as on geography and historic place names. Since only a few, sometimes only a single one of such experts exist, it is highly desirable for them to dedicate their work solely to the transcription of labels from their field of expertise.

In order to facilitate this, an automated classification process of newly digitised specimens would be desirable. New images would regularly undergo batch processing; if patterns of certain collectors are detected, the specimens would be tagged accordingly. These tags could be used later to present only certain specimens to experts, allowing them to fully benefit from their knowledge, for example using a crowdsourcing platform.

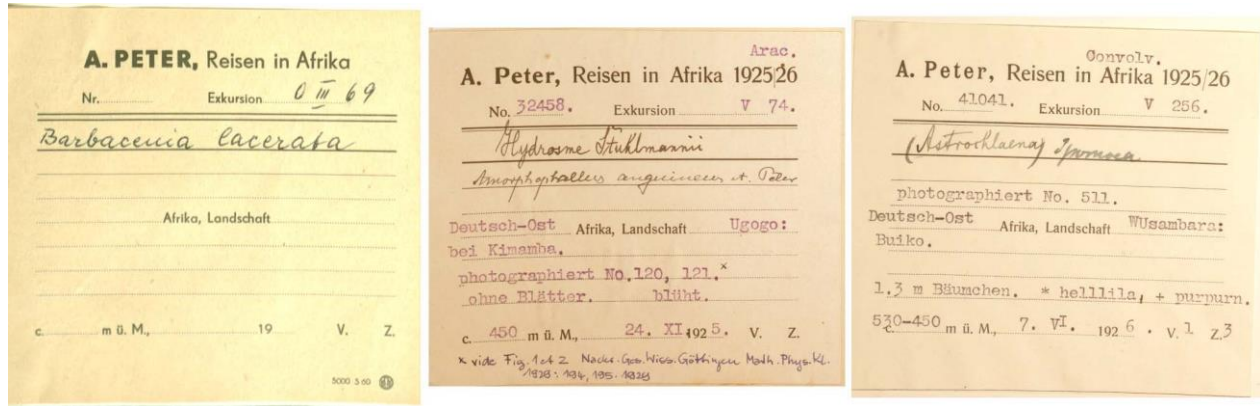


Figure 2: Sample collector labels for A. Peter

For historic expeditions undertaken by certain collectors usually special labels were used. Figure 2 shows three different labels that can be found on specimens gathered by German botanist Dr. Albert Peter during expeditions in Africa in 1913-19 and 1925-26, figure 3 shows three different label variations for the Herbarium Berolinense. Clearly they use different fonts, one of the main challenges beside the usual issues like angular mounting and partial overlap with other labels and parts of the specimen.

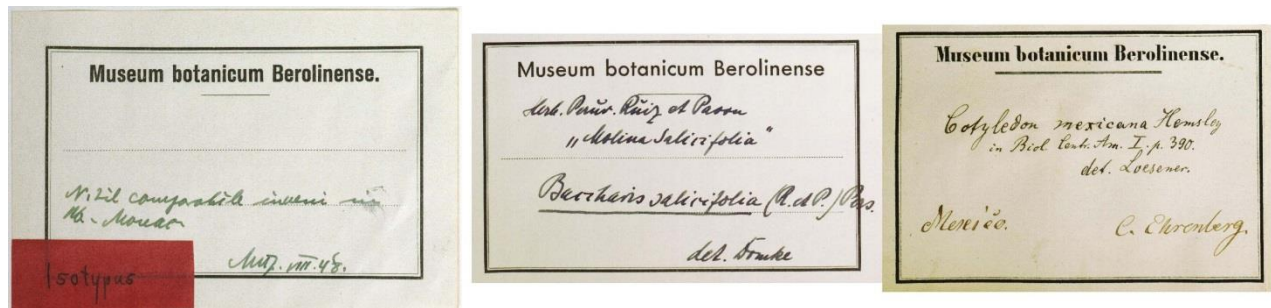


Figure 3: Sample labels for specimens of Museum botanicum Berolinense

The set aim for the studies at the BGBM was to investigate the potential for (semi) automated classification of specimens (specimen images) with regard to certain collections or collectors. Classification (aka tagging) in this context will be the marking of all specimens which are considered to be part of a certain collection.

A first test challenge was the discovery of specimens with the label “Herbarium botanicum Berolinense” in 465 randomly selected images. After that, the lessons learned should be applied to a set of 916 images, trying to identify all specimens of Dr. Albert Peter. The characteristic feature of the labels used for identification will be their label heading “Museum botanicum Berolinense” or “A. Peter, Reisen in Afrika”.

Aim of studies

Feature Detection in the Herbar-Digital Project

Herbar-Digital was a joint project of BGBM and University of Applied Sciences Hannover with the goal of “Rationalizing the virtualization of botanical document material and their usage by process optimization and automation”. Its long-term aim was the digitization of the more than 3.5 million specimens of Herbarium Berolinense. After the project finished in 2011, this work was continued by the follow-up project StanDAP-Herb (a standardised and optimised process for data acquisition from digital images of herbarium specimens).

One subtask of StanDAP-Herb was the implementation of feature detection algorithms for high-resolution scans of the herbarium sheets, which resulted in a software package “Linienextraktor”, the name referring to one of its first purposes, the extraction of trails from hand writing. It bundles several functions, amongst others algorithms for feature detection. For the studies, focus was on the template matching functionality shown below. It allows the definition of templates from images, which can be searched later in a batch of files.

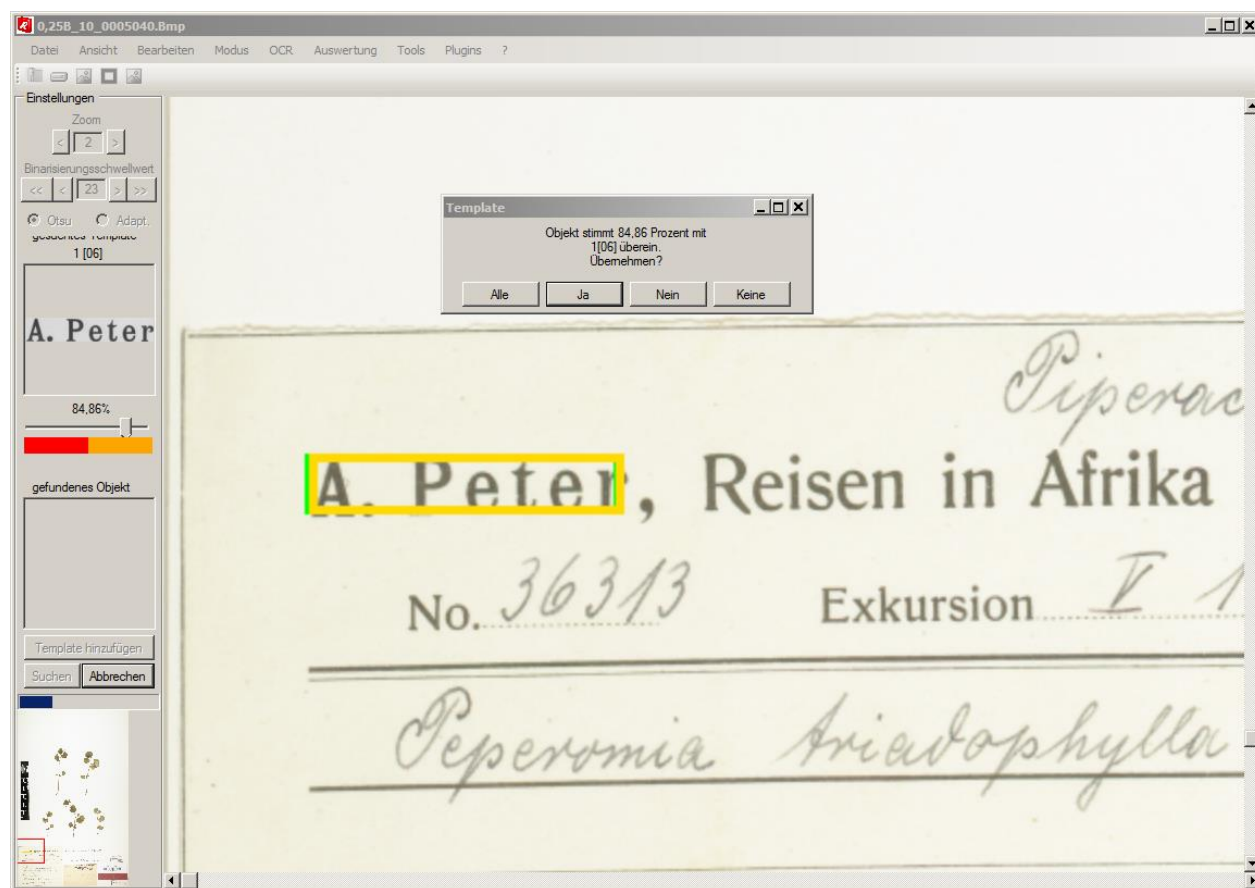


Figure 4: Template matching feature of Linienextraktor

Generally, the objects on a specimen sheet can be classified according to their variability in respect to size and orientation:

- Stable: Colour scales, ruler

- Less stable: Stamps, labels with printed titles
- Variable: Type designation mark, barcode
- More variable: Envelopes
- Extremely variable: Hand-written annotations, plant.

For all but the last of these classes it is possible to create a collection of templates, i.e. digital examples representing the typical colours, shapes etc. of these objects. A template matching algorithm can then be used to find these templates in a given batch of specimen images and, depending on the templates found, classify and tag the image accordingly.

Template Matching feature of Linienextraktor

Before the template matching algorithm can be run on a set of images, one or several templates need to be created for the objects to be searched on the specimens. They will be defined based on different occurrences and variants of the label heading in question and will represent its different flavours. The algorithm will then try to locate the templates in these images according to several settings and, if successful, store regions of the images that contain potential hits and the likelihood calculated by the algorithm. This likelihood would be the criterion for classifying the specimens as containing or not the object(s) in question.

The image below shows a sample template and the parameters for the template matching algorithm that can be set for each template individually. The template was used in test runs to find images that contain labels of collector Dr. A. Peter:

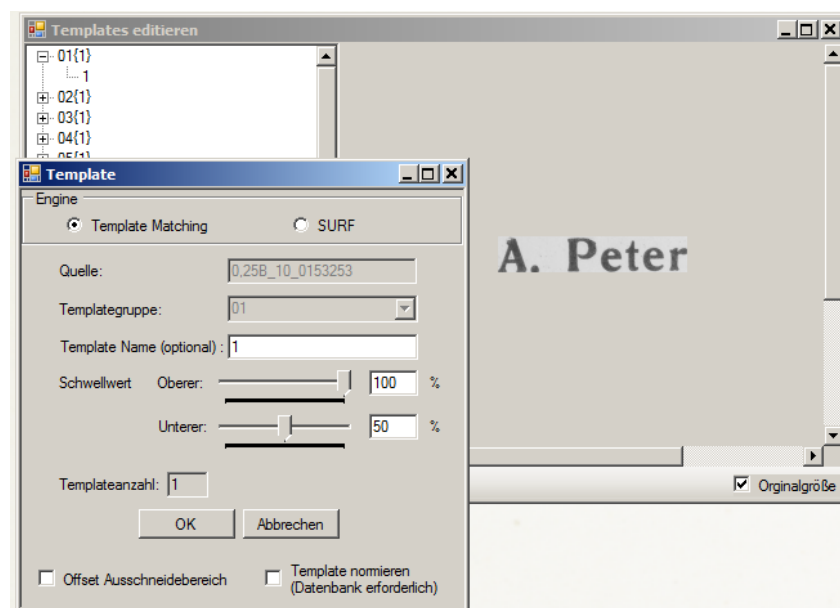


Figure 5: Sample template for finding specimens of A. Peter

Source (Quelle) specifies the image file the template was created from by simply selecting the region of the pattern that is to be searched for. Templates are organised in template groups (Templategruppe), with all templates that represent one feature to be found placed in one group. Once one template of a group is found in a given image, all other templates of that group will be skipped. In production use, this will save computing time by skipping tests for already successfully identified objects.

The software might find a given template several times on an image. For each of these identified objects, the algorithm returns the region of the image, which will be saved as an image, and the likelihood that the object within this region is the same as the object in the template. The thresholds (Schwellwerte) that can be set for a template determine the range for which the algorithm considers found objects as false hits, potential hits or positive hits. Objects with likelihood below the lower (unterer) threshold will be discarded and ignored; objects with likelihood above the upper (oberer) threshold are considered to be a hit. Likelihoods in between are potential hits and would require for manual inspection.

Approach

As no documentation on results of real-world test cases for the software is available so far, a rather long training and testing period on the first challenge was necessary before usable results could be produced (with “training” referring to the person, not the algorithm). By varying different factors – the number and layout of templates, image resolutions, and parameters of the algorithm – the usability of the algorithm’s results for automatic classification of specimen images was evaluated.

This training was done on a limited number of images (465) that have been selected and classified manually. Subsequent runs with changed resolutions, templates, template settings and different contrast levels were done, with only one of the factors varied in a test row. The potential hits identified by the algorithm and their computed likelihood were inspected manually and assessed with respect to the question whether they would allow an automatic classification at an acceptable error rate.

For this evaluation, it was desirable for all templates to be tested in one run. The software’s behaviour of skipping the remaining templates in a group once one template has been found (i.e. identified with a likelihood greater than or equals the upper threshold) had to be avoided. For this reason, each template was placed in a separate template group, forcing each template to be processed.

Aim of training was to find reasonable settings for the different factors that lead to a maximum number of correct classifications. Correct classifications would be true positives (correctly tagged) and true negatives (correctly non-tagged); whereas incorrect classification would be false positives (specimens tagged erroneously) and false negatives (specimens incorrectly untagged).

Factors investigated were

- Image resolution for source images and templates: Full or reduced resolution?
- File formats: TIFF (i.e. lossless compression) or JPG?
- Templates: Objects of interest completely in template or only partly?
- Number of templates (with slightly differing versions of the labels) required?
- Contrast: Does altering the contrast improve the results?

As described before, the threshold parameters are used for separating potential hits that are discarded because they are considered to be false hits (below lower threshold) from the potential hits that require manual inspection (in between lower and upper threshold) and the potential hits that are considered to be true hits. In order to reduce the number of manual inspections, this range

should be as small as possible. In production, different scenarios are possible, depending on the use case:

1. The classification result needs to be as accurate as possible, i.e. neither false positives nor false negatives are desirable or acceptable. In this case, both thresholds must be used, with the upper as low as possible without getting false positives and the lower threshold as high as possible without getting false negatives (i.e. missing specimens). Hits with a likelihood in-between would require manual inspection.
2. If false positives are acceptable, the range for manual inspection could be narrowed by lowering the upper threshold. This could reduce the number of manual inspections at the cost of few false positives.
3. In case false negatives are acceptable, the same effect could be achieved by raising the lower threshold. The number of images for manual inspection would be reduced at the cost of a few missed specimens.
4. In case the variety of objects to be searched for is rather small and image quality good, the results of the algorithm would be pretty good, the range between upper and lower threshold narrow. If this coincides with the acceptability of a few errors, upper and lower threshold could be set to the same value, completely eliminating the need for manual inspection. If this common value would be rather high (i.e. the lower threshold is raised to the upper threshold), it would be at the cost of several missed specimens. If the common value would be rather low (i.e. the upper threshold is lowered to the lower threshold), it would be at the cost of false positives.

The use case at the BGBM – classification of digitised specimens in respect to being part of a certain expedition or a certain collector's collection with subsequent channelling of specimens to the appropriate expert for deciphering the handwriting – is characterized by scenario 4. For one thing, the scans created are high-quality and the objects in question (printed headings of labels) pretty stable. Moreover, false positives are acceptable: If the expert encounters them, he or she can easily mark them as incorrectly classified. For that reason, focus was laid on the lower threshold, on finding values that result in no or almost no false negatives (missed specimens). The same value would be used as an upper threshold, accepting a few false positives.

For these tests, 465 selected images were used. Manual inspection of all images showed that 82 of them contained a label with the caption "Museum botanicum Berolinense" (in fact, the initial manual count was 80, with 2 more images found by the algorithm during one of the first runs). In order to investigate the effect of changing the factors mentioned above on the outcome of the algorithm, the following procedure was repeated with different settings:

1. Running the algorithm with one or more templates.
2. Evaluating the results: Counting false negatives and false positives.
3. Adaption of one factor: Changing/adding/removing templates, changing the threshold value, shifting to another resolution or image format.

The following section documents the findings of playing with the algorithm. Even though they're documented factor by factor, experiments were not necessarily done in that order. For example,

some runs trying to find the number of templates required were done first, then on the effects of image resolution and file formats, which required the tests on the number of templates to be repeated. Likewise, the good final result of finding all occurrences but one was not achieved by just increasing the number of templates, but after having played with all of the factors.

Findings

Number of templates

As can be seen in figure 3, the labels use different fonts for a headline. Since the algorithm does not do OCR to recognize the text, but examines the headlines for their appearance, separate templates should be created for each font. Surprisingly enough, some templates did result in true positives for labels which actually use another font. But this effect was not sufficient to identify all occurrences.

For the first runs, 11 templates were created for six different label fonts. When choosing the source images for the templates, straight aligned labels were preferred over labels mounted askew. The algorithm tolerates small inclinations, so labels not mounted straight would, to a certain extent, also be found. The image below shows some of the templates for the different fonts:

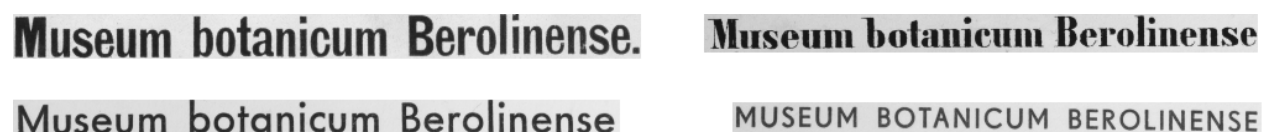


Figure 6: Sample templates used

11 templates resulted in 17 missed specimens (false negatives). This was due to some of the templates having a low contrast. Adding a high-contrast template for five of the fonts increased the true positives to 81, with one specimen missed. (As manual inspection showed, the label headline was partly covered by a leaf, so it was missed in all of the test runs.) After gathering some experience, it turned out that two, at highest three good-quality templates per font are required. They should be straight aligned and taken from high-contrast images.

The Linienextraktor software allows creating an averaging image from several images. This can be used to combine several templates for one font (for example one with a straight aligned heading, one slightly inclined to one direction, one inclined to the other direction) into one template, which will then be used for template matching. The results for this have not been investigated for this report, but if processing time is of high concern, this option should be considered.

Image resolution

Images are scanned at the BGBM with 600ppi resolution and saved as TIFF files. Typical scans have up to 8,192 x 12,000 pixels, resulting in files with about 250M. Running the algorithm on thousands of images of this size would take a rather long time, even on good hardware. Apart from that, the software at its current development stage crashes for unknown reasons when run on the full-size images, even if sufficient memory is available.

For that reason, resolution was reduced by 50%, images were saved as bitmaps (reduction and conversion was done with the batch processor of Linienextraktor). Other test runs included resolution reduced to 25% and even to 12.5%. For each of these resolutions, the same templates (i.e. consisting of exactly the same region of the same original image) were re-created.

It showed that the results of the algorithm were quite good and almost identical on half-resolution and quarter-resolution images: 81 true positives (1 false negative) with no false positive for a 25% resolution at a threshold of 60%. With a 50% resolution and a threshold of 60%, there were two false negative with no false positive; the better resolution apparently allows unwanted particles on the scans like stains and crumbled leaves to disturb the algorithm. Therefore, a higher resolution requires the threshold to be lowered: with a threshold of 55%, the result was 81 true positives (1 false negative) with no false positives, the same as with 25% resolution. The 12.5% resolution turned out to be unusable, the number of false positives skyrocketed. Raising the threshold reduced their number, but also eliminated a high percentage of the true positives.

It seems that an image resolution of 600ppi (original scan resolution) is too much for the current implementation of the algorithm to handle. Resolutions of 300ppi and 150ppi are both usable, with higher resolutions requiring slightly lower thresholds.

Image format

At the BGBM, in addition to the original scan files in lossless TIFF format, the images exist as JPG files with a 33% resolution, about 2MB of size, with no visible compression artefacts. So the question was whether these could be used also, sparing the effort for creating 25% or 50% resolution images from the original TIFFs.

The answer is simply no. With the same templates created from the JPGs and a threshold of 60%, the algorithm returned no false positives, but also 17 false negatives (misses). At a threshold of 55%, the number of misses was down to 13 with no false positives. With a threshold of 50%, still 11 specimens were missed, but the 19 false positives had come up. So apparently JPG format is unusable due to compression losses. Looking at the images with a 500% zoom (see picture below), artefacts especially around the edges of letters can be seen that diminish the algorithm's result.

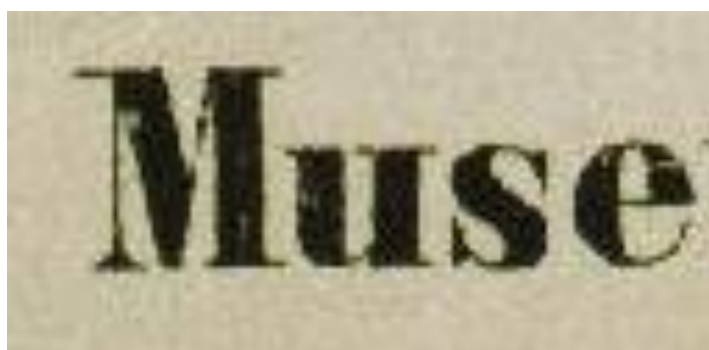


Figure 7: Compression artefacts around characters in JPG format

Full/partial templates

Since the labels to be found have the heading “Museum botanicum Berolinense”, the first test runs were done with these words in the templates. As written before, two or three templates per font used were created.

At a threshold of 53%, 13 specimens have been missed (false negatives) with 2 false positives. Moving the threshold in either direction would decrease the number of one error type at the cost of the other, but since false positives are more acceptable than false negatives, a lower threshold is advisable. At a threshold of 50%, the number of misses was down to 7, with 8 false positives. Reducing the threshold even more produces an unacceptable large number of false positives (> 100), while not increasing the number of true positives.

Having a non-negligible number of both error types is not satisfactory. A possible way to overcome that is not to include the full headline into the template, but just only one or two of the words. This would make the templates smaller and could increase the number of true positives. So, in consequent runs, not the full heading was included in the templates, but only “Museum botanicum” or “Berolinense”. Again, the same number of templates was created from the same files at the same resolution.

Unsurprisingly, smaller templates caused the likelihood value for positives to increase. At the same threshold as before, 53%, templates containing “botanicum Berolinense” resulted in 79 true positives (meaning only 3 specimens had been missed) and 8 false positives. At a threshold of 55%, there were still 3 false negatives, and the number of false positives down to 5. Reducing templates to the word “Berolinense” yielded even better results: At a threshold of 55%, the number of false negatives was down to 0, with 14 false positives; at a threshold of 60%, there were one false negative and 9 false positives.

So it turns out that smaller templates yield better results: Likelihood values increase for all hits, unchanged thresholds will cause the number of both true and false negatives to go up. Raising thresholds moderately will cause the number of false negatives to go down while leaving false positives at an acceptable level.

Contrast, Image manipulations

When examining the number of true/false positives per template after the first test runs, it showed that there were quite huge differences between the different templates. Some of them produced a relatively huge number of false positives while resulting in only a few true positives at the same time, whereas other templates accounted for a much higher number of true positives and no false positives, a discrepancy that remained even after the different frequency of the different fonts had been taken into account.

The reason for this was that no attention had been paid to the quality of the source images when the templates had been created. Apart from the fact that some of the labels apparently had a poorer printing quality, the scans were different in contrast, probably due to changes in the scanning equipment, procedures or settings. For subsequent test runs, this was taken into account: After reviews of a few random specimen images, the ones with the best contrast were chosen for creating the templates.

