Grant Agreement Number: 777483 / **Acronym:** ICEDIG
Call: H2020-INFRADEV-2017-1 / **Type of Action:** RIA
Start Date: 01 Jan 2018 / **Duration:** 27 months

REFERENCES:

Deliverable **D4.1** / R / PU
Work package 4 / Lead: APM
Delivery date [13]

ICEDIG.EU

Innovation and consolidation for large scale digitisation of natural heritage

Methods for Automated Text Digitisation

Deliverable 4.1

Authors: David Owen[1], Quentin Groom[2], Alex Hardisty[1], Thijs Leegwater[3], Myriam van Walsum[3], Noortje Wijkamp[3], Irena Spasić[1].

Contributors: Mathias Dillen[2], Laurence Livermore[4], Sarah Phillips[5], Zhengzhe Wu[6]

1. School of Computer Science & Informatics, Cardiff University, Cardiff, United Kingdom
2. Meise Botanic Garden, Meise, Belgium
3. Picturae BV, Heiloo, Netherlands
4. The Natural History Museum, London, United Kingdom
5. Royal Botanic Gardens, Kew, United Kingdom
6. Finnish Museum of Natural History, LUOMUS, Helsinki, Finland

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

ICEDIG.EU

1. Executive Summary

In this document we describe an effective approach to automated text digitisation with respect to specimen labels. These labels contain much useful data about the specimen including its collector, country of origin and collection date. Our approach to automatically extracting these data takes the form of a pipeline. Recommendations are made for the pipeline's component parts based on some of the state-of-the-art technologies.

Optical Character Recognition (OCR) can be used to digitise text on images of specimens. However, recognising text quickly and accurately from these images can be a challenge for OCR. We show that OCR performance can be improved by prior segmentation of specimen images into their component parts. This ensures that only text-bearing labels are submitted for OCR processing as opposed to whole specimen images, which inevitably contain non-textual information that may lead to false positive readings. In our testing Tesseract OCR version 4.0.0 offers promising text recognition accuracy with segmented images.

Not all the text on specimen labels is printed. Handwritten text varies much more and does not conform to standard shapes and sizes of individual characters, which poses an additional challenge to OCR. Recently, deep learning has allowed for significant advances in this area. Google's Cloud Vision, which is based on deep learning, is trained on large-scale datasets, and is shown to be quite adept at this task. This may take us some way towards negating the need for humans to routinely transcribe handwritten text.

Determining the countries and collectors of specimens has been the goal of previous automated text digitisation research activities. Our approach also focuses on these two pieces of information. An area of Natural Language Processing (NLP) known as Named Entity Recognition (NER) has matured enough to semi-automate this task. Our experiments demonstrated that existing approaches can accurately recognise location and person names within the text extracted from segmented images via Tesseract version 4.0.0. Potentially, NER could be used in conjunction with other online services, such as those of the Biodiversity Heritage Library to map the named entities to entities in the biodiversity literature (https://www.biodiversitylibrary.org/docs/api3.html).

We have highlighted the main recommendations for potential pipeline components. The document also provides guidance on selecting appropriate software solutions. These include automatic language identification, terminology extraction and integrating all pipeline components into a scientific workflow in order to automate the overall digitisation process.

2. Contents

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

3. Introduction

3.1 Background

We do not know how many specimens are held in the world's museums and herbaria. However, estimates of three billion seem reasonable (Wheeler et al., 2012). These specimens are irreplaceable and contribute to a diverse range of scientific fields (Suarez & Tsutsui, 2004; Pyke & Ehrlich, 2010). Their labels hold data on species distributions, scientific names, traits, people and habitats. Among those specimens are nomenclatural types that underpin the whole of formal taxonomy and define the species concept. These specimens span more than 200 years of biodiversity research and are an important source of data on species populations and environmental change. This enormous scientific legacy is largely locked into the handwritten or typed labels mounted with the specimen or in associated ledgers and field notebooks. It is a significant challenge to extract these data digitally, particularly without introducing errors. Furthermore, the provenance of these data must be maintained so that they can be verified against the original specimen.

Perhaps, the method most widely used today to extract these data from labels is for expert technicians to type the specimen details into a dedicated collection management system. They might, at the same time, georeference specimens where coordinates are not already provided on the specimen. Volunteers have also been recruited to help with this process and in some cases transcription has been outsourced to companies specializing in document transcription (Engledow et al., 2018; Ellwood et al., 2018).

Nevertheless, human transcription of labels is slow and requires both skill to read the handwritten labels and knowledge of the names of places, people and organisms. These labels are written in many languages often in the same collection and sometimes on the same label. Furthermore, abbreviations are frequently used and there is little standardisation on where each datum can be found on the label.

Full or partial automation of this process is desirable to improve the speed and accuracy of data extraction and reduce the associated costs. Automating even the simplest tasks such as triaging the labels by language and/or writing method (handwritten vs typed) stands to improve the overall efficiency of the human-in-the-loop approach.

OCR and NLP proved effective for extracting data from biodiversity literature (Thessen, Cui & Mozzherin, 2012; Hoehndorf et al., 2016). However, specimen labels pose additional problems compared to formally structured text such as that found in literature. The context of individual words is often difficult to determine; specimens that overlap with the label may obscure some words; the orientation of labels typically varies; typed and handwritten text may coexist within the same label and the handwriting on the same specimen may come from different people (Figure 1). Therefore, the task of digitising the text found in specimen labels is far from simple and requires different approaches to standard text recognition.

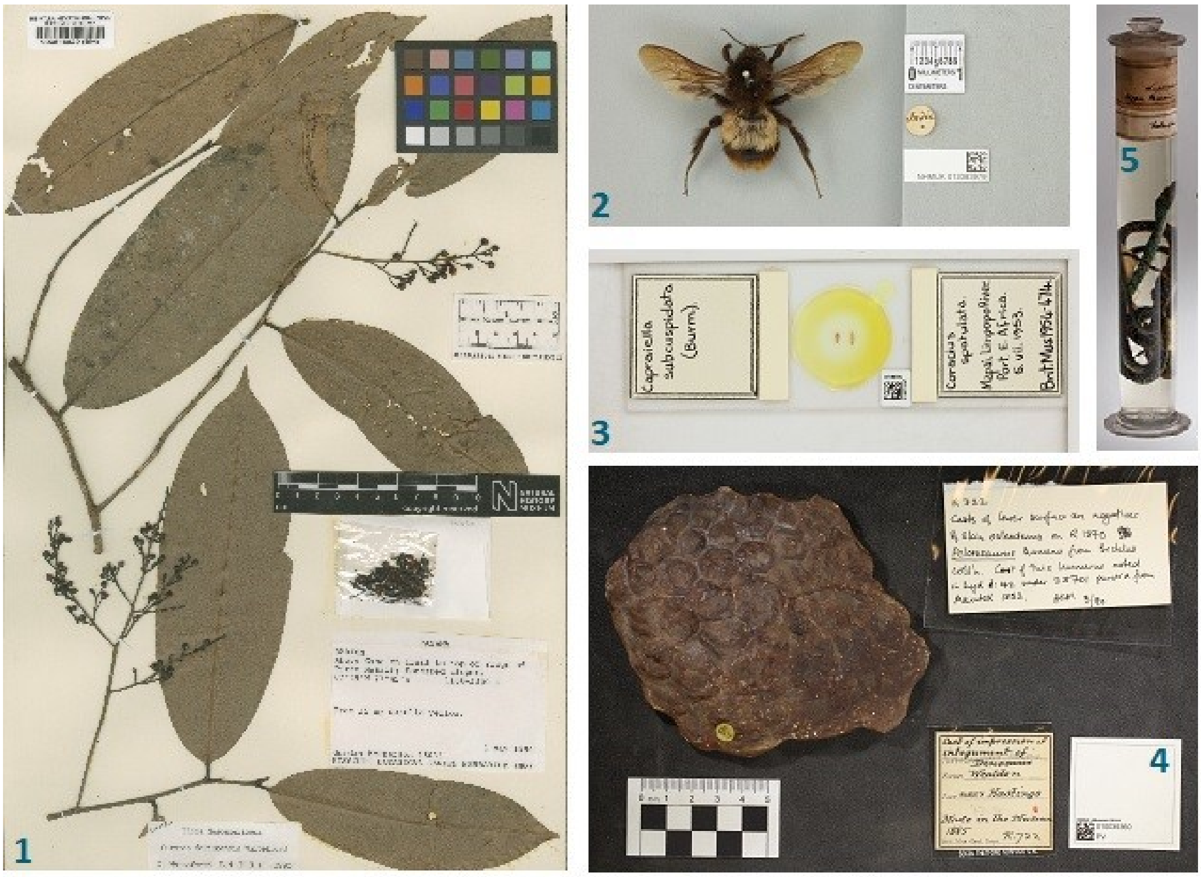


Figure 1: A range of sample specimens. 1=Herbarium specimen, 2=Pinned insect, 3=Microscope slide, 4=Fossilized animal skin, 5=Liquid preserved specimen. These examples demonstrate the wide taxonomic range of specimens encountered in collections, but also the diversity of label types, which include handwritten, typed and printed labels. Note the presence of various barcodes, rulers and a colour chart in addition to labels describing the origin of the specimen and its identity. Specimen source: NHM Data Portal (Natural History Museum London, 2018).

This document examines the state of the art in automated text digitisation with respect to specimen images. The recommendations within are designed to enhance the digitisation and transcription pipelines that exist at partner institutions. They are also intended to provide guidance towards a proposed centralised specimen enrichment pipeline that could be created under a pan-European Research Infrastructure for biodiversity collections (DiSSCo, https://dissco.eu/). This pipeline would provide state-of-the-art label digitisation services to institutions that need them.

In this document we focus mainly on herbarium specimens, even though similar data extraction problems exist for pinned insects, liquid collections and animal skins. Herbarium specimens are among the most difficult targets and we know from recent successful pilot studies for large-scale digitisation such as Herbadrop (EUDAT, 2017) that they provide a good test of the technology. Furthermore, herbaria have been among the first to mass image their collections, so there are a vast number of specimen images available for testing.

3.2 Digitisation Workflow

We now outline a potential digitisation workflow, which is designed to process specimens in order to extract targeted data from them (see Figure 2). Starting with the original specimen, it is initially converted into a digital image. Though a digital object itself, the image does not immediately contain digitised text. In other words, though readable by humans, the image of the text is not yet searchable. The role of OCR is to convert text images into searchable text documents. To make these text documents searchable by the type of information they contain, another layer of information (metadata) is required on top of the original text. This step requires deeper analysis of the textual content, which is performed using NLP including language identification, NER and terminology extraction. The role of language identification here is twofold. If the labels are to be transcribed manually, then language identification can help us direct transcription tasks to the transcribers with suitable language skills. Similarly, if the labels were to be processed automatically, then the choice of tools will also depend on the given language. NER will support further structuring of the text by interpreting relevant portions of the text, such as those referring to people and locations. In addition to the extracted data and the associated metadata, the digitised collection should also incorporate a terminology that facilitates the interpretation of the scientific content described in the specimens. Many specimen labels contain either obscure or outdated terminology. Therefore, standard terminologies need to be supplemented by terminology extracted from the specimens. Finally, the performance of both OCR and NLP can be improved by restricting their view only to the labels on the specimen. This can be achieved by segmenting images prior to processing to identifying the areas of the image that relate to individual labels. However, there are trade-offs between the time it takes to segment images compared to the improved performance of OCR and NLP. In a production environment processing time is limited because of the need to ingest images into storage from a production line through a pipeline that includes quality control, the creation of image derivatives and image processing.

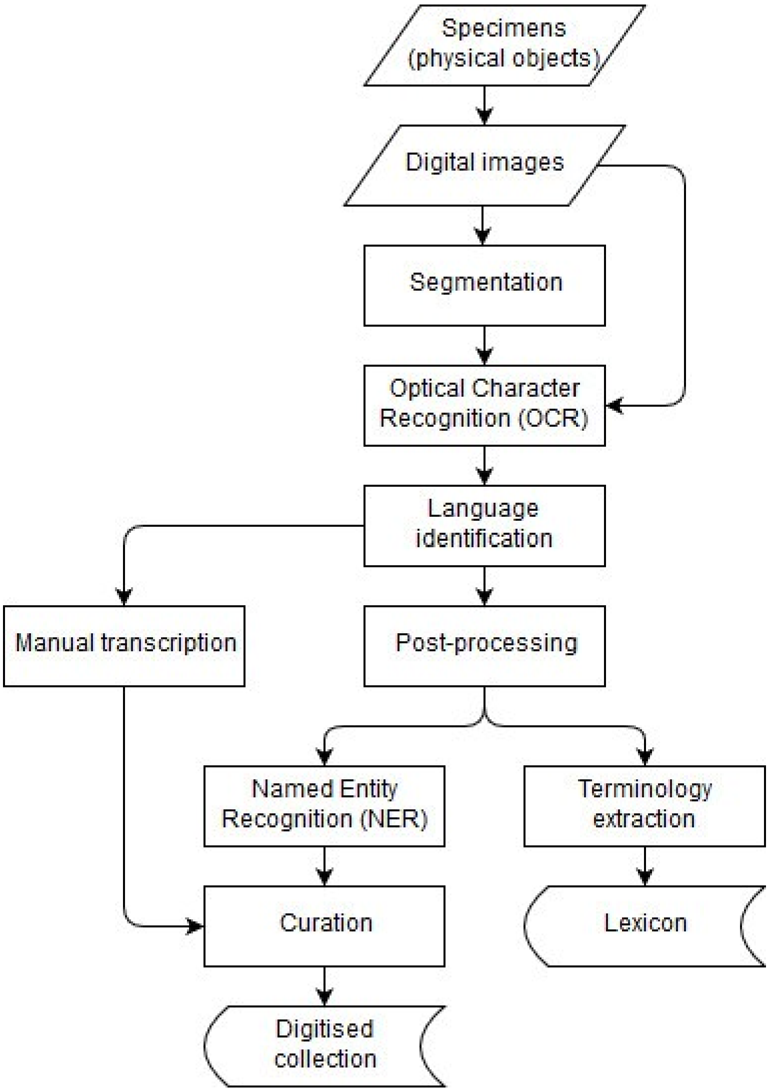


Figure 2: A possible semi-automatic digitisation workflow to extract data from the labels of collection specimens.

To help determine the subsequent steps in the pipeline it may be necessary to establish the language of the text recognised in the OCR step. This next step may be the deployment of language-specific NLP tools for identifying useful information in the target specimen. Or it may be the channelling of the text for manual transcription. A number of software solutions exist for performing language identification and are explored in section 5.3.

An approach to automatic identification of data from OCR recognised text might include Named Entity Recognition. This is an NLP task that identifies categories of information such as people and places. This approach may be suitable for finding a specimen's collector and collection country from text. Section 5.4 investigates this possibility using an NER tool.

4. Data

4.1 Data Collection

As noted above there is a large body of digitised herbarium specimens available for experimentation. A herbarium is a collection of pressed plant specimens and associated data (see Item 1 of Figure 1 for an example). As indicated in Figure 2, the first step in digitisation of these specimens is to produce a digital image. This requires physical manipulation of specimens, which is beyond the scope of the present task. Instead of gaining access to the original specimens, we collected their images in JPEG format from the partner institutions (Dillen et al., 2019). The choice of images sampled from these collections was based on the requirement to test OCR on a representative sample of the specimens in terms of their temporal and spatial coverage, because the age and origin of specimens may present different OCR challenges. For example, specimens can include printed, typed and/or handwritten labels, which may be partially obscured or have different orientations.

Each partner herbarium contributed 200 images containing a geographical and temporal cross-section of nomenclatural type and non-type herbarium specimens (see Figure 3), where a type specimen is the one used to name a newly identified species.

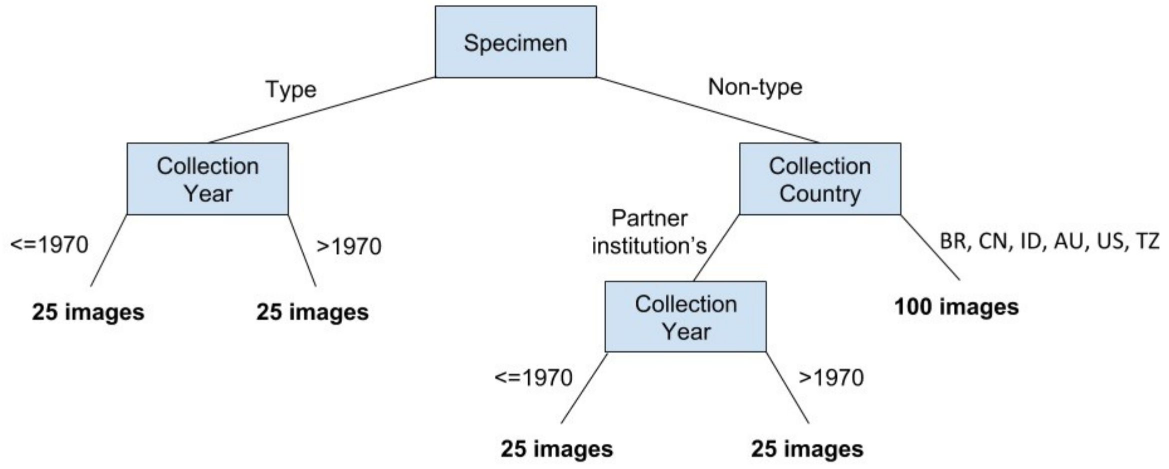


Figure 3: The criteria used by partner institutions to compile a test set of herbarium specimens. We did not attempt global coverage, but instead aimed at a representative sample from BR=Brazil, CN=China, ID=Indonesia, AU=Australasia, US=United States of America and TZ=Tanzania.

