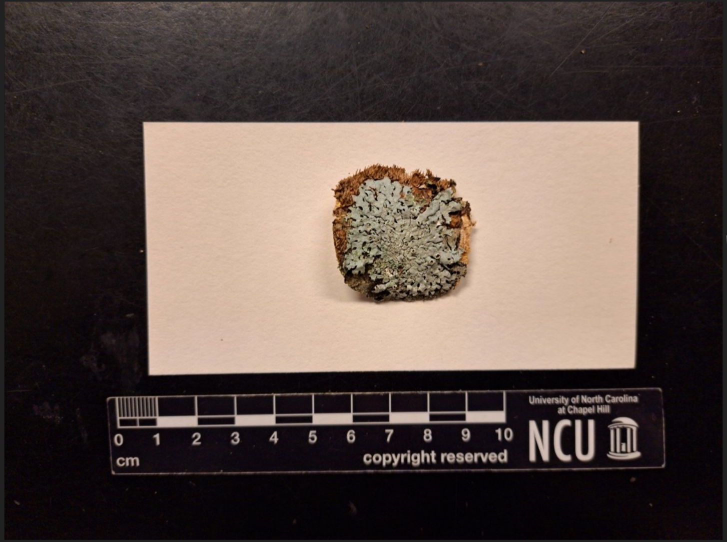
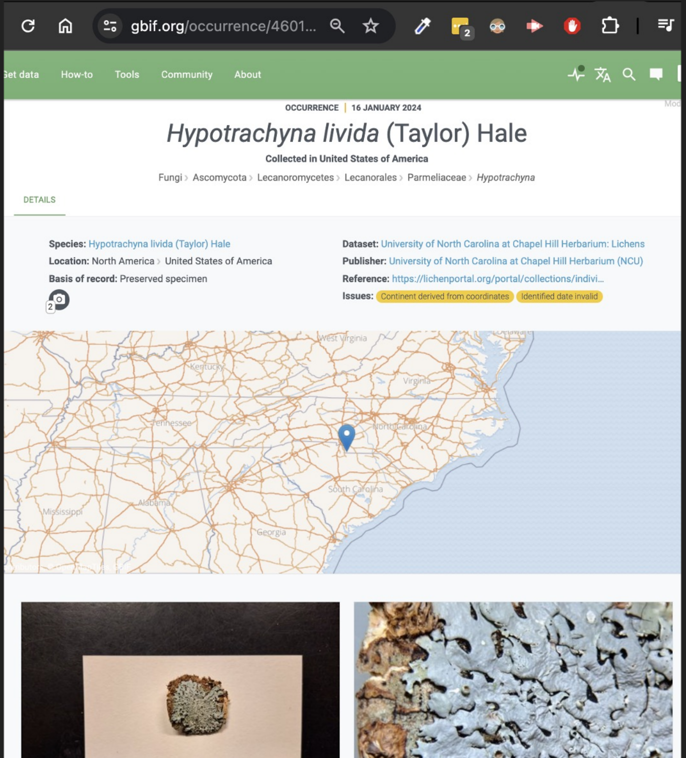
AI digitisation overview