Additionally, some experiments were done with image manipulations. After the templates had been created, several image parameters were changed that could potentially emphasize the characteristic headline on the templates. After the algorithm was re-run, the likelihoods were compared to those for the original templates for true and false positives. The hope was to see that the effect of the manipulations on true positives was different from the effect on false positives. If the increase/decrease would be significantly different, these manipulations could be used to enhance the results of the algorithm.

As it showed, image manipulations did not have such an effect, or it was too small to be effectively used. Some even had a negative effect: The likelihood value increased more for false positive than for true positives.

The manipulations done in Photoshop that did not have a recognisable effect were

- Image/Adjustments/Levels/Increase contrast 1/2/3
- Image/Adjustments/Curves/Increase contrast

Manipulations with a negative effect:

- Image/Adjustments/Levels/Midtones darker

Manipulations with a slight, but not significant effect

- Image/Adjustments/Contrast +50, +70, +90, +100
- Image/Adjustments/Curves/Linear contrast
- Image/Adjustments/Curves/Medium contrast
- Image/Adjustments/Curves/Strong contrast

Results for Peter Reisen

After sufficient training, the findings on useful settings and practices of template creation were used on a second challenge. This batch consisted of 916 images of specimens gathered by Albert Peter during his expeditions to Africa. As in the “training challenge”, all of these specimens do have labels with a custom-printed headline (see figure 2). But unlike the test challenge, the batch did not contain images of other collections.

This was due to a lesson learnt in the training: Measuring the quality of the outcome can be done by counting false results of the algorithm. Once the algorithm is working pretty well, true hits outnumber false hits by far. A true result can be seen as the expected result of the algorithm, while false results need to be counted. Rates for both error types can be evaluated with the setup mentioned above, since both will occur: False negatives are easy to identify by just counting the specimens without any hits. But also the rate for false positives can be evaluated, even though all specimens in the batch are positives.

In addition to the custom-printed “A. Peters” label with the handwriting of the collector, the specimens contain a multitude of other printed labels and objects that will trigger false positives. If a given specimen image gets one or more positive hits and is tagged correctly, it can still have false positives, that is words, phrases or even parts of the plant and shadows cast by envelopes misidentified as the “A. Peters” label heading. These false positives wouldn’t have an effect on the classification result, which would be still correct for the true positives. For the assessment of the outcome of the algorithm, not the classification result for the specimen as a whole, but the quality of individual positives identified is much more appropriate.

This was done for this second challenge: The algorithm was run on image specimens of the A. Peters. The numbers of specimens that have been correctly and incorrectly tagged are counted (true positives and false negatives, summing up to the total of images). In addition, the number of false positives was determined, with positives referring to false regions of an image identified a hit. These false hits would, in production use, cause incorrect classification results.

According to the findings of the training, the resolution of the original TIFF images was reduced to 25%, the resulting images saved as bitmaps. The algorithm was run with a threshold (both lower and upper) of 50%, and only a part of the label heading was used in the templates (“A. Peter”).

In a first try, 13 templates with different variations of the labels were created, resulting in a huge number of false positives. It showed that two of the templates accounted for almost all of these false hits. These two templates had been created for a variation of the label with a very tiny font, resulting in templates less than 100 pixels wide and less than 25 pixels in height. So there seems to be a certain limit lower limit in template size. Removing these two templates reduced the number of false hits to 3, with 834 specimens correctly tagged and 82 specimens missed.

In a second run, additional 11 templates were created, resulting in 21 templates. This resulted in 907 correct classifications and only 9 missed specimens. Five of these missed specimens did have the tiny font that was too small for the algorithm to cope with; the other four had poorly printed labels. The number of false positives was at acceptable 21.

Interestingly, most of them were caused by one template. The font used in this variation has a peculiarity that the other fonts didn’t have: The characters get very thin at the baseline and the x/cap height, resulting in horizontal streaks only one or two pixels wide, while the vertical streaks of the letters are more dominant. Some specimens did have parts of the plant very similar to these vertical pattern, with no horizontal links (see left of image below). Apparently, these missing streaks reduced the likelihood computed by the algorithm only by a few percent, so it was still higher than the 50% threshold, resulting in false positives:



Figure 8: Font with very thin horizontal streaks at baseline and x/cap height (left); false hits on specimen parts (right)

Despite these few errors, the result seems to be pretty good – only 9 missed specimens out of 916 images. But also a downside of the algorithm showed up: For objects smaller than 100 pixels, the number of false hits increases unacceptably, and fonts with very thin horizontal streaks are problematic.

Conclusions

Apparently, the template matching feature of Linienextraktor can be used for semi-automatic classification of specimen images based on printed label captions – under certain preconditions:

1. The label headings are stable, meaning only a small variation of fonts is used, the quality of label headings is good, and the labels are mounted well-adjusted with only small inclinations.
2. Size of the headings is not too small. Depending on the image resolution, templates must not get smaller than 25px in height.
3. A small number of wrong classifications are acceptable. Depending on the use case, the number of errors can be balanced between false positives and false negatives using the thresholds, but in any case both error types can occur. If no errors can be accepted, using template matching is not an option.
4. The fact that several templates will be used and the algorithm will compute the likelihood for all of them allows more complex rules for tagging the specimens. Instead of basing the decision on the likelihood of just one template (the one with the highest likelihood value), the values for all templates could be taken into account. A specimen could be marked as a hit if either
 - (a) one (or more) template has a likelihood value larger than 60% or
 - (b) two (or more) templates have a value larger than 55% or
 - (c) three (or more) templates have a value larger than 50%.This would require some tool to be developed that analyses the output of Linienextraktor's result files and allows to apply such a rule automatically for all images, but it seems a promising approach to eliminate both false positives and false negatives.
5. Templates need to be created manually before the algorithm can be used. A good overview of the existing variations of label headings is required in order to create efficient templates and keep error rates low.
6. The fonts with very thin horizontal streaks are problematic and might cause quite a few false positives.

Guidelines for “good” templates

- Resolution: Usable results can be expected with resolutions of 300ppi or 150ppi. If resolution of specimen images is higher, it should be reduced to allow Linienextraktor to handle image size. This can be done using the batch mode of Linienextraktor, and must be done before any templates are created from the images.
- Two or three templates per font variation are recommended. The labels chosen for the template creations should be of high printing quality, good contrast values and placed well-adjusted on the specimen. They shouldn't be covered partially by other labels or the plant, and labels with a lot of noise (for example crumbled parts of the specimen) should be avoided.
- If label the heading consists of a longer phrase, only two characteristic words should be part of the template.

- When cutting out the templates, care should be taken to include both baseline and cap height into the template, see image below. Cropping them will increase false positives unnecessarily.



Figure 9: Baseline part of the template (left) and cropped (right)

- Good values for both upper and lower thresholds are between 50 and 60%. If false positives are more acceptable than false negatives, the threshold should be slightly lower (-5%); if false negatives are preferable over false positives, higher (+5%). The smaller the templates created (i.e. the fewer words of the template are included), the higher the threshold value can be set. The higher the resolution used, the smaller the thresholds to be used.
- Linienextraktor allows configuring the number of threads started in parallel. This should be set to the number of physical processor cores or lower.

Outlook

Even though Linienextraktor can be used in its current implementation, it still has some issues and imposes a few restrictions that should be tackled in future versions:

- Certain operations cause runtime errors and access violations, even though sufficient resources are available (free memory & processor time). Such errors occurred when the template matching was run with full resolution images (600ppi), even with just one thread, and sometimes also when template matching was run with smaller resolutions (300ppi, 150ppi) in 2 parallel threads. Also, converting large TIFF files to Bitmaps and reducing image resolution in batch mode failed for some images for no apparent reason: When re-run on the problematic images, it succeeded.
- The batch mode can be used to run template matching, resolution reduction and image format conversion on a huge number of images. However, only files from one individual directory can be added to the batch list in one step. At the BGBM, files are stored in a hierarchy based on the barcode, with about 15 images per directory. Allowing adding a whole directory tree to the batch list would be desirable.
- As explained before, templates are organised in template groups. If one template is found in an image, all other templates of the same group are skipped. This might be useful, but it would be nice to have a system preference to change that behaviour into performing the template matching for all templates.
- Even though the batch mode can be used to run operations on a list of files, this can only be done through the user interface after launching the application manually. To allow a productive use, for example running the template matching automatically once a week on

all newly digitised specimens, starting these batch operations with command line parameters would be required.

Appendix: Terminology

This small glossary defines the terms and their meaning in the context of this report.

Feature – The piece of information in an image which, by its presence or absence, can be used to decide whether or not an object does exist on a specimen

Feature Detection – The process of finding features in a set of images

Image/Specimen classification – Marking images/specimens as considered being part of a certain collection/expedition by means of feature detection

Object – Item that can be found on a herbarium specimen: The plant organism, label(s), type marker, ruler, colour chart, annotation slips, stamps, and barcodes

Tagging – Synonym for classification

Template – Prototypic representation (i.e. image) of an object

PART 2: REVIEW AND TRIALS OF HANDWRITTEN TEXT RECOGNITION (HTR)

INTRODUCTION

The vast majority of the labels on the historic specimens and a large proportion of more recent specimens are either entirely or partially handwritten. This has been problematic for manual transcription since the start of record keeping, particularly for some of the handwriting that is difficult to decipher, and for languages where a significant change in script has occurred such as in Germany. The exciting possibilities which are now being seen and utilised in OCR technology cannot yet be used on most handwriting, only succeeding in transcribing occasional clearly written capitals or numbers.

The discovery of an EU-funded project working on the automatic transcription of handwriting in historical documents was therefore of great interest to the natural history collections community. The project, tranScriptorium, part of the FP7 programme, is a collaboration of six institutes across Europe

- Universitat Politècnica de València – UPV (Spain)
- University of Innsbruck – UIBK (Austria)
- National Center for Scientific Research “Demokritos” – NCSR (Greece)
- University College London – UCL (UK)
- Institute for Dutch Lexicology – INL (Netherlands)
- University London Computer Centre – ULCC (UK)

The tranScriptorium project partners have developed tools which incorporate Handwritten Text Recognition (HTR) technology and which are now available for more general use. The collaboration has resulted in the testing of one of the tools from the tranScriptorium project, Transkribus, being carried out by three partners within the SYNTHESYS3 project.

A protocol for using Transkribus for natural history collections was written and is provided in Appendix 3.

MATERIALS AND METHODS

RBGE HERBARIUM SPECIMENS

The Royal Botanic Garden Edinburgh (RBGE) has a collection of three million herbarium specimens. One of the most important collectors represented in the herbarium is George Forrest (1873–1932) who collected plants and seed in China from 1905 until his death in 1932. He collected significant collections for the horticultural trade resulting in the description of more than 1,200 new species. He collected more than 31,000 specimens in total, the top set of which are held at RBGE. Several duplicate sets were distributed among the sponsors of his expeditions and other botanic gardens including the Royal Botanic Gardens, Kew. His early labels were preprinted with the country, partial date, collector name, and two headings (Alt. and Locality). He would then handwrite the additional collection information such as the month, the altitude, the locality, the species name if known, and a description of the plant and habitat.

There are currently 9,688 specimens already databased (3,559 with images), with approximately 20,000 still to database. RBGE is currently digitising the herbarium collections, and have developed a process for large-scale digitisation which uses minimal data capture and imaging. The minimal data captured are Filing Name, Filing Region and Barcode. The minimal data being entered does not include the collector name. All specimen images are routinely processed using optical character recognition software (OCR). Forrest's use of preprinted labels which include his name allows us to search and pull the records and images of his specimens which have not yet been databased. This search resulted in 750 specimen records.

All specimens digitised at RBGE have been scanned on an Epson 10,000XL flatbed scanner within the HerbScan framework at 600dpi or photographed with a Leaf Aptus II-10 digital back with Mamiya camera and Schneider lens at 300ppi.

HTR TRAINING DATASET

HTR: FORREST_COLLECTION

A training dataset was compiled in two stages at RBGE. The first stage was a small sample of 61 specimens to test the process in order to decide whether the results were sufficiently successful to justify the continuation of the trial.

HTR: FORREST_COLLECTION_2

This preliminary training dataset was then expanded to include an additional 75 specimens resulting in a second training dataset of 136 specimens.

The specimen images were marked up and transcribed by RBGE herbarium staff. Questions arose around consistency of marking up and transcribing, including inclusion of punctuation in marking up and transcribing.

RBGE TEST DATASET

The specimens in the test dataset were selected by searching the OCR output for the minimally databased specimens for the term “Forrest”. The results included specimens collected by Alan Forrest and Laura Forrest. These images were not marked up. The dataset also included a small number of specimens collected by George Forrest which have typewritten labels. These images were also not marked up.

RBGK HERBARIUM SPECIMENS

The Royal Botanic Gardens Kew (RBGK) has a collection of approximately 7 million herbarium specimens. The collections have their origins in the amalgamation, in 1853, of two large pre-existing private collections, namely those of George Bentham and William Hooker. During the Victorian era, the collection grew further, fostered by these two men and by Joseph Hooker. Amongst the collection RBGK have specimens collected by several famous collectors including William Burchell, George Forrest, George Gardner, Arthur Kerr and Richard Spruce amongst others. Currently Kew has >780,000 digital herbarium specimen records of which >430,000 have an accompanying image.

Specimens collected by Arthur Kerr and George Forrest were chosen for the Transkribus Trial. Unlike Royal Botanic Gardens Edinburgh, Kew has not yet implemented OCR within digitisation workflows. Therefore it is not until the specimens labels are transcribed, post minimal data capture and imaging, that the collector of the specimen becomes known. Hence for the trial RBGK used specimens already transcribed. It is planned to investigate OCR integration into digitisation workflows in the future and include this step in any potential Transkribus workflow.

George Forrest collections were chosen so they could be included in the trials being conducted by RBGE, the aim was to investigate if the HTR model developed for RBGE could be successfully applied to RBGK specimens. Arthur Francis George Kerr collections were chosen as he is one of the top five collectors within the Herbarium Specimen Catalogue with over 7377 specimens currently digitised and therefore example images readily available. Similarly to Forrest collections the Kerr collections labels tend to have a standard format with preprinted Country info e.g. Flora of Siam and preprinted headings for Collector number, Locality, Date, Local Name and Notes. He would then handwrite the information for these headings alongside them.

All specimens included in the trial at RBGK have been scanned on an Epson 10,000XL flatbed scanner within the HerbScan framework at 600dpi, photographed with a Leaf Aptus-II I2 digital back (80MP) with Cambo sliding back and bellows with a 90mm Rodenstock lens and Schneider electronic shutter or photographed with a PhaseOne iXR camera setup with Credo 80 back (80MP).

RBGK FORREST TEST DATASETS

A batch of 250 Forrest specimens images were chosen at random and uploaded to test against the HTR model “Forrest_Collection_2”. The images were converted from the original tiff format to Jpeg but the resolution was kept at 600ppi resulting in a large 6MB file. Only those files with handwritten text were marked up.

MRAC TERVUREN SPECIMEN REGISTRIES

MRAC tested the Transkribus program specifically for transcribing specimen registries. The registers of MRAC are handwritten books where specimens were documented as they came to museum. The information on the original labels, if they were kept, came from these books and are not as detailed. Over time, some curators have had the unfortunate idea to replace some labels, losing original information. The data stored in these registries are therefore, for many specimens, the only original documentation from their acquisition. Curators often refer to these valuable documents when there is a doubt either to the numbering or the provenance of a specimen.

These registers are being scanned as part of the DIGIT03 project with Book scanner placed at the Botanical Garden of Meise (Figure 3). The books have big format (39.5×55.5 cm). Unlike most books, one line (one record) runs over two pages and the pages are not numbered, therefore a complete book spread was scanned, instead of each page separately. Images are of high resolution (600 dpi) and are in tiff format.



Figure 3

The headers of the columns are printed and written in Dutch or French. The content is handwritten, composed of Latin names, collection numbers, dates, collector and location names, and abbreviations such as ♀ and ♂ to denote the gender of the specimen. Additional remarks are mostly in French or Dutch, but sporadically other languages can appear. The columns of these registers are described below.

1. specimen number
2. scientific name (sometimes with corrections when an error was made or has been redefined, in the same column you may find a sign if it is a male, female or juvenile)
3. locality
4. date of collection which can be a range or a specific date
5. the name of the determiner
6. on the opposite page and lists the collector
7. date of reception at the museum
8. additional information about the observation.

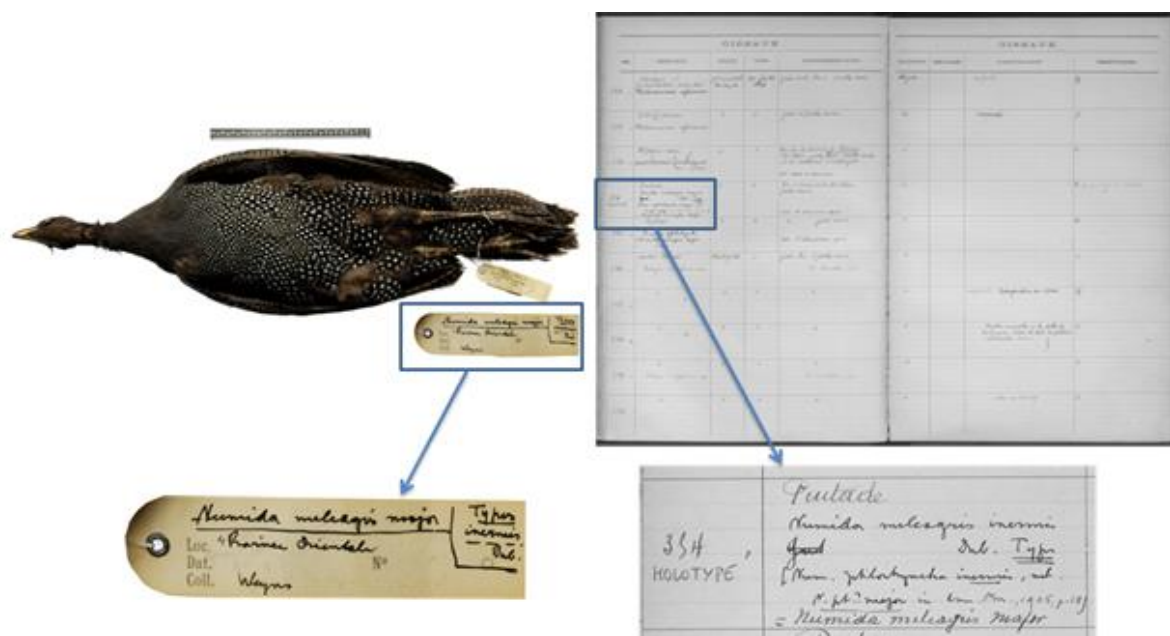


Figure 4. An example of a specimen and its associated entry in the register of birds.

The work on transcribing the registries is ongoing.

RESULTS

RBGE HERBARIUM SPECIMENS

It took approximately 1 min/specimen to mark up and transcribe the training set. This was a total of approximately 2.5 hours for all 136 specimens. The tranScriptorium team then processed these records to create an HTR model.

COMPARISON OF HTR MODELS

The first HTR model was trained using 61 specimens as a test. This HTR model “Forrest_Collection” showed some success and a decision was made to continue with the trial.

The second HTR model was trained using 136 specimens.

The results from this HTR model “Forrest_Collection_2” were compared with the results from “Forrest_Collection”. The metrics used for comparison were the Word Error Rate (WER) and the Character Error Rate (CER) which have been incorporated into the software. For each page the output from each of the HTR models was compared to the manual transcription and the Error Rates calculated.

Doc No	Page	HTR1 (based on 61 specimens)	WER	CER	HTR2 (based on 136 specimens)	WER	CER
771	1	Forrest_Collection	60.0	6.667	Forrest_Collection_2	44.0	6.667
771	2	Forrest_Collection	82.609	14.474	Forrest_Collection_2	60.870	15.789
771	3	Forrest_Collection	60.714	6.936	Forrest_Collection_2	53.571	17.341
771	4	Forrest_Collection	33.333	3.315	Forrest_Collection_2	36.667	3.867
771	5	Forrest_Collection	47.059	12.963	Forrest_Collection_2	52.941	12.963
771	6	Forrest_Collection	na	na	Forrest_Collection_2	na	na
771	7	Forrest_Collection	77.778	10.778	Forrest_Collection_2	59.260	10.778
771	8	Forrest_Collection	73.077	10.180	Forrest_Collection_2	57.692	8.982
771	9	Forrest_Collection	47.059	4.566	Forrest_Collection_2	35.294	4.566
771	10	Forrest_Collection	na	na	Forrest_Collection_2	na	na
774	1	Forrest_Collection	63.415	10.811	Forrest_Collection_2	51.220	10.811
774	2	Forrest_Collection	na	na	Forrest_Collection_2	na	na
774	3	Forrest_Collection	na	na	Forrest_Collection_2	na	na
774	4	Forrest_Collection	53.659	12.157	Forrest_Collection_2	41.463	13.725
774	5	Forrest_Collection	na	na	Forrest_Collection_2	na	na
774	6	Forrest_Collection	na	na	Forrest_Collection_2	na	na
774	7	Forrest_Collection	55.000	18.966	Forrest_Collection_2	75.000	29.310
774	8	Forrest_Collection	75.758	9.906	Forrest_Collection_2	57.576	10.377
774	9	Forrest_Collection	95.000	9.483	Forrest_Collection_2	95.000	27.586
774	10	Forrest_Collection	na	na	Forrest_Collection_2	na	na

Table 21. The results of the comparison between the two HTR models.

RBGE TEST DATASET

It took approximately 36 secs/specimen to mark up the test set. This was a total of approximately 7.5 hours for all 750 specimens.

RBGK HERBARIUM SPECIMENS

RBGK experienced some issues running the Transkribus software within Kew as It was not possible to log in to the application on the Kew network although it was possible to log in via the Wi-Fi network, however this was slow and did not allow the upload of images. It was suspected by Transkribus that the issue was a firewall problem which was blocking access, this issue was reported to Kew IT department however the IT department was not clear how to solve the problem without further details. After a few weeks with little progress a new version of the software was downloaded and tested v.0.6.3 through which it was then possible to login using the Kew network. It was unclear if a change in the Transkribus software or a change in the settings in the Kew network solved the log in issue.

RBGK FORREST TEST DATASET

Uploading images was very slow a batch of 200-250 images took many hours so the upload was left to run overnight. It was also found that after marking up all the images as more data and versions of the data were created when transcribing or running the HTR model the images stopped loading up properly and they appeared blank with only the mark up visible. The software began to run slowly with Java error messages appearing. Transkribus advised that this was due to the large size of images it was suggested that this could be resolved by sending the images to them to upload, splitting the images into several documents, or using a higher compression e.g. 50% is fine for the HTR since it removes just the number of colours and colours are more or less irrelevant for HTR (pers com. Günter Mühlberger).

The 250 Forrest specimens were marked up, skipping any labels without handwritten labels and the HTR model created by RBGE "Forrest_Collection_2". run on a few specimens. It was immediately noticed that the results were very poor compared to those obtained by RBGE.

Doc No.	Page	HTR2 (based on 136 specimens)	WER	CER
1993	31	Forrest_Collection_2	128.889	64.591
1993	70	Forrest_Collection_2	110.000	63.462
1993	246	Forrest_Collection_2	97.778	47.348

Table 22. Example of results using 600ppi RBGK images.

Doc No	Page	HTR1 (based on 61 specimens)	WER	CER	HTR2 (based on 136 specimens)	WER	CER
2016	1	Forrest_Collection	81.481	31.361	Forrest_Collection_2	74.074	27.219
2016	2	Forrest_Collection	87.500	51.724	Forrest_Collection_2	71.875	37.931
2016	3	Forrest_Collection	73.333	40.467	Forrest_Collection_2	64.444	31.518
2016	4	Forrest_Collection	87.234	57.627	Forrest_Collection_2	76.596	47.458
2016	5	Forrest_Collection	94.000	78.967	Forrest_Collection_2	84.000	64.945
2016	6	Forrest_Collection	100.000	78.365	Forrest_Collection_2	92.500	72.115
2016	7	Forrest_Collection	92.105	46.606	Forrest_Collection_2	86.842	36.652
2016	8	Forrest_Collection	68.571	32.275	Forrest_Collection_2	68.571	28.571
2016	9	Forrest_Collection	88.571	41.436	Forrest_Collection_2	71.429	36.464
2016	10	Forrest_Collection	86.667	70.076	Forrest_Collection_2	91.111	70.833

Table 23. Example of results using images resized to match RBGE images.