A total of nine herbaria, described in Table 11, contributed 200 specimen images each to form a dataset used in this study, giving a total of 1,800 images.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

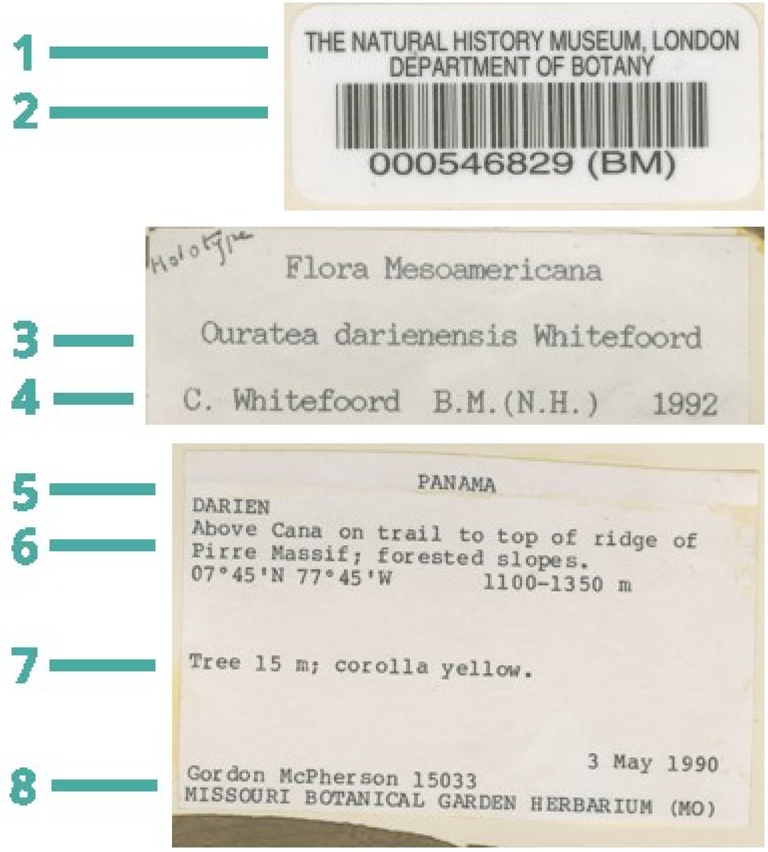
4.2 Data Properties

To illustrate the textual content of these images and better understand the challenges posed to the OCR, Figure 4 provides an example of labels attached to a specimen shown in Item 1 of Figure 1. In general, the labels can contain the following information:

1. **Title**: Organisation that owns the specimen.
2. **Barcode**: The specimen's machine readable identifier.
3. **Species name**: Scientific or common name of the species.
4. **Determined by and date**: The person who identified the specimen and the date of identification.
5. **Locality**: The geographical location where the specimen was collected.
6. **Habitat and altitude**: The habitat in which the specimen was collected and its altitude.
7. **Notes**: Additional notes written by the collector, often related to the characters of the species.
8. **Collector name, specimen number and collection date**: The name of the person(s) who collected the specimen, identifier they used to record and manage specimens and the date that the specimen was collected.

The above list is non-exhaustive and more or less information may be recorded by the collector or determiner.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483



THE NATURAL HISTORY MUSEUM, LONDON
DEPARTMENT OF BOTANY
000546829 (BM)

Holotype

Flora Mesoamericana

Ouratea darienensis Whitefoord

C. Whitefoord B.M.(N.H.) 1992

PANAMA
DARIEN
Above Cana on trail to top of ridge of
Pirre Massif; forested slopes.
07°45'N 77°45'W 1100-1350 m

Tree 15 m; corolla yellow.

3 May 1990
Gordon McPherson 15033
MISSOURI BOTANICAL GARDEN HERBARIUM (MO)

Figure 4: An example of specimen labels. 1=Title, 2=Barcode, 3=Species name, 4=Determined by and date, 5=Locality, 6=Habitat and altitude, 7=Notes, 8=Collector name, species number and collection date.

The properties of textual content of the given herbarium have been extrapolated from a random sample of 10 specimen per institution (see Table 1).

Contributor	Words per specimen	Handwritten content
BR	47	49.0%
H	77	21.3%
P	45	42.3%
L	64	22.0%
BM	59	32.8%
B	61	50.1%
E	54	68.0%
K	79	17.8%
TU	26	62.2%
Average	**57**	**40.6%**

Table 1: A summary of specimen properties. The Names and Index Herbariorum codes for the contributing herbaria are listed in Table A1

A subset of 250 images with labels written in English has been selected to test the performance of image segmentation and its effects on OCR and NER. For the purposes of these tests these images were manually divided into a total of 1,837 label segments, which were then processed separately. The images have also been used for testing and refining semantic segmentation methods as part of image quality control practices (Nieva et. al., 2019). The sematic segmentation methods tested support the creation of a segmentation service which could be later incorporated into automated processing workflows.

The segments effectively separate labels, barcodes and colour charts (see Figure 5 for examples). Item 1 is a label containing the species name, the collection location and the collector's name. Some of the information is printed while other is handwritten. In contrast, the label shown as Item 2 contains printed text only. However, its vertical orientation may cause additional difficulties. The label seen in Item 3 contains printed text that states the organisation that owns the specimen together with a barcode that identifies the specimen locally. However, the barcode stripes can sometimes be misinterpreted as text by overzealous OCR. A colour chart, such as the one shown in Item 4, contains no text, so it does not need to be processed further. Finally, Item 5 presents a ruler, which is accompanied by text that is not specific to the specimen and therefore, does not need to be considered. A machine learning classifier can be trained on segmented images to differentiate between different classes of labels in order to triage them ahead of the subsequent steps in the digitisation workflow.

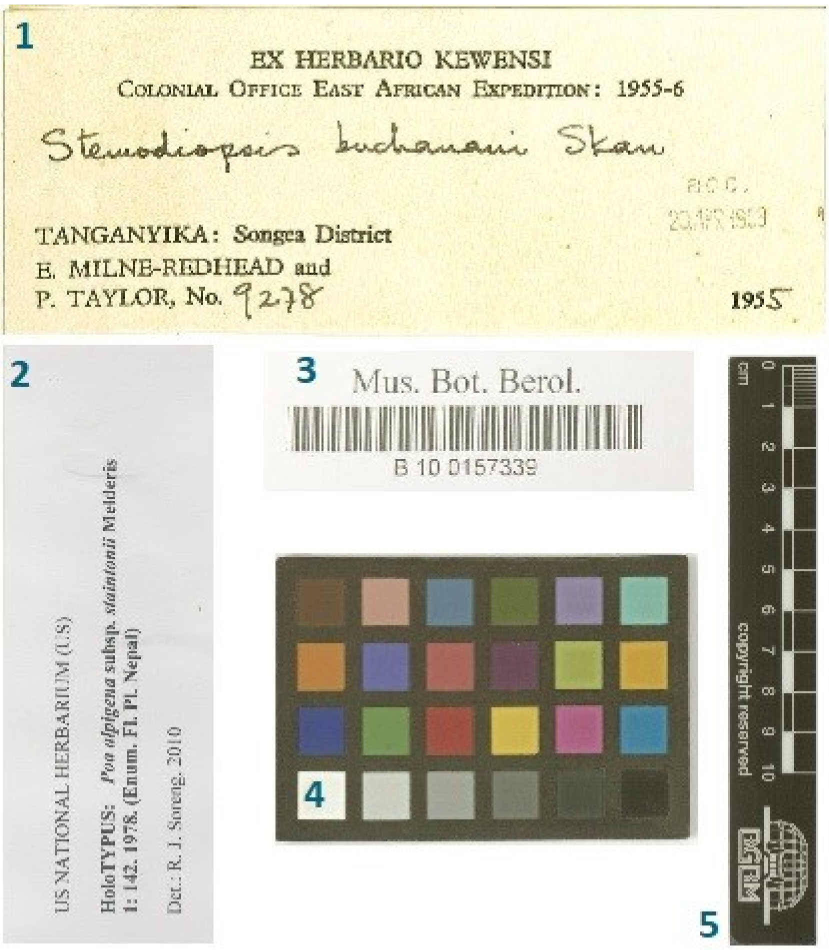

1
EX HERBARIO KEWENSI
COLONIAL OFFICE EAST AFRICAN EXPEDITION: 1955-6
Stemodiopsis buchananii Skan
TANGANYIKA: Songea District
E. MILNE-REDHEAD and
P. TAYLOR, No. 9278
1955
2
US NATIONAL HERBARIUM (US)
HoloTYPUS: Poa alpigena subsp. staintonii Melderis
1: 142. 1978. (Enum. Fl. Pl. Nepal)
Det.: R. J. Soreng. 2010
3
Mus. Bot. Berol.
B 10 0157339
4
cm 0 1 2 3 4 5 6 7 8 9 10
copyright reserved
BGBM
5

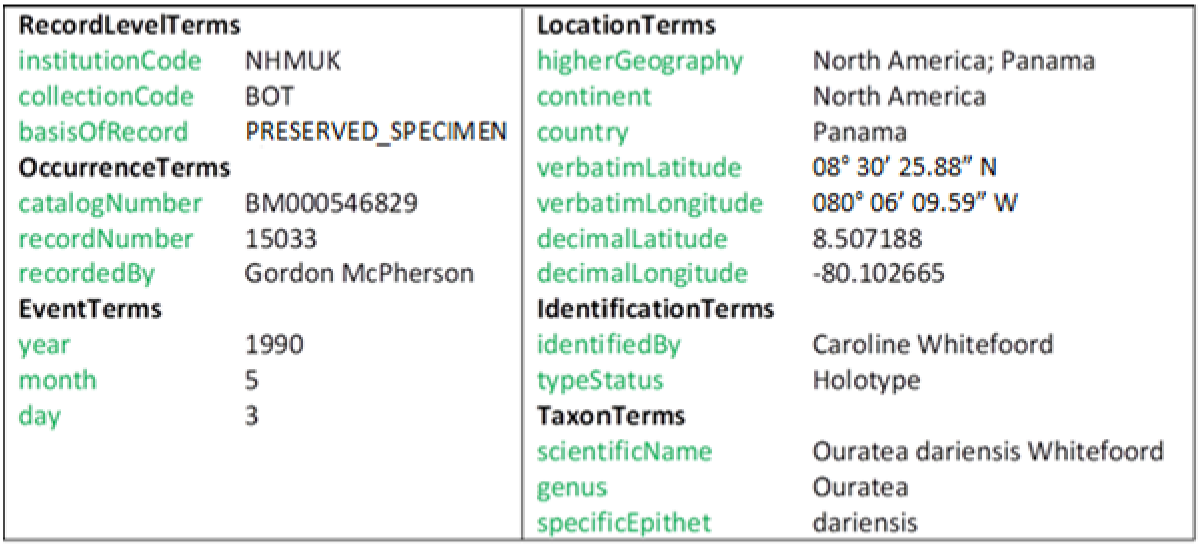
Figure 5: An impression of the different challenges presented by specimen image segments. 1=A label with both printed and handwritten text, 2=A printed label placed non-horizontally, 3=A barcode composed of unhelpful characters, 4=A colour chart containing no text, 5=A ruler containing no useful text.

4.3 Metadata

The role of OCR is to convert text image into searchable text. To make this text searchable by the type of information they contain, another layer of information (metadata) is required on top of the original text. The term *metadata* simply means data about data (Weibel, 1997). We can differentiate between three types of metadata (Riley, 2017):

1. *Descriptive* metadata facilitate searching using descriptors that qualify their content. For example, digitised specimens can be accessed by a species name, its collection location, collector, etc.
2. *Structural* metadata describe how the components of the data object are organised thereby facilitating navigation through its content. For example, labelling each segment of a digitised specimen by its type (see Figure 5 for examples) can facilitate their management, e.g. colour chart, ruler, barcode, collector's label, determination.
3. *Administrative* metadata convey technical information that can be used to manage data objects, e.g. time of creation, digital format, software used, etc.

While metadata can take many forms, it is important to comply with a common standard to improve accessibility to the data. Darwin Core (Wieczorek et al., 2012) is one such standard maintained by the Darwin Core Maintenance Group of the Biodiversity Information Standards organisation (TDWG). It includes a glossary of terms intended to facilitate the sharing of information on biological diversity by providing global identifiers, labels and definitions. Darwin Core is primarily based on taxa, their occurrence in nature as documented by observations, specimens, samples and related information. Figure 6 shows how the text content of the specimen shown in Figure 4 could be structured using Darwin Core standard, version 2014 (Darwin Core Maintenance Group, 2014; https://dwc.tdwg.org/). Once structured, the data can be stored in a database allowing for complex queries and efficient retrieval. For example, the geographic coordinates can be used to retrieve data referring to specimens collected within a given radius, further restricted by a time period, etc.



RecordLevelTerms		**LocationTerms**	
institutionCode	NHMUK	higherGeography	North America; Panama
collectionCode	BOT	continent	North America
basisOfRecord	PRESERVED_SPECIMEN	country	Panama
OccurrenceTerms		verbatimLatitude	08° 30’ 25.88” N
catalogNumber	BM000546829	verbatimLongitude	080° 06’ 09.59” W
recordNumber	15033	decimalLatitude	8.507188
recordedBy	Gordon McPherson	decimalLongitude	-80.102665
EventTerms		**IdentificationTerms**	
year	1990	identifiedBy	Caroline Whitefoord
month	5	typeStatus	Holotype
day	3	**TaxonTerms**	
		scientificName	Ouratea dariensis Whitefoord
		genus	Ouratea
		specificEpithet	dariensis

Figure 6: An example of an instantiated Darwin Core record instantiated

The problem of populating a predefined template such as the one defined by Darwin Core with information found in free text is an area of NLP known as information extraction (IE). The complexity of the template usually requires a bespoke IE system to be developed, which is beyond the scope of this feasibility study. Therefore, we will be focusing on information that could be extracted using NER, a subtask of IE, which can be supported using off-the-shelf software. Here, we focus on two commonly supported named entities, namely location and person names. Specifically, in the context of Darwin Core, we aim to automatically extract a specimen’s country and collector name, which have been associated with an increase of over 50% in the speed of semi-automatic digitisation (Drinkwater et al., 2014).

5. Digitisation Experiments

This section describes a selection of software tools that can be used to automate the steps of the digitisation workflow shown in Figure 2 together with the test results obtained using the data described in section 4.

5.1 Optical Character Recognition

OCR is a technology that allows the automatic recognition of characters through an optical mechanism or computer software (Mori et al., 1999). OCR can be used to convert image-borne characters to text documents, which are machine readable in the sense that the text can then be indexed, searched, edited or processed by NLP software.

We tested three off-the-shelf OCR software tools, described in Table 2. Tesseract is the most accurate open-source OCR software whose development is sponsored by Google (Google Open Source, 2018). It has the ability to recognise more than 100 languages out of the box . We originally considered version 3.0.51, but later extended our experiments to version 4.0.0, which was released in the meantime and was reported to offer significantly higher accuracy than than its earlier version (Ooms, 2018). The software development kit ABBYY FineReader Engine 12.0 allows software developers to integrate OCR functionality into their applications to extract textual information from paper documents, images or displays (ABBYY, 2018).

Microsoft's OneNote is a note taking and management application for collecting, organising and sharing digital information (Microsoft Corporation, 2018). It contains native OCR functionality whose performance had not been evaluated in another recent investigation into automating data capture from natural history specimens (Haston et al., 2015). Unlike Tesseract and ABBYY FineReader Engine, OneNote is a stand-alone software application whose OCR functionality cannot readily be integrated into other software.

	Founded year	Latest stable version	License	Windows	Macintosh	Linux
Tesseract	1985	4.0.0	Apache	Windows 10	Mac OS X 10.14.x	Ubuntu 18.04, 18.10
ABBYY FineReader Engine	1989	12.0	Proprietary	Windows 10, 8.1, 8, 7-SP1	Mac OS X 10.12.x, 10.13.x	Ubuntu 17.10, 16.04.1, 14.04.5
Microsoft OneNote	2012	17.10325.20049	Proprietary	Windows 10, 8.1	Mac OS X, 10.12 or later	Ubuntu 18.04, 18.10

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

Table 2: Comparison of OCR software libraries and applications

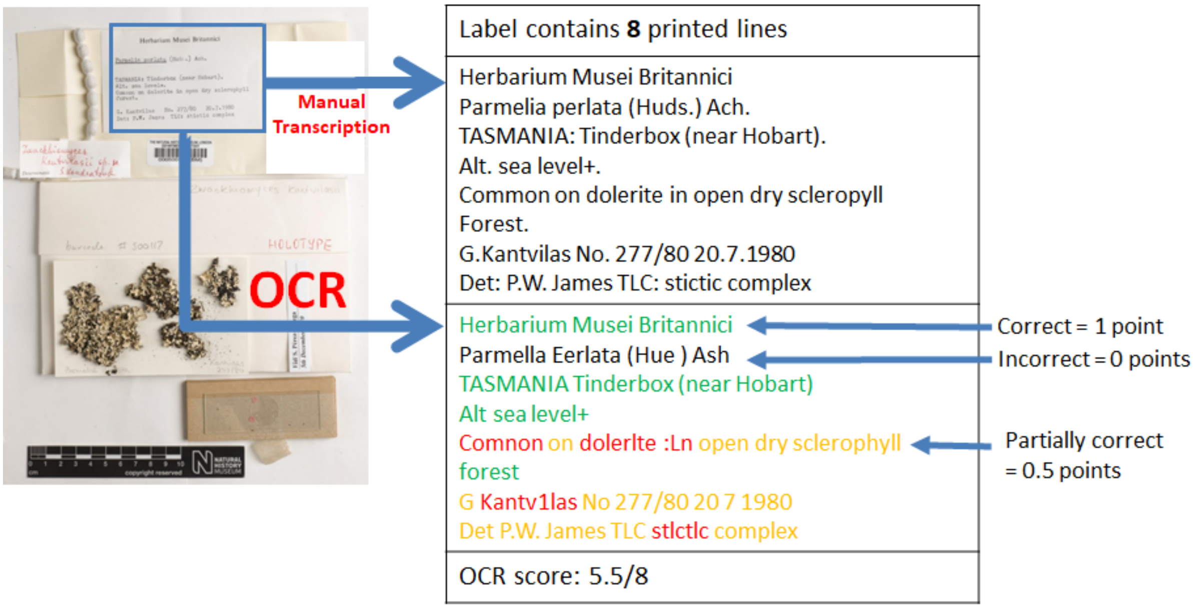
To evaluate the OCR performance of the aforementioned software tools, we ran two sets of experiments, one against the whole digital images of specimens and the other against the segmented images with an expectation that the latter would result in shorter processing time and higher accuracy. Indeed, the results shown in Table 3 demonstrate that the processing time was reduced by 49% on average when images were segmented prior to undergoing OCR. Out of the three batch processing software tools considered, Tesseract 3.0.51 was the fastest in both scenarios. All experiments were performed using the following configuration: a desktop computer containing an Intel i5-4590T 2.00GHz 4 Core CPU, 8.00 GB RAM and Microsoft Windows 10 Education Version 10.0.17134.