AI digitisation

How to get from specimen…



…to published record, using AI



OCCURRENCE | 16 JANUARY 2024

Hypotrachyna livida (Taylor) Hale

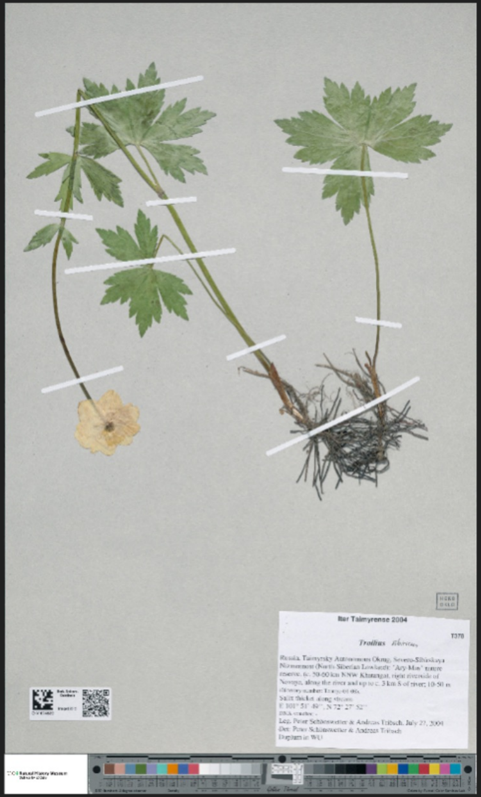
Collected in United States of America

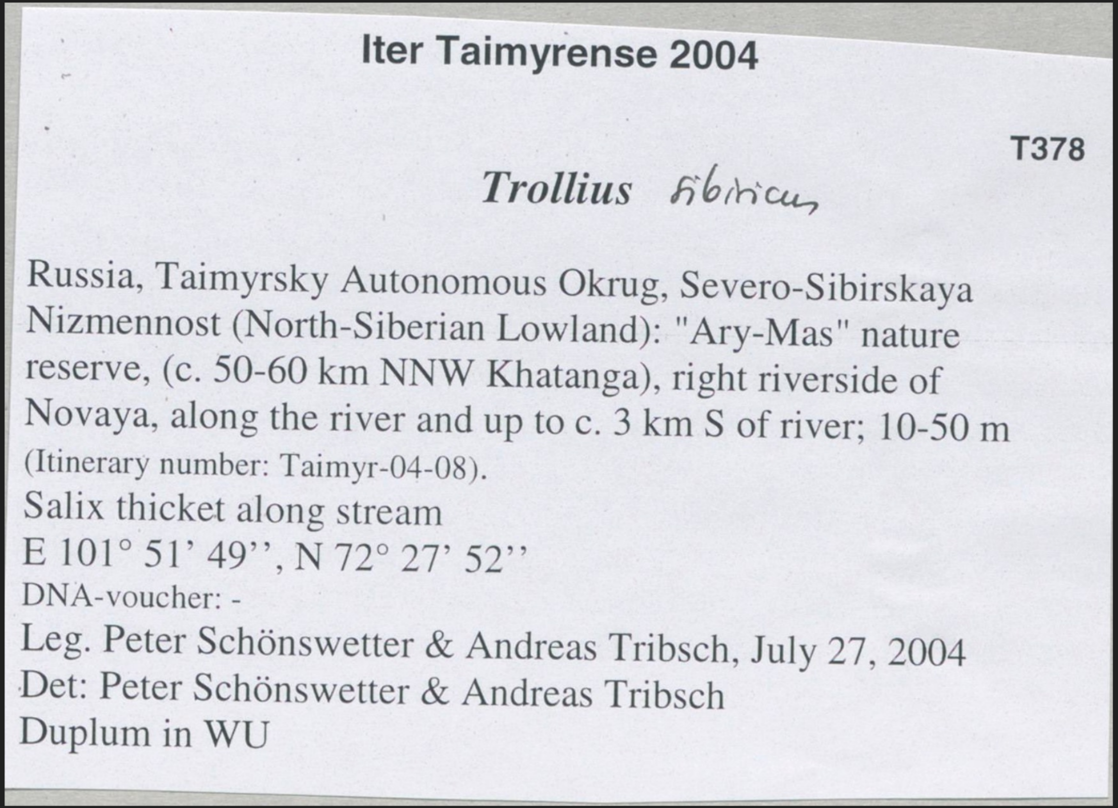
Fungi › Ascomycota › Lecanoromycetes › Lecanorales › Parmeliaceae › *Hypotrachyna*

Species: Hypotrachyna livida (Taylor) Hale
Location: North America › United States of America
Basis of record: Preserved specimen

Dataset: University of North Carolina at Chapel Hill Herbarium: Lichens
Publisher: University of North Carolina at Chapel Hill Herbarium (NCU)
Reference: https://lichenportal.org/portal/collections/indivi…
Issues: Continent derived from coordinates | Identified date invalid

1. Use AI (OCR) to get text from image





Iter Taimyrense 2004

T378

Trollius sibiricus

Russia, Taimyrsky Autonomous Okrug, Severo-Sibirskaya Nizmennost (North-Siberian Lowland): "Ary-Mas" nature reserve, (c. 50-60 km NNW Khatanga), right riverside of Novaya, along the river and up to c. 3 km S of river; 10-50 m (Itinerary number: Taimyr-04-08).

Salix thicket along stream