RBGK KERR TEST DATASETS

A batch of 200 Kerr specimens images were chosen at random and uploaded for label Mark up. Only those files with handwritten text by Kerr were marked up. The tranScriptorium team then processed these 840 transcribed lines to create the HTR model. Kerr_Collection_1.

FLORA OF SIAM

No. 1658

Locality Doi them, Chiang mai

Altitude 1,100 ft.

Date January 23rd. 1911.

Notes Flowers white with purple markings on lip; by stream.

A. F. G. Kerr

Line based | Region: 1/1 | HTR suggestions CATTI

- 6
- Doi them Chiangmai
- 1100 ft
- January 2nd 1911
- Flowers white with purple among
- on lip ; by stream

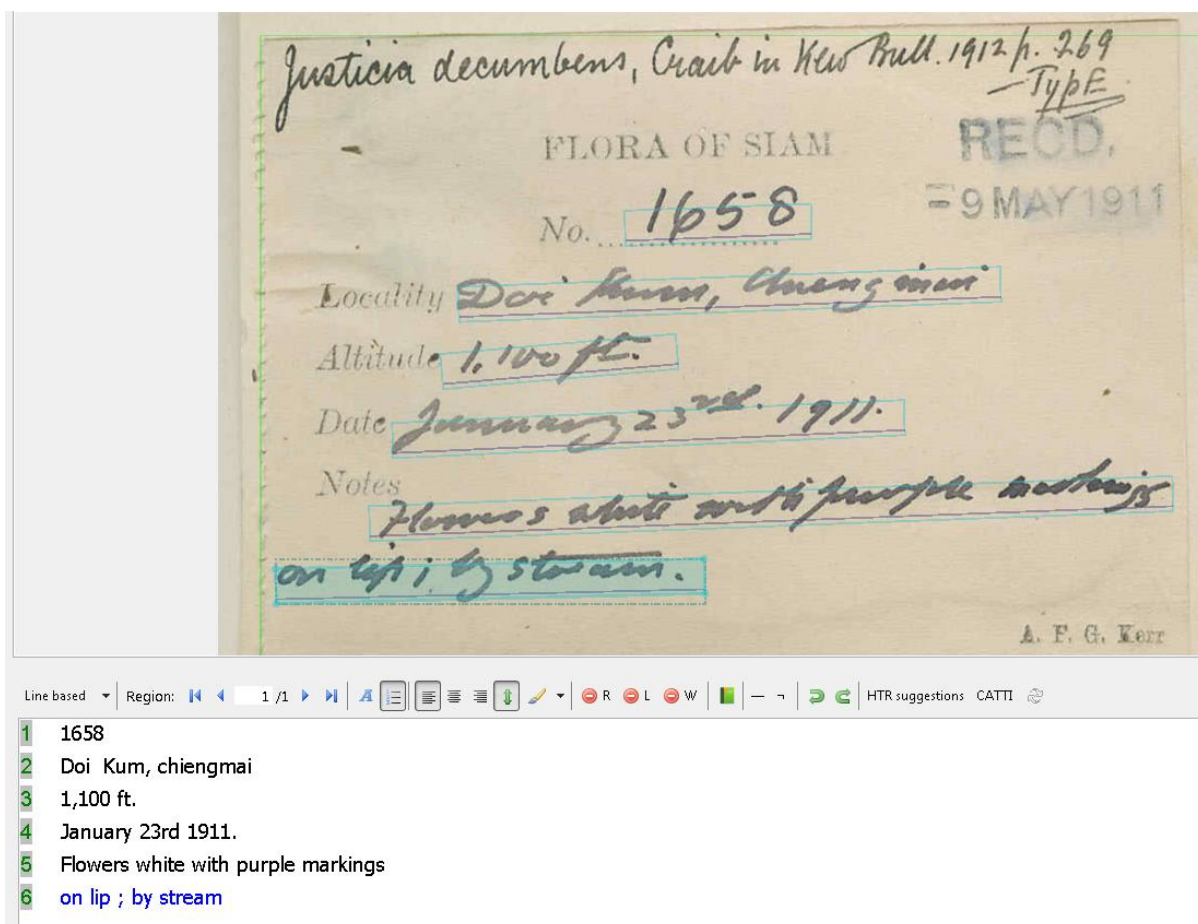


Figure 5. An example of the HTR result using the Kerr_Collection_1 model from a Kerr specimen label. HTR result below first image, actual transcription below second image.

Doc No.	Page	HTR 2 based on 840 transcribed lines)	WER	CER
2547	1	Kerr_Collection_1	66.667	46.296
2547	2	Kerr_Collection_1	77.277	23.478
2547	3	Kerr_Collection_1	42.105	17.822
2547	4	Kerr_Collection_1	63.636	37.705
2547	5	Kerr_Collection_1	50.000	30.588
2547	6	Kerr_Collection_1	105.882	95.789
2547	7	Kerr_Collection_1	46.154	22.973
2547	8	Kerr_Collection_1	69.23	47.1443
2547	9	Kerr-collection_1	54.16667	24.545
2547	10	Kerr-collection_1	66.67	23.076

Table 24. Example of results of Kerr_Collection_1 model.

A specimen label was also run using Forrest_Collection_3 but as expected the results were poor as the handwriting and vocabulary especially localities are very different.

First there was an issue connecting to Transkribus from the museum. It was only possible if the user was in a group that allows complete internet access. Because initially the member of staff testing the tool was in another group, it was not possible to log in with their username and password and even with the ICT help the problem with login procedure could not be identified. The program indicated: login failed. There may be a port connection issue but this has not yet been resolved. The program has therefore only been possible to use with the Firewall down which is not generally advisable.

For such unusual type of document the layout is very important and it took some time to determine how it should be arranged. The time required to transcribe a test set of 100 pages and the time required to mark up additional pages, combined with the login issues resulted in the test not continuing.

DISCUSSION

For the George Forrest dataset, the comparison of the two HTR models built from the smaller and the larger training sets showed the benefit of the additional transcriptions in the HTR model. The main difference was apparent in the Word Error Rate where on average the rate of errors was reduced. Where the initial error rates were low, a slight increase of errors were seen using the larger training set.

George Forrest generally used a relatively narrow vocabulary on his specimen labels. He used standard localities, his plant descriptions tend to follow a consistent format and his habitat descriptions use a limited lexicon. His handwriting is also generally neatly aligned and does not appear to change significantly over time.

The results obtained from the Kerr-Collection_1 model are encouraging and it would be interesting to see if improvements could be made by increasing the number specimens used in the training set. Like Forrest Kerr uses limited Vocabulary on his labels and often localities are recurring. *Reliquiae Kerrianae* (Blumea Vol. XI, NO.2, 1962 pp.427-493) includes a list of published material by Kerr, his collections and localities, as well as a detailed itinerary of places he visited and collected in. It might be interesting to see if any of this material could somehow be incorporated into the training model to improve it.

Although promising it is unlikely that Kew will currently incorporate Transkribus into its workflow as it is still quicker to transcribe the label manually rather than mark up the label, run the HTR model and then correct the output that it produced. However as the technology improves this might change. Further investigation on archive material, diaries and letters would be worth exploring.

SECTION 4: REVIEW OF AUTOMATIC CAPTURE OF CHARACTER INCLUDING COLOUR, SHAPE AS WELL AS EXIF DATA.

PART 1: COMPUTER VISION FOR SPECIMEN CLASSIFICATION

This report has been produced as a separate document and is inserted here.

Computer Vision for Specimen Classification

Project: Synthesis of systematic resources

Project acronym: SYNTHESYS3

Grant Agreement number: 312253

Workpackage: Work Package 4 Moving from physical to digital collections

Deliverable number: 4.2

Deliverable title: Optimal automated metadata capture

Deliverable authors: James Durrant, Laurence Livermore and Lawrence Hudson

Date: 21 July 2015

Table of Contents

[Summary](#)

[Tools Used](#)

[Software Prototypes](#)

[Specimen segmentation](#)

[Method](#)

[Morphological feature detection](#)

[Calculating physical dimensions](#)

[Colour analysis](#)

[Heat maps for regions of interest](#)

[Dissemination](#)

[Links](#)

[References](#)

Summary

SYNTHESYS is a European Union-funded Integrated Activities grant which aims to create an accessible, integrated European resource for researchers in the Natural Sciences. The Joint Research Activity (JRA) is one of its three main activities and aims to improve the quality of and increase access to digital collections and data within natural history institutions' virtual collections.

One of the JRA objectives was to support and develop technology that automated data collection from digital images. As part of the NHM's contribution to this objective we have developed a series of open source prototypes that do the following: 1) segment specimens from their backgrounds and segment regions of interest (e.g. particular body parts); 2) detect morphological features to be used for classification (e.g. markings that indicate gender); 3) calculate of physical dimensions from images (e.g. wing length); 4) colour analysis to be used for classification (e.g. wing colours); 5) heat maps for regions of interest.

Tools Used

All of the external libraries used to develop the software prototypes are open source. They are continually being updated and improved but are also free to use and work on commonly used operating systems (Windows, OSX, Linux). Wherever possible built-in methods from these libraries were used since they well-supported, generally well-documented and tested.

Main programming language: **Python** <https://www.python.org/>

External libraries used:

- **OpenCV**: Open-source computer vision library. <http://opencv.org/>
- **SciPy**: Scientific computing in Python. <http://www.scipy.org/>
- **scikit-image**: Image processing. <http://scikit-image.org/>

Code repository: **GitHub** https://github.com/NaturalHistoryMuseum/insect_analysis

Software Prototypes

Specimen segmentation

Specimen segmentation acts as utility function and is required to perform more advanced metadata extraction. The aim was to extract specimens from their surrounding image and exclude features of non-interest e.g. labels and rulers.

Method

1. Generate a saliency map.

- This is essentially a map of which parts of the image are more interesting or relevant.
- Maps can be created in different ways for different applications, but for this set of butterfly images this was done using a weighted combination of the Saturation and Brightness channels of the image.
- This gets a very high response from the specimen with not much elsewhere except in this case the label.
- The whole image is thresholded such that any pixel above a certain saliency value is set to white and any pixel below is set to black
- By analysing contiguous areas of white we can find the largest such region, which will be the specimen

2. Separating specimen from background to get a silhouette

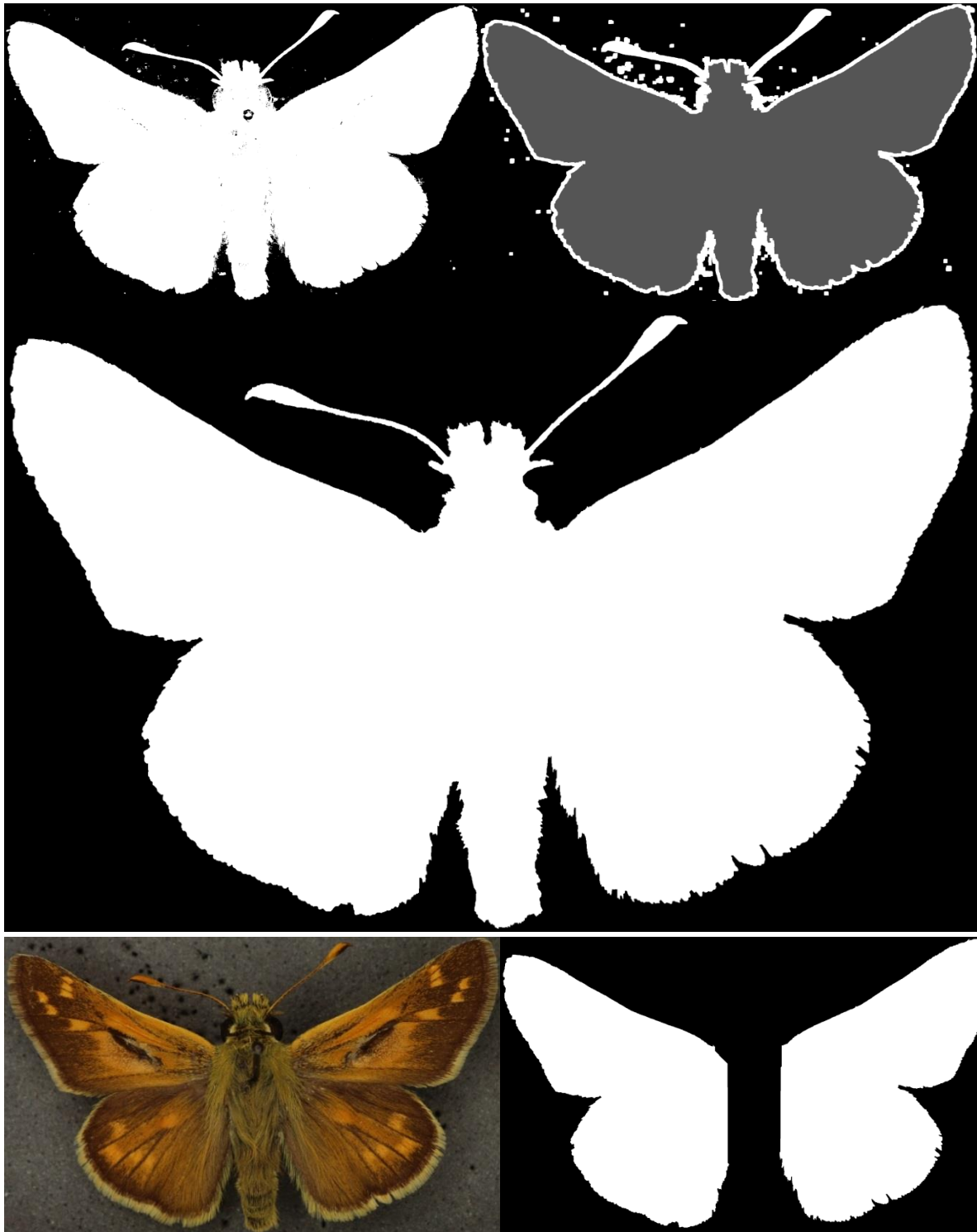
- At this point the image is cropped so that it contains only this region
- There are some false positives in the classified pixels, however, and also some holes in the silhouette that shouldn't be there
- To correct this, first the holes are filled in in the largest connected component. This is done by finding the complete contour with the largest area and filling in any black pixels
- Next, we shrink the black and white regions away from the edge where they join, since we can assume that the real edge is somewhere inside this new

region (white area in the diagram) Using a graph cut algorithm (inside OpenCV) we can find the single cut that partitions these two regions, given the constraints. This results in a clean mask with minimal errors

3. Segmenting the wings away from the abdomen

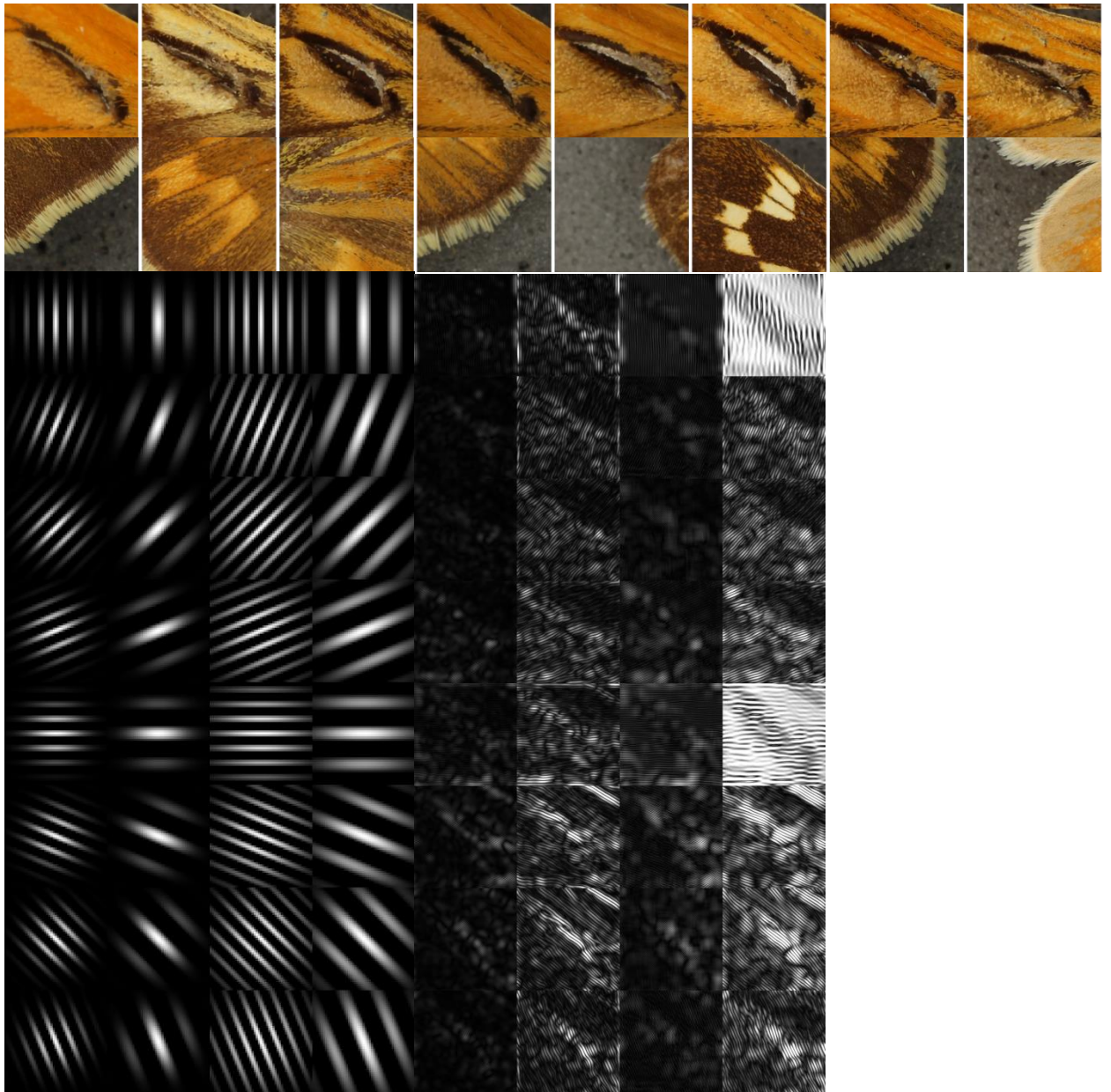
- Found by looking for the shortest path between a point above the specimen and a point below the specimen, directly underneath the abdomen. This goes between the wing and the body since it is almost invariably the shortest distance
- Using this path we partition the image such that the middle segment is the abdomen and the outer segments are the two wings
- In computing the shortest path we assume that moving through an area of black (background) costs only as much as the distance travelled, whereas going across an area of white pixels (foreground) incurs additional costs

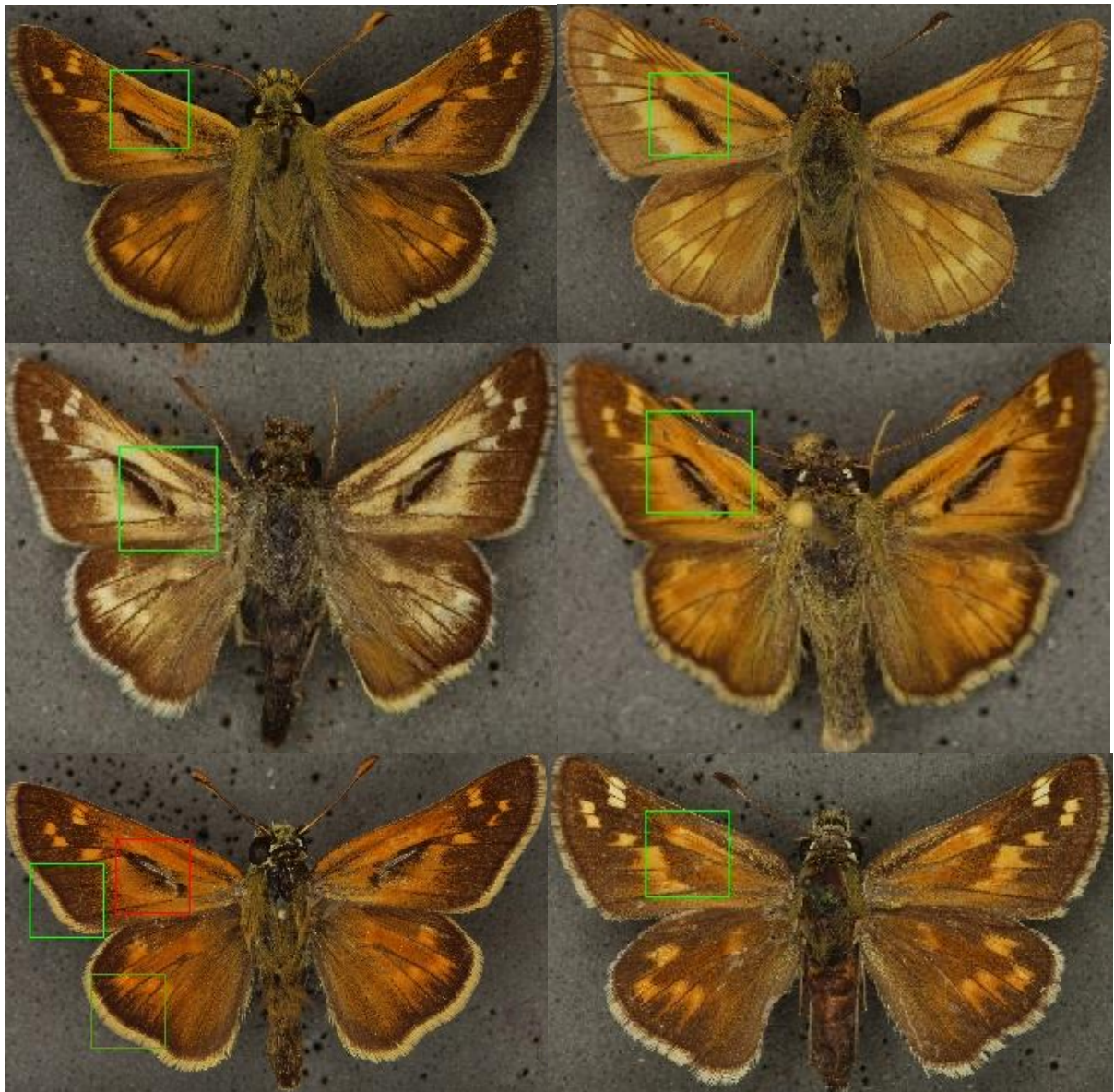




Morphological feature detection

- Manually extract examples of the object you are trying to detect, in this case a gland on the wing of the butterfly that is found only on the males of this species. These are the 'positive' samples
- Generate 'negative' samples, which are chosen as anything where the specified object is not observed. This is done automatically on the set of female specimens since no sample could contain the object
- Need a way to describe a particular region:
 - Pixel values?
 - Lots of information
 - Not much structure
 - Lots of noise
 - Feature descriptors
 - Meaningful
 - Compact (generally)
- Run a filter over all of the samples. This produces results as seen below and is a good descriptor for textured regions
- Using the filtered results from the example regions need to produce a model that can differentiate between them.
 - Support Vector Machines:
 - Finds a Hyperplane in the vector space of the feature descriptor that minimizes
- For any new image we can apply the same set of filters again and use the response together with the model to determine whether that image is representative or not
- For a full specimen, this needs to be done for every possible sub-image, though for efficiency it is done on a coarse grid, this is still good enough for reasonable accuracy however.
- This may turn up multiple possible locations, which need to be handled:
 - If many possible matches are overlapping, only select the one with the highest predicted probability and remove all others