	Processing Time (h:m:s)			
	250 whole images	1,837 segments	Difference	Difference (Percentage % saving)
Tesseract 4.0.0	01:06:05	00:45:02	-00:21:03	-31.9%
Tesseract 3.0.51	00:50:02	00:23:17	-00:26:45	-53.5%
ABBYY FineReader Engine 12	01:18:15	00:29:24	-00:48:51	-62.4%

Table 3: Processing times for OCR programs using whole images and segments

The accuracy of OCR will be measured in terms of **line correctness** as described by Haston et al. (2015). To create a gold standard, the text from a digital image is manually transcribed verbatim and the number of original lines counted. The lines from the OCR output are then compared against the gold standard and classified into one of three classes: correct, partially (in)correct and incorrect and scored 1, 0.5 and 0, respectively (see Figure 7 for an example). The line scores are then aggregated into overall accuracy. This method considers only printed text and not handwritten text.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483



Label contains **8** printed lines

Herbarium Musei Britannici
Parmelia perlata (Huds.) Ach.
TASMANIA: Tinderbox (near Hobart).
Alt. sea level+.
Common on dolerite in open dry scleropyll
Forest.
G.Kantvilas No. 277/80 20.7.1980
Det: P.W. James TLC: stictic complex

Herbarium Musei Britannici
Parmella Eerlata (Hue) Ash
TASMANIA Tinderbox (near Hobart)
Alt sea level+
Comnon on dolerlte :Ln open dry sclerophyll
forest
G Kantv1las No 277/80 20 7 1980
Det P.W. James TLC stlctlc complex

OCR score: 5.5/8

Figure 7: Measuring OCR accuracy

Bearing in mind the time and effort involved in creating the gold standard, only a subset of the data available for testing was used to evaluate the correctness of the OCR. Five herbarium sheet images, their segments and manual transcriptions and OCR text used in these experiments can be found in appendix 9.2. A summary of results is given in Table 4.

	5 whole images Mean line correctness (%)	22 segments Mean line correctness (%)	Difference
Tesseract 4.0.0	72.8	75.2	+2.4
Tesseract 3.0.51	44.1	63.7	+19.6
ABBYY FineReader Engine 12	61.0	77.3	+16.3
Microsoft OneNote 2013	78.9	65.5	-13.4

Table 4: Line correctness for OCR using whole images and their segments

Apart from ABBYY FineReader Engine, all other tools recorded an accuracy around 70%, with Tesseract 4.0.0 proving to be the most robust with respect to image segmentation. Its performance could be improved by further experiments focusing on its configuration parameters.

5.2 Handwritten Text Recognition

Another separate test was conducted on the test dataset with Google's Cloud Vision v1 (Google Cloud, 2018), which can recognise both typed and handwritten text. Cloud Vision currently supports 56 languages, but language setting can be used to improve speed and accuracy of the text recognition. This is a paid service and has a limit of 20MB and 20M pixels per image.

The results from Cloud Vision were compared against baseline data known for that specimens and the output of ABBYY FineReader 11 using Levenshtein distance (Levenshtein, 1966). The Levenshtein distance measures the minimum difference between two strings by counting the number of insertions, deletions and substitutions needed to change one word into another. Note that this metric is not case sensitive. Every field from the test dataset was compared to the text obtained through OCR.

One must be cautious when comparing interpreted baseline data. For example, is the catalog number is “BM000521570”, where Cloud Vision OCR finds “000521570 (BM)”. Technically, the OCR engine has found the correct string, but because the baseline contains an interpreted value, it looks like the OCR is not correct. Another example is that the baseline data contains fields with abbreviations, such as country code. In the case of Australia, as long as “AU” is found anywhere in the OCR the Levenshtein distance for this field resolves as 0.

Segmentation is not available for this OCR engine, so the whole image was used with Cloud Vision and ABBYY FineReader to compare results. However, it is an option to perform segmentation before OCR through another script. Like other scriptable OCR engines, it is possible to obtain coordinates of where the result is located in the image.

Owing to the limitations outlined above, only specific fields were included for further analysis: catalogNumber, genus, specificEpithet, country, recordedBy, typeStatus, verbatimLocality, verbatimRecordedBy. Verbatim coordinates are likely too complex or too often interpreted to be comparable in this analysis. VerbatimEventDate was ignored because it is not technically verbatim: “3/8/59” on label, “1959-08-03” in database (Finnish Biodiversity Info Facility, 2018). Instead, Year was used while acknowledging the lesser relevance. Because of the minimal Levenshtein distance to achieve the correct year, all results for year with a distance of greater than 0 were ignored for further analysis. Please note that type status is not always present in the image as (printed) text but interpreted, however it was included because of its importance to taxonomy.

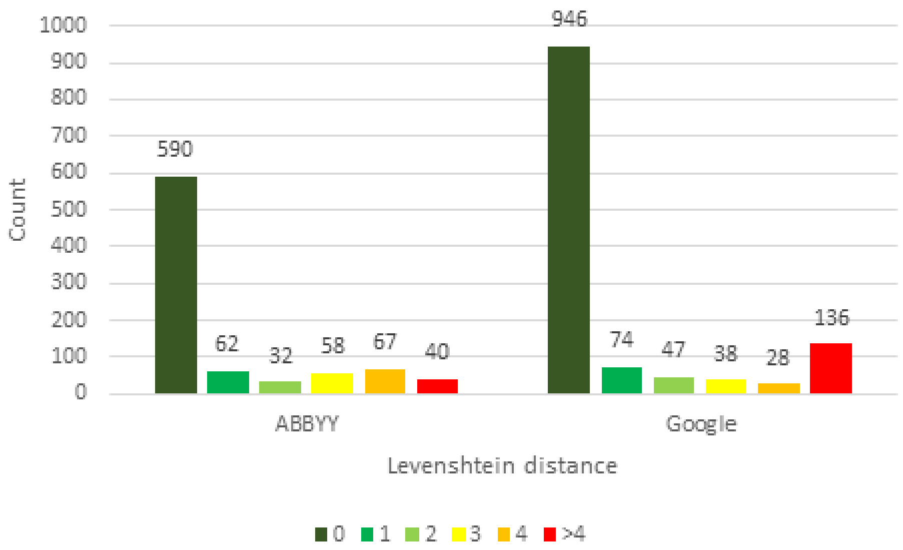
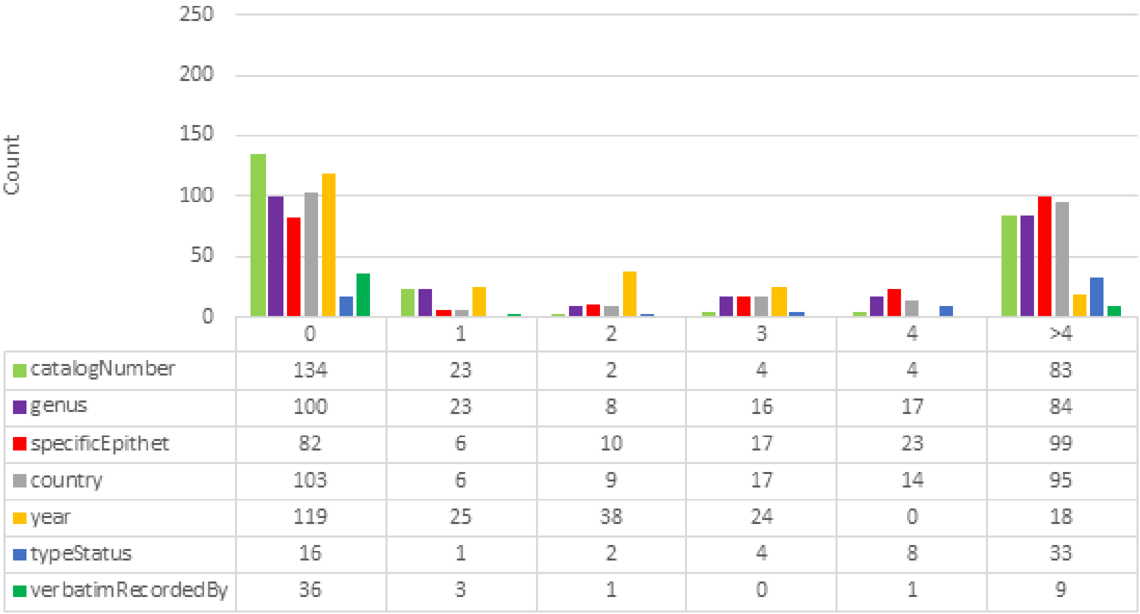


Figure 8: Comparison of Levenshtein distance scores for Google Cloud Vision OCR and ABBYY for selected fields, Levyear>0 excluded.

Figure 8 shows the count of Levenshtein distance scores for all selected fields combined, $\mathrm{Lev}_{\mathrm{year}}$>0 excluded. Google Cloud Vision OCR scores better, considering that the decrease of bad results with a distance of greater than 4 results in an increase of results with a distance of 4. The high number of results with a distance greater than 4 is partly caused by the interpretation of field, such as type status.

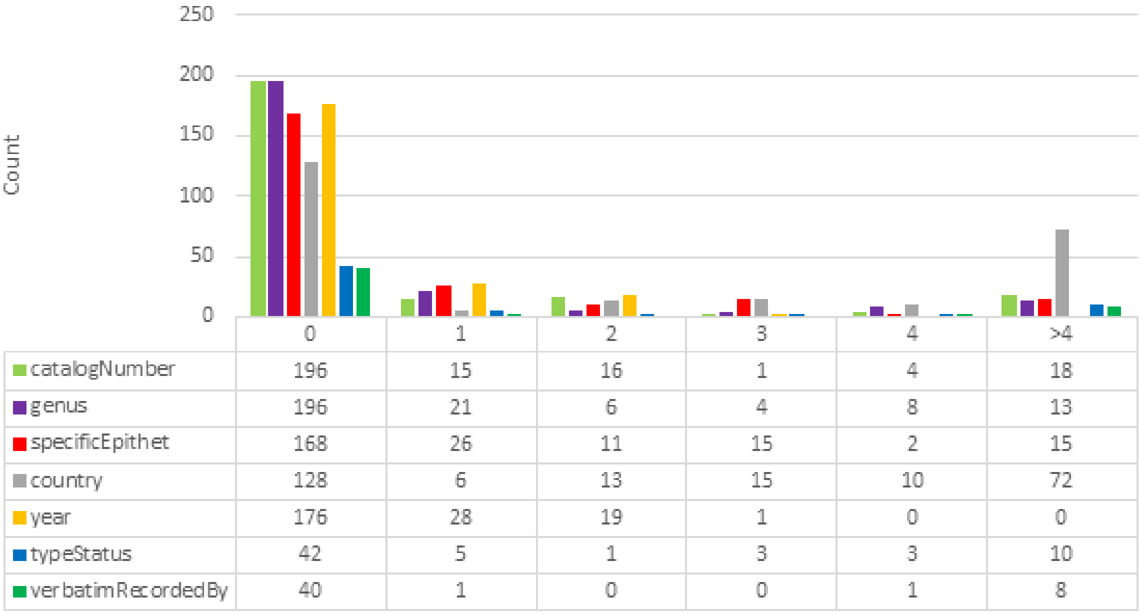
Comparing the results in Figure 8 it shows that the scores of Google Cloud Vision are higher for the 3 best distances. Comparing the results in Figure 9 and Figure 10 show that Google Cloud Vision has more results in the best category for each field, while ABBYY has a higher count of Lev≥4 for each field. Distances greater than 4 can be considered as low quality results. When these results are excluded as well as $\mathrm{Lev}_{\mathrm{year}}$>0, Google Cloud Vision OCR obtained 1133 results while ABBYY OCR obtained 809. When the results are weighted for accuracy (5 for distance=0, 1 for distance≥4, $\mathrm{Lev}_{\mathrm{year}}$>0 excluded) ABBYY’s weighted score is 4689 versus Google’s weighted score of 6540.

In conclusion, this comparative test indicates that the results from Google Cloud Vision OCR are of higher quality and even of higher quantity when the lowest category are excluded. This result demonstrates that handwriting recognition adds a considerable amount of data of high quality. Handwriting recognition should no longer by dismissed because it has already become a viable technique.



	0	1	2	3	4	>4
🟩 catalogNumber	134	23	2	4	4	83
🟪 genus	100	23	8	16	17	84
🟥 specificEpithet	82	6	10	17	23	99
◼ country	103	6	9	17	14	95
🟨 year	119	25	38	24	0	18
🟦 typeStatus	16	1	2	4	8	33
🟩 verbatimRecordedBy	36	3	1	0	1	9

Figure 9: Results per field from ABBYY.



	0	1	2	3	4	>4
🟩 catalogNumber	196	15	16	1	4	18
🟪 genus	196	21	6	4	8	13
🟥 specificEpithet	168	26	11	15	2	15
◼ country	128	6	13	15	10	72
🟨 year	176	28	19	1	0	0
🟦 typeStatus	42	5	1	3	3	10
🟩 verbatimRecordedBy	40	1	0	0	1	8

Figure 10: Results per field from Google Cloud Vision OCR.

5.3 Language Identification

Language identification is the task of determining the natural language that a document is written in. It is a key step in automatic processing of real-world data where a multitude of languages exist (Lui & Baldwin, 2012). Languages used on specimen labels can vary across a collection as can be seen in Figure 11. In the context of digitisation workflow knowing the languages that specimen labels are written in allows us to inform the subsequent steps including NLP as well as improving manual curation of the results by forwarding them to people with the required language skills.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

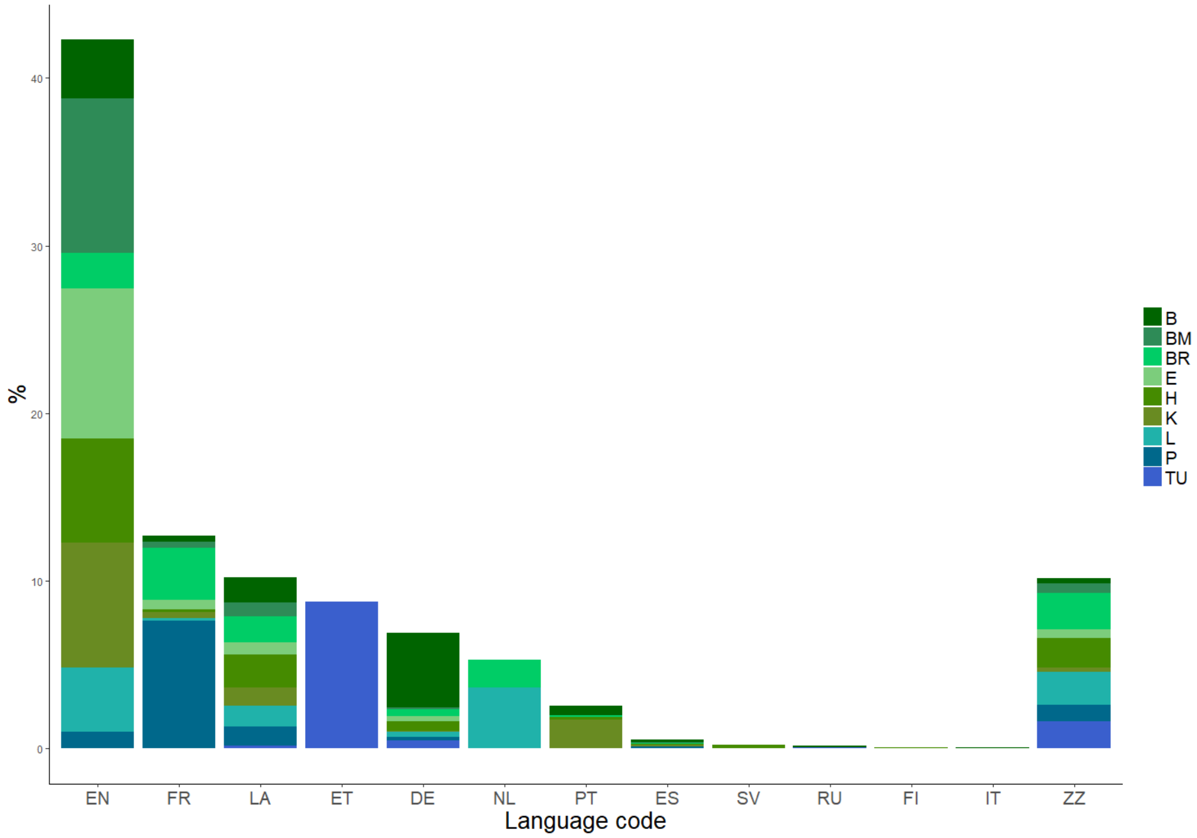


Figure 11: The distribution of languages across the specimen and herbaria. EN=English, FR=French, LA=Latin, ET=Estonian, DE=German, NL=Dutch, PT=Portuguese, ES=Spanish, SV=Swedish, RU=Russian, FI=Finnish, IT=Italian, ZZ=Unknown. The codes for the contributing herbaria are listed in Table 11 (from Dillen et al., 2019).