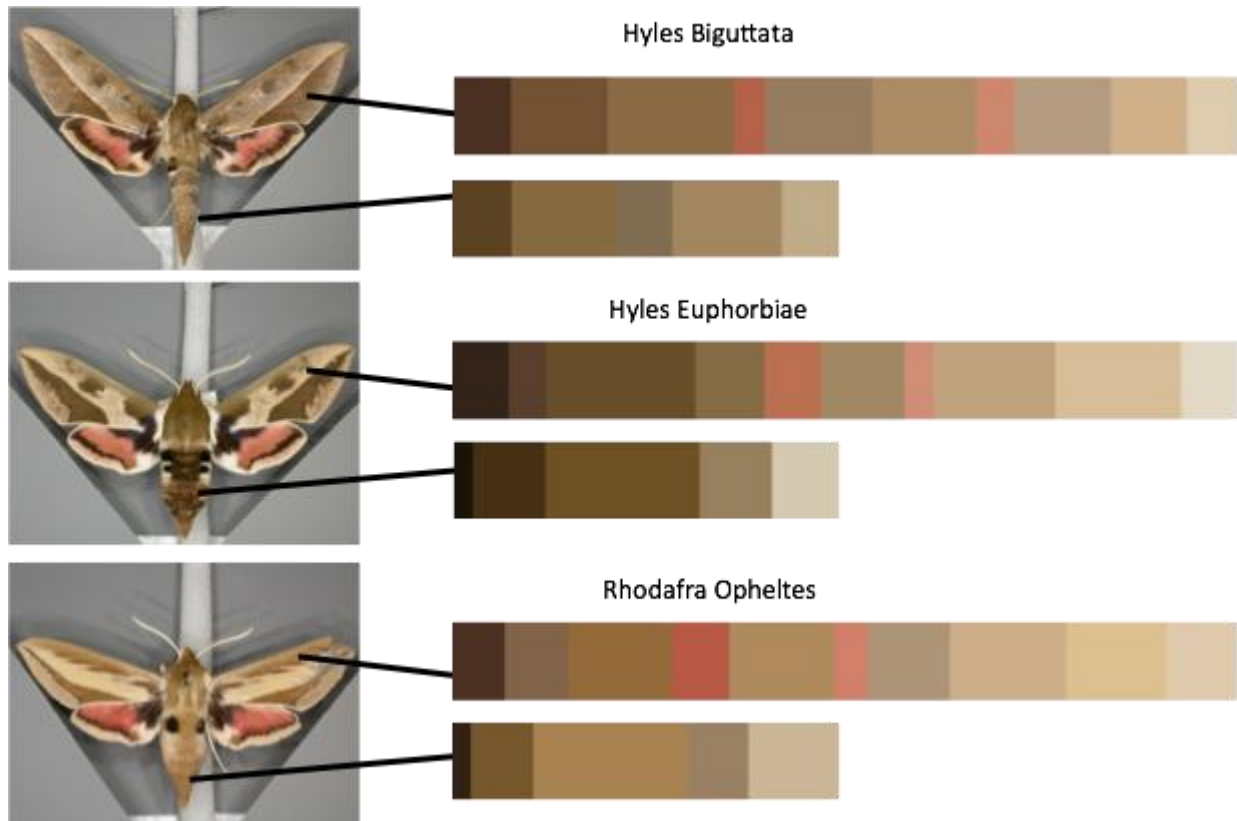
Calculating physical dimensions



- Also using the segmentation
- Calculate length from the body up to the wing tip
- First step is to find the point where the wing meets the body, which is done by analysing the partition between the two segments
- Then, considering only the region above and to the right (or left, depending on the wing), the distance to every point on the wing is measured and the point furthest away is chosen as the wing tip, and the wing length is recorded as the distance between these two points
- This measurement is recorded as a distance in pixels, and so needs to be converted into real world units. This can be done by analysing the ruler at the bottom of the specimen images, however there was not enough time in the project to implement this
- For the images of the *Hesperia comma* we have hand measured wing lengths that can be compared against and so for 20 images the scale conversion was done by hand to see how well the automated measurements compared. With the exception of 1 outlier, all measurements were within 0.45mm and the average of the left and right wings was within 0.21mm of the average of the actual recorded lengths
- It appears from the images that some of the wings are angled away from the camera such that they appear shorter in the image than they would be in real life. If there was a second view of the specimen from the front then this angled could be estimated and accounted for in the length calculation

Colour analysis

Moths



- Used segmentation to separate abdomen from wings so that they can be analysed independently
- Want to find a way to compare the colours of different moths to see whether they could be from the same species
- Pixel values have a lot of noise and variation so rather than comparing them directly, instead use descriptive metrics
- Tried using the mean saturation and hue of each segment, but this cannot take into account the wing markings and is very approximate
- We can instead find the most dominant colours in the segment and compare these
- This cannot take into account the spatial distribution, but it does compare the proportions of the colours with respect to each other and appears to be descriptive enough to differentiate between species in the same family
- L*a*b* colour space - designed to be perceptually uniform:
 - Pairs of colours that are the same distance apart in Lab colour space should appear to be of equal similarity to a human
 - In contrast to RGB which is nonlinear since human eyes are more receptive to green wavelengths than the other two channels

Heat maps for regions of interest

- Tried initially to find areas of higher chlorophyll using two methods:
 - Firstly, using just the green channel of the RGB images
 - Secondly using the negative of the 'a' channel of the LAB images, that is, a lower value of 'a' indicates a higher concentration of chlorophyll
- Neither of these had a very strong response in the areas where a higher concentration was expected
- Hu et al (2015) suggested some other metrics that seemed less intuitive but have a much higher correlation with measured values. The one used here is an average of the negation of the Red and Green channels



Dissemination

Research Presentation: NHM Science Seminar (July 2015)

James Durrant. From Pixels to Species: Computer Vision for Specimen Classification

<https://www.youtube.com/watch?v=JnkfYtg7ipQ>

Public Outreach: Science Uncovered (September 2015 - planned)

“Looking at our specimens in a different way” - Interactive stand staffed by scientists to discuss computer vision work with the general public.

Links

Insect analysis repository: https://github.com/NaturalHistoryMuseum/insect_analysis

References

Hu, Liu, Zhang, Zhu, Yao, Zhang & Zheng. 2010. Assessment of chlorophyll content based on image color analysis, comparison with SPAD-502. Information Engineering and Computer Science (ICIECS), 2010 2nd International Conference on.
DOI: 10.1109/ICIECS.2010.5678413

PART 2: CORRELATION OF LEAF COLOUR AND DNA QUALITY

INTRODUCTION

The use of herbarium specimens as a source of molecular data for species and populations has potential to enable researchers to access data for species not available anywhere else. In addition, herbarium specimens represent a verifiable entity, often identified by world experts. The variables which can affect the quality of the DNA extracted include: age of specimen; method of collection and drying; taxonomic group; conditions of storage; subjection to pest treatment including freezing, heating or chemicals; amount of material available; and leaf colour.

Research undertaken on the quality of DNA in herbarium specimens has previously found that the method of drying has a large impact on DNA quality. In this study, we aimed to control for age, collection & drying method, region and taxon in order to increase the strength of signal from colour variable.

MATERIALS AND METHODS

A total of 96 specimens were selected for DNA extraction. The following criteria for specimen selection were used:

- Sets of specimens collected by a single collector for whom the collection and drying method are known
- Specimens which were collected within particular periods of time as much as possible
- Specimens from two geographical regions, although vegetation type was not controlled for
- Sets of specimens from a single family or genus

For each specimen, a section of leaf was sampled. Both upper and lower leaf surface of the sample was imaged using a Leaf Aptus II-10 digital back with a colour chart. The samples were then transferred for DNA extraction. The protocol for this can be found in Appendix 4.

RESULTS

The results are yet to be analysed and this work will now feed into the NA2 Task 2.3.

REFERENCES

Smith, V., & Blagoderov, V. (2012). Bringing collections out of the dark. *ZooKeys*, 209(0), 1–6. doi:10.3897/zookeys.209.3699

Ariño, A. (2010). Approaches to estimating the universe of natural history collections data. *Biodiversity Informatics*, 7, 81–92. Retrieved from <https://journals.ku.edu/index.php/jbi/article/viewArticle/3991>

Duckworth, W.D., Genoways, H.H., Rose, C.L. (1993). Preserving natural science collections: chronicle of our environmental heritage. National Institute for the Conservation of Cultural Property 140pp.

Haston, E., Cubey, R., & Harris, D. J. (2012). Data concepts and their relevance for data capture in large scale digitisation of biological collections. *International Journal of Humanities Arts Computing*, 6(1/2), 111–119. Retrieved from <https://libproxy.library.unt.edu:9443/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=hlh&AN=71714722&site=ehost-live&scope=site>

SOFTWARE AND PROJECTS

ABBYY FineReader. <http://www.abbyy.com/finereader/>

ABBYY Recognition Server. <http://www.abbyy.com/recognition-server/>

Cuneiform. <http://www.filesriver.com/app/107/openocr>

Custom OCR. <http://www.customocr.com/>

Free OCR. <http://www.free-ocr.com/>

Free online OCR. <http://www.free-online-ocr.com/>

GImageReader. <http://dottech.org/21372/gimagereader-open-source-google-powered-ocr-optical-character-recognition-program-that-actually-works/>

Global Plants. <https://plants.jstor.org/>

I2OCR. <http://www.i2ocr.com/>

Newocr.com. <https://www.newocr.com/>

OCR Convert. <http://www.ocrconvert.com/>

OCRextrACT. <http://www.cvisiontech.com/ocr/best-ocr/best-ocr-extract.html>

Ocrgeek.com. <http://ocrgeek.com/>

OCRonline.net. <http://www.onlineocr.net/>

OCRonline.com. <http://www.ocronline.com/>

OmniPage. <http://www.nuance.co.uk/for-business/by-product/omnipage/standard/index.htm>

Presto!OCR. http://us.newsoft.com.tw/company/news_style.php?NT_Id=1&N_Id=313

Pumanet. <http://pumanet.codeplex.com/>

Salix. <http://daryllafferty.com/salix/>

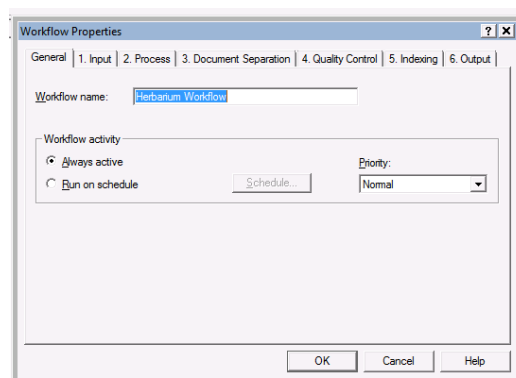
Scanitto. <https://www.scanitto.com/>

Simple OCR. <http://www.simpleocr.com/>
Symbiota. <http://symbiota.org/docs/>
TopOCR. <http://www.topocr.com/>
tranScriptorium. <http://transcriptorium.eu/>
Transkribus. <http://transcriptorium.eu/transkribus/>
TypeReader. <http://www.expervision.com/ocr-software>
WeOCR. <http://weocr.ocrgid.org/>

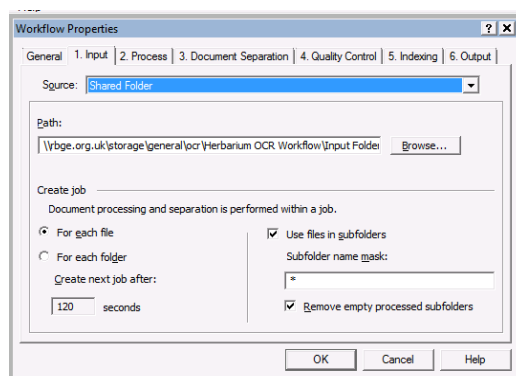
APPENDIX 1A: SETTINGS FOR ABBYY RECOGNITION SERVER V3 AT RBGE

The following screenshots show the settings used.

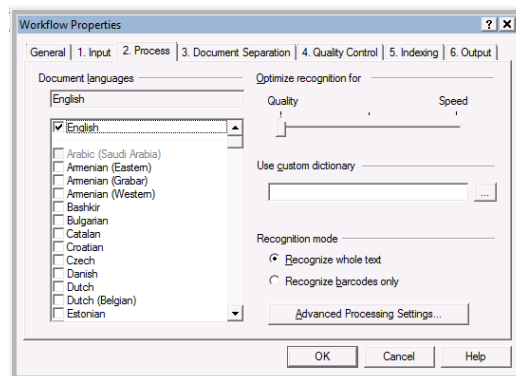
1. General settings: Option used [other options available]
 - a. Workflow Name: Herbarium Workflow
 - b. Workflow activity: Always active [Run on schedule]
 - c. Priority: Normal [High, Above normal, Below normal, Low]



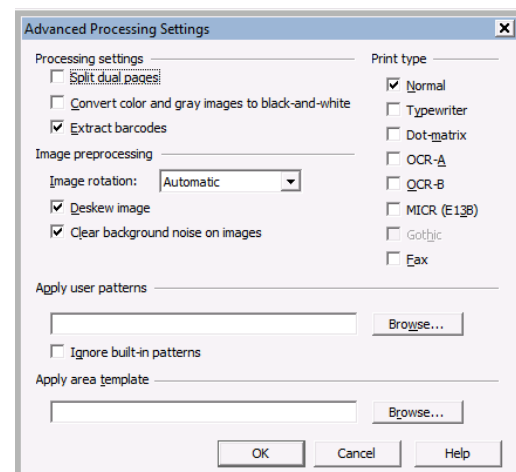
2. Input settings: Option used [other options available]
 - a. Source: Shared Folder
 - b. Path: url of Input Folder
 - c. Create job: For each file [For each folder]
 - d. Create next job after: 120 seconds
 - e. Use files in subfolders
 - f. Remove empty processed subfolders



3. Process settings: Option used [other options available]
 - a. Document languages: English [many]
 - b. Optimize recognition for: Quality [Speed]
 - c. Use custom dictionary: none selected
 - d. Recognition mode: Recognize whole text [Recognize barcodes only]

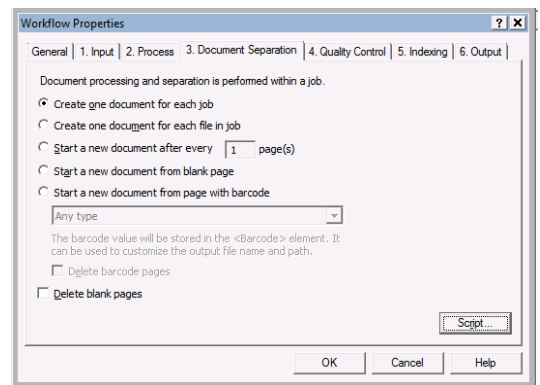


4. Advanced Processing Settings: Option used [other options available]
 - a. Processing settings: Extract barcodes [Split dual pages, Convert color and gray images to black-and-white]
 - b. Image preprocessing: Automatic image rotation [No rotation, Clockwise, Counterclockwise, Upside-down]
 - c. Deskew image: selected
 - d. Clear background noise on images: selected
 - e. Apply user patterns: none selected
 - f. Apply area template: none selected



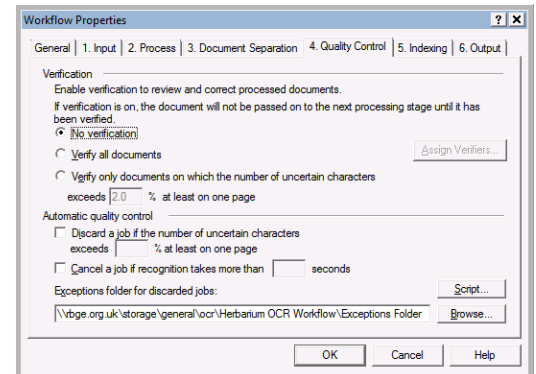
5. Document Separation settings: Option used [other options available]

- Document processing and separation is performed within a job: Create document for each job [Create one document for each file in job, Start a new document after every x page(s), Start a new document from blank page, Start a new document from page with barcode]
- Delete blank pages: not selected
- Script: none selected



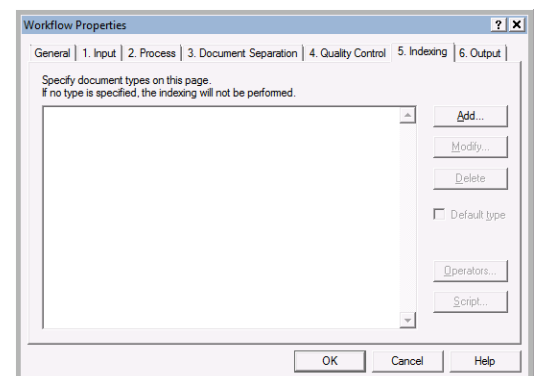
6. Quality Control: Option used [other options available]

- Verification: no verification [Verify all documents, Verify only documents on which the number of uncertain characters exceeds x% at least on one page]
- Discard a job if the number of uncertain characters exceeds x% at least on one page: not selected
- Cancel a job if recognition takes more than x seconds: not selected
- Script: none selected
- Exceptions folder for discarded job: url of Exceptions Folder



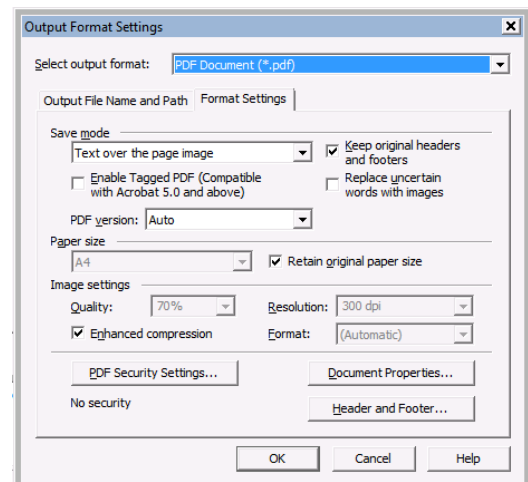
7. Indexing settings: Option used [other options available]

- none specified



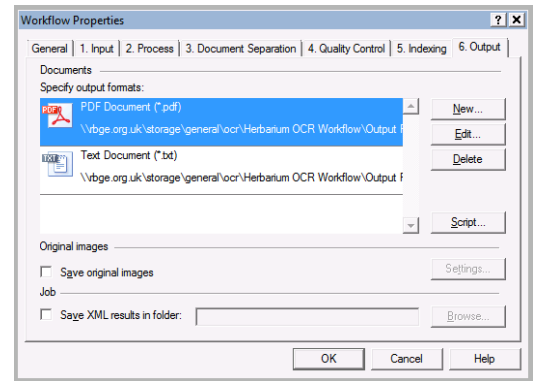
8. Output settings: Option used [other options available]

- Specify output formats: .pdf, .txt [.doc, .docx, .xls, .xlsx, .rtf, .csv, .htm, .tiff, .jpg, .j2k, .jb2, .epub]
- Script: none selected
- Save original images: not selected
- Save XML results in folder: not selected



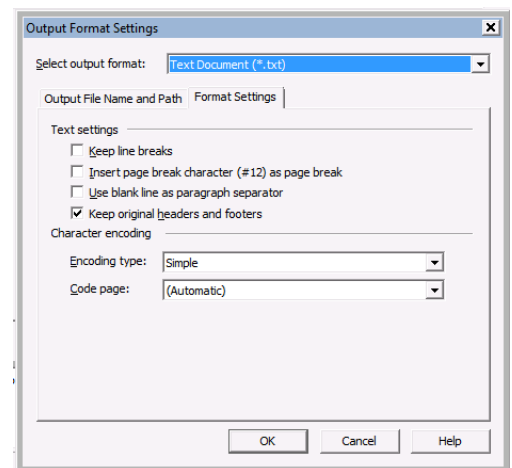
9. PDF output settings: Option used [other options available]

- a. Save mode: Text over the page image [Text only, Page image over the text, Image only]
- b. Keep original headers and footers: selected
- c. Replace uncertain words with images: not selected
- d. Retain original paper size: selected
- e. Quality: 70%
- f. Resolution: 300dpi
- g. Enhanced compression: selected
- h. Format: automatic
- i. PDF security settings: No security
- j. Document properties: none selected
- k. Header and footer: none specified



10. Text document settings: Option used [other options available]

- a. Keep line breaks: not selected
- b. Insert page break character (#12) as page break: not selected
- c. Use blank line as paragraph separator: not selected
- d. Keep original headers and footers: not selected
- e. Character encoding type: Simple [Unicode UTF-16, Unicode UTF-8]
- f. Code page: Automatic [multiple options available]



APPENDIX 1B: TRIAL 2 - SUMMARY OF OCR OUTPUT FOR ONE SPECIMEN FROM EACH INSTITUTE.

RBGE: E00037202



ACTUAL /15
Det Utrecht 19 Herb. Hort. Bot. Reg. Edin. Flora of Lebanon Acer Araya, Nahr Beyrouth Alt. 1,500ft In 'maquis' thickets, small tree 10-15ft. high; Flowers greenish yellow 4.4.1959 Coll. O. Polunin No. 5204 Royal Botanic Garden Edinburgh E00015007 E00015007
NY ORIGINAL 13.5/15
hIUZS >llì g°sl E o o IN ROYAL BOTANIC GARDEN EDINBURGH E00015007 £ Det. F?Co*-* Je'-j S'UMs^ 3. Utrecht / « -/- 19 9^ ffL.HA OF LEBANON A?,?£ ġienSS 'f if Araya, Nahr Beyrouth Alt. 1,500 ft. In 'maquis* thickets, small tree 10-15 ft., high; flowers greenish yellow 4.4.1959 Coll. O. Polunin No. 5204 HERB. HORT. BOT. REG. EDIN. csv-xäj copyright reserved E00015007
EDINBURGH ORIGINAL 12/15
f~] cJisf- S ~Ý Utrecht /1 -2- 19 9 é T :S@S '.. !N ; « I mmmm !w- / Kips % ' •:' i'SP&i'v' ' ^ *e (N 80 ^ ^Cc^.\JCL_ • - r n '/'• / f- /" •• -' . V" <-.- HERB. HORT. BOT. REG. EDÝN.

'<r\{'/'/'S?L}£"

"|'•|fifth' ' /&<|'•-.

ROYAL BOTANIC GARDEN EDINBURGH

ELQRA OF LEBANON

|||||•||•|||>||>><•<||•|>i- k«M««i |>||"»«inidi iHMimwa

Acer &p t ct CA^k^

Araya, Nahr Beyrouth

Alt. 1,500 ft.