A number of off-the-shelf software libraries can be used to perform language identification (see Table 5). The given libraries can all be integrated into larger software applications.

Software	**Licence**	**Organisation**
langid.py	Open Source	University of Melbourne
langdetect	Apache License Version 2.0	N/A
language-detection	Apache License Version 2.0	Cybozu Labs, Inc.

Table 5: Language identification software tools and their properties

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

Input: “Unangwa Hill about 6 km. E. of Songea in crevices in vertical rock faces”
Output: English [99%]

Input: “Herbier de Jardin botanique de l'Etat”
Output: French [99%]

Input: “Tartu olikooli juures oleva loodusuurijate seltsi botaanika sekstsiooni”
Output: Estonian [99%]

Input: “Arbusto de ca. 2 m, média ramificação.”
Output: Portuguese [100%]

Table 6: Example of langid.py usage with fragments of OCR text. Output lines denote the language identified in the input text and the probability estimate for the language.

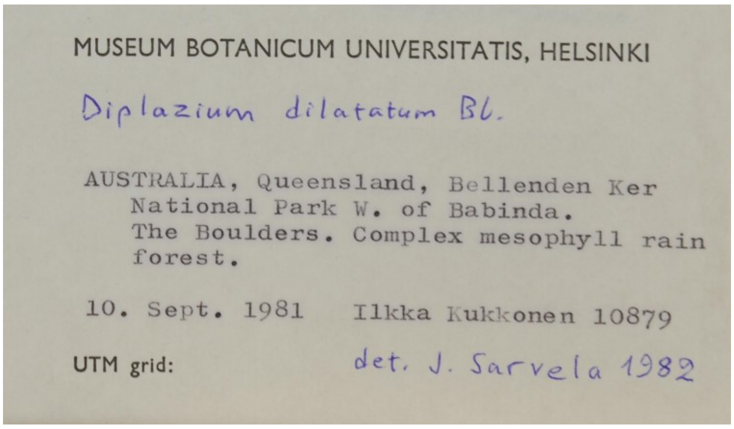
Table 6 provides output obtained by langid.py from a sample of our test data. The automatically identified language is quantified with a probability estimate. langid.py is able to identify 97 different languages without requiring any special configuration. It generally outperforms langdetect (Danilák, 2018) in terms of accuracy. langid.py is also reportedly the faster of the two (Lui & Baldwin, 2012). The corpus used in the evaluation contained government documents, online encyclopedia entries and software documentation (Lui & Baldwin, 2012; Baldwin & Lui, 2010).

The program language-detection (Shuyo, 2014) provides a third option for language detection. Unlike langid.py and langdetect, no evaluation of its performance appears to have been published. It advertises 99% precision over 53 languages although texts of 10 to 20 words are recommended to support accurate detection. This may prove problematic when used with short fragments of OCR text obtained from specimen images.

5.4 Named Entity Recognition

NER is commonly used in information extraction to identify text segments that refer to entities from predefined categories. The state-of-the-art approaches use conditional random fields trained on data manually labelled with these categories to learn automatically how to extract named entities from text. Traditionally, these categories include persons, organisations and locations. Therefore, pre-trained models for these categories are readily available, e.g. Stanford NER (The Stanford Natural Language Processing Group, 2018).

As mentioned earlier in section 4.3, in this study we are interested in two categories of named entities: country (subcategory of location) and collector (subcategory of person) Pre-trained NER software can only identify names of locations and persons, but cannot verify that a location is a country or that a person is a collector. Therefore, we will generalise our NER problem into that of recognising persons and locations in general and will accordingly measure the performance of Stanford NER on our dataset. A subset of specimen labels was manually transcribed and annotated with person and location labels to create a gold standard against which to evaluate Stanford NER (see Figure 12 and Figure 13 for an example).



MUSEUM BOTANICUM UNIVERSITATIS, HELSINKI

Diplazium dilatatum Bl.

AUSTRALIA, Queensland, Bellenden Ker National Park W. of Babinda. The Boulders. Complex mesophyll rain forest.

10. Sept. 1981 Ilkka Kukkonen 10879

UTM grid: det. J. Sarvela 1982

Figure 12: An example of a specimen label

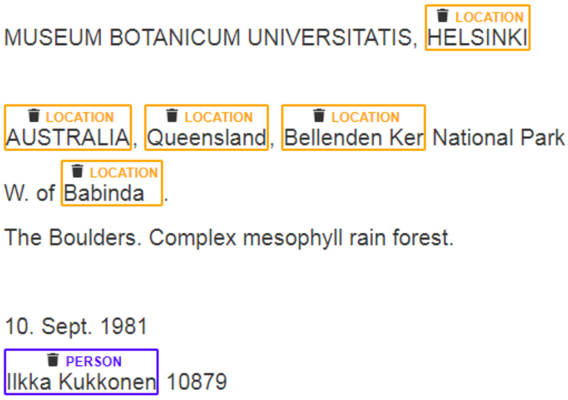
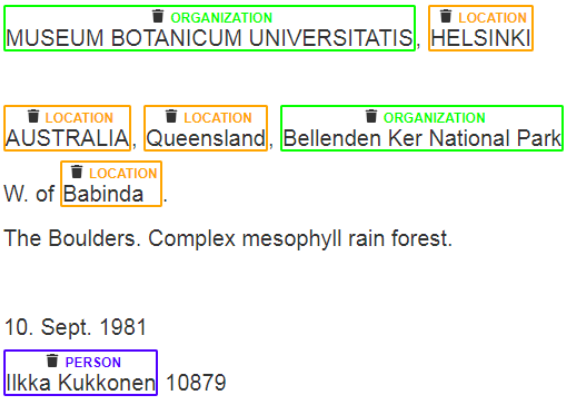
Gold standard	NER output
 MUSEUM BOTANICUM UNIVERSITATIS, HELSINKI [LOCATION] AUSTRALIA [LOCATION], Queensland [LOCATION], Bellenden Ker [LOCATION] National Park W. of Babinda [LOCATION]. The Boulders. Complex mesophyll rain forest. 10. Sept. 1981 Ilkka Kukkonen [PERSON] 10879	 MUSEUM BOTANICUM UNIVERSITATIS [ORGANIZATION], HELSINKI [LOCATION] AUSTRALIA [LOCATION], Queensland [LOCATION], Bellenden Ker National Park [ORGANIZATION] W. of Babinda [LOCATION]. The Boulders. Complex mesophyll rain forest. 10. Sept. 1981 Ilkka Kukkonen [PERSON] 10879

Figure 13: Gold standard vs. NER output

According to Jiang et al. (2016) a named entity is recognised correctly if either of the following criteria are met:

1. Both boundaries of a named entity and its type match. For example, the segment “Ilkka Kukkonen” in Figure 13 is recognised fully and correctly as a person.
2. Two text segments overlap partially and match on the type.

Either way, the NER results are usually evaluated using the three most commonly used measures in NLP, precision, recall and F1 score. **Precision** is the fraction of automatically recognised entities that are also correct. **Recall** is the fraction of manually annotated named entities that were successfully recognised by the NER system. **F1 score** is a measure that combines precision and recall - it is the harmonic mean of the two.

Table 7 and Figure 14 show how these might be calculated. An example follows that explains the terms used.

		Predicted (NER)	
		Negative	**Positive**
Actual (gold standard)	**Negative**	True Negative	False Positive
	Positive	False Negative	True Positive

Table 7: Confusion matrix for predicted and actual labels

$$Precision = \frac{True\ Positive}{True\ Positive\ +\ False\ Positive}$$

$$Recall = \frac{True\ Positive}{True\ Positive\ +\ False\ Negative}$$

$$F1\ Score\ = 2 * \frac{Precision\ *\ Recall}{Precision\ +\ Recall}$$

Figure 14: Formulae for Precision, Recall and F1 Score

To evaluate the performance of NER on our dataset, we selected a subset of five herbarium sheet images and their segments, which are to be found in Appendix 9.3. These are the same images and segments used to calculate line correctness in section 5.1. The OCR output used is that obtained using Tesseract 4.0.0.

Table 8 and Table 9 show the results of Stanford NER performance.

Measure/Entity	PERSON	LOCATION	Overall
Precision	0.81	0.38	0.69
Recall	0.71	0.21	0.53
F1	0.76	0.27	0.60

Table 8: NER performance on OCR text from whole images

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

Measure/Entity	PERSON	LOCATION	Overall
Precision	0.85	0.43	0.74
Recall	0.74	0.50	0.69
F1	0.79	0.46	0.71

Table 9: NER performance on OCR text from image segments

An improvement across all measures can be observed when using OCR text from segmented images. This is consistent with the increased line correctness observed described in section 5.1.

5.5 Terminology Extraction

To improve the accessibility of a specimen collection, its content needs to be not only digitised but also organised in alphabetical or any other systematic order. This is naturally expected to be done by the species name. The problem with old specimens is that the content of their labels is not likely to comply with today's standards. Therefore, matching them against existing taxonomies will fail to recognise non-standard terminology. To automatically extract species names together with other relevant terminology, we propose an unsupervised data-driven approach to terminology extraction. FlexiTerm is a method developed in-house at Cardiff University. It has been designed to automatically extract multi-word terms from a domain-specific corpus of text documents (Spasić et al., 2013; Spasić, 2018).

OCR text extracted from specimens in a given herbarium fits a description of a domain-specific corpus, therefore FlexiTerm can exploit linguistic and statistical patterns of language use within a specific herbarium to automatically extract relevant terminology. Appendix 9.6 shows the multi-word terms extracted from the text recognised using Tesseract 4.0.0 on the segmented images. The result show that the majority of extracted terminology refers to organisations (herbaria) that host the specimens, e.g. “Royal Botanic Gardens Edinburgh” or “Nationaal Herbarium Nederland”. There are also mentions of collectors, e.g. “Ilkka Kukkonen” that was also recognised as person by NER. In that respect, there is some overlap between NER and terminology extraction. Regardless of their type, the multi-word terms extracted by FlexiTerm will represent the longest repetitive phrases found in a collection. Therefore, their recognition can facilitate transcription or curation of a digital collection should these activities be crowdsourced.

6. Putting It All Together

Many scientific disciplines are increasingly data driven and new scientific knowledge is often gained by scientists putting together data analysis and knowledge discovery “pipelines” (Ludäscher et al., 2006). These “pipelines” are known as scientific workflows. Interpreting data and attaching meaning to it creates information. Interpreting information in the context of prior knowledge, experience and wisdom can lead to new knowledge.

A scientific workflow consists of a series of analytical steps. These can involve data discovery and access, data analysis, modelling/simulation and data mining. Steps can be computationally intensive and therefore are often carried out on high-performance computing clusters. Herbadrop, a pilot study of specimen digitisation using OCR, demonstrated successful use of high performance digital workflows (EUDAT, 2017). In this section, we review workflow management systems that can be used to automate the workflow presented in Figure 1.

The tools that allow scientists to compose and execute scientific workflows are generally known as workflow management systems, of which Apache Taverna and Kepler are among the most well-known and best established examples.

Apache Taverna is open-source and domain-independent (The Apache Software Foundation, 2018). It is designed for use in any scientific discipline and is supported by a large community of users.

Taverna has been successfully deployed within the domain of biodiversity via BioVeL - a virtual laboratory for data analysis and modelling in biodiversity (Hardisty et al., 2016). BioVeL allows the building of workflows through the selection of a series of data processing services and can process large volumes of data when the services needed to do that are distributed among multiple service providers.

Taverna supports BioVeL users by allowing them to create workflows via a visual interface as opposed to writing code. Users are presented with a selection of processing steps and can “drag and drop” them to create a workflow. They can then test the workflow by running it on their desktop machine before deploying it to more powerful computing resources.

Kepler is a scientific workflow application also designed for creating, executing and sharing analyses across a broad range of scientific disciplines (The Kepler Project, 2018). Application areas include bioinformatics, particle physics and ecology.

Like Taverna, Kepler provides a graphical user interface to aid in the selection of analytical components to form scientific workflows (Barseghian et al., 2010). It also offers data provenance features that allows users to examine workflow output in detail for diagnostic purposes (Liew et al., 2017). This supports the reliability and reproducibility of evidence from data, which is necessary for the presentation of conclusions in research publications.

Tools like Apache Taverna and Kepler can be used for creating workflows for OCR, NER and IE, like that depicted in Figure 1. When managed and executed in virtual research environments such as BioVeL, the data and results can be collated, managed and shared appropriately. Such workflows can be run repeatedly, reliably and efficiently with the possibility to process many tens of thousands of label images in parallel within a single workflow run.

7. Conclusions

We designed a modular approach for automated text digitisation with respect to specimen labels (see Figure 1). To minimise implementation overhead, we proposed implementing this approach as a scientific workflow using off-the-shelf software to support individual components. An additional advantage of this approach is an opportunity to run the workflow in a distributed environment, thus supporting large-scale digitisation as well as an optimal use of resources across multiple institutions. Based on the local experience and expertise associated with both development and applications, we recommend the use of Apache Taverna for implementing and executing the workflow. We evaluated off-the-shelf software that can support specific modules within the workflow. Our recommendations are summarised in Table 10. Further research is needed with respect to image segmentation, which has shown to have significant effect on the performance across all tasks listed in Table 10.

Task	Software	Comment
Optical character recognition	Tesseract 4.0.0	robust with respect to segmentation
Handwritten text recognition	Cloud Vision	supports 56 languages
Language identification	langid.py	supports 97 languages
Named entity recognition	Stanford NER	a wide variety of entities including location, organisation, date, time, person
Terminology extraction	FlexiTerm	robust with respect to orthographic variation (e.g. introduced by OCR)

Table 10: A summary of recommendations

8. References

ABBYY (2018) *AI-powered OCR SDK for Windows, Linux & Mac OS | ABBYY OCR API* [Online]. Available at: https://www.abbyy.com/en-gb/ocr-sdk/ [Accessed: 21st November 2018].

Baldwin, T. and Lui, M. (2010) Language identification: The long and the short of the matter. In *Proceedings of NACL HLT 2010*, pages 229–237, Los Angeles, USA.

Barseghian, D., Altintas, I., Jones, M., Crawl, D., Potter, N., Gallagher, J., Cornillon, P., Schildhauer, M., Borer, E., Seabloom, E. and Hosseini, P. (2010) Workflows and extensions to the Kepler scientific workflow system to support environmental sensor data access and analysis. *Ecological Informatics*, 5(1): 42–50. https://doi.org/10.1016/j.ecoinf.2009.08.008

Danilák, M. (2018) *langdetect* [Online]. GitHub. Available at: https://github.com/Mimino666/langdetect [Accessed: 31st October 2018].

Darwin Core Maintenance Group, Biodiversity Information Standards (TDWG) 2014. *Darwin Core*. Zenodo. Available at: https://doi.org/10.5281/zenodo.592792 [Accessed: 24th November 2018].

Dillen, M., Groom, Q., Chagnoux, S., Güntsch, A., Hardisty, A., Haston, E., Livermore, L., Runnel, V., Schulman, L., Willemse, L., Wu, Z. and Phillips, S. (2019) A benchmark dataset of herbarium specimen images with label data. *Biodiversity Data Journal*. In press.

Drinkwater, R., Cubey, R. and Haston, E. (2014) The use of Optical Character Recognition (OCR) in the digitisation of herbarium specimen labels. *PhytoKeys*, 38: 15–30. https://dx.doi.org/10.3897%2Fphytokeys.38.7168

Ellwood, E.R., Kimberly, P., Guralnick, R., Flemons, P. Love, K., Ellis, S., Allen, J.M. et al. (2018) Worldwide Engagement for Digitising Biocollections (WeDigBio): The Biocollections Community's Citizen-Science Space on the Calendar. *BioScience*, 68(2): 112–124. https://doi.org/10.1093/biosci/bix143

Engledow, H., De Smedt, S., Bogaerts, A. and Groom, Q. (2018) An Evaluation of In-house versus Out-sourced Data Capture at the Meise Botanic Garden (BR). *Biodiversity Information Science and Standards* 2: e26514. https://doi.org/10.3897/biss.2.26514

EUDAT (2017) *EUDAT & Herbadrop Collaboration* [Online]. Available at: https://www.eudat.eu/eudat-herbadrop-collaboration [Accessed: 8th October 2018].

Finnish Biodiversity Info Facility (2018) *Suomen Lajitietokeskus* [Online]. Available at: http://id.luomus.fi/EIG.6494 [Accessed: 22nd December 2018].