In 'maquis1 thickets, small tree 10-15 ft., high;

flowers greenish yellow

4.4.1959

Coll. 0. Polunin

JUZjt >*È Si O o s ! »S S3 O S II ° -O > °> s d) . > h- a o o

No. 5204

E00015007

PARIS ORIGINAL 1/15

A,AAA"AQV*•A'Av->=(A'~AVLD{AAAVA,_A'A»*i.;:A-VVA—5*_A-
.A1p_ _'~"V.AI'A'AA.._A"#V/AAHY'vAIsrlat'/AA;/ 'A...:—,A_V·VAnv3.—
€gr.AvAVA"\$VAA`~—A".AI}.Vi>iifyx~_1A._5}A·f\$—
,""Aff'hwif>6*_ `(4%QA;'A,(isi`M`IQAAA;UA_ A'AJx—\$V"A.Av.-Ag.;/r.VV,;A-
·.,_A"AA·\A'»,·GAVAVAAA6,AAJ/\$;I;»:_AV'~·AVA'Vff_VA··A,Vr,_,A;»V
g,""E'-:)\A.rYAr._·VV`EyI;ijYAA-j"VJAVA.7};-
IV`2A`**7AVF\$;',M·TA'wp·~'V. 'A"VAV.V~—

[large amount of characters removed]

AAfgwzyAI`O`Llh·AEAA',z.A`A4*%.;·"s.A··`_Q,/AAV`VTij-*2.av.-
¢A*!{€.%é*.:;%i*·?L\$~"Aα·V'AIV'»%»';<AA°1500ftT"AA'°~Y'. ". 'V""*""V""—
'·.V**V·V%;·AA·.xr-
A·A·AA'.m`?'jiα···"W'Hgg·~AV=¢£~»»*A·»' {AA·.., 'AAA·A"·A·¢CA{·i,··..—
"I.A~.+:?'·:A:in·~x*\$,-YA,A·q_V:,,·α;,{")L..·{·A··;·A·`~'·;A*!i'!'-
5A>"Af:T~Arl?}l`_?'jL.é,"

Rom Bommc GAR A II1&CT\l1S thick AAAA A

DE `··' A, et if .A·+~.=V·.V—· sz

EDINBURGH N fi]-Owrrer S} tl'€€ 5 ft fAAA"A; 'VA· A A

'sv nj 1 A. 'V' AA AVk'A·A_v;AA,

A AV I _ 4;U

IIIIIIHHIIIIHHIIIIIIIIIIIIIIIIIIIIIIII! Ai A, AA 88 S °"u°" AA °£A

E0001 5007 A JAH 1 A A

) _

~—» —AAA —»AAA—A. _____

Coll 0 . 0-.. · . Polunzm

V -..- _ ~. _ . 0• 520;,

NYBG: 01295628



ACTUAL /17
<p>New York Botanical Garden 01295628 01295628 Jardin Botanico Nacional "Dr Rafael M Moscoso" Santo Domingo, republica Dominicana 40763 Theaceae Arbusto. 1.5m. de alto; postrado; Haz de la hoja verde mediano con brillo, enves verde claro; Sepalos rojos; fr. Verde. Republica Dominicana: Sierra de Baoruco: Prov. Pedernales-Independencia limite: Cerca del paso en el camino forestal entre Aceitillar (de Pedernales) y Puerto Escondido: Pinar de Pinus occidentalis abierto, con hierbas & arbustos por abajo. 18°12'N, 71°33' Oeste, alt. 1800m 10 abril, 1988 T. Zanoni, J. Pimentel, R. Garcia</p>
PARIS ORIGINAL 13/17
<p>o JARDIN BOTANICO NACIONAL "DR. RAFAEL M. MOSCOSO" SANTO DOMINGO, REPUBLICA DOMINICANA 40763 Theaceae "for't* o- Arbusto. 1.5m. de alto; postrado; haz de la hoja verde mediano con brillo, enves verde claro; sepalos rojos; fr. verde. Republica Domlnlcana: Sierra de Baoruco: Prov. Pedernales-Independencia llmite: cerca del paso en el camino forestal entre Aceitillar (de Pedernales) y Puerto Escondido: pinar de Pinus occidentalis abierto, con hierbas & arbustos por abajo. 18°12'N, 71°33'Oeste, alt. 1800m. 10 abril, 1988 T. Zanoni, J. Pimentel, R. Garcxa</p>
EDINBURGH ORIGINAL 12/17
<p>The New York copyright reserved botanical Garden BOTANICAL \gardeN/ JARDIN BOTANICO NACIONAL "DR. RAFAEL M. MOSCOSO' SANTO DOMýNGO. BEPUBLICA DOMINICANA 40763 Theaceae o- Arbusto. 1.5m. de alto; postrado; haz de la hoja verde mediano con brillo, envés verde claro; sépalos rojos; fr. verde. República Dominicana: Sierra de Baoruco: Prov. Pedernales-Independencia limite:</p>

cerca del paso en el **carnino** forestal entre Aceitillar (de Pedernales) y Puerto Escondido:

pinar de *Pinus occidentalis* abierto, con hierbas & arbustos por abajo.

18° 12'N, 71°33'Oeste,

ait. 1800m.

10 abril, 1988

T. Zanoni, J. Pimentel, R. Garcia

NEW YORK ORIGINAL 13/17

o

JARDIN BOTANICO NACIONAL "DR. RAFAEL M. MOSCOSO"

SANTO DOMINGO, REPUBLICA DOMINICANA

40763

Theaceae

"'for'ft'* o-

Arbusto. 1.5m. de alto; postrado;

haz de la hoja verde mediano con brillo, envés verde claro;

sepalos rojos; fr. verde.

Republica Domlnlcana: Sierra de Baoruco: Prov. Pedernales-Independencia llimite:

cerca del paso en el camino forestal entre Aceitillar (de Pedernales) y Puerto

Escondido:

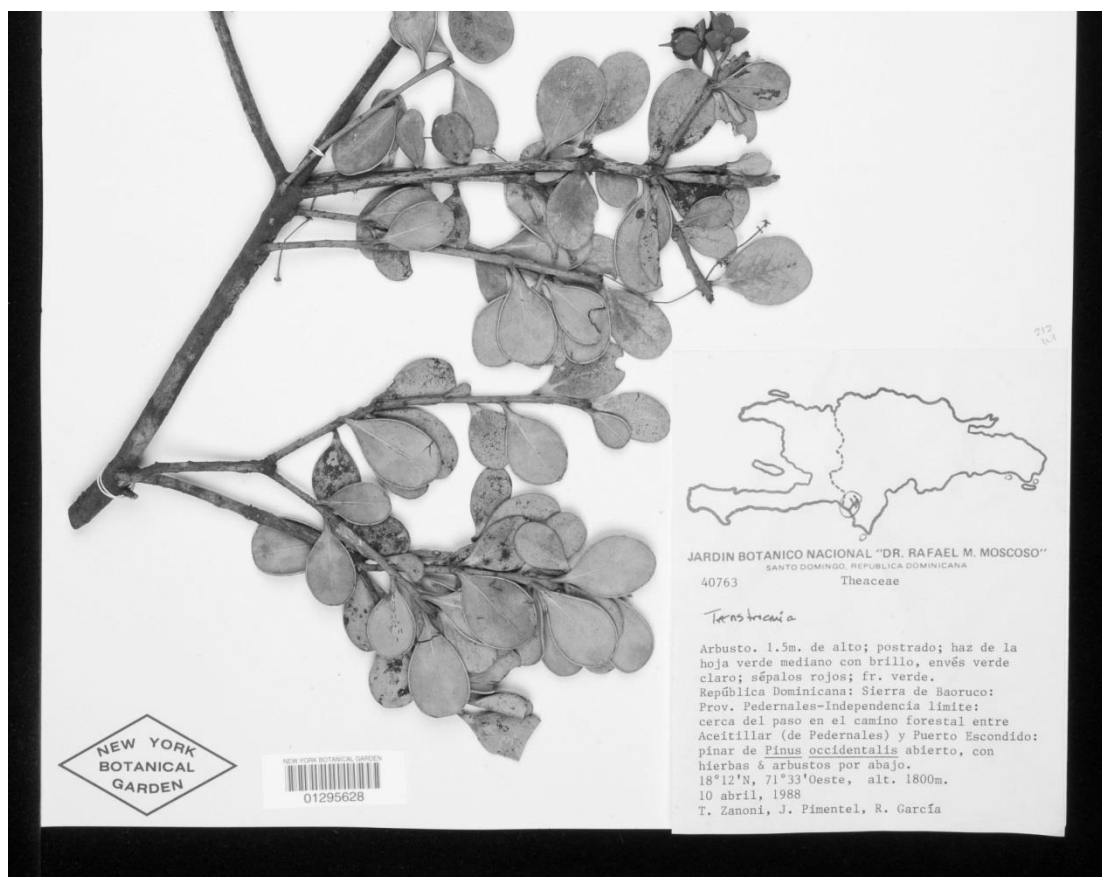
pinar de *Pinus occidentalis* abierto, con hierbas & arbustos por abajo.

18°12'N, 71°33'Oeste,

alt. 1800m.

10 abril, 1988

T. Zanoni, J. Pimentel, R. Garca



EDINBURGH FORMATTED 13/17

BOTANICAL GARDEN

Arbusto. 1.5m. de alto; postrado;
haz de la hoja verde mediano con brillo, envés verde claro;
sépalos rojos; fr. verde.

Republica Dominicana: Sierra de Baoruco: Prov. Pedernales-Independencia límite:
cerca del paso en el camino forestal entre Aceitillar (de Pedernales) y Puerto
Escondido:

pinar de *Pinus occidentalis* abierto, con hierbas & arbustos por abajo.

18°12'N, 71°33'Oeste,

ait. 1800m.

10 abril, 1988

T. Zanoni, J. Pimentel, R. García

JARDIN BOTANICO NACIONAL "DR. RAFAEL M. MOSCOSO"

SANTO DOMINGO. REPUBLICA DOMINICANA

40763

Theaceae

NEW YORK FORMATTED 12.5/17

JARDIN BOTANICO NACIONAL "DR. RAFAEL M. MOSCOSO"

SANTO DOMINGO. REPUBLICA DOMINICANA

40763

Theaceae

O~

Arbusto. 1.5m. de alto; postrado;

haz de la hoja verde mediano con brillo, envés verde claro;

sepalos rojos; fr. verde.

Republica Dominicana: Sierra de Baoruco: Prov. Pedernales-Independencia limlte:

cerca del paso en el camino forestal entre Aceitillar (de Pedernales) y Puerto

Escondido:

pinar de *Pinus occidentalis* abierto, con hierbas & arbustos por abajo.

18°12'N, 71°33'Oeste,

alt. 1800m.

10 abril» 1988

T. Zanoni, J. Pimentel, R. Garcia



ACTUAL TEXT /34
<p>PL00073444 PL00073444</p> <p>Coll c. Sauvageau Guethary 12-J-35 Herbier Museum Paris Cryptogamie PC0559997 Enteromorpha Linza J.AG. var crispata (Bertol) Recueilli par M C Sauvageau A Guethary (Basses-Pyrenees) Du 10 Juillet au 30 Aout 1896</p> <p>Coll c. Sauvageau Guethary 12-J-35 Herbier Museum Paris Cryptogamie PC0559998 Enteromorpha Linza J.AG. var crispata (Bertol) Recueilli par M C Sauvageau A Guethary (Basses-Pyrenees) Du 10 Juillet au 30 Aout 1896</p> <p>Coll c. Sauvageau Guethary 12-J-35 Herbier Museum Paris Cryptogamie PC0559999 Enteromorpha Linza J.AG. var crispata (Bertol) Recueilli par M C Sauvageau A Guethary (Basses-Pyrenees) Du 10 Juillet au 30 Aout 1896</p> <p>Coll c. Sauvageau Guethary 12-J-35 Herbier Museum Paris Cryptogamie PC0560000 Enteromorpha Linza J.AG. var crispata (Bertol) Recueilli par M C Sauvageau A Guethary (Basses-Pyrenees) Du 10 Juillet au 30 Aout 1896</p>
ORIGINAL (P Process) 19.5/34
<p>GÖLL .C .SAU VAGE AU (;1 KTII \UV 12-.)<;;</p> <p>COÎ.LC.SAUVAGEAU tiUETII\«Y tä-l-Jil HKjJ; Herbier Muséum Jff L Paris Cryptogamie J£Ü! PC0559997</p>

ENTE'OMORPH\ IJXZA J.Af». V Ait CRISPA TA (BERTÜL,)

Recueilli par M. C. Sauvageau

à GUÉTHARY j 'Basses-PyrénéesJ

(lu 10 Juillet au 30 Août 1800

H iifa

ENTEROMOMV 1JNZ\ JAfr VAIL CRISPAT A ("SüTOU

Recueilli par M C. Sauvageau

« GUÉTHARY Basses-Pijrénées/

du 10 Juillet au 30 Août 1800

.■IV.“.

Herhier Muséum Paris Cryptogamie

PC0560000

GinrriiAiîY

ll-UJj

E\TE: !OM0n?H\ LTNZA J.Afi. VAU CRiSPATA (BKRTOL,)

Recueilli par M. C. Sauvageau

à GUÉTHARY ('Basses-PyrénéesJ

du 10 Juillet au 30 Août 1800

Herbier Muséum Paris Cryptogamie

PC0559999

COLL.C.SAUVAGEAU

Recueilli par M-. C. Sauvageau

(i GUÉTHARY CBasses-PyrénéesJ

du 10 Juillet au 30 Août 1800

COLL.C.SAP VAGÉAU

ENTE'ïOMORPH \ T.IN7>\ J MS. VAU CRISPAT A (BE8TÖL,)

(ililTII ARY

4HH; Herbier Muséum ■^F'iî Paris Cryptogamie

■!IJÜ

PC0559998

FORMATTED (P Process) /34

ORIGINAL (E Process) 28/34

GOIL.C.SAUVAGEAU

ETII\R\

ü-J-J,;

COL[G.SAUVAGEAÜ

liIJETIIVHY

Herbier Muséum Paris Cryptogamie

PC0559997

Herbier Muséum Paris Cryptogam i e

PC0559998

EN*TE'!OM0RPH\ UNZA J.Afi. VAIL CfîÝSPATA (BERTOL,)

Recueilli par M C. Sauvageau

à GUÉTHARY (Basses-Pyrénées)

(lu 10 Juillet au 30 Août 1896

sami

ENTEROMORPH \ 1,TN2\ JAG, VAR CRISP VTA (BERTOÜ

Recueilli par M C. Sauvageau
à GUÉTHARY (^Basses-Pyrénées
/ du 10 .Juillet au 30 Août 189(5
COLL.C.SAUVAGÉAU
(ililililARY
L'-.l-iJo
<

Herbier Muséum Paris Cryptogamie
PC0559999

E\TE;!OM0nPH\ I.LİNZA J.Afi. VAU CRISPA TA (BERTOL,)

Recueilli par M. C. Sauvageau
à GUETHARY ('Basses-Pyrénées)
du 10 Juillet au 30 Août 1896

5A

COLL.C.SAUVAGEAU

GlÉTHAI Y

Herbier Muséum ParisCryptogamie
PC0560000

ENTE'OMORPH \ 1 MI. VAR CRISPAT A (BERTOL,)

Recueilli par M-. C. Sauvageau
à GUÉTHARY ('Basses-Pyrénées)
du 10 Juillet au 30 Août 1896

PL00073444

PC0559999

PC0559997

PC0559998

PC0560000

FORMATTED (E Process) 16.5/34

iýinýý \ i! v

il K l II\I5V

!jOu.C.SAUVAGtAU

boli.g.sauvagea;;

I-»- -Jlw

2-J-JS

Herbier Muséum Pans Ciyptogamle

PC0559997

r

EN7E'IOWRPH\ I.IN7.A JAG. VAH CRÝSPATA (BERTUL,)

Recueilli par M C. Sauvageau

rf ('•VETU A II Y i '!\{< i s sirs- Pyi'én éeyl

«lu 10 Juillet au 30 Août 18%

Herbier Muséum Paris Ciyploaamii

PC0559998

m

;

ENTEROIORPH \ MNU J '.fr VAH CRISPVTA (BERTO»,.)

Recueilli par M C. Sauvageau
ci GfÊTIARY ^Basses-Pyrénées>
du 10 Juillet au 30 Août 1896

cdil.c.sauvagéau

V

\ V

s7§

E\TE:îrtUO?IPa \ I.TXZA J.A6. VAH CRISPA l'A (BEHTOL,)

Recueilli par M C. Sauvageau

" CI'ÉTIAUY l'Basses-Pyrénées'

du 10 Juillet au 30 Août 1890

GUtëillART L'-l-Ilj

cdil.c.sauvageau

y

Herbier Muséum Paris Cryptogair.:.

PC0559999

\V.

EN'TE'IOMORPH \ MN7A UG VA» CRISPAT A (BEBTOL.)

Recueilli par M C. Sauvageau

,i GVÉTHARY l'Basses-Pyrénées'

«du 10 Juillet au 30 Août 1 896

GUBTIARY 12-1-Ji

Hfrbier Muséum Paris Cryptqgamie

«S* PC0560000

91. foi"

PC0559998

ORIGINAL (NY Process) 18/34

GÖLL .C .SAU VAGE AU

(j1 KTII \UV 12-.)«,,;

COÎ.LC.SAUVAGEAU

tiUETII\«Y

tä-l-Jil

HKjJ;

Herbier Muséum Jff L Paris Cryptogamie

J£Ü!

PC0559997

ENTE'IOMORPH\ IJXZA J.Af». VAt CRISPA TA (BERTÜL,)

Recueilli par M. C. Sauvageau

à GUÉTHARY j 'Basses-PyrénéesJ

(lu 10 Juillet au 30 Août 1800H iïfa

ENTEROMOMV 1JNZ\ JAfr VAl CRISPAT A ("SüTOU

Recueilli par M C. Sauvageau

« GUÉTHARY Basses-Pijrénées/

du 10 Juillet au 30 Août 1800

.■IV. “. Herhier Muséum Paris Cryptogamie

PC0560000

GinrriiAiîY

Il-UJj

E\TE: !0M0n?H\ LTNZA J.Afi. VAU CRiSPATA (BKRTOL,)

Recueilli par M. C. Sauvageau

à GUÉTHARY ('Basses-PyrénéesJ

du 10 Juillet au 30 Août 1800

Herbier Muséum Paris Cryptogamie

PC0559999

COLL.C.SAUVAGEAU

Recueilli par M-. C. Sauvageau

(i GUÉTHARY CBasses-PyrénéesJ

du 10 Juillet au 30 Août 1800

COLL.C.SAP VAGÉAU

ENTE'ïOMORPH \ T.IN7>\ J MS. VAU CRISPAT A (BE8TÖL,)

(ililiTII ARY

4HH;

Herbier Muséum ■^F'iî Paris Cryptogamie

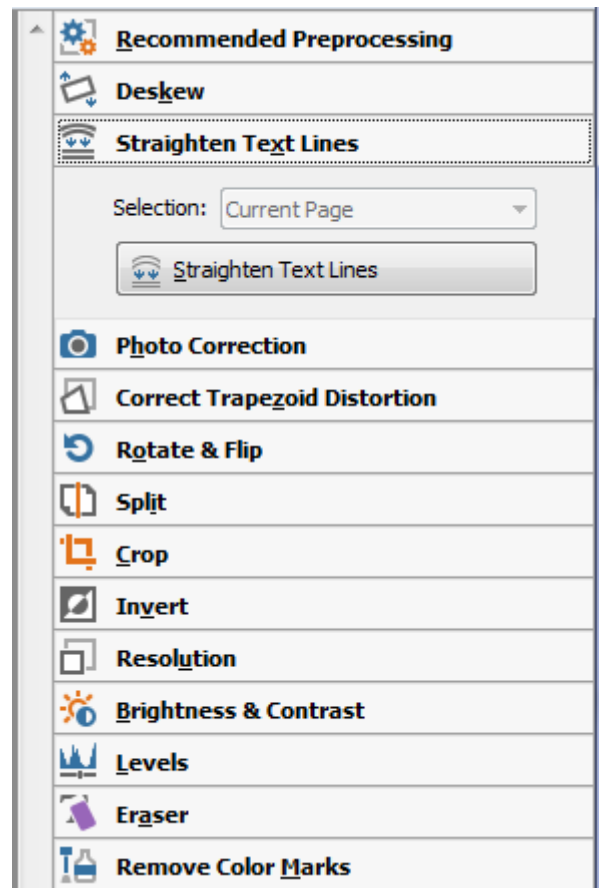
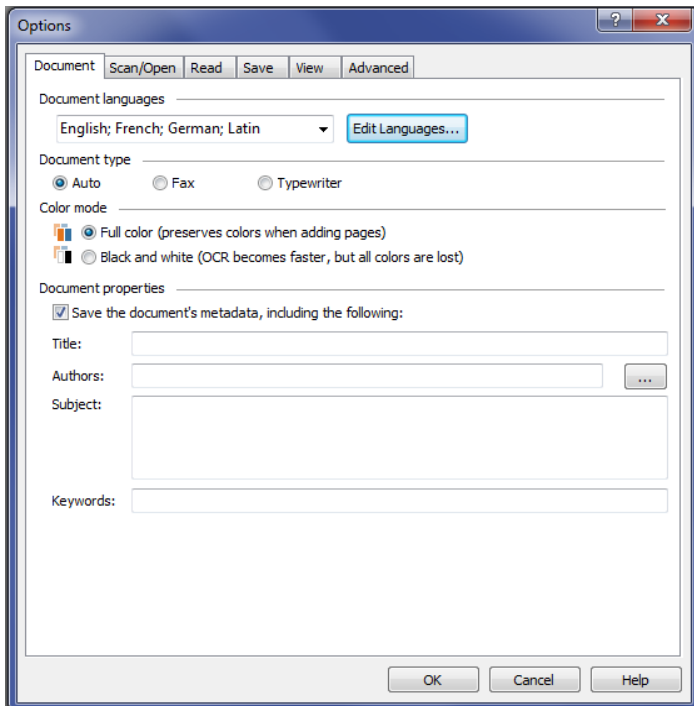
■!IJÛ

PC0559998

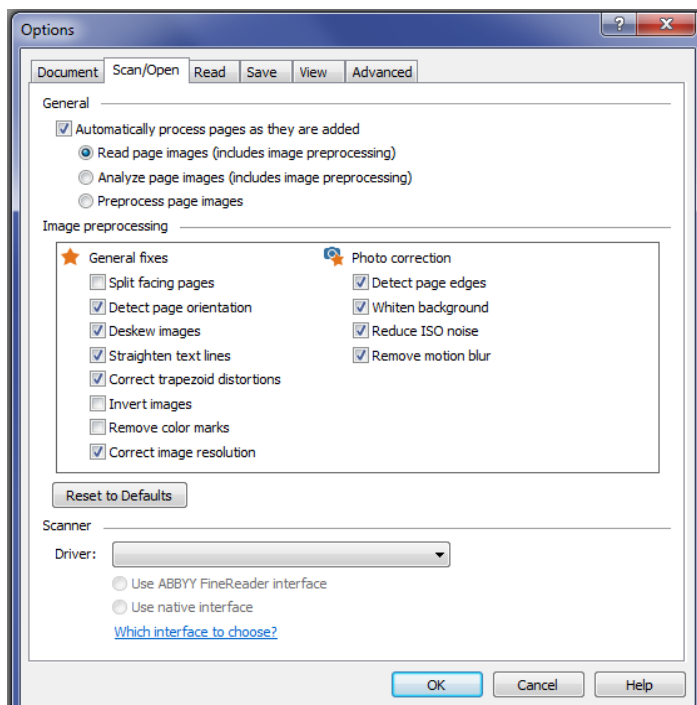
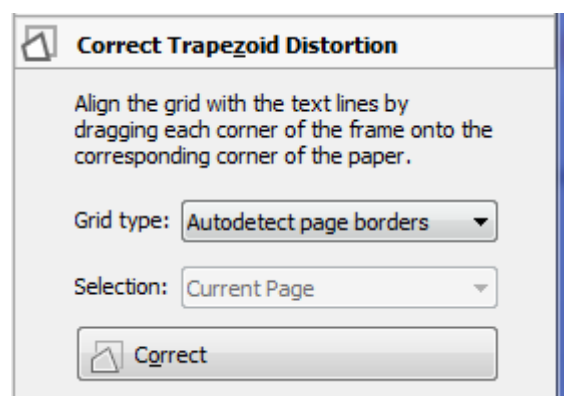
FORMATTED (NY Process) /34

APPENDIX 1C: SETTINGS FOR ABBYY FINEREADER V12 PROFESSIONAL AT RBGK

The following screenshots show the settings used.