Google Cloud (2018) *Detect Text (OCR)* [Online]. Available at: https://cloud.google.com/vision/docs/ocr [Accessed: 22nd December 2018].

Google Open Source (2018) *Tesseract OCR* [Online]. Available at: https://opensource.google.com/projects/tesseract [Accessed: 22nd October 2018].

Hardisty, A.R., Bacall, F., Beard, N., Balcázar-Vargas, M. P., Balech, B., Barcza, Z., et al. (2016). BioVeL: a virtual laboratory for data analysis and modelling in biodiversity science and ecology. *BMC Ecology*, 16(1): 49. https://doi.org/10.1186/s12898-016-0103-y

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. and Vacek, F. (2015) *D4.2 - Automating data capture from natural history specimens | SYNTHESYS3* [Online]. Available at: http://synthesys3.myspecies.info/node/695 [Accessed: 21st October 2018].

Hoehndorf, R., Alshahrani, M., Gkoutos, G. V., Gosline, G., Groom, Q., Hamann, T., Kattge, J., Mota de Oliveira, S., Schmidt, M., Sierra, S., Smets, E., Vos, R.A. and Weiland, C. (2016) The flora phenotype ontology (FLOPO): tool for integrating morphological traits and phenotypes of vascular plants. *Journal of Biomedical Semantics*, 7(1), 65. https://doi.org/10.1186/s13326-016-0107-8

Jiang, R., Banchs, R.E. and Li, H., 2016. Evaluating and combining name entity recognition systems. In *Proceedings of the Sixth Named Entity Workshop* (pp. 21–27).

Levenshtein, V.I. (1966) Binary codes capable of correcting deletions, insertions and reversals. *Soviet physics doklady*, 10(8): 707–710.

Liew, C.S., Atkinson, M.P., Galea, M., Ang, T.F., Martin, P. and van Hemert, J. (2016) Scientific Workflows: Moving Across Paradigms. *ACM Computing Surveys*, 49(4): 1–39. http://dx.doi.org/10.1145/3012429

Ludäscher, B. , Altintas, I. , Berkley, C. , Higgins, D. , Jaeger, E. , Jones, M. , Lee, E. A., Tao, J. and Zhao, Y. (2006) Scientific workflow management and the Kepler system. Concurrency Computation: *Practice and Experience*, 18: 1039–1065. https://doi.org/10.1002/cpe.994

Lui, M. and Baldwin, T. (2012) langid. py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 system demonstrations* (pp. 25-30). Association for Computational Linguistics.

Microsoft Corporation (2018) *Microsoft OneNote* [Online]. Available at: http://www.onenote.com/?404&public=1 [Accessed: 22nd November 2018].

Mori, S., Nishida, H. and Yamada, H. (1999) *Optical character recognition*. John Wiley & Sons, Inc.

Natural History Museum London (2018) *Data Portal* [Online]. Available at: http://data.nhm.ac.uk/ [Accessed: 22nd December 2018].

Riley, J. (2017) Understanding metadata. *National Information Standards Organization* https://www.niso.org/publications/understanding-metadata [Accessed: 27th January 2019]

Nieva de la Hidalga, A., Rosin, P., Sun, X., Wijers, A. and Groom, Q. (2019). MS 20 Interim report on Quality Control in imaging. ICEDIG Project. Task 3.4. 2019/01/31.

Ooms, J. (2018) *Tesseract 4 is here! State of the art OCR in R! - rOpenSci - open tools for open science* [Online] Ropensci.org. Available at: https://ropensci.org/technotes/2018/11/06/tesseract-40/ [Accessed: 20th December 2018].

Pyke, G.H. and Ehrlich, P.R. (2010) Biological collections and ecological/environmental research: a review, some observations and a look to the future. *Biological reviews*, 85(2): 247–266.

Shuyo, N. (2014) *language-detection* [Online]. GitHub. Available at: https://github.com/shuyo/language-detection/ [Accessed: 31st October 2018].

Spasić, I., Greenwood, M., Preece, A., Francis, N. and Elwyn, G. (2013) FlexiTerm: a flexible term recognition method. *Journal of Biomedical Semantics*, 4(1): 27. https://doi.org/10.1186/2041-1480-4-27

Spasić, I. (2018) Acronyms as an Integral Part of Multi-Word Term Recognition – A Token of Appreciation. *IEEE Access*, 6: 8351–8363. https://doi.org/10.1109/ACCESS.2018.2807122

Suarez, A.V. and Tsutsui, N.D. (2004) The Value of Museum Collections for Research and Society. *BioScience*, 54(1): 66–74. https://doi.org/10.1641/0006-3568(2004)054[0066:TVOMCF]2.0.CO;2

The Apache Software Foundation (2018) *Apache Taverna* [Online]. Available at: https://taverna.incubator.apache.org/ [Accessed: 21st October 2018].

Thessen, A.E., Cui, H. and Mozzherin, D. (2012) Applications of natural language processing in biodiversity science. *Advances in Bioinformatics*, 2012: e391574. https://doi.org/10.1155/2012/391574

The Stanford Natural Language Processing Group (2018) *Stanford Named Entity Recogniser (NER)* [Online]. Available at: https://nlp.stanford.edu/software/CRF-NER.shtml [Accessed: 22nd October 2018].

Weibel, S. (1997) The Dublin Core: A Simple Content Description Model for Electronic Resources. *Bulletin of the American Society for Information Science and Technology*, 24: 9–11. https://doi.org/10.1002/bult.70

Wheeler, Q.D., Knapp, S., Stevenson, D.W., Stevenson, J., Blum, S.D., Boom, B.M., et al. (2012) Mapping the biosphere: exploring species to understand the origin, organization and sustainability of biodiversity. *Systematics and Biodiversity*, 10(1): 1-20. https://doi.org/10.1080/14772000.2012.665095

Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., Robertson, T. and Vieglais, D. (2012) Darwin Core: an evolving community-developed biodiversity data standard. *PLoS ONE*, *7*(1): .e29715. https://doi.org/10.1371/journal.pone.0029715

9. Appendices

9.1 Institutions

Institution	Index Herbariorum code	ICEDIG Partner
Naturalis Biodiversity Center, Leiden, Netherlands	L	Yes
Meise Botanic Garden, Belgium	BR	Yes
University of Tartu, Estonia	TU	Yes
The Natural History Museum, London, United Kingdom	BM	Yes
Muséum National D'Histoire Naturelle (MNHN), Paris, France	P	Yes
The Royal Botanic Gardens Kew, RGBK, Richmond, United Kingdom	K	Yes
Finnish Museum of Natural History, Helsinki	H	Yes
Botanic Garden and Botanical Museum, Berlin	B	No
Royal Botanic Garden, Edinburgh	E	No

Table 11: Contributing institutions and their codes from Index Herbariorum (http://sweetgum.nybg.org/science/ih/)

9.2 OCR Software Settings

Settings are provided for the OCR programs used in section 5. This is to aid reproducibility of results.

9.2.1 Tesseract 4.0.0 and Tesseract 3.0.51

Page segmentation mode: “3” (Fully automatic page segmentation)

To not create accompanying xml file of results

- tessedit_create_hocr "0"

To remove noise in the input image

- textord_max_noise_size "45"

To specify that the space between word is variable

- textord_space_size_is_variable "1"

To not run in parallel

- tessedit_parallelize "0"

To load Directed Acyclic Word Graphs

- load_system_dawg "0"

9.3 ABBYY Finereader Engine 12

Load predefined profile: "TextExtraction_Accuracy"

Plain text export format

- FileExportFormatEnum.FEF_TextUnicodeDefaults

Rich text export format

- FileExportFormatEnum.FEF_RTF

9.4 Line Correctness analysis

Column headings denote image file name and number of printed lines contained within. The bracketed number in each column denotes the number of lines of printed text. For example, “B 10 0002520 (/24)” concerns an image named “B 10 0002520” containing 24 lines of printed text.

Segmented image naming convention:

IMAGENAME+TYPE+NUMBER
IMAGENAME= the original file from which the fragment was extracted
TYPE = The type of segment detected (lbl=label, bcd=barcode, clc=colour chart)
NUMBER = sequential item number within herbarium sheets
Example:
L.2168354_lbl02.jpg
IMAGENAME = L.2168354_lbl02.jpg (original file: L.2168354.jpg)
TYPE = lbl (label)
Number = 2

9.4.1 Line Correctness with Whole Images

Whole images	B 10 0002520 (/24)	BM000500117 (/14)	E00015443 (/10)	EIG.2770 (/15)	K000025814 (/32)
Tesseract 4.0.0	12	10.5	7.5	10.5	30
Tesseract 3.0.51	9	7	3	4	24.5
ABBYY FineReader Engine 12	19.5	3.5	9	10.5	30
Microsoft OneNote 2013	15	10.5	7.5	12.5	31.5

Table 12: Line correctness results for OCR programs using whole images

9.4.1.1 Image “B 10 0002520”

Manual transcription of **B 10 0002520** | Number of lines: 24

Mus. Bot. Berol.
B 10 0002520
5cm
EX HERBARIO KEWENSI
COLONIAL OFFICE EAST AFRICAN EXPEDITION: 1955-6
TANGANYIKA: Songea District
E. MILNE-REDHEAD and
P.TAYLOR, No. 9278
1955
acc.
20.APR.1959
Unangwa Hill about 6 km. E. of Songea
in crevices in vertical rock faces:
1140 m.
Perennial, roots very difficult to
remove from crevices; stems reddish
below, pale green above; petioles pale
green; lamina pale green above, paler
beneath with prominent nerves and
purple tinge; calyx pale green; corolla
white with mauve blotch on callus.
M.-R. & T. 9278. 22.3.1956.
Image 2001
Mus. bot. Berol.

Tesseract 4.0.0 | **B 10 0002520** | Score: 12/24

```
|

|

| ;

|

|
```

|

CL t——

Unangwa Hill about 6 = Ee of Songea

in crevices in verilesl. rock Faosg; :

1140 m,

Perennial, roots very aiffioult to

remove from crevices; stems redd:

below, pale green eve; Poth

green; Lauila pale green a

EX HERBARIO KEWENSI

CoLoNIAL OFFICE EAST AFRICAN EXPEDITION: 1955-6

a0

TANGANYIKA: Songea District

E. MILNE-REDHEAD and

| P. TAYLOR, No. 1 2%

of 1004509

GANT 13

1955”

A

Mus. Bot. Berol.

HL

15cm

Mus. bot. Berol.

Tesseract 3.0.51 | **B 10 0002520** | Score: 9/24

\

\

\

i

V

\

M

13mm Hill ahead: 6 km. E of Song“

11 Meet! in Wind rock fauna;

1m 3.

Perennial, roots very Wm“

rm: rm ormimm; atom

balm, palm: yam mum; m1

wan; lamina mum

EX HERBARIO KEWENSI

COLONIAL OFFICE EAST AFRICAN Expnnmon 1955 6

mm“

TANGANYIKA Songea District

E MILNE REDHEAD and

P TAYLOR No ‘12,}?

r“1}

Lv‘ M.

1955’

'i

Mus Bot Berol

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

HIIHIJmllliIHIIMIILIIQMMQIMIIIINIIIIIIIIWIH

IScm

(

Mus bot Berol

ABBYY FineReader Engine 12 | **B 10 0002520** | Score: 19.5/24

I

Perennii

purple tinge; calyx pale green; corolla

white with mauve blotch on callus.

22.3.1956.

ku-cX^a-^_a-v<Jv <SV<^vv

Hmafle fOOf |

ocw(k

1

Mus. Bot. Berol.

Mus. bot. Berol.

HniRIIIIIIIWIIIIIRIKIIIIIHnH

195.S-

B 10 0002520

§

m

EX HERBARIO KEWENSI

Colonial Office East African Expedition: 1955-6

Unangwa Hill about 6 Ion. E. of Songea

in crevices in vertical rock faces:

1140 m.

TANGANYIKA: Songea District

E. MILNE-REDHEAD and

P. TAYLOR, No.

__________, roots very difficult to

remove from crevices; stems reddish

petioles pale

„ above, paler

sneath with prominent nerves and

acc.

2GMi9L9

J®1?1

JS*",

Lal,

fror

below, pale green above;

green; lamina pale green

beneath with prominent nt

Microsoft OneNote | **B 10 0002520** | Score: 15/24

Unangwa Eli 11 about 6 E. of Songea

in crevices in vertical rock faces:

Perennial 9 roots very difficult to

remove from crevices; sterns reddish

below, pale green above; petioles pale

green; la.rl)ina pale green above, paler

beneath with prominent nerves and

purple tinge; calyx pale green; corolla

white with mauve blotch on callus.

& TO 927B.

22.3.1956,

EX HERBARIO KEWENSI

COLONIAL OFFICE EAST AFRICAN EXPEDITION: 1955-6

TANGANYIKA: Songea District

E. MILNE-REDHEAD and

P. TAYLOR, No.

ma e 2

Mus. Bot. Berol.

B 10 0002520

E

10

MUS. boto Berol.

9.4.1.2 Image “BM000500117”

Manual transcription of **BM000500117** | Number of lines: 14
THE NATURAL HISTORY MUSEUM, LONDON DEPARTMENT OF BOTANY 000500117 (BM) Vidi S. Pérez-Ortega 5th December 2010 Herbarium Musei Britannici Parmelia perlata (Huds.) Ach. TASMANIA: Tinderbox (near Hobart). Alt. sea level+. Common on dolerite in open dry sclerophyll Forest. G.Kantvilas No. 277/80 20.7.1980 Det: P.W. James TLC: stictic complex Determinavit

Tesseract 4.0.0 | **BM000500117** | Score: 10.5/14
Herbarium Musei Britannici Parmelia perlats (Huds.) Ach. TASMANTA: Tinderbox (near Hobart). Alt. sea level+. Common on dolerite in open dry sclerophyll forest.

G. Kantvilas No. 277/80 20.7.1980 Det: P.W. James TLC: stictic complex THE NATURAL HISTORY MUSEUM, LONDON Tx cep lu oA 4 DEPARTMENT OF BOTANY | | wr im $ leuwdoadputt Determinavit 10 Lye ; TEN copyright reserved MUSEUM 5s 213 ES | 2% 3% S| § 2 2 4 Sis = |S

Tesseract 3.0.51 | **BM000500117** | Score: 7/14
Herbarium Musei Britannici

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

Parmella Ecrlata (hm) Acb

TASMANIA Tinderbox (near hobart)

Alt sea level+

Common on dolerlte 10 open dry sclelopnvll

forest

G Kantv1las No 277/80 20 71380

Det P W James TLC stlctlc complex

THE NATURAL HISTORY MUSEUM LONDON

W \ “IL IL L n i DEPARTMENTOFBDTANV

] ‘

72:“ 7 ,7 HlmmwmwmwW

< Wilcdzgx [;

NLHVHHUd

1O NATURAL

HISTORY

copyright reserved

MUSEUM

ma

B‘N

*s.

5H,.

:73

b E

&‘ §

[”3

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

§\

>1u

ABBYY FineReader Engine 12 | **BM000500117** | Score: 3.5/14

Herbarium Musei Britannici

Parmelia perlata (Huds.) Ach.

(n

Common on dolerito In

tXi:

kjLU^XitlL

"7^.. (w-

MOLOWt

$0O((7

I

8 910 r^n history

jht reserved

U MUSEUM

gF'

<<

■ ■■

■ ■

(near Hobart).

n open dry sclerophyll

M

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

7 8

copyright reserved

h

il

jf

Microsoft OneNote 2013 | **BM000500117** | Score: 10/14

Herbarium Musei Britannici

Parmelia oerlata (Huds) Ach.

TASMANIA: Tinderbox (near Hobart) .

Alt. sea level+.

Comnon on dolerite in open dry sclerophyll

forest.

G. Kantvilas No. 277/80 20.7.1980

Det: P.W. James TLC: stictic complex

Sic . m

THE NATURAL HISTORY MUSEUM, LONDON

DEPARTMENT OF BOTANY

0005001 i i (8M)

Determinavit

9.4.1.3 Image “E00015443”

Manual transcription of **E00015443** | Number of lines: 10

ROYAL BOTANIC GARDEN

EDINBURGH

E00015443

HERB. HORT. EDINB.

Flora from Kunming Institute, China

Corydalis

Fumariaceae

B & L 12,228

Gang-Ho Ba, Lijiang

September 1987

Tesseract 4.0.0 | **E00015443** | Score: 7.5/10

HERB. HORT. EDINB.

TTT TET

dh

copyright reserved

ROYAL BOTANIC GARDEN

yg

Flora from Kunming Institute, China

" & a }]

Corydalis sp Smithiana Rede 4,

Fumariaceae gid

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

B&L 12228

’ y Gang-Ho Ba, Liii

- September 1987 ang

Tesseract 3.0.51 | **E00015443** | Score: 3/10
HERB HORT EDINB I i l I | IL” IIH Hil H \umuumfifififlfiifwumum Flora from Kunming Institute, China a a) N i ‘ Calydalzs sp 5 m ,Uw. a my RM; 1 Fumarmceae { Ly B&L12228 ' Gang H0 Ba, L j September 1987 lllang

ABBYY FineReader Engine 12 | **E00015443** | Score: 9/10
HERB. HORT. EDINB. V ____ > 9 A

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

V

1 "*

I**

\

♦-'

I

niiTnnT

/

/

i

A

s

r>

:

'A

K

«

I I

EOOO15443

//&

/ tn ,7Afa « ReA£L

Kc

Gang-Ho Ba, Lijiang

September 1987

LULU

r

||||||||||||||||

CD

CO

:

J

E

o o

'A

4

H

ROYAL BOTANIC GARDEN

EDINBURGH

I

r

1

i

■

■

■HL

: Y|

L

Flora from Kunming Institute, China

Corydalis sp

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

Fumariaceae

B & L 12,228

!

i

I

/A

£

-nrof

CO 2

O)

-I

o

o

21

r

■

-

<2“l

sits

773ft

- .

I /

Microsoft OneNote 2013 | **E00015443** | Score: 7.5/10

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

HERB. HORTO EDINB.

Luu-I

ROYAL BOTANIC GARDEN

EDINBURGH

EOOOI 5443

Flora from Kunming Institute, China

Col)'dalis sp-

Fumariaceae

B & L 12,228

September 1987

Gang-Ho Ba, Lijiang

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

9.4.1.4 Image “EIG.2770”

Manual transcription of **EIG.2770** | Number of lines: 15

Digitarium

http://id.luomus.fi/

EIG.2770

2013-04-02

MUSEUM BOTANICUM UNIVERSITATIS, HELSINKI

AUSTRALIA, Queensland, Bellenden Ker

National Park W. of Babinda.

The Boulders. Complex mesophyll rain

Forest.

10. Sept. 1981 Ilkka Kukkonen 10879

UMT grid:

MUSEUM BOTANICUM

UNIV. (H). HELSINKI

1459257

OUABR.

Tesseract 4.0.0 | **EIG.2770** | Score: 10.5/15

MUSEUM BOTANICUM UNIVERSITATIS, HELSINKI

Diplaziunn dilatatem 5.

AUSTRALIA, Queensland, Bellenden Ker

MUSEUM BOTANICUM

National Park W. of Babinda.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

UNIV. {H). HELSINKI is Fa

Complex mesophyll rain

ores .

um

EE http://id.luomus. Fil

145925

L270 EIG.2770

10. Sept. 1981 Ilkka Kukkonen 10879

[m]CTs. 2013-04-02

QUASR.

tar

i.

igi

D

UTM grid:

Tesseract 3.0.51 | **EIG.2770** | Score: 4/15
MUSEUM BOTANICUM UNIVERSITATIS HELSINKI B'p‘ul‘um d"‘\\"£\rurn DC AUEJLHQ‘U LS, kglloenslgmd, 1h llenaen ‘el MUSEUM BOTANICUM Aatlondl idllx u. of 3001:1511 UNIV NH HLLSINKI #119 Hcgulders DONDMR mesopiy‘l rain ores .

lum

El'fiEi htth/idluomusfi/ 11;,925'7

J- r'l' , E|G2770 iU.)eit. 1051 1111c: u} omen 108 0

E15 4 2013 04 02 OUABR

tar

' \

Igl

D

UTMgrid 01W J 3M \\1U 1 ‘32

ABBYY FineReader Engine 12 | **EIG.2770** | Score: 10.5/15

T

in

o-

\

1 -

1

2

3

4

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

5

\m

6

7

8

9

10

11

\\^

12

13

kV

<4/

I

r

o

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

1

W/

2

3

4

5

6

7

MUSEUM BOTANICUM UNIVERSITATIS, HELSINKI

8

0 i p I tx Z. (U rvx J i I ex t t U R b ,

9

10

forest.

11

10. Sept. 1981

12

OUASR.

2013-04-02

UTM grid:

13

i

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

II

v

Ilkka Kukkonen IO879

det. J. S\rvel a

1

AUSTRALIA, Queensland, Bellenden Ker

National Park W. of Babinda.

The Boulders. Complex mesophyll rain

MUSEUM BOTANICUM

UNIV. (H). HELSINKI

1459257

I

1

®!

I HS|B http://id.luomus.fi/

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

jffiH BG2'"

yl

1

Microsoft OneNote 2013 | **EIG.2770** | Score: 12.5/15

10

11

12

13

10

11

12

13

MUSEUM BOTANICUM UNIVERSITATIS, HELSINKI

AUSTRALIA, Queensland, Bellenden Ker

National Park W. of Babinda

http://id.luomus.fi/

EIG.2770

2013-04-02

MUSEUM BOTANICUM

H EL SIN Ki

UNIV.

1459237

OUAbR.

10.

UTM

The Boulders

forest.

Septe 1981

grid:

Complex mesophyll rain

Ilkka 10879

Jet, J. s

9.4.1.5 Image “K000025814”

Manual transcription of **K000025814** | Number of lines: 32

ROYAL BOTANIC GARDENS KEW

K000025814

THE PLANTS OF WESTERN

CAMEROON

Recorded on Database (K) 1992-

Psychotira geophylax Cheek & Sonké

(Psych. Sp. B aff gabonica)

Cited in protologue

DET Cheek Jan 2007

Det… 20…

HERB. HORT. KEW.

FLORA OF WESTERN CAMEROON

RBG Kew & Herbier National du Cameroun

support by Darwin Initiative & Earthwatch

Rubiaceae

Psychotria

Division: Kupe-Muanenguba South West

Gazette: Nyasoso

LongLat: N; E Alt: 1190m

Above Nyasoso on Max's trail up Mount Kupe. Montane forest with canopy to 35cm tall. Many stands of Marantaceae. Volcanic soils.

Large shrub to 2.5cm tall. Leaves shiny, nerves depressed above. Stipules green, very broad; flowers orange, clustered in dense creamy coloured heads. Buds orange, only a few flowers open per head. Sepals brown; corolla fleshy orange below, cream above. Stigma white. Fruit green when immature, turning orange.

Sidwell K. 416 26/Oct/1995

With: Etuge. Schoenengerger & Takele

Duplicates at:

New dets to M.Cheek at RBG Kew and HNC BP 1601 Yaounde.

Tesseract 4.0.0 | **K000025814** | Score: 30/32

HERB. HORT. KEW.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

ROYAL BOTANIC GAR

MINER

A

& PR

Av " He

3 z

. iY

L O 5

) r

p LL g N

{

. $HrP < ’

#

y Ll CLA ¥

J :

i"

()

>

Q |

G8

(0)

a.

=

Oo) |

= |

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

p_

-

- 2

(&)

THE PLANTS OF WESTERN

CAMEROON

Recorded on Database (K) 1992-

Psychofifa geophylax Cheek & Sonké

(Psych. sp. B aff gabonica)

Cited in protologue

DET Cheek

Jan 2007

cesssesasesanse

FLORA OF WESTERN CAMEROON

RBG Kew & Herbier National du Cameroun

support by Darwin Initiative & Earthwatch

Rubiaceae

Psychotria

Division: Kupe-Muanenguba South West

Gazette: Nyasoso

LonglLat: N; E Alt: 1190m

Above Nyasoso on Max's trail up Mount Kupe. Montane

forest with canopy to 35m tall. Many stands of

Marantaceae. Volcanic soils.

Large shrub to 2.5m tall. Leaves shiny, nerves depressed

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

above. Stipules green, very broad: flowers orange, clustered

In dense creamy coloured heads. Buds orange, only a few

flowers open per head. Sepals brown: corolla fleshy orange

below, cream above. Stigma white. Fruit green when

Immature, turning orange.

Sidwell K. 416 26/0ct/1995

With: Etuge, Schoenengerger & Takele

Duplicates at:

New dets to M.Cheek at RBG Kew and HNC BP 1601 Yaounde.

Tesseract 3.0.51 | **K000025814** | Score: 24.5/32

HERB HORT KEW

ROYAL BOTANIC GAR

11111111111111111’1111115

W.

‘ 7 ‘9.

x . ,

,

I»

. 4'. .

‘ I

‘1 4' ’ .

1

9 “x . ‘- /

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

I

I F A '4 ‘ »

{fl ‘

.0

d)

>

L

(D

m

(D

L

4.:

.C

CD

':

>~.

O.

O

0

THE PLANTS OF WESTERN

CAMEROON

Recorded on Database (K) 1992

Psychofla qeophvlax Cheek & Sonke

(Psych ép B aff gabonica)

Cited in protologue

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

DET Cheek Jan 2007

H OLOTVPL

1457 61101140 geophylax Cheek +§oml<e

DET 20

I

FLORA OF WESTERN CAMEROON

RBG Kew & Herbler Natlonal du Cameroun

support by Darwm Inltlatlve & Earthwatch

Rubiaceae HOLOTV 1715:

PSVC/zotiza 91> B 4‘1 'LLCA

bet C&MCL VYDOL

Division Kupe Muanenguba South West

Gazette Nyasoso

LongLat N E Alt 1190m

Above Nyasoso on Max 3 trail up Mount Kupe Montana

forest with canopy to 35m tall Many stands of

Marantaceae Volcanic soils

Large shrub to 2 5m tall Leaves shiny nerves depressed

above Stipules green very broad flowers orange clustered

m dense creamy coloured heads Buds orange only a few

flowers open per head Sepals brown corolla fleshy orange

below cream above Stigma white Fruit green when

immature turning orange

Sidwell K 416 26/Oct/1995

With htuge Sulloencngugel 6L lakele

Dupliutus u k Lu‘ol YA' S(k(NAG! BR; r10. P

EKLU arm 1AA EA C&Jfi

.\u\ dCLS 10 M Check it RBU Ixew 1nd HNC BP 1601 \ wounds

ABBYY FineReader Engine 12 | **K000025814** | Score: 30/32

HERB. HORT. KEW.

ROYAL BOTANIC GARDENS KEW

<000025814

I

Cheek & Sonke

Jan 2007

Cheek -t-Sonke

OX

DET ....

20 ....

!

1

Rubiaceae

f.

Division:

South West

Gazette:

26/Oct/1995

i

eV

*

c

■

CO

co

/

co

04

4 .

r

E

o o

0)

£

o

cn

CD

Kupe-Muanenguba

Nyasoso

♦

£

Q5<^

LongLat: N; E   Alt: 1] 90m

Above Nyasoso on Max’s trail up Mount Kupe. Montane

Microsoft OneNote 2013 | **K000025814** | Score: 31.5/32

HERB. HORT. KEW

ROYAL BOTANIC GARDENS KEW

K000025814

SidlL'e(l

THE PLANTS OF WESTERN

CAMEROON

Recorded on Database (K) 1992-

Ps ch01a eo h lax Cheek & Sonké

(Psych. p. B aff gabonica)

Cited in protologue

DET Cheek

clnö+riQ geo?

lax

DET........

Jan 2007

Chee14 +

20 ....

FLORA OF WESTERN CAMEROON

RBG Kew & Herbier National du Cameroun

support by Darwin Initiative & Earthwatch

Rubiaceae

Psychotria s

Kupe-Muanenguba

Division.

Nyasoso

Gazette:

LongLat: N; E

A OCOTY/E

CUCA

South West

Alt: 11 90m

Above Nyasoso on Max's trail up Mount Kupe. Montane forest with canopy to 35m tall. Many stands of Marantaceae. Volcanic soils.

Large shrub to 2.5m tall. Leaves shiny, nerves depressed above. Stipules green, very broad; flowers orange, clustered in dense creamy coloured heads. Buds orange, only a few flowers open per head. Sepals brown; corolla fleshy orange below, cream above. Stigma white. Fruit green when immature, turning orange.

Sidwell K. 416

With: Etuge, Schoenengerger & Takele

26/Oct/1995

Duplicates at: WAG t

New to M.Cheek at RBG Kew and HNC BP 1601 Yaounde.

9.4.2 Line Correctness with Segmented Images

9.4.2.1 Segments of image “B 10 0002520”

	bcd01 (/2)	lbl01 (/1)	lbl02 (/8)	lbl03 (/11)	lbl04 (/1)	lbl06 (/1)

Tesseract 4.0.0	1	1	4.5	10	0	1
Tesseract 3.0.51	1	1	4.5	7	0	1
ABBYY FineReader Engine 12	2	0	6	9.5	0	1
Microsoft OneNote 2013	1.5	0	5	8	0	1

9.4.2.2 Segment “B 10 0002520_bcd01”

Manual transcription of **B 10 0002520_bcd01** | Number of lines: 2
Mus. Bot. Berol. B 10 0002520

Tesseract 4.0.0 | **B 10 0002520_bcd01** | Score: 1/2
Mus. Bot. Berol. HL

Tesseract 3.0.51 | **B 10 0002520_bcd01** | Score: 1/2
Mus Bot Berol mummmmnlnygigugmgwmmmlwm

ABBYY FineReader Engine 12 | **B 10 0002520_bcd01** | Score: 2/2
Mus. Bot. Berol. B 10 0002520

Microsoft OneNote 2013 | **B 10 0002520_bcd01** | Score: 1.5/2
Muş. Bot. Beroı. B 10 0002520

9.4.2.3 Segment “B 10 0002520_lbl01”

Manual transcription of **B 10 0002520_lbl01** | Number of lines: 1
5cm

Tesseract 4.0.0 | **B 10 0002520_lbl01** | Score: 1/1
1 5cm

Tesseract 3.0.51 | **B 10 0002520_lbl01** | Score: 1/1
-5cm

ABBYY FineReader Engine 12 | **B 10 0002520_lbl01** | Score: 0/1
in E o

Microsoft OneNote 2013 | **B 10 0002520_lbl01** | Score: 0/1

9.4.2.4 Segment “B 10 0002520_lbl02”

Manual transcription of **B 10 0002520_lbl02** | Number of lines: 8
EX HERBARIO KEWENSI COLONIAL OFFICE EAST AFRICAN EXPEDITION: 1955-6 TANGANYIKA: Songea District E. MILNE-REDHEAD and P.TAYLOR, No. 195 acc. 20.APR.1959

Tesseract 4.0.0 | **B 10 0002520_lbl02** | Score: 4.5/8
EX HERBARIO KEWENSI CoLoNIAL Orrice EAST AFRICAN EXPEDITION: 1955-6 TANGANYIKA: Songea District E. MILNE-REDHEAD and P. TAYLOR, No. 123%

Tesseract 3.0.51 | **B 10 0002520_lbl02** | Score: 4.5/8
EX HERBARIO KEWENSI COLONIAL OFFICE EAST AFRICAN Expanmon 1955 6 TANGANYIKA Songea District E MILNE REDHEAD and

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

P TAYLOR No fizq‘af

ABBYY FineReader Engine 12 | **B 10 0002520_lbl02** | Score: 6/8

a

195S"

EX HERBARIO KEWENSI

Colonial Office East African Expedition: 1955-6

TANGANYIKA: Songea District

E. MILNE-REDHEAD and

P. TAYLOR, No. ZSfif

a c c.

2G.W9C9

Microsoft OneNote 2013 | **B 10 0002520_lbl02** | Score: 5/8

EX HERBARIO KEWENSI

COLONIAL OFFICE EAST AFRICAN EXPEDITION: 1955-6

TANGANYIKA: Songea District

E. MILNE-REDHEAD and

P. TAYLOR, No.

9.4.2.5 Segment “B 10 0002520_lbl03”

Manual transcription of **B 10 0002520_lbl03** | Number of lines: 11

Unangwa Hill about 6 km, E. of Songea

in crevices in vertical rock faces:

1140 m,

Perennial, roots very difficult to

remove from crevices; stems reddish

below, pale green above; petioles pale

green; lamina pale green above, paler

beneath with prominent nerves and :

purple tinge; calyx pele green; corolla

white with mauve blotch on callus

Me=Re & T. 9278. 22,3+1956,

Tesseract 4.0.0 | **B 10 0002520_lbl03** | Score: 10/11

Unangwa Hill about 6 km, E. of Songea

in crevices in vertical rock faces:

1140 m,

Perennial, roots very difficult to

remove from crevices; stems reddish

below, pale green above; petioles pale

green; lamina pale green above, paler

beneath with prominent nerves and :

purple tinge; calyx pele green; corolla

white with mauve blotch on callus

Me=Re & T. 9278. 22,3+1956,

Tesseract 3.0.51 | **B 10 0002520_lbl03** | Score: 7/11

Unangwa Hill about 6 km. E of Songea

1n crevices in vertieal rock faces

1140 m

Perennial, roots very difficult to

remove from crevices, stems reddish

below, pala green above; patioles pale

green, lamina. pale green above, yaler

beneath with prominent harms and

purple tinge, calyx pale green, mmlla

white with mauve blotch 0n callus.

u, R. & T 9278 22 3 1956.

ABBYY FineReader Engine 12 | **B 10 0002520_lbl03** | Score: 9.5/11

22.3.1956.

9278.

M.-R. & T.

Perennial, roots very difficult to

remove from crevices; stems reddish

below, pale green above; petioles pale

green; lamina pale green above, paler

beneath with prominent nerves and

purple tinge; calyx pale green; corolla

white with mauve blotch on callus.

Unangwa Hill about 6 Ion. E. of Songea

in crevices in vertical rock faces:

1140 m.

Microsoft OneNote 2013 | **B 10 0002520_lbl03** | Score: 8/11

TJnangwa Eli 11 about 6 lcm. E. of Songea

in crevices in vertical rock faces;

Perenniäl, roots very difficult to

remove from crevices; sterns reddish

below, pale green above; petioles pale

green; laxnina pale green above, paler

beneath with prominent nerves and

purple tinge; calyx pale green; corolla

white with mauve blotch on callus.

& TO 927B.

22.3.1956.