Suggested further settings for MfN and MRAC images:



APPENDIX 1D: FILE PREPARATION AT RBGK

Several different file types and file processing techniques were investigated in order to achieve the best results within Finereader. The following adjustments were tried:

Original

Tif images – no photoshop adjustments, no preprocessing or adjustments in Finereader

Formatted

1. Tif images – no photoshop adjustments, no preprocessing, English and Latin determined as language in Finereader
2. Tif images – no photoshop adjustments, no preprocessing, English, German and Latin determined as language in Finereader
3. Tif images – no photoshop adjustments, no preprocessing, English, German and Latin determined as language, ISO adjustment in Finereader
4. Tif images – no photoshop adjustments, white background adjustment, English, German and Latin determined as language in Finereader
5. Tif images – no photoshop adjustments, English, German and Latin determined as language and Automatic selection template in Finereader
6. Tif images – no photoshop adjustments, Automatic preprocessing, English, German and Latin determined as language in Finereader
7. Tif images – no photoshop adjustments, Black and White setting chosen, English, German and Latin determined as language in Finereader
8. Tif images - desaturated in Photoshop, English, German and Latin determined as language in Finereader
9. Tif images - greyscaled in photoshop, English, German and Latin determined as language in Finereader
10. Jpeg full res images - English, German and Latin determined as language in Finereader
11. Jpeg images – Res reduced in Photoshop (1566 pixels wide and 150 dpi.), English, German and Latin determined as language in Finereader
12. Jpeg images – greyscaled, colourchart cropped out and Res reduced in Photoshop (1566 pixels wide and 150 dpi.), English, German and Latin determined as language in Finereader
13. Jpeg images – greyscaled, colourchart cropped out and Res reduced in Photoshop (1566 pixels wide and 150 dpi.), English, German and Latin determined as language and Straighten Text lines function used in Finereader
14. Jpeg images – greyscaled, colourchart cropped out and Res reduced in Photoshop (1566 pixels wide and 150 dpi.), English, German and Latin determined as language and Brightness and Contrast adjusted to 5 and 50 respectively in Finereader
15. Jpeg images – greyscaled, colourchart cropped out and Res reduced in Photoshop (1566 pixels wide and 150 dpi.), English, German and Latin determined as language, Straighten Text lines function used and Brightness and Contrast adjusted to 5 and 50 respectively in Finereader
16. Jpeg images – greyscaled, colourchart cropped out and Res reduced in Photoshop (1566 pixels wide and 150 dpi.), English, German and Latin determined as language, Black and White setting chosen and Straighten Text lines function used in Finereader

APPENDIX 1E: SCORES FOR EACH SPECIMEN FROM EACH INSTITUTE BY WORD

MFN SPECIMENS

Scores by word for each specimen:

Image: A001 – 20150304_132859

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBYY Recognition Server v3	26	25	17	65.4
ABBYY FineReader v12 Suggested protocol	26	Not read		n/a
Onlineocr.net	26	31	21	80.8
Newocr.com	26	18	6	23.1
Ocrgeek.com	26	10	5	19.2
Ocrconvert.com	26	28	7	26.9

Image: A001 – 20150304_132859

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBYY Recognition Server v3	12	4	0	0
ABBYY FineReader v12 Suggested protocol	12	4	0	0
Onlineocr.net	12	30	5	41.7
Newocr.com	12	16	0	0
Ocrgeek.com	12	6	0	0
Ocrconvert.com	12	4	0	0

Image: A001 – 20150420_095955

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBYY Recognition Server v3	26	4	1	3.8
ABBYY FineReader v12 Suggested protocol	26	28	16	61.5
Onlineocr.net	26	56	16	61.5
Newocr.com	26	15	6	23.1
Ocrgeek.com	26	6	0	0
Ocrconvert.com	26	18	0	0

Image: A001 – 20150512_153102

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBYY Recognition Server v3	25	30	8	32
ABBYY FineReader v12 Suggested protocol	25	7	0	0
Onlineocr.net	25	32	0	0
Newocr.com	25	15	0	0

Ocrgeek.com	25	10	0	0
Ocrconvert.com	25	20	1	4

Image: A001 – 20150529_141828

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBYY Recognition Server v3	21	1	1	4.8
ABBYY FineReader v12 Suggested protocol	21	5	0	0
Onlineocr.net	21	38	9	42.9
Newocr.com	21	17	2	9.5
Ocrgeek.com	21	6	0	0
Ocrconvert.com	21	14	0	0

Image: A001 – 20150603_121251

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBYY Recognition Server v3	13	13	1	7.7
ABBYY FineReader v12 Suggested protocol	13	Not read		n/a
Onlineocr.net	13	33	3	23.1
Newocr.com	13	10	2	15.4
Ocrgeek.com	13	11	1	7.7
Ocrconvert.com	13	11	0	0

Image: A001 – 20150615_101235

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBYY Recognition Server v3	15	10	1	6.7
ABBYY FineReader v12 Suggested protocol	15	Not read		n/a
Onlineocr.net	15	37	0	0
Newocr.com	15	26	0	0
Ocrgeek.com	15	4	0	0
Ocrconvert.com	15	0	0	0

Image: A001 – 20150617_115733

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBYY Recognition Server v3	18	6	1	5.6
ABBYY FineReader v12 Suggested protocol	18	6	0	0
Onlineocr.net	18	No text to extract	0	0
Newocr.com	18	10	0	0
Ocrgeek.com	18	5	0	0
Ocrconvert.com	18	16	0	0

Image: A001 – 20150702_142034

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBY Recognition Server v3	19	6	0	0
ABBY FineReader v12 Suggested protocol	19	Not read		n/a
Onlineocr.net	19	41	0	0
Newocr.com	19	15	1	5.3
Ocrgeek.com	19	9	0	0
Ocrconvert.com	19	20	0	0

Image: A001 – 20150708_144237

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBY Recognition Server v3	21	0	0	0
ABBY FineReader v12 Suggested protocol	21	7	0	0
Onlineocr.net	21	No recognized text		n/a
Newocr.com	21	15	1	4.8
Ocrgeek.com	21	6	0	0
Ocrconvert.com	21	1	0	0

Image: A002 – 20150415_101503

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBY Recognition Server v3	24	6	1	4.2
ABBY FineReader v12 Suggested protocol	24	Not read		n/a
Onlineocr.net	24	28	0	0
Newocr.com	24	12	3	12.5
Ocrgeek.com	24	15	2	8.3
Ocrconvert.com	24	6	0	0

MNHN SPECIMENS

Scores by word for each specimen:

Image: P01523160

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBY Recognition Server v3	32	Tiff: 52	Tiff: 24	Tiff: 75
ABBY FineReader v12	32	JPEG: 48	JPEG: 24	JPEG: 75
Onlineocr.net	32	Tiff: 21 JPEG: 21	Tiff: 7 JPEG: 13	Tiff: 21.9 JPEG: 40.6
Newocr.com	32	JPEG: 72	JPEG: 5	JPEG: 15.6
Ocrgeek.com	32	Tiff: 51 JPEG: 30	Tiff: 10 JPEG: 3	Tiff: 31.3 JPEG: 9.4
Ocrconvert.com	32	Tiff: 63 JPEG: 45	Tiff: 0 JPEG: 0	Tiff: 0 JPEG: 0

Image: P01583356

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBYY Recognition Server v3	132	Tiff: 133	Tiff: 114	Tiff: 86.4
ABBYY FineReader v12	132	JPEG: 128	JPEG: 111	JPEG: 84.1
Onlineocr.net	132	Tiff: 115 JPEG: 114	Tiff: 43 JPEG: 43	Tiff: 32.6 JPEG: 32.6
Newocr.com	132	JPEG: 129	JPEG: 94	JPEG: 71.2
Ocrgeek.com	132	Tiff: 102 JPEG: 127	Tiff: 4 JPEG: 13	Tiff: 3 JPEG: 9.8
Ocrconvert.com	132	Tiff: 7 JPEG: 8	Tiff: 0 JPEG: 1	Tiff: 0 JPEG: 0.8

Image: P01583601

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBYY Recognition Server v3	25	Tiff: 48	Tiff: 10	Tiff: 40
ABBYY FineReader v12	25	JPEG: 39	JPEG: 11	JPEG: 44
Onlineocr.net	25	Tiff: 13 JPEG: 13	Tiff: 6 JPEG: 5	Tiff: 24 JPEG: 20
Newocr.com	25	JPEG: 72	JPEG: 5	JPEG: 20
Ocrgeek.com	25	Tiff: 23 JPEG: 12	Tiff: 2 JPEG: 5	Tiff: 8 JPEG: 20
Ocrconvert.com	25	Tiff: 22 JPEG: 21	Tiff: 3 JPEG: 2	Tiff: 12 JPEG: 8

Image: P01596658

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBYY Recognition Server v3	35	Tiff: 37	Tiff: 23	Tiff: 65.7
ABBYY FineReader v12	35	JPEG: 43	JPEG: 26	JPEG: 75.3
Onlineocr.net	35	Tiff: 1 JPEG: 14	Tiff: 0 JPEG: 2	Tiff: 0 JPEG: 5.7
Newocr.com	35	JPEG: 104	JPEG: 0	JPEG: 0
Ocrgeek.com	35	Tiff: 12 JPEG: 0	Tiff: 0 JPEG: 0	Tiff: 0 JPEG: 0
Ocrconvert.com	35	Tiff: 9 JPEG: 4	Tiff: 0 JPEG: 0	Tiff: 0 JPEG: 0

Image: PC0559998

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBYY Recognition Server v3	130	Tiff: 123	Tiff: 85	Tiff: 65.4
ABBYY FineReader v12	130	JPEG: 114	JPEG: 61	JPEG: 46.9
Onlineocr.net	130	Tiff: 54 JPEG: 107	Tiff: 10 JPEG: 24	Tiff: 7.7 JPEG: 18.5
Newocr.com	130	JPEG: 336	JPEG: 0	JPEG: 0
Ocrgeek.com	130	Tiff: 116 JPEG: 92	Tiff: 5 JPEG: 8	Tiff: 3.8 JPEG: 6.2
Ocrconvert.com	130	Tiff: 194 JPEG: 201	Tiff: 0 JPEG: 0	Tiff: 0 JPEG: 0

Image: EL10000.tiff

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBY Recognition Server v3	4	6	2	50
ABBY FineReader v12	4	n/a	n/a	n/a
Onlineocr.net	4	n/a	n/a	n/a
Newocr.com	4	n/a	n/a	n/a
Ocrgeek.com	4	n/a	n/a	n/a
Ocrconvert.com	4	n/a	n/a	n/a

Image: EL10001.tiff

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBY Recognition Server v3	5	2	1	20
ABBY FineReader v12	5	n/a	n/a	n/a
Onlineocr.net	5	n/a	n/a	n/a
Newocr.com	5	n/a	n/a	n/a
Ocrgeek.com	5	n/a	n/a	n/a
Ocrconvert.com	5	n/a	n/a	n/a

Image: EL10004.tiff

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBY Recognition Server v3	4	5	1	25
ABBY FineReader v12	4	n/a	n/a	n/a
Onlineocr.net	4	n/a	n/a	n/a
Newocr.com	4	n/a	n/a	n/a
Ocrgeek.com	4	n/a	n/a	n/a
Ocrconvert.com	4	n/a	n/a	n/a

Image: EL10005.tiff

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBY Recognition Server v3	4	5	1	25
ABBY FineReader v12	4	n/a	n/a	n/a
Onlineocr.net	4	n/a	n/a	n/a
Newocr.com	4	n/a	n/a	n/a
Ocrgeek.com	4	n/a	n/a	n/a
Ocrconvert.com	4	n/a	n/a	n/a

Image: EL10032.tiff

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBY Recognition Server v3	1	6	1	100
ABBY FineReader v12	1	n/a	n/a	n/a
Onlineocr.net	1	n/a	n/a	n/a
Newocr.com	1	n/a	n/a	n/a
Ocrgeek.com	1	n/a	n/a	n/a
Ocrconvert.com	1	n/a	n/a	n/a

Image: EL10033.tiff

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBY Recognition Server v3	1	3	1	100
ABBY FineReader v12	1	n/a	n/a	n/a
Onlineocr.net	1	n/a	n/a	n/a
Newocr.com	1	n/a	n/a	n/a
Ocrgeek.com	1	n/a	n/a	n/a
Ocrconvert.com	1	n/a	n/a	n/a

Image: EL10034.tiff

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBY Recognition Server v3	1	11	1	100
ABBY FineReader v12	1	n/a	n/a	n/a
Onlineocr.net	1	n/a	n/a	n/a
Newocr.com	1	n/a	n/a	n/a
Ocrgeek.com	1	n/a	n/a	n/a
Ocrconvert.com	1	n/a	n/a	n/a

Image: EL10035.tiff

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBY Recognition Server v3	1	11	1	100
ABBY FineReader v12	1	n/a	n/a	n/a
Onlineocr.net	1	n/a	n/a	n/a
Newocr.com	1	n/a	n/a	n/a
Ocrgeek.com	1	n/a	n/a	n/a
Ocrconvert.com	1	n/a	n/a	n/a

Image: EL10036.tiff

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBY Recognition Server v3	1	11	1	100
ABBY FineReader v12	1	n/a	n/a	n/a
Onlineocr.net	1	n/a	n/a	n/a
Newocr.com	1	n/a	n/a	n/a
Ocrgeek.com	1	n/a	n/a	n/a
Ocrconvert.com	1	n/a	n/a	n/a

MRAC SPECIMENS

Scores by word for each specimen:

Image: anomala_Aspidifrontia_AT_RMCA.jpg

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBY Recognition Server v3	11	13	7	63.6
ABBY FineReader v12	11	26	8	72.7
Onlineocr.net	11	n/a	n/a	n/a
Newocr.com	11	n/a	n/a	n/a
Ocrgeek.com	11	n/a	n/a	n/a
Ocrconvert.com	11	n/a	n/a	n/a

Image: aurantiipennis_Apaegocera_A_RMCA_02.JPG

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBY Recognition Server v3	13	20	12	92.3
ABBY FineReader v12	13	16	12	92.3
Onlineocr.net	13	n/a	n/a	n/a
Newocr.com	13	n/a	n/a	n/a
Ocrgeek.com	13	n/a	n/a	n/a
Ocrconvert.com	13	n/a	n/a	n/a

Image: basilewskyi_Mentaxya_HT_RMCA.jpg

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBY Recognition Server v3	9	24	7	77.8
ABBY FineReader v12	9	31	6	66.7
Onlineocr.net	9	n/a	n/a	n/a
Newocr.com	9	n/a	n/a	n/a
Ocrgeek.com	9	n/a	n/a	n/a
Ocrconvert.com	9	n/a	n/a	n/a

Image: caloxantha_Anaphosia_HT_RMCA_01.jpg

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBY Recognition Server v3	10	13	5	50
ABBY FineReader v12	10	16	6	60
Onlineocr.net	10	n/a	n/a	n/a
Newocr.com	10	n/a	n/a	n/a
Ocrgeek.com	10	n/a	n/a	n/a
Ocrconvert.com	10	n/a	n/a	n/a

Image: distalis_Micraxyliia_HT_RMCA_02.jpg

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBY Recognition Server v3	10	18	7	70
ABBY FineReader v12	10	13	8	80
Onlineocr.net	10	n/a	n/a	n/a
Newocr.com	10	n/a	n/a	n/a

Ocrgeek.com	10	n/a	n/a	n/a
Ocrconvert.com	10	n/a	n/a	n/a

Image: fasciata_Hypocoela_PT_RMCA_01.jpg

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBY Recognition Server v3	10	10	7	70
ABBY FineReader v12	10	18	8	80
Onlineocr.net	10	n/a	n/a	n/a
Newocr.com	10	n/a	n/a	n/a
Ocrgeek.com	10	n/a	n/a	n/a
Ocrconvert.com	10	n/a	n/a	n/a

Image: gabunica_Nudaurelia_PT_RMCA_02.jpg

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBY Recognition Server v3	13	2	2	15.4
ABBY FineReader v12	13	17	5	38.5
Onlineocr.net	13	n/a	n/a	n/a
Newocr.com	13	n/a	n/a	n/a
Ocrgeek.com	13	n/a	n/a	n/a
Ocrconvert.com	13	n/a	n/a	n/a

Image: IMG_1746.tif

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBY Recognition Server v3	55	78	47	85.5
ABBY FineReader v12	55	53	35	63.6
Onlineocr.net	55	n/a	n/a	n/a
Newocr.com	55	n/a	n/a	n/a
Ocrgeek.com	55	n/a	n/a	n/a
Ocrconvert.com	55	n/a	n/a	n/a

Image: IMG_1750.tif

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBY Recognition Server v3	68	55	49	72.1
ABBY FineReader v12	68	58	40	58.8
Onlineocr.net	68	n/a	n/a	n/a
Newocr.com	68	n/a	n/a	n/a
Ocrgeek.com	68	n/a	n/a	n/a
Ocrconvert.com	68	n/a	n/a	n/a

Image: IMG_1909.tif

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBY Recognition Server v3	25	32	12	48
ABBY FineReader v12	25	15	8	32

Onlineocr.net	25	n/a	n/a	n/a
Newocr.com	25	n/a	n/a	n/a
Ocrgeek.com	25	n/a	n/a	n/a
Ocrconvert.com	25	n/a	n/a	n/a

Image: IMG_1913.tif

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBYY Recognition Server v3	27	38	10	37
ABBYY FineReader v12	27	42	11	40.7
Onlineocr.net	27	n/a	n/a	n/a
Newocr.com	27	n/a	n/a	n/a
Ocrgeek.com	27	n/a	n/a	n/a
Ocrconvert.com	27	n/a	n/a	n/a

Image: vocata_Epaena_HT_RMCA_01.jpg

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBYY Recognition Server v3	14	22	10	71.4
ABBYY FineReader v12	14	14	2	14.3
Onlineocr.net	14	n/a	n/a	n/a
Newocr.com	14	n/a	n/a	n/a
Ocrgeek.com	14	n/a	n/a	n/a
Ocrconvert.com	14	n/a	n/a	n/a

Image: anomala_Aspidifrontia_AT_RMCA.jpg

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBYY Recognition Server v3	11	13	7	63.6
ABBYY FineReader v12	11	26	8	72.7
Onlineocr.net	11	n/a	n/a	n/a
Newocr.com	11	n/a	n/a	n/a
Ocrgeek.com	11	n/a	n/a	n/a
Ocrconvert.com	11	n/a	n/a	n/a

Image: anomala_Aspidifrontia_AT_RMCA.jpg

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBYY Recognition Server v3	11	13	7	63.6

ABBYY FineReader v12	11	26	8	72.7
Onlineocr.net	11	n/a	n/a	n/a
Newocr.com	11	n/a	n/a	n/a
Ocrgeek.com	11	n/a	n/a	n/a
Ocrconvert.com	11	n/a	n/a	n/a

Image: anomala_Aspidifrontia_AT_RMCA.jpg

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBYY Recognition Server v3	11	13	7	63.6
ABBYY FineReader v12	11	26	8	72.7
Onlineocr.net	11	n/a	n/a	n/a
Newocr.com	11	n/a	n/a	n/a
Ocrgeek.com	11	n/a	n/a	n/a
Ocrconvert.com	11	n/a	n/a	n/a

NMP SPECIMENS

Scores by word for each specimen:

Image: F372

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBYY Recognition Server v3	26	4	1	3.8
ABBYY FineReader v12	26	25	17	65.3
Onlineocr.net	26	23	0	0
Newocr.com	26	22	12	46.2
Ocrgeek.com	26	19	9	34.6
Ocrconvert.com	26	154	10	38.5

Image: F382

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBYY Recognition Server v3	24	9	2	8.3
ABBYY FineReader v12	24	20	14	58.3
Onlineocr.net	24	2	0	0
Newocr.com	24	12	10	41.7
Ocrgeek.com	24	6	2	8.3
Ocrconvert.com	24	91	8	33.3

Image: F384

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)

ABBY Recognition Server v3	46	54	18	39.1
ABBY FineReader v12	46	37	12	26.1
Onlineocr.net	46	37	19	41.3
Newocr.com	46	41	20	43.5
Ocrgeek.com	46	9	1	2.2
Ocrconvert.com	46	45	9	19.6

Image: L12619

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBY Recognition Server v3	61	53	31	50.8
ABBY FineReader v12	61	24	21	34.4
Onlineocr.net	61	50	34	55.7
Newocr.com	61	57	23	37.7
Ocrgeek.com	61	55	18	29.5
Ocrconvert.com	61	51	20	32.8

Image: L12635

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBY Recognition Server v3	98	60	46	46.9
ABBY FineReader v12	98	54	42	42.9
Onlineocr.net	98	42	37	37.8
Newocr.com	98	57	9	9.2
Ocrgeek.com	98	99	23	23.5
Ocrconvert.com	98	85	26	26.5

Image: L12648

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBY Recognition Server v3	51	86	41	80.4
ABBY FineReader v12	51	62	43	84.3
Onlineocr.net	51	79	44	86.3
Newocr.com	51	90	10	19.6
Ocrgeek.com	51	81	28	54.9
Ocrconvert.com	51	104	35	68.6

Image: F373

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBY Recognition Server v3	28	4	2	7.1
ABBY FineReader v12	28	41	13	46.4
Onlineocr.net	28	n/a	n/a	n/a
Newocr.com	28	n/a	n/a	n/a
Ocrgeek.com	28	n/a	n/a	n/a
Ocrconvert.com	28	n/a	n/a	n/a

Image: F384

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBY Recognition Server v3	31	54	18	58
ABBY FineReader v12	31	37	11	35.5
Onlineocr.net	31	n/a	n/a	n/a
Newocr.com	31	n/a	n/a	n/a
Ocrgeek.com	31	n/a	n/a	n/a
Ocrconvert.com	31	n/a	n/a	n/a

IDigBio SPECIMENS

Scores by word for each specimen:

Image: EMEC609636_Cerceris_compar_compar.jpg

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBY Recognition Server v3	25	45	23	92
ABBY FineReader v12	25	36	18	72
Onlineocr.net	25	n/a	n/a	n/a
Newocr.com	25	n/a	n/a	n/a
Ocrgeek.com	25	n/a	n/a	n/a
Ocrconvert.com	25	n/a	n/a	n/a

Image: EMEC609740_Cerceris_convergens.jpg

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBY Recognition Server v3	17	51	9	52.9
ABBY FineReader v12	17	27	7	41.2
Onlineocr.net	17	n/a	n/a	n/a
Newocr.com	17	n/a	n/a	n/a
Ocrgeek.com	17	n/a	n/a	n/a
Ocrconvert.com	17	n/a	n/a	n/a

Image: EMEC609742_Cerceris_convergens.jpg

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBY Recognition Server v3	13	25	12	92.3
ABBY FineReader v12	13	27	12	92.3
Onlineocr.net	15	n/a	n/a	n/a
Newocr.com	13	n/a	n/a	n/a
Ocrgeek.com	13	n/a	n/a	n/a
Ocrconvert.com	13	n/a	n/a	n/a

Image: EMEC609853_Cerceris_convergens.jpg

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBYY Recognition Server v3	19	15	14	93.3
ABBYY FineReader v12	19	36	11	57.9
Onlineocr.net	19	n/a	n/a	n/a
Newocr.com	19	n/a	n/a	n/a
Ocrgeek.com	19	n/a	n/a	n/a
Ocrconvert.com	19	n/a	n/a	n/a

Image: EMEC609879_Cerceris_convergens.jpg

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBYY Recognition Server v3	15	19	16	
ABBYY FineReader v12	15	59	14	
Onlineocr.net	15	n/a	n/a	n/a
Newocr.com	15	n/a	n/a	n/a
Ocrgeek.com	15	n/a	n/a	n/a
Ocrconvert.com	15	n/a	n/a	n/a

Image: EMEC609885_Cerceris_convergens.jpg

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBYY Recognition Server v3	12	201	10	83.3
ABBYY FineReader v12	12	20	10	83.3
Onlineocr.net	12	n/a	n/a	n/a
Newocr.com	12	n/a	n/a	n/a
Ocrgeek.com	12	n/a	n/a	n/a
Ocrconvert.com	12	n/a	n/a	n/a

Image: EMEC609952_Stigmus_sp.jpg

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBYY Recognition Server v3	28	55	24	85.7
ABBYY FineReader v12	28	53	22	78.6
Onlineocr.net	28	n/a	n/a	n/a
Newocr.com	28	n/a	n/a	n/a
Ocrgeek.com	28	n/a	n/a	n/a
Ocrconvert.com	28	n/a	n/a	n/a

Image: NY01075765_lg.jpg

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBYY Recognition Server v3	32	34	30	93.75
ABBYY FineReader v12	32	32	32	100%
Onlineocr.net	32	n/a	n/a	n/a
Newocr.com	32	n/a	n/a	n/a
Ocrgeek.com	32	n/a	n/a	n/a

Ocrconvert.com	32	n/a	n/a	n/a
----------------	----	-----	-----	-----

Image: NY01075766_lg.jpg

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBYY Recognition Server v3	53	58	45	84.9
ABBYY FineReader v12	53	57	46	86.8
Onlineocr.net	53	n/a	n/a	n/a
Newocr.com	53	n/a	n/a	n/a
Ocrgeek.com	53	n/a	n/a	n/a
Ocrconvert.com	53	n/a	n/a	n/a