9.4.2.6 Segment “B 10 0002520_lbl04”

Manual transcription of **B 10 0002520_lbl04** | Number of lines: 1

Image 2001

Tesseract 4.0.0 | **B 10 0002520_lbl04** | Score: 0/1

Tesseract 3.0.51 | **B 10 0002520_lbl04** | Score: 0/1

ABBYY FineReader Engine 12 | **B 10 0002520_lbl04** | Score: 0/1

■minimi

Microsoft OneNote 2013 | **B 10 0002520_lbl04** | Score: 0/1

ma e

9.4.2.7 Segment “B 10 0002520_lbl06”

Manual transcription of **B 10 0002520_lbl06** | Number of lines: 1
Mus. bot. Berol.

Tesseract 4.0.0 | **B 10 0002520_lbl06** | Score: 1/1
Mus. bot. Berol.

Tesseract 3.0.51 | **B 10 0002520_lbl06** | Score: 1/1
Mus bot Berol

ABBYY FineReader Engine 12 | **B 10 0002520_lbl06** | Score: 1/1
Mus. bot.Berol.

Microsoft OneNote 2013 | **B 10 0002520_lbl06** | Score: 1/1
MUS. bot. Berol.

9.4.2.8 Segments of image “BM000500117”

Image segments [BM000500117]	bcd01 (/3)	lbl03 (/2)	lbl04 (/8)	lbl07 (/1)
Tesseract 4.0.0	1	0	8	1
Tesseract 3.0.51	1	0	5.5	1
ABBYY FineReader Engine 12	3	0	7.5	1
Microsoft OneNote 2013	3	0	7.5	0

9.4.2.9 Segment “BM000500117_bcd01”

Manual transcription of **BM000500117_bcd01** | Number of lines: 3
THE NATURAL HISTORY MUSEUM, LONDON DEPARTMENT OF BOTANY 000500117 (BM)

Tesseract 4.0.0 | **BM000500117_bcd01** | Score: 1/3
THE NA Alii in) LONDON wii iif ll 0500117 (|

Tesseract 3.0.51 | **BM000500117_bcd01** | Score: 1/3
THE NATUEAL WHISEDRVFBO MUSEUM YLONDON HIIHIHIEMMH WIENMILAHWW 05001 1 7 (E:

ABBYY FineReader Engine 12 | **BM000500117_bcd01** | Score: 3/3
111111111!!! Ill III II HI II 000500117 (BM) THE NATURAL HISTORY MUSEUM, LONDON DEPARTMENT OF BOTANY

Microsoft OneNote 2013 | **BM000500117_bcd01** | Score: 3/3

THE NATURAL HISTORY MUSEUM, LONDON

DEPARTMENT OF BOTANY

I Ill I I Il Il I I I I I I Il Il I I I I I I I Il

0005001 17 (8M)

9.4.2.10 Segment “BM000500117_lbl03”

Manual transcription of **BM000500117_lbl03** | Number of lines: 2
Vidi S. Pérez-Ortega 5th December 2010

Tesseract 4.0.0 | **BM000500117_lbl03** | Score: 0/2
0102 42quia2ad YI§ en ©30310-23.19d °S PIA Ce ——————

Tesseract 3.0.51 | **BM000500117_lbl03** | Score: 0/2
Sam Luefimumfi 5m \\ awatO N2?— m :2: Ill“

ABBYY FineReader Engine 12 | **BM000500117_lbl03** | Score: 0/2
'' -CJ s Cl

■s

I s

£

I

Oh

I C/5

3

&

Microsoft OneNote 2013 | **BM000500117_lbl03** | Score: 0/2

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

9.4.2.11 Segment "BM000500117_lbl04"

Manual transcription of **BM000500117_lbl04** | Number of lines: 8

Herbarium Musei Britannici

Parmelia perlata (Huds.) Ach.

TASMANIA: Tinderbox (near Hobart).

Alt. sea level+.

Common on dolerite in open dry scleropyll

Forest.

G.Kantvilas No. 277/80 20.7.1980

Det: P.W. James TLC: stictic complex

Tesseract 4.0.0 | **BM000500117_lbl04** | Score: 8/8

Herbarium Musei Britannici

Parmelia perlata (Huds.) Ach.

TASMANIA: Tinderbox (near Hobart).

Alt. sea level+.

Common on dolerite in open dry sclerophyll

forest.

G. Kantvilas No. 277/80 20.7.1980

Det: P.W. James TLC: stictic complex

Tesseract 3.0.51 | **BM000500117_lbl04** | Score: 5.5/8

Herbarium Musei Britannici

Parmella Eerlata (Hue) Ash

TASMANIA Tinderbox (near Hobart)

Alt sea level+

Comnon on dolerlte :Ln open dry sclerophyll

forest

G Kantv1las No 277/80 20 7 1980

Det P.W. James TLC stlctlc complex

ABBYY FineReader Engine 12 | **BM000500117_lbl04** | Score: 7.5/8

Herbarium Musei Britannici

Parmelia perlata (Hudt..) Ach.

TASMANIA: Tinderbox (near Hobart).

Alt. sea level+.

Common on dolerite in open dry sclerophyll

forest.

G. Kantvilas No. 277/80 20.7-1980

Det: P.W. James TLC: stictic complex

Microsoft OneNote 2013 | **BM000500117_lbl04** | Score: 7.5/8

Herbarium Musei Britannici

Parmelia per Iata (Huds) Ach.

TASMANIA: Tinderbox (near Hobart) .

Alt. sea level+.

Com.non on dolerite in open dry sclerophyll

forest.

g. Kantvilas No. 277/80 20.7. 1980

Det: P.W. James TLC: stictic complex

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

9.4.2.12 Segment “BM000500117_lbl07”

Manual transcription of **BM000500117_lbl07** | Number of lines: 1
Determinavit

Tesseract 4.0.0 | **BM000500117_lbl07** | Score: 1/1
Determinavit

Tesseract 3.0.51 | **BM000500117_lbl07** | Score: 1/1
Determinavit

ABBYY FineReader Engine 12 | **BM000500117_lbl07** | Score: 1/1
Determinavit t cuu^£ $ ll

Microsoft OneNote 2013 | **BM000500117_lbl07** | Score: 0/1
Determinavzt

9.4.2.13 Segments of image “E00015443”

Image segments [E00015443]	bcd01 (/3)	lbl01 (/1)	lbl03 (/6)
Tesseract 4.0.0	2	1	5
Tesseract 3.0.51	0	1	3.5
ABBYY FineReader Engine 12	2	1	6
Microsoft OneNote 2013	2	1	5

9.4.2.14 Segment “E00015443_bcd01”

Manual transcription of **E00015443_bcd01** | Number of lines: 3
ROYAL BOTANIC GARDEN EDINBURGH E00015443

Tesseract 4.0.0 | **E00015443_bcd01** | Score: 2/3
ROYAL BOTANIC GARDEN EDINBURGH

Tesseract 3.0.51 | **E00015443_bcd01** | Score: 0/3
AAAAAAAAAAAAAAAAAA EEEEEEEEE

ABBYY FineReader Engine 12 | **E00015443_bcd01** | Score: 2/3
EOOO15443 ROYAL BOTANIC GARDEN EDINBURGH

Microsoft OneNote 2013 | **E00015443_bcd01** | Score: 2/3
EOOO15443 ROYAL BOTANIC GARDEN

EDINBURGH

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

9.4.2.15 Segment “E00015443_lbl01”

Manual transcription of **E00015443_lbl01** | Number of lines: 1
HERB. HORT. EDINB.

Tesseract 4.0.0 | **E00015443_lbl01** | Score: 1/1
HERB. HORT. EDINB

Tesseract 3.0.51 | **E00015443_lbl01** | Score: 1/1
HERB HORT EDINB {

ABBYY FineReader Engine 12 | **E00015443_lbl01** | Score: 1/1
HERB. HORT. EDINB.

Microsoft OneNote 2013 | **E00015443_lbl01** | Score: 1/1
HERB. HORT. EDINB

9.4.2.16 Segment “E00015443_lbl03”

Manual transcription of **E00015443_lbl03** | Number of lines: 6
Flora from Kunming Institute, China Corydalis Fumariaceae B & L 12,228 Gang-Ho Ba, Lijiang September 1987

Tesseract 4.0.0 | **E00015443_lbl03** | Score: 5/6
Flora from Kunming Institute, China Corydalis sp smithrana Rede 4, Fumariaceae we, B&1L12228 Gang-Ho Ba, Lijiang September 1987

Tesseract 3.0.51 | **E00015443_lbl03** | Score: 3.5/6
Flora from Kunming Institute, China COlj’dallS Spy J m IU’h {4 H» Rd; 1,; Fumaridceae (L B & L 12 228 Gang H0 Ba, Lijiang

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

September 1987

ABBYY FineReader Engine 12 | **E00015443_lbl03** | Score: 6/6
Ara ReAIL Gang-Ho Ba, Lijiang September 1987 Flora from Kunming Institute, China Corydalis sp'' Fumariaceae B & L 12,228

Microsoft OneNote 2013 | **E00015443_lbl03** | Score: 5/6
Flora from Kunming Institute, China Col)'dalis sf S m Fumariaceae B & L 12,228 September 1987 Gang-Ho Ba, Lijiang

9.4.2.17 Segments of image "EIG.2770"

Image segments [EIG.2770]	bcd01 (/4)	lbl02 (/7)	lbl03 (/4)
Tesseract 4.0.0	3	7	3
Tesseract 3.0.51	2	6	2.5
ABBYY FineReader Engine 12	3	6.5	3
Microsoft OneNote 2013	2.5	6.5	2.5

9.4.2.18 Segment "EIG.2770_bcd01"

Manual transcription of **EIG.2770_bcd01** | Number of lines: 4
Digitarium http://id.luomus.fi/ EIG.2770 2013-04-02

Tesseract 4.0.0 | **EIG.2770_bcd01** | Score: 3/4
http://id.luomus.fi/ | EIG.2770 2013-04-02 Fi o £ 3 To ji] -— = o

Tesseract 3.0.51 | **EIG.2770_bcd01** | Score: 2/4
http //id luomus fi/

EIG 2770 2013 04 02 533 E! E g L- 1‘5 9.- g: o

ABBYY FineReader Engine 12 | **EIG.2770_bcd01** | Score: 3/4
http://id.luomus.fi/ EIG.2770 2013-04-02

Microsoft OneNote 2013 | **EIG.2770_bcd01** | Score: 2.5/4
http://id.luomus.fi/ ElG.2770 2013-04-02

9.4.2.19 Segment “EIG.2770_lbl02”

Manual transcription of **EIG.2770_lbl02** | Number of lines: 7
MUSEUM BOTANICUM UNIVERSITATIS, HELSINKI AUSTRALIA, Queensland, Bellenden Ker National Park W. of Babinda. The Boulders. Complex mesophyll rain Forest. 10. Sept. 1981 Ilkka Kukkonen 10879 UMT grid:

Tesseract 4.0.0 | **EIG.2770_lbl02** | Score: 7/7
MUSEUM BOTANICUM UNIVERSITATIS, HELSINKI AUSTRALIA, Queensland, Bellenden Ker National Park W. of Babinda. The Boulders. Complex mesophyll rain forest, 10. Sept. 1981 Ilkka Kukkonen 10879 UTM grid: det

Tesseract 3.0.51 | **EIG.2770_lbl02** | Score: 6/7
MUSEUM BOTANICUM UNIVERSITATIS HELSINKI Dip [02(um <11 latafurn BL AUSTRALIA, Queensland, Bcllenden ker

National Park W. of Babinda.

The Boulders Complex mesophyll rain

forest.

10 bept 1981 Illka Lullonen 10879

UTM grid clef J SuFVQIQ {)82

ABBYY FineReader Engine 12 | **EIG.2770_lbl02** | Score: 6.5/7

MUSEUM BOTANICUM UNIVERSITATIS, HELSINKI

J i I CK L,

C* I p i a Zi U 4<V\

10. Sept. 1981

Ilkka Kukkonen IO879

J. WvgJa 4382

UTM grid:

AUSTRALIA, Queensland, Bellenden Ker

National Park W. of Babinda.

The Boulders. Complex mesophyll rain

forest.

Microsoft OneNote 2013 | **EIG.2770_lbl02** | Score: 6.5/7

MUSEUM BOTANICUM UNIVERSITATIS, HELSINKI

AUSTRALIA, Queensland, Bellenden Ker

National Park W. of Babinda

The Boulders

forest.

10. Septe 1981

UTM grid:

Complex mesophy11 rain

Ilkka Ikukkonen 10879

Jet, J.

9.4.2.20 Segment "EIG.2770_lbl03"

Manual transcription of **EIG.2770_lbl03** | Number of lines: 4
MUSEUM BOTANICUM UNIV. (H). HELSINKI 1459257 OUABR.

Tesseract 4.0.0 | **EIG.2770_lbl03** | Score: 3/4
MUSEUM BOTANICUM UNIV. {H). HELSINKI }438257 QUABR.

Tesseract 3.0.51 | **EIG.2770_lbl03** | Score: 2.5/4
MUSEUM BOTANICUM UNIV 'H) iiELSIHKI l4 39237 OUABR

ABBYY FineReader Engine 12 | **EIG.2770_lbl03** | Score: 3/4
OUAtt. MUSEUM BOTANICUM

UNIV. (H). HELSINKI

1459257

Microsoft OneNote 2013 | **EIG.2770_lbl03** | Score: 2.5/4

MUSEUM BOTANICUM

HELSINKI

UNIV. (H).

OUAbR.

9.4.2.21 Segments of image “K000025814”

Image segments [K000025814]	K000025814_bcd01 (/2)	K000025814_lbl01 (/3)	K000025814_lbl03 (/4)	K000025814_lbl04 (/1)	K000025814_lbl06 (/1)	K000025814_lbl07 (/21)
Tesseract 4.0.0	0.5	3	4	1	1	21
Tesseract 3.0.51	0.5	3	2.5	1	1	18
ABBYY FineReader Engine 12	2	3	3.5	0.5	1	20.5
Microsoft OneNote 2013	2	3	3	0	1	20

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

9.4.2.22 Segment "K000025814_bcd01"

Manual transcription of **K000025814_bcd01** | Number of lines: 2
ROYAL BOTANIC GARDENS KEW K000025814

Tesseract 4.0.0 | **K000025814_bcd01** | Score: 0.5/2
ROYAL BOTANIC GARDE LL] I i I

Tesseract 3.0.51 | **K000025814_bcd01** | Score: 0.5/2
ROYAL BOTANIC GARDE IIIIIIIIIIIIJLIIILIIIIIICIIIIIJLIIIIIIILIIIIIIIIIIIIIIIIIw

ABBYY FineReader Engine 12 | **K000025814_bcd01** | Score: 2/2
ROYAL BOTANIC GARDENS KEW K000025814

Microsoft OneNote 2013 | **K000025814_bcd01** | Score: 2/2
ROYAL BOTANIC GARDENS KEW K000025814

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

9.4.2.23 Segment “K000025814_lbl01”

Manual transcription of **K000025814_lbl01** | Number of lines: 3
THE PLANTS OF WESTERN CAMEROON Recorded on Database (K) 1992-

Tesseract 4.0.0 | **K000025814_lbl01** | Score: 3/3
THE PLANTS OF WESTERN CAMEROON Recorded on Database (K) 1992-

Tesseract 3.0.51 | **K000025814_lbl01** | Score: 3/3
THE PLANTS OF WESTERN CAMEROON Recorded on Database (K) 1992

ABBYY FineReader Engine 12 | **K000025814_lbl01** | Score: 3/3
THE PLANTS OF WESTERN CAMEROON Recorded on Database (K) 1992-

Microsoft OneNote 2013 | **K000025814_lbl01** | Score: 3/3

* THE PLANTS OF WESTERN

CAMEROON

Recorded on Database (K) 1992-

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

9.4.2.24 Segment “K000025814_lbl03”

Manual transcription of **K000025814_lbl03** | Number of lines: 4
Psychotira geophylax Cheek & Sonké (Psych. Sp. B aff gabonica) Cited in protologue DET Cheek Jan 2007

Tesseract 4.0.0 | **K000025814_lbl03** | Score: 4/4
Psychofita geophylax Cheek & Sonké (Psych. 'sp. B aff gabonica) Cited in protologue DET Cheek Jan 2007

Tesseract 3.0.51 | **K000025814_lbl03** | Score: 2.5/4
Psvchofla qeophvlax Cheek & Sonke (Psych 'ép B aff gabonica) Cited in protologue DET Cheek Jan 2007

ABBYY FineReader Engine 12 | **K000025814_lbl03** | Score: 3.5/4
4 Jan 2007

Psycho|ira qeophylax Cheek & Sonke (Psych, sp. B aff gabonica) Cited in protologue DET Cheek

Microsoft OneNote 2013 | **K000025814_lbl03** | Score: 3/4
Ps chota eo h lax Cheek & Sonké (Psych. p. B aff gabonica) Cited in protologue DET Cheek Jan 2007

9.4.2.25 Segment “K000025814_lbl04”

Manual transcription of **K000025814_lbl04** | Number of lines: 1
Det… 20…

Tesseract 4.0.0 | **K000025814_lbl04** | Score: 1/1
DET icici sinister 20

Tesseract 3.0.51 | **K000025814_lbl04** | Score: 1/1
DET ... 20 ...

ABBYY FineReader Engine 12 | **K000025814_lbl04** | Score: 0.5/1
DET...

Microsoft OneNote 2013 | **K000025814_lbl04** | Score: 0/1
… ⊥ ∃ CI ⊥ 010H

A3.2.25 Segment “K000025814_lbl06”

Manual transcription of **K000025814_lbl06** | Number of lines: 1
HERB. HORT. KEW.

Tesseract 4.0.0 | **K000025814_lbl06** | Score: 1/1
HERB. HORT. KEW;

Tesseract 3.0.51 | **K000025814_lbl06** | Score: 1/1
HERB HORT KEW

ABBYY FineReader Engine 12 | **K000025814_lbl06** | Score: 1/1
HERB. HORT. KEW.

Microsoft OneNote 2013 | **K000025814_lbl06** | Score: 1/1
HERB. HORT. KEW.

9.4.2.26 Segment “K000025814_lbl07”

Manual transcription of **K000025814_lbl07** | Number of lines: 21

FLORA OF WESTERN CAMEROON

RBG Kew & Herbier National du Cameroun

support by Darwin Initiative & Earthwatch

Rubiaceae

Psychotria

Division: Kupe-Muanenguba South West

Gazette: Nyasoso

LongLat: N; E Alt: 1190m

Above Nyasoso on Max’s trail up Mount Kupe. Montane

forest with canopy to 35cm tall. Many stands of

Marantaceae. Volcanic soils.

Large shrub to 2.5cm tall. Leaves shiny, nerves depressed

above. Stipules green, very broad; flowers orange, clustered

in dense creamy coloured heads. Buds orange, only a few

flowers open per head. Sepals brown; corolla fleshy orange

below, cream above. Stigma white. Fruit green when

immature, turning orange.

Sidwell K. 416 26/Oct/1995

With: Etuge. Schoenengerger & Takele

Duplicates at:

New dets to M.Cheek at RBG Kew and HNC BP 1601 Yaounde.

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

Tesseract 4.0.0 | **K000025814_lbl07** | Score: 21/21

FLORA OF WESTERN CAMEROON

RBG Kew & Herbier National du Cameroun

support by Darwin Initiative & Earthwatch

Rubiaceae HOLOT wre

Psychotria Sp. BR df’ naa

Det. Clack (2) V.2002

Division: Kupe-Muanenguba South West

Gazette: Nyasoso

LonglLat: N; E Alt: 1190m

Above Nyasoso on Max's trail up Mount Kupe. Montane

forest with canopy to 35m tall. Many stands of

Marantaceae. Volcanic soils.

Large shrub to 2.5m tall. Leaves shiny, nerves depressed

above. Stipules green, very broad: flowers orange, clustered

in dense creamy coloured heads. Buds orange, only a few

flowers open per head. Sepals brown: corolla fleshy orange

below, cream above. Stigma white. Fruit green when

immature, turning orange.

Sidwell K. 416 26/0ct/1995

With: Etuge, Schoenengerger & Takele

Duplicates at: “J halo, Ya! Seal Wh | Br, no, FP

oe BRLYV Sua ya! EA, Cad

New dets to M.Cheek at RBG Kew and HNC BP 1601 Yaounde.

Tesseract 3.0.51 | **K000025814_lbl07** | Score: 18/21

FLORA OF WESTERN CAMEROON

RBG Kew & Herbler N atlonal du Cameroun

support by Darwm Inmatlve & Earthwatch

Rubiaceae H O LOT ‘7 176:

Psychotria g? B 0%., 'L; CA

M GAQQlLCL V1001

Division Kupe Muanenguba South West

Gazette Nyasoso

LongLat N E Alt 1190m

Above Nyasoso on Max 3 trail up Mount Kupe Montane

forest with canopy to 35m tall Many stands of

Marantaceae Volcanic soils

Large shrub to 2 5m tall Leaves shiny nerves depressed

above Stipules green very broad flowers orange Clustered

in dense creamy coloured heads Buds orange only a few

flowers open per head Sepals brown corolla fleshy orange

below cream above Stigma white Fruit green when

immature, turning orange

Sidwell K 416 26/Oct/1995

With Etuge Schoenengergel & Takele

Dupliuatesat k kn‘ol Yh‘ S(OJ NA&' BK; r10. P

B&Lv Pawn 1M EA C&«J'S

New dets to M Cheek at RBG Kew and HNC BP 1601 Yaounde

ABBYY FineReader Engine 12 | **K000025814_lbl07** | Score: 20.5/21

Uocot we

Rubiaceae

Psych otria Sr>. £>.

r

W_. CUcHCE) V.7DO2-

Division:

Kupe-Muanenguba

South West

Gazette:

Nyasoso

Alt: 11 90m

26/Oct/1995

i

ax co.

Large shrub to 2.5m tall. Leaves shiny, nerves depressed

above. Stipules green, very broad; flowers orange, clustered

in dense creamy coloured heads. Buds orange, only a few

flowers open per head. Sepals brown; corolla fleshy orange

below, cream above. Stigma white. Fruit green when

immature, turning orange.

Sidwell K. 416

With: Etuge, Schoenengerger & Takele

LongLat: N; E

Above Nyasoso on Max’s trail up Mount Kupe. Montane

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

forest with canopy to 35m tall. Many stands of

Marantaceae. Volcanic soils.

FLORA OF WESTERN CAMEROON

RBG Kew & Herbier National du Cameroun

support by Darwin Initiative & Earthwatch

Duplicates at:

New dets to M.Cheek at RBG Kew and HNC BP 1601 Yaounde.

Microsoft OneNote 2013 | **K000025814_lbl07** | Score: 20/21

FLORA OF WESTERN CAMEROON

RBG Kew & Herbier National du Cameroun

support by Darwin Initiative & Earthwatch

Rubiaceae

Psychotria s . B

cueucc

Division:

Kupe-Muanenguba

Gazette:

Nyasoso

LongLat: N; E

HO Co TYIIE

South West

Alt: 11 90m

Above Nyasoso on Max's trail up Mount Kupe. Montane

forest with canopy to 35m tall. Many stands of

Marantaceae. Volcanic soils.

Large shrub to 2.5m tall. Leaves shiny, nerves depressed

above. Stipules green, very broad; flowers orange, clustered

in dense creamy coloured heads. Buds orange, only a few

flowers open per head. Sepals brown; corolla fleshy orange

below, cream above. Stigma white. Fruit green when

immature, turning orange.

Sidwell K. 416

With: Etuge, Schoenengerger & Takele

26/0ct/1995

Duplicates at: k. (o (SC A (k'A& MO

New deus to M.Cheek at RBG Kew and HNC BP 1601 Yaounde.

9.5 Named Entity Recognition analysis

9.5.1 Named Entity gold standards

9.5.1.1 Image “B 10 0002520”

LOCATION	PERSON
Songea District Tanganyika Unangwa Hill Songea	E. Milne-Redhead P. Taylor

9.5.1.2 Image segment “B 10 0002520_lbl02”

LOCATION	PERSON
Tanganyika Songea District	E. Milne-Redhead P. Taylor

9.5.1.3 Image segment “B 10 0002520_lbl03”

LOCATION	PERSON
Songea Unangwa Hill	

9.5.1.4 Image “BM000500117”

LOCATION	PERSON
Tasmania	E. Kantvilas

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

Hobart	P.W. Jones

9.5.1.5 Image segment “BM000500117_lbl04”

LOCATION	PERSON
Tasmania Hobart	E. Kantvilas P.W. Jones

9.5.1.6 Image “E00015443”

LOCATION	PERSON
China Lijiang	

9.5.1.7 Image segment “E00015443_lbl03”

LOCATION	PERSON
China Lijiang	

9.5.1.8 Image “EIG.2770”

LOCATION	PERSON
Australia Helsinki (x2) Babinda Bellenden Ker	Ilkka Kukkonen

Queensland	

9.5.1.9 Image segment “EIG.2770_lbl02”

LOCATION	PERSON
Australia Helsinki Babinda Bellenden Ker Queensland	Ilkka Kukkonen

9.5.1.10 Image segment “EIG.2770_lbl03”

LOCATION	PERSON
Helsinki	

9.5.1.11 Image “K000025814”

LOCATION	PERSON
Kew Western Cameroon (x2) Cameroun Mount Kype Yaoundé Kupe-Muanenguba Nyasoso (x2)	Cheek (x2) Sonké Etuge Schoenengerger Takele M.Cheek Max

	Sidwell

9.5.1.12 Image segment “K000025814_lbl01”

LOCATION	PERSON
Western Cameroon	

9.5.1.13 Image segment “K000025814_lbl03”

LOCATION	PERSON
	Cheek (x2) Sonké

9.5.1.14 Image segment “K000025814_lbl07”

LOCATION	PERSON
Western Cameroon Kew Cameroun Kupe-Muanenguba Nyasoso (x2) Mount Kype Yaoundé	Sidwell Etuge Schoenengerger Takele M.Cheek Max

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

9.5.2 Stanford NER using OCR text from Whole Images

9.5.2.1 Image “B 10 0002520”

LOCATION	Lauila
LOCATION	Songea
LOCATION	Songea District
LOCATION	Unangwa Hill
PERSON	Poth

LOCATION		
TP	FP	FN
3	1	1

PERSON		
TP	FP	FN
0	1	2

9.5.2.2 Image “BM000500117”

LOCATION	Hobart
LOCATION	LONDON
PERSON	P.W. James TLC

LOCATION		
TP	FP	FN
2	0	1

PERSON		
TP	FP	FN
1	0	1

9.5.2.3 Image “E00015443”

LOCATION	B&L
LOCATION	China
LOCATION	Flora
PERSON	HERB

LOCATION		
TP	FP	FN
1	2	1

PERSON		
TP	FP	FN
0	1	0

9.5.2.4 Image “EIG.2770”

LOCATION	AUSTRALIA
LOCATION	Babinda
LOCATION	Bellenden Ker MUSEUM BOTANICUM National Park
LOCATION	HELSINKI
LOCATION	HELSINKI
LOCATION	Queensland
LOCATION	UNIVERSITATIS
PERSON	Ilkka Kukkonen

LOCATION		
TP	FP	FN
6	1	0

PERSON		
TP	FP	FN
1	0	0

9.5.2.5 Image “K000025814”

LOCATION	CAMEROON
LOCATION	CAMEROON
LOCATION	Cameroun
LOCATION	Mount Kupe
LOCATION	Yaounde
PERSON	Cleo
PERSON	HERB
PERSON	Max
PERSON	Montane

LOCATION		
TP	FP	FN
5	0	4

PERSON		
TP	FP	FN
1	3	8

9.5.3Stanford NER using OCR text from Segmented Images

9.5.3.1 Segments of Image “B 10 0002520”

B 10 0002520_lbl02.jpg	LOCATION	Skaw TANGANYIKA
B 10 0002520_lbl02.jpg	LOCATION	Songea District
B 10 0002520_lbl03.jpg	LOCATION	Songea
B 10 0002520_lbl03.jpg	LOCATION	Unangwa Hill

LOCATION		
TP	FP	FN
3	1	0

PERSON		
TP	FP	FN
0	0	2

9.5.3.2 Segments of Image “BM000500117”

BM000500117_bcd01.jpg	LOCATION	LONDON
BM000500117_lbl04.jpg	LOCATION	Hobart
BM000500117_lbl04.jpg	LOCATION	TASMANIA
BM000500117_lbl04.jpg	PERSON	P.W. James TLC

LOCATION		
TP	FP	FN
3	0	0

PERSON		
TP	FP	FN
1	0	1

9.5.3.3 Segments of Image “E00015443”

E00015443_bcd01.jpg	LOCATION	ROYAL BOTANIC GARDEN EDINBURGH
E00015443_lbl01.jpg	PERSON	HERB
E00015443_lbl02.jpg	PERSON	INCE
E00015443_lbl03.jpg	LOCATION	Lijiang
E00015443_lbl03.jpg	PERSON	Flora

LOCATION		
TP	FP	FN
1	1	1

PERSON		
TP	FP	FN
0	0	0

9.5.3.4 Segments of Image “EIG.2770”

EIG.2770_lbl02.jpg	LOCATION	AUSTRALIA
EIG.2770_lbl02.jpg	LOCATION	Babinda
EIG.2770_lbl02.jpg	LOCATION	Bellenden Ker National Park
EIG.2770_lbl02.jpg	LOCATION	HELSINKI
EIG.2770_lbl02.jpg	LOCATION	Queensland
EIG.2770_lbl02.jpg	LOCATION	UNIVERSITATIS

EIG.2770_lbl02.jpg	PERSON	Ilkka Kukkonen
EIG.2770_lbl03.jpg	LOCATION	HELSINKI

LOCATION		
TP	FP	FN
5	1	1

PERSON		
TP	FP	FN
1	0	0

9.5.3.5 Segments of Image “K000025814”

K000025814_lbl01.jpg	LOCATION	CAMEROON
K000025814_lbl06.jpg	LOCATION	KEW
K000025814_lbl06.jpg	PERSON	HERB
K000025814_lbl07.jpg	LOCATION	Cameroun
K000025814_lbl07.jpg	LOCATION	Mount Kupe
K000025814_lbl07.jpg	LOCATION	Yaounde
K000025814_lbl07.jpg	PERSON	BRLYV Sua
K000025814_lbl07.jpg	PERSON	Clack
K000025814_lbl07.jpg	PERSON	Max
K000025814_lbl07.jpg	PERSON	Montane

LOCATION		
TP	FP	FN
5	0	4

PERSON		
TP	FP	FN
1	4	0

TOTAL

LOCATION		
TP	FP	FN
17	3	6

PERSON		
TP	FP	FN
3	4	3

9.6 Non-standard Terminology Extraction analysis

9.6.1 FlexiTerm output using OCR text from Segmented Images

Coalesced OCR text output from segmented images listed in A3.2. OCR was performed using Tesseract 4.0.0.

Rank	Term representative	Score	Frequency
1	royal botanic gardens	63.1336	60
2	royal botanic	39.6827	61
3	royal botanic garden edinburgh	35.5238	27
4	MO	31.075	31
5	museum botanicum	20.7944	31
6	horti bot	18.0218	27
7	nationaal herbarium	12.8232	20
8	nationaal herbarium nederland	11.9016	12
9	mo at bm	9.7041	7
10	natural history	8.7337	14
11	royal botanic gardens ke	8.3178	6
12	herbarium nederland	7.8227	13
13	j. sarvela	7.6246	12
14	center herb	7.278	12
15	imaged gpi	6.9315	11
16	tvaoy royal botanic garden edinburgh	6.4378	4
17	herbier du jardin botanigue	6.317	7
18	herbarium bogoriense	6.065	10
19	imaged african plant initiative	5.5452	4
19	nn mo	5.5452	5
19	ethiopian flora project expedition	5.5452	4
19	missouri botanical garden herbarium	5.5452	5

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483

20	york botanical garden	5.4931	6
20	bogor herbarium of new identification	5.4931	5
21	b. bartholomew	4.852	7
21	m. watson	4.852	7
21	m. gilbert	4.852	7
22	flora of china sino-british qinghai expedition	4.8283	3
23	sino-british qinghai expedition	4.3944	7
23	national science foundation	4.3944	4
23	herbier musaum paris	4.3944	5
24	c.r. fraser-jenkins	4.1589	6
24	ilkka kukkonen	4.1589	7
24	royal botanic gardens kew	4.1589	4
24	american museum of natural history	4.1589	4
24	herbier musaum paris p02733867	4.1589	4
25	harvard university	4.0203	7
26	herbarium vadense	3.4657	5
26	west china	3.4657	5
26	george forrest	3.4657	6
26	plants of qinghai	3.4657	6
26	paasasai jybuadoo	3.4657	6
26	university of helsinki	3.4657	5
26	rain forest	3.4657	6
26	utm grid	3.4657	6
26	paris herbier	3.4657	6
26	department of botany	3.4657	6
27	universiteit leiden branch	3.2958	3

27	c. g. pringle	3.2958	3
27	thomas b. croat	3.2958	3
28	paniasal jybuuadoo	3.2347	6
29	alpine meadow	2.7726	5
29	arnold arboretum	2.7726	4
29	e.j. strangman	2.7726	4
29	r. steele	2.7726	4
29	voucher specimen	2.7726	5
29	t. reichstein	2.7726	4
29	flora of ethiopia	2.7726	8
29	e.j. weeda	2.7726	4
29	a. gray	2.7726	5
29	type specimen	2.7726	5
29	herbier du jardin botauique de eta	2.7726	3
30	paniasal ybuadoo ol	2.1972	3
30	line for computer entry	2.1972	3
31	yushu xian	2.0794	4
31	e side	2.0794	4
31	consolidated scree	2.0794	3
31	institute of botany	2.0794	4
31	academia sinica	2.0794	3
31	r.j.d. mcbeath	2.0794	3
31	flora of taiwan	2.0794	3
31	national taiwan	2.0794	3
31	jaakko sarvela	2.0794	3
31	western australian	2.0794	4
31	guinea expedition	2.0794	3

31	richard archbold	2.0794	3
31	det. vern	2.0794	3
31	east kalimantan	2.0794	4
31	herbarium wanariset	2.0794	3
31	bukit raya	2.0794	4
31	flora malesiana	2.0794	4
31	east african	2.0794	4
31	garden belgium	2.0794	4
32	perennial herb	1.7329	4
32	jybuadoo ol	1.7329	4
33	forest shade	1.3863	3
33	ll lll	1.3863	3
33	river styx	1.3863	4
33	rock crevices	1.3863	3
33	mixed forest	1.3863	3
33	montane forest	1.3863	3
33	herbario bogoriensi	1.3863	4
33	joint expedition	1.3863	3
33	h.p. nooteboom	1.3863	3
33	herbario kewensi	1.3863	3
33	flora of yemen	1.3863	3
33	r. melville	1.3863	3
34	flora of china	0.6931	4
34	kuo date	0.6931	3
34	g. pringle	0.6931	4

Funded by the Horizon 2020 Framework of the European Union
H2020-INFRADEV-2016-2017
Grant Agreement No 777483