Image: TENN-L-0000007_lg.jpg

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBYY Recognition Server v3	32	33	31	96.9
ABBYY FineReader v12	32	31	29	90.6
Onlineocr.net	32	n/a	n/a	n/a
Newocr.com	32	n/a	n/a	n/a
Ocrgeek.com	32	n/a	n/a	n/a
Ocrconvert.com	32	n/a	n/a	n/a

Image: TENN-L-0000009_lg.jpg

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBYY Recognition Server v3	51	51	48	94.1
ABBYY FineReader v12	51	51	50	98
Onlineocr.net	51	n/a	n/a	n/a
Newocr.com	51	n/a	n/a	n/a
Ocrgeek.com	51	n/a	n/a	n/a
Ocrconvert.com	51	n/a	n/a	n/a

Image: TENN-L-0000073_lg.jpg

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBYY Recognition Server v3	46	43	33	71.7
ABBYY FineReader v12	46	43	30	65.2
Onlineocr.net	46	n/a	n/a	n/a
Newocr.com	46	n/a	n/a	n/a
Ocrgeek.com	46	n/a	n/a	n/a
Ocrconvert.com		n/a	n/a	n/a

Image: TENN-L-0000077_lg.jpg

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
---------	-------------------	-------------------------	---------------	-----------------------------------

ABBY Recognition Server v3	36	50	13	36.1
ABBY FineReader v12	36	41	19	52.8
Onlineocr.net	36	n/a	n/a	n/a
Newocr.com	36	n/a	n/a	n/a
Ocrgeek.com	36	n/a	n/a	n/a
Ocrconvert.com	36	n/a	n/a	n/a

Image: TENN-L-0000080_lg.jpg

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBY Recognition Server v3	37	44	35	94.5
ABBY FineReader v12	37	68	32	86.5
Onlineocr.net	37	n/a	n/a	n/a
Newocr.com	37	n/a	n/a	n/a
Ocrgeek.com	37	n/a	n/a	n/a
Ocrconvert.com	37	n/a	n/a	n/a

Image: WIS-L-0012029_lg.jpg

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBY Recognition Server v3	45	54	43	95.5
ABBY FineReader v12	45	58	40	88.88
Onlineocr.net	45	n/a	n/a	n/a
Newocr.com	45	n/a	n/a	n/a
Ocrgeek.com	45	n/a	n/a	n/a
Ocrconvert.com	45	n/a	n/a	n/a

RBGE SPECIMENS

Scores by word for each specimen:

Image: E00000219.tiff

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBY Recognition Server v3	58	147	50	86.2
ABBY FineReader v12	58	98	52	89.7
Onlineocr.net	58	n/a	n/a	n/a
Newocr.com	58	n/a	n/a	n/a
Ocrgeek.com	58	n/a	n/a	n/a
Ocrconvert.com	58	n/a	n/a	n/a

Image: E00000522.tiff

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)

ABBY Recognition Server v3	52	83	46	88.5
ABBY FineReader v12	52	74	48	92.3
Onlineocr.net	52	n/a	n/a	n/a
Newocr.com	52	n/a	n/a	n/a
Ocrgeek.com	52	n/a	n/a	n/a
Ocrconvert.com	52	n/a	n/a	n/a

Image: E00000534.tiff

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBY Recognition Server v3	54	75	48	88.9
ABBY FineReader v12	54	73	50	92.6
Onlineocr.net	54	n/a	n/a	n/a
Newocr.com	54	n/a	n/a	n/a
Ocrgeek.com	54	n/a	n/a	n/a
Ocrconvert.com	54	n/a	n/a	n/a

Image: E00001044.tiff

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBY Recognition Server v3	38	73	30	78.9
ABBY FineReader v12	38	49	27	71.1
Onlineocr.net	38	n/a	n/a	n/a
Newocr.com	38	n/a	n/a	n/a
Ocrgeek.com	38	n/a	n/a	n/a
Ocrconvert.com	38	n/a	n/a	n/a

Image: E00002764.tiff

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBY Recognition Server v3	52	82	41	78.8
ABBY FineReader v12	52	62	21	40.4
Onlineocr.net	52	n/a	n/a	n/a
Newocr.com	52	n/a	n/a	n/a
Ocrgeek.com	52	n/a	n/a	n/a
Ocrconvert.com	52	n/a	n/a	n/a

Image: E00012183.tiff

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBY Recognition Server v3	50	85	45	90
ABBY FineReader v12	50	86	41	82
Onlineocr.net	50	n/a	n/a	n/a
Newocr.com	50	n/a	n/a	n/a
Ocrgeek.com	50	n/a	n/a	n/a
Ocrconvert.com	50	n/a	n/a	n/a

Image: E00012185.tiff

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBYY Recognition Server v3	51	80	48	94.1
ABBYY FineReader v12	51	80	46	90.2
Onlineocr.net	51	n/a	n/a	n/a
Newocr.com	51	n/a	n/a	n/a
Ocrgeek.com	51	n/a	n/a	n/a
Ocrconvert.com	51	n/a	n/a	n/a

Image: E00014366.tiff

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBYY Recognition Server v3	67	100	65	97
ABBYY FineReader v12	67	84	62	92.5
Onlineocr.net	67	n/a	n/a	n/a
Newocr.com	67	n/a	n/a	n/a
Ocrgeek.com	67	n/a	n/a	n/a
Ocrconvert.com	67	n/a	n/a	n/a

Image: E00015007.TIF

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBYY Recognition Server v3	33	127	31	93.9
ABBYY FineReader v12	33	12	5	15.2
Onlineocr.net	33	n/a	n/a	n/a
Newocr.com	33	n/a	n/a	n/a
Ocrgeek.com	33	n/a	n/a	n/a
Ocrconvert.com	33	n/a	n/a	n/a

Image: E00015451.tiff

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBYY Recognition Server v3	59	74	58	98.3
ABBYY FineReader v12	59	81	47	79.6
Onlineocr.net	59	n/a	n/a	n/a
Newocr.com	59	n/a	n/a	n/a
Ocrgeek.com	59	n/a	n/a	n/a
Ocrconvert.com	59	n/a	n/a	n/a

Image: E00262827.TIF

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBYY Recognition Server v3	40	78	30	75
ABBYY FineReader v12	40	51	30	75
Onlineocr.net	40	n/a	n/a	n/a

Newocr.com	40	n/a	n/a	n/a
Ocrgeek.com	40	n/a	n/a	n/a
Ocrconvert.com	40	n/a	n/a	n/a

Image: E00262858.TIF

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBYY Recognition Server v3	26	92	20	76.9
ABBYY FineReader v12	26	44	23	88.5
Onlineocr.net	26	n/a	n/a	n/a
Newocr.com	26	n/a	n/a	n/a
Ocrgeek.com	26	n/a	n/a	n/a
Ocrconvert.com	26	n/a	n/a	n/a

Image: E00314438.tiff

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBYY Recognition Server v3	54	63	53	98.1
ABBYY FineReader v12	54	48	39	72.2
Onlineocr.net	54	n/a	n/a	n/a
Newocr.com	54	n/a	n/a	n/a
Ocrgeek.com	54	n/a	n/a	n/a
Ocrconvert.com	54	n/a	n/a	n/a

Image: E00448970.TIF

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBYY Recognition Server v3	50	109	48	96
ABBYY FineReader v12	50	60	44	88
Onlineocr.net	50	n/a	n/a	n/a
Newocr.com	50	n/a	n/a	n/a
Ocrgeek.com	50	n/a	n/a	n/a
Ocrconvert.com	50	n/a	n/a	n/a

RBGK SPECIMENS

Scores by word for each specimen:

Image: K000809768.tif

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBYY Recognition Server v3	68	72	66	97.1
ABBYY FineReader v12	68	69	66	97.1
Onlineocr.net	68	n/a	n/a	n/a
Newocr.com	68	n/a	n/a	n/a

Ocrgeek.com	68	n/a	n/a	n/a
Ocrconvert.com	68	n/a	n/a	n/a

Image: K000823582.tif

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBYY Recognition Server v3	57	153	32	56.1
ABBYY FineReader v12	57	56	29	50.9
Onlineocr.net	57	n/a	n/a	n/a
Newocr.com	57	n/a	n/a	n/a
Ocrgeek.com	57	n/a	n/a	n/a
Ocrconvert.com	57	n/a	n/a	n/a

Image: K001142491.tif

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBYY Recognition Server v3	64	124	57	89.1
ABBYY FineReader v12	64	65	60	93.8
Onlineocr.net	64	n/a	n/a	n/a
Newocr.com	64	n/a	n/a	n/a
Ocrgeek.com	64	n/a	n/a	n/a
Ocrconvert.com	64	n/a	n/a	n/a

Image: K001142499.tif

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBYY Recognition Server v3	62	104	61	98.4
ABBYY FineReader v12	62	64	61	98.4
Onlineocr.net	62	n/a	n/a	n/a
Newocr.com	62	n/a	n/a	n/a
Ocrgeek.com	62	n/a	n/a	n/a
Ocrconvert.com	62	n/a	n/a	n/a

Image: K001142502.tif

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBYY Recognition Server v3	100	141	89	89
ABBYY FineReader v12	100	100	84	84
Onlineocr.net	100	n/a	n/a	n/a
Newocr.com	100	n/a	n/a	n/a
Ocrgeek.com	100	n/a	n/a	n/a
Ocrconvert.com	100	n/a	n/a	n/a

Image: K001142504.tif

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBY Recognition Server v3	52	133	48	92.3
ABBY FineReader v12	52	53	50	96.2
Onlineocr.net	52	n/a	n/a	n/a
Newocr.com	52	n/a	n/a	n/a
Ocrgeek.com	52	n/a	n/a	n/a
Ocrconvert.com	52	n/a	n/a	n/a

Image: K001142505.tif

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBY Recognition Server v3	69	171	66	95.7
ABBY FineReader v12	69	68	58	84.1
Onlineocr.net	69	n/a	n/a	n/a
Newocr.com	69	n/a	n/a	n/a
Ocrgeek.com	69	n/a	n/a	n/a
Ocrconvert.com	69	n/a	n/a	n/a

Image: K001148043.tif

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBY Recognition Server v3	54	221	50	92.6
ABBY FineReader v12	54	57	53	98.1
Onlineocr.net	54	n/a	n/a	n/a
Newocr.com	54	n/a	n/a	n/a
Ocrgeek.com	54	n/a	n/a	n/a
Ocrconvert.com	54	n/a	n/a	n/a

Image: K001148047.tif

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBY Recognition Server v3	64	222	62	96.9
ABBY FineReader v12	64	74	62	96.9
Onlineocr.net	64	n/a	n/a	n/a
Newocr.com	64	n/a	n/a	n/a
Ocrgeek.com	64	n/a	n/a	n/a
Ocrconvert.com	64	n/a	n/a	n/a

Image: K001148070.tif

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBY Recognition Server v3	78	84	75	96.2
ABBY FineReader v12	78	83	74	94.9
Onlineocr.net	78	n/a	n/a	n/a
Newocr.com	78	n/a	n/a	n/a

Ocrgeek.com	78	n/a	n/a	n/a
Ocrconvert.com	78	n/a	n/a	n/a

Image: K001148104.tif

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBYY Recognition Server v3	68	227	65	95.6
ABBYY FineReader v12	68	68	62	91.2
Onlineocr.net	68	n/a	n/a	n/a
Newocr.com	68	n/a	n/a	n/a
Ocrgeek.com	68	n/a	n/a	n/a
Ocrconvert.com	68	n/a	n/a	n/a

Image: K001148819.tif

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBYY Recognition Server v3	76	223	76	100
ABBYY FineReader v12	76	77	75	98.7
Onlineocr.net	76	n/a	n/a	n/a
Newocr.com	76	n/a	n/a	n/a
Ocrgeek.com	76	n/a	n/a	n/a
Ocrconvert.com	76	n/a	n/a	n/a

Image: K001148829.tif

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBYY Recognition Server v3	62	149	38	61.3
ABBYY FineReader v12	62	73	43	69.4
Onlineocr.net	62	n/a	n/a	n/a
Newocr.com	62	n/a	n/a	n/a
Ocrgeek.com	62	n/a	n/a	n/a
Ocrconvert.com	62	n/a	n/a	n/a

Image: K001148830.tif

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBYY Recognition Server v3	50	109	48	96
ABBYY FineReader v12	50	60	44	88
Onlineocr.net	50	n/a	n/a	n/a
Newocr.com	50	n/a	n/a	n/a
Ocrgeek.com	50	n/a	n/a	n/a
Ocrconvert.com	50	n/a	n/a	n/a

Image: K001148964.tif

Service	Actual word count	Total Output word count	Correct words	% of ocr correct (Correct/Actual)
ABBYY Recognition Server v3	20	123	13	65
ABBYY FineReader v12	20	33	14	70

Onlineocr.net	20	n/a	n/a	n/a
Newocr.com	20	n/a	n/a	n/a
Ocrgeek.com	20	n/a	n/a	n/a
Ocrconvert.com	20	n/a	n/a	n/a

APPENDIX 1F: OCR SOFTWARE RESULTS FROM RBGK TESTING OF DIFFERENT FORMATTING OPTIONS

Barcode	Adjustment settings	Score	Score as %
K000823582	Original	3/8	38
K000823582	Formatted 1	3/8	38
K000823582	Formatted 2	3/8	38
K000823582	Formatted 3	3/8	38
K000823582	Formatted 4	1/8	13
K000823582	Formatted 5	0/8	0
K000823582	Formatted 6	3/8	38
K000823582	Formatted 7	0/8	0
K000823582	Formatted 8	1/8	13
K000823582	Formatted 9	3/8	38
K000823582	Formatted 10	1/4	25
K000823582	Formatted 11	1/4	25
K000823582	Formatted 12	1/4	25
K000823582	Formatted 13	1/4	25
K000823582	Formatted 14	1/8	13
K000823582	Formatted 15	1/8	13
K000823582	Formatted 16	1/4	25
K001142491	Original	16/19	84
K001142491	Formatted 1	16/19	84
K001142491	Formatted 2	16/19	84
K001142491	Formatted 3	14/19	74
K001142491	Formatted 4	15/19	79
K001142491	Formatted 5	4 /19	21
K001142491	Formatted 6	16/19	84
K001142491	Formatted 7	15/19	79
K001142491	Formatted 8	15/19	79
K001142491	Formatted 9	13/19	68

K001142491	Formatted 10	12/19	63
K001142491	Formatted 11	17/19	89
K001142491	Formatted 12	17/19	89
K001142491	Formatted 13	17/19	89
K001142491	Formatted 14	17/19	89
K001142491	Formatted 15	16/19	84
K001142491	Formatted 16	17/19	89
K001142499	Original	10/13	77
K001142499	Formatted 1	/	/
K001142499	Formatted 2	10/13	77
K001142499	Formatted 3	8/13	62
K001142499	Formatted 4	9/13	69
K001142499	Formatted 5	3/13	23
K001142499	Formatted 6	10/13	77
K001142499	Formatted 7	11/13	85
K001142499	Formatted 8	10/13	77
K001142499	Formatted 9	10/13	77
K001142499	Formatted 10	9/13	69
K001142499	Formatted 11	12/13	92
K001142499	Formatted 12	12/13	92
K001142499	Formatted 13	11/13	85
K001142499	Formatted 14	12/13	92
K001142499	Formatted 15	11/13	85
K001142499	Formatted 16	12/13	92
K001142502	Original	6/10	60
K001142502	Formatted 1	/	/
K001142502	Formatted 2	7/10	70
K001142502	Formatted 3	6/10	60
K001142502	Formatted 4	5/10	50
K001142502	Formatted 5	1/10	10
K001142502	Formatted 6	7/10	70
K001142502	Formatted 7	4/10	40

K001142502	Formatted 8	4/10	40
K001142502	Formatted 9	5/10	50
K001142502	Formatted 10	/	/
K001142502	Formatted 11	7/10	70
K001142502	Formatted 12	7/10	70
K001142502	Formatted 13	8/10	80
K001142502	Formatted 14	7/10	70
K001142502	Formatted 15	4/10	40
K001142502	Formatted 16	/	/
K001142504	Original	9/12	75
K001142504	Formatted 1	/	/
K001142504	Formatted 2	9/12	75
K001142504	Formatted 3	9/12	75
K001142504	Formatted 4	10/12	83
K001142504	Formatted 5	4/12	33
K001142504	Formatted 6	9/12	75
K001142504	Formatted 7	9/12	75
K001142504	Formatted 8	9/12	75
K001142504	Formatted 9	11/12	92
K001142504	Formatted 10	10/12	83
K001142504	Formatted 11	/	/
K001142504	Formatted 12	5/6	83
K001142504	Formatted 13	10/12	83
K001142504	Formatted 14	10/12	83
K001142504	Formatted 15	10/12	83
K001142504	Formatted 16	9/12	75
Average	Original		66.73
	Formatted 1		60.86
	Formatted 2		68.73
	Formatted 3		61.54
	Formatted 4		58.80

	Formatted 5		17.49
	Formatted 6		68.73
	Formatted 7		55.71
	Formatted 8		56.67
	Formatted 9		64.90
	Formatted 10		60.18
	Formatted 11		72.02
	Formatted 12		72.02
	Formatted 13		72.48
	Formatted 14		69.52
	Formatted 15		60.93
	Formatted 16		70.45

Table 13. *The scores for different image formatting approaches for a subset of RBGK specimens*

APPENDIX 2: SCREENSHOTS OF PORTALS USING

Lichen Portal: <http://lichenportal.org/portal/>

The screenshot shows the homepage of the Consortium of North American Lichen Herbaria (CNALH). At the top, there is a header with the CNALH logo and the text "NORTH AMERICAN LICHEN HERBARIA". Below the header, there is a "Main Menu" on the left with links to "Search Collections", "Map Search", "Exsiccati", "Image Browser", "Search Images", "About CNALH", and "Data Usage Policy". The main content area has a "Welcome to the Consortium of North American Lichen Herbaria" section with a paragraph about the consortium's purpose. Below this, there is a "Flora Projects" section with a list of regions: Arizona, California, Colorado, Florida, Massachusetts, North Carolina, Wisconsin, Arctic Flora, and Southern Subpolar Region. To the right of the list is a photograph of lichen. Further right, there is a "Join the Consortium" section with contact information for CNALHAdmin@asu.edu. At the bottom right, there is a "News and Events" section with a bullet point about an NSF Press Release.

Bryophyte Portal: <http://bryophyteportal.org/portal/>

The screenshot shows the homepage of the Consortium of North American Bryophyte Herbaria (CNABH). At the top, there is a header with the CNABH logo and the text "NORTH AMERICAN BRYOPHYTE HERBARIA". Below the header, there is a navigation bar with links to "Home", "Explore", "About", "Data Usage", "Crowdsourcing", "Flora Projects", and "Other Resources". The main content area has a "Welcome to the Consortium of North American Bryophyte Herbaria" section with a paragraph about the consortium's purpose. Below this, there is a "News and Events" section with a bullet point about an NSF Press Release. To the right of the list is a photograph of a bryophyte. Further right, there is a "Join the Consortium" section with contact information for CNABHAdmin@asu.edu. At the bottom right, there is a "News and Events" section with a bullet point about an NSF Press Release.

SERNEC (Southeast Regional Network of Expertise and Collections) Portal:

<http://sernecportal.org/portal/index.php>

SERNEC

Southeast Regional Network of Expertise and Collections

[Home](#) [Search Collections](#) [Map Search](#) [State Floras](#) [Dynamic Tools](#) [Images](#) [Log In](#) [New Account](#) [Sitemap](#)

Welcome to SERNEC

Herbaria are not simply repositories of plant specimens, they are repositories of a tremendous amount of information. Current technologies provide an opportunity to access this information at an unprecedented scale. The real power of herbaria as research tools can be fully realized when both large and small collections within a broad geographic region are electronically available and searchable.


SERNEC (SouthEast Regional Network of Expertise and Collections) is designed to facilitate this process, by building partnerships, encouraging the utilization of the collective expertise of the network, and assisting herbaria in providing information to the public.

SERNEC is 1) networking the 230 herbaria in 14 states in southeastern North America, 2) developing a strategy for advancing each state's ongoing databasing effort, and 3) working to publish online botanical resources that will be available to scientists, land managers, state and federal agencies, educators and the general public. These data will provide a greater understanding of one of the most botanically diverse regions of the earth and will lead to better research, better management planning and a more well-informed public.

Development of a searchable collective database at a regional scale will provide a powerful research tool, and by combining 150 years of botanical information housed in herbaria in the Southeast with models of past plant migrations and current ecological parameters, we can revolutionize studies in biodiversity, evolution, ecology and systematics. We are also working to link our efforts with those of other regional herbarium groups through the US Virtual Herbarium and with the national biodiversity informatics effort, iDigBio.


Search Collections

[General Data Usage Policy](#)



This project made possible by National Science Foundation Award 1410069

Plant of the Day



What is this plant?
[Click here to test your knowledge](#)

APPENDIX 3: PROTOCOL FOR USING TRANSKRIBUS FOR NATURAL HISTORY COLLECTIONS

Transkribus Manual: <https://transkribus.eu/Transkribus/docs/How%20to%20use%20TRANSKRIBUS-0.1.6.pdf>

INTRODUCTION

This is a modified manual for Natural History collections based on the original Transkribus manual. Modifications have been made based on the experiences of the Royal Botanic Garden Edinburgh (RBGE), the Royal Botanic Gardens Kew (RBGK), the Botanic Garden and Botanical Museum Dahlem-Berlin (BGBM) and the Royal Museum for Central Africa (RMCA).

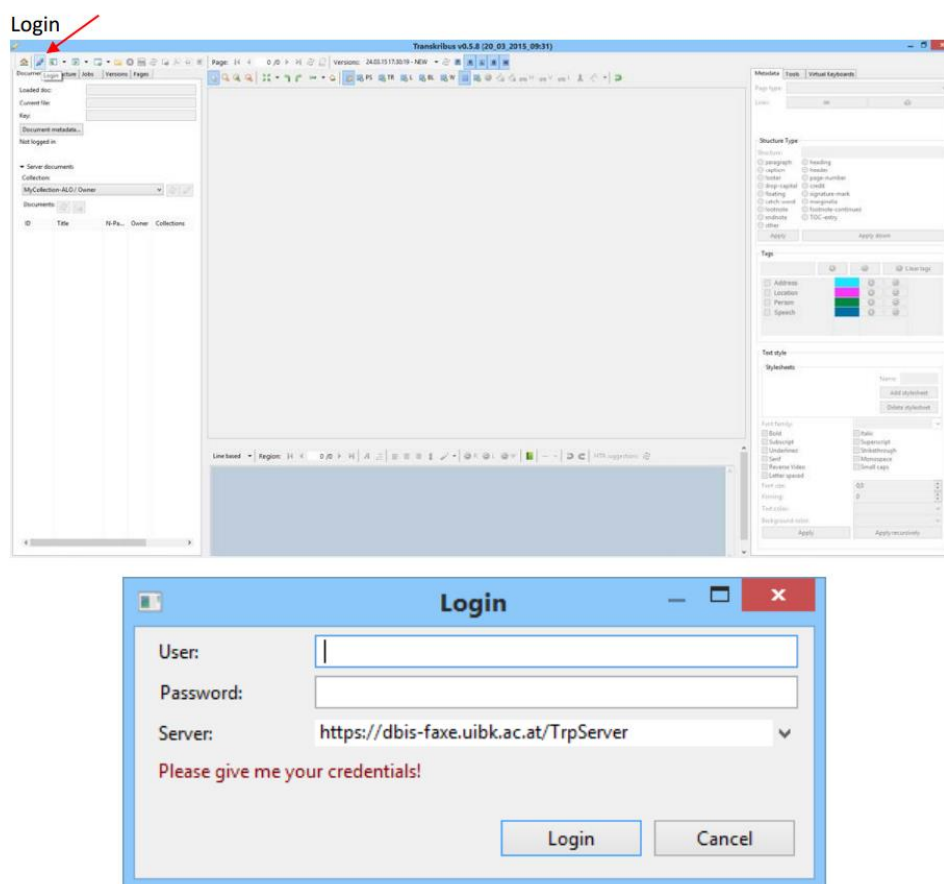
Transkribus is an expert tool. As with other feature-rich software it is designed to meet the needs of users who “know what to do and how”.

Be aware that it will take you some time until you explore all options and get familiar with the behaviour of Transkribus. Of course we are happy to support you in the best way we can (don't be shy in contacting us, or use the bug report and feature request button within Transkribus).

STEP 1: REGISTER AND DOWNLOAD SOFTWARE

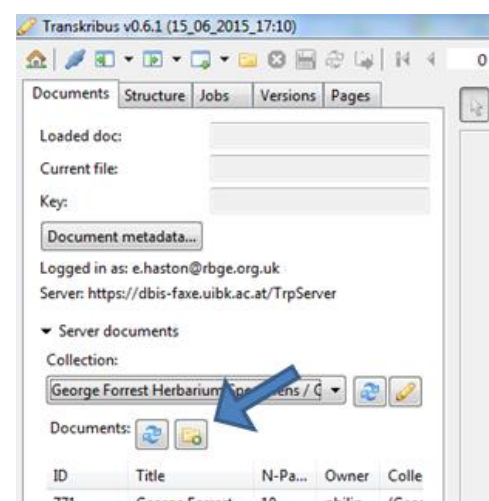
1. <http://transkribus.eu/>
2. The University of Innsbruck is offering this service as a research infrastructure.
3. Read our User agreement (will soon come in English!) – we will respect your privacy and use the data only to improve our services and support research in humanities and computer science!
4. Activate your account when you receive an e-mail from Transkribus
5. Download the tool – it will run on Windows, MacOS and Linux. Attention: Unzip the file – you cannot start the tool from the zip File.
6. Start the tool with
 - a. i. Transkribus.exe (as Windows user)
 - b. ii. Transkribus.command (as Mac user)
 - c. iii. Transkribus.sh (as Linux user).
7. The Transkribus exe file must stay in the folder with the other files, so make a shortcut if you want to open it without going into the folder.

STEP 2: LOG IN



STEP 3: UPLOAD DOCUMENTS TO YOUR PRIVATE COLLECTION

1. Images should be in a standard resolution. Variations in resolution, particularly between the training dataset and documents for transcription will affect the performance of the HTR. This may occur more commonly if multiple institutes are collaborating on material from a single collector.
2. Use the "Upload" button in Transkribus to transfer the images from your computer to the platform.
3. Note, the images have to reside in a separate folder!
4. Be aware that the upload of several hundred megabytes may be difficult with a wireless connection or from at home.
5. You may use a file sharing system, such as Dropbox or WeTransfer and afterwards contact Transkribus directly, sending the link. Your documents will be uploaded into your private collection.
6. When uploading a document create your own collection. Only users who are authorised by you will be able to access your documents, it is you who has full control on all your documents.



7. Use the “Collection Manager” to add users (who need to be registered in Transkribus) to your collection.

Upload dialog

Local folder:

Title on server:

Collection: **MyCollection-ALO**

Create collection:

Upload **Cancel**

CollectionManager

Collection Manager

Collections

ID	Name	Description	Role
4	Transkribus CL...		Transcriber
5	DHd Worksho...		Transcriber
169	MyCollection-...	created by alo...	Owner

Users in collection

Username	Name	Role
alo@uibk.ac.at	alo literat...	Owner

Documents in collection

ID	Title	N-Pa...	Owner	Collections
437	Papyrus Exam...	1	alo@...	(MyCollection-ALO,

My documents

ID	Title	N-Pa...	Owner	Collections
437	Papyrus Exam...	1	alo@...	(MyCollection-ALO,

Find users

Username / E-Mail:

First name:

Last name:

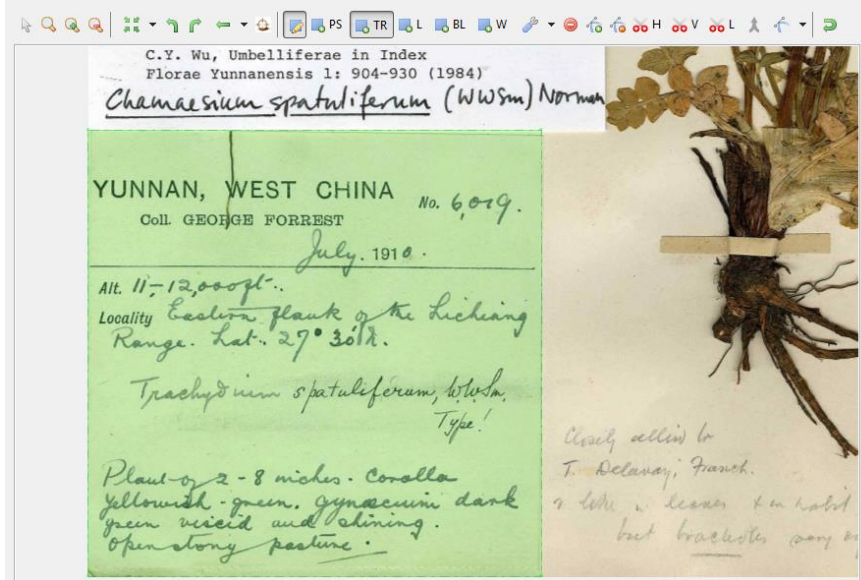
Find users

STEP 4: SEGMENT YOUR DOCUMENT INTO TEXT BLOCKS AND BASELINES

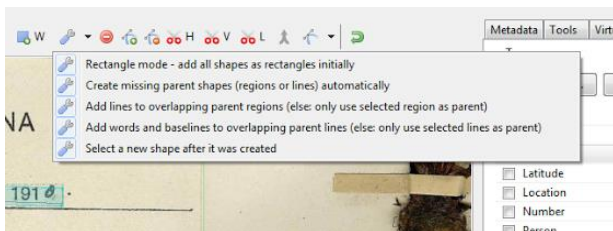
1. TRANSKRIBUS allows you to make a direct link between the images of your document and the text (actually it is not possible to create a transcription without this direct link).
2. You need first to draw the text blocks and afterwards to draw the baselines of lines.
 - a. Select +TR. Decide if you want to draw a rectangle (usually the sufficient) or a polygon (might be helpful in some cases).

- b. Use a single click to start the box in one corner. Place the mouse where the opposite corner should be and use a single click to create the text region.
- c. Text blocks can overlap, no problem!

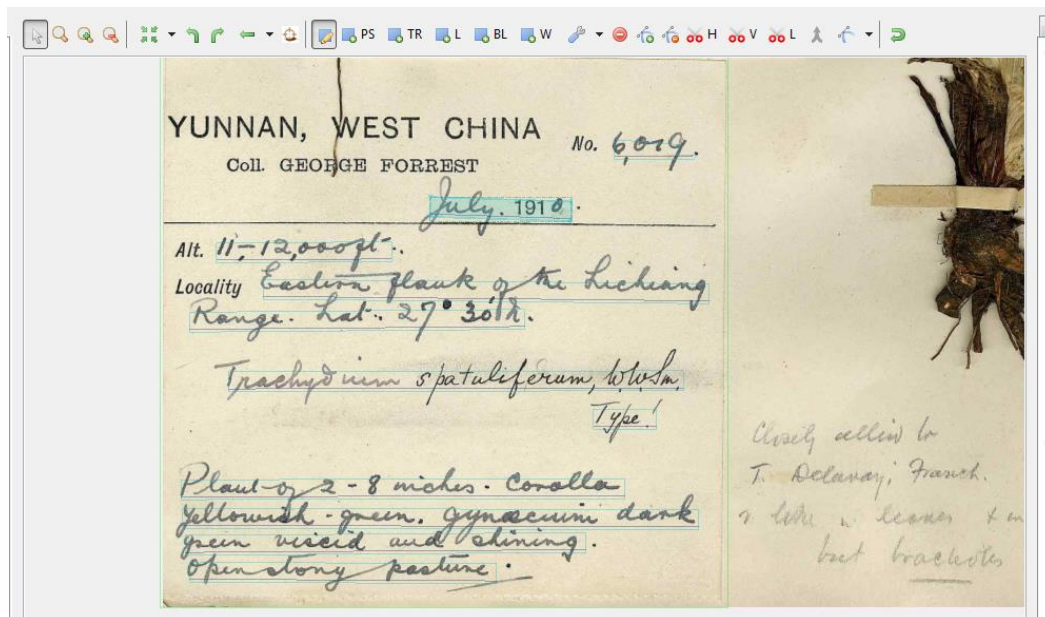
A text region for a herbarium specimen



The options in the settings tool



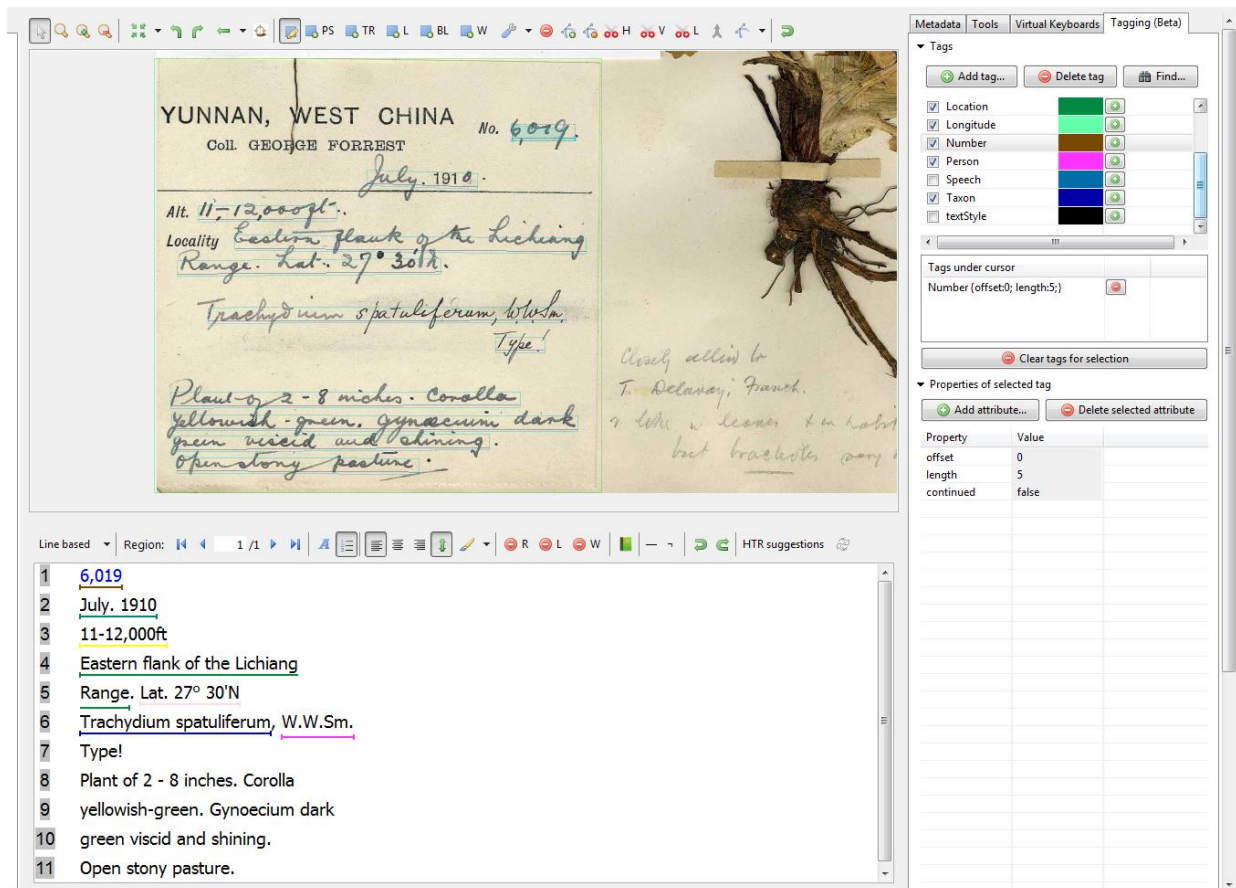
3. Afterwards add the baselines and line regions.
 - a. In the settings tool, all the options should be selected. This will automatically create a line region when the baseline is drawn.
 - b. Select the baseline tool (BL) and manually draw in the baselines for each line of handwritten text. The baseline should sit just below the text.
 - c. Do not include typewritten text on the baseline - it confuses the HTR tool.
 - d. Baselines cannot be seen on the document, but they are important: It is the
 - e. invisible line on which the characters are „sitting“.
 - f. It is possible to automatically detect baselines and lines, but for specimen labels this function does not work successfully at present.
 - g. If a text region has not been created first, TRANSKRIBUS will create a unique text region for each baseline drawn.



STEP 5: MANUALLY TRANSCRIBE A TRAINING DATASET OF 100 PAGES.

1. Start your transcription.
 - a. Once there are text blocks and baselines visible on your image you are able to write text into the text field.
 - b. Display of image and text are synchronised this will make it easier for your eyes when transcribing the text.
 - c. Use the Structure tab on the left hand side to navigate through the page, one click highlights the element, double-click zooms the element.
 - d. Special characters can be found in the „Virtual Keyboard“ on the right hand side (you may add specific characters in the custom section).
 - e. Tagging
 - i. You may also be interested to tag parts of the text with specific tags, such as „person“ or „date“, or „Description of landscape“.
 - ii. Use the tags buttons to add tags or create your own tags.

An example of the tags used for a specimen at RBGE.



2. Save and export your transcription

- Press the "Save" button to save the document at the platform. You will see that a new version is generated. This gives you the chance to start from a step before – if you have experienced some problems.
- You are able to export the whole document at any time of the process.
- You may also be interested in working versions, e.g. download the documents as RTF (Rich Text Format), or as PDF (with the text in the background) or TEI (Text Encoding Initiative). This can be done at any time. Note: RTF and TEI are currently very basic – this will be refined during the next months.

3. At the "Tools" Tab on the right hand side you will see several tools for segmentation

However, these tools are not optimised for natural history specimens so their use is not recommended at present. Note: All tools are currently applied only to the selected page or the selected region, not to the whole document.

a. Tools: Layout Analysis

i. Detect regions

- Detects blocks of handwritten or printed text on a page. Runs well with simple layout, has problems with more sophisticated layout.
- Usually it is better to just draw the region by hand.

ii. 2. Detect lines and baselines

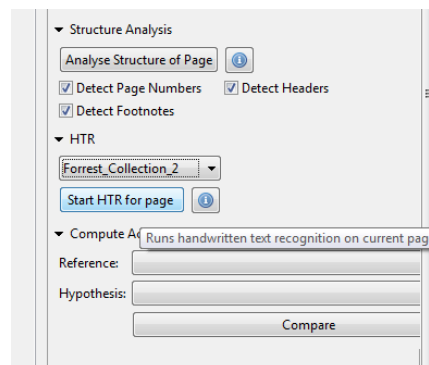
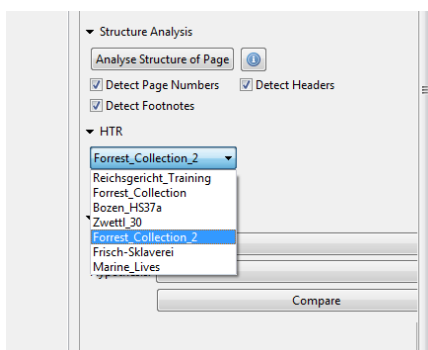
1. Detects lines and baselines of a text region in one step. Note that for transcription only the baseline is needed – so forget the line regions.
 2. Runs well with straight lines, may cause problems with short lines and long ascenders and descenders. Benefits a lot from good text regions (manually drawn)
- iii. 3. Detect baselines
1. In some rare cases there might be already line regions, with this tool the baselines are added.
 2. Usually not necessary.

STEP 6: TRAINING THE HTR MODEL.

1. When you have completed the manual transcription of 100 pages to form a training dataset, inform the team at TRANSKRIBUS.
2. They will then process the training dataset and build the HTR model.
3. Once this is completed and available you will need to search for updates (an option in the main menu) and restart TRANSKRIBUS.

STEP 7: RUNNING THE HTR MODEL.

1. Upload the documents to be automatically transcribed. This should be at the same resolution as the training dataset since any variation will cause the HTR not to work effectively.
2. All pages to be automatically transcribed should be marked up with text regions and baselines as above.
3. Once this is completed you can then run the HTR on each page individually by selecting the relevant HTR model from the drop-down list and clicking Start HTR for page.
4. An HTR job will be created and you can see the progress in the Jobs tab.



Transkribus v0.6.4.3-SNAPSHOT (03_08_2015_15:50), Loaded doc: George Forrest

Doc-Id	User-Id	Page	Description	ID
115 09:35:10	1989	hartm...	-1	Job was cancell...
115 10:01:57	1990	hartm...	-1	Job was cancell...
115 10:16:14	1991	lei.pa...	-1	Job was cancell...
115 10:56:30	1168	lei.pa...	4	HTR finished!
115 11:11:03	1168	lei.pa...	4	HTR finished!
115 12:07:47	1024	lei.pa...	1	HTR finished!
115 12:52:48	1992	hansu...	-1	DONE
115 13:18:21	1168	lei.pa...	2	HTR finished!
115 13:25:25	1992	hansu...	-1	OCR job finish...
115 15:20:56	1118	e.hast...	301	HTR finished!
115 04:21:55	1993	sarah...	-1	DONE
115 08:23:16	1994	hartm...	-1	Uploaded ima...
115 14:00:07	904	hansu...	-1	Job was cancell...
115 14:00:57	904	hansu...	1	HTR process...
115 13:17:36	1995	lei.pa...	-1	DONE
115 13:20:13	1995	lei.pa...	1	HTR process...
115 13:27:37	1995	lei.pa...	1	HTR finished!
115 13:30:14	1995	lei.pa...	3	HTR finished!
115 13:58:06	1995	lei.pa...	3	HTR finished!
115 14:07:13	1995	lei.pa...	3	HTR finished!
115 14:19:26	1995	lei.pa...	3	HTR finished!
115 12:09:56	1996	a.kirc...	-1	Uploaded ima...
115 12:17:58	1997	a.kirc...	-1	Uploaded ima...
115 12:23:11	1998	a.kirc...	-1	Uploaded ima...
115 12:35:49	1999	a.kirc...	-1	Uploaded ima...
115 12:51:40	2000	a.kirc...	-1	Uploaded ima...
115 15:10:56	1993	e.hast...	6	Job was cancell...
115 15:57:47	1993	e.hast...	10	HTR finished!
115 16:40:02	1993	sarah...	6	HTR finished!
115 16:46:14	1993	sarah...	11	HTR finished!
115 16:50:49	1993	sarah...	30	HTR finished!
115 16:54:56	1993	sarah...	31	HTR finished!
115 17:26:07	1993	e.hast...	210	HTR finished!
115 17:38:01	1993	e.hast...	210	HTR finished!
115 17:42:37	1993	e.hast...	6	HTR finished!
115 17:59:21	1993	e.hast...	30	HTR finished!
115 18:07:16	1993	e.hast...	30	HTR finished!
115 21:26:36	2001	lei.pa...	-1	DONE
115 21:37:46	2001	lei.pa...	5	HTR finished!
115 21:42:03	2001	lei.pa...	5	HTR finished!
115 22:18:24	2001	lei.pa...	5	HTR finished!
115 22:34:02	1168	lei.pa...	2	HTR finished!
115 22:40:23	1168	lei.pa...	4	HTR finished!
115 22:43:06	1168	lei.pa...	4	HTR finished!
115 08:59:34	1119	e.hast...	213	HTR finished!

APPENDIX 4: PROTOCOLS FOR SAMPLING AND EXTRACTING DNA FROM HERBARIUM SPECIMENS AT RBGE

PROTOCOL FOR CAPTURING SPECIMEN COLOUR METADATA AND COMPARING IT TO DNA VIABILITY

1. Select specimens
2. If the specimen is not databased, attach a barcode to the specimen and create at least a minimal database record.
3. Fold a silica dried collection (SDC) capsule and write on in pencil the collector's name and number and the barcode in the top left corner.
4. Using tweezers which have been cleaned on a tissue wetted with ethanol, but dry, remove material from the capsule if present, or from the specimen itself. Record if the material came from the capsule or the specimen. Gloves are not necessary, and could have static which could cause other issues. The main issue to be careful about is cross-contamination from fragments, therefore care should be taken to ensure that no fragments are on your hands or on the tweezers. Hands do not have to be washed in between but if there are any fragments these should be wiped off.
5. The amount of material should be at least equal to the size of the lid of an eppendorff tube. If possible, it should be more than two of these. The idea would be to aim for a high concentration of DNA to reduce the need to go back and resample, and some excess material could go into the silica dried collection.
6. The material should then be imaged. Place the empty capsule on a blank herbarium sheet with the leaf sample above it face up. Take the image and save it as barcode_y. Then turn the leaf sample upside-down so the lower surface is visible and taken a second image, saving it as barcode_z.
7. All the raw images should be processed as normal to get tiffs, but not sent to the image poller. Instead, they should be transferred here:
8. Once this is done, the appropriate amount of material can be taken for extraction. For this study, the amount should be equal to a single eppendorff lid to aim for a consistent amount as much as possible.
9. The 96-tube plate should be prepared beforehand. A full plate and an empty plate should be laid out. A strip of 8 tubes should be transferred from the full place and placed into the empty plate, replacing the lid on the full plate to reduce risk of contamination. A strip of tape should be placed along the top of the strip of tubes. As you place the leaf sample in the tube, you should pull back the tape to open a single tube, place the leaf sample in and then close the first lid on a tape of lids. Then pull back the tape to open the second tube, place in the leaf sample, and cover with the second lid. Continue for all eight tubes on the strip. Replace them in the full plate and take out the next strip of eight tubes and repeat.

10. All samples should be listed in a spreadsheet and the number of the tube, eg 8D, should be recorded.
11. All herbarium specimens should be placed in a box and sent to Suzanne Cubey who is in charge of destructive sampling of herbarium specimens. She will record the sampling event and annotate the specimens accordingly.

DNA EXTRACTION METHODOLOGY: USING THE QIAGEN AUTOMATED QIAXTRACTOR

1. Samples were first ground using a QIAGEN TissueLyser II (for set-up see manufacturer's instructions). Tube caps were covered by a sheet of sticky film (supplied with the QIAxtractor consumables). The samples were shaken for 30 seconds at 20Hz. The adaptors containing the tubes were disassembled; the tube inserts turned 180° and the samples shaken again for 30 seconds at 20Hz. The grinding process was repeated until a fine powder was obtained (see Figure 1).
Figure 1 Colour variation of ground leaf samples from unidentified (at the taxonomic rank of family) RBGE herbarium specimens. (See spreadsheet for specimen information for each sample).
2. Samples were centrifuged for 10 minutes at 6,000RPM.
3. The lysis/digestion buffer was made up in a trough: 400µl RNase, 400µl DX enzyme and 40.4ml of the DXT reagent. (All reagents supplied by QIAGEN and stored as per manufacturer's instructions. If there are precipitates in the DXT reagent, it should be incubated at 37°C with gentle shaking).
4. Using a multi-channel pipette, 420µl of lysis/digestion buffer was pipetted into the samples. (The buffer was added to the samples within 10 minutes of being made up to avoid enzyme degradation.)
5. The powdered sample material was loosened and mixed with the reagents by tapping the tubes on the bench.
6. Samples were placed in a Thermo mixer for 1 hour, at 65°C, at 800RPM.
7. QIAxtractor is prepared for samples (for set-up see manufacturer's instructions).
8. The lysed/digested samples were spun at 2,000RPM for 10 minutes.
9. 220µl of the uppermost clear liquid was pipetted into the QIAxtractor white sample plate (see Figure 2). To ensure no solid material was transferred across this stage was carried out in two steps: 110µl being extracted at a time, resting the ridge of the pipette tips on top of the sample tubes to avoid contact with the solid material.

Figure 2 Colour variation of the leaf samples after lysis/digestion in preparation for DNA extraction using the QIAGEN automated QIAxtractor. Samples were taken from unidentified (at the taxonomic rank of family) RBGE herbarium specimens. (See spreadsheet for specimen information for each sample).

10. The 96 well plate with the lysed samples was placed into the extractor and the program set to run (see manufacturer's instructions for details).
11. The elution tubes were removed and the strip caps attached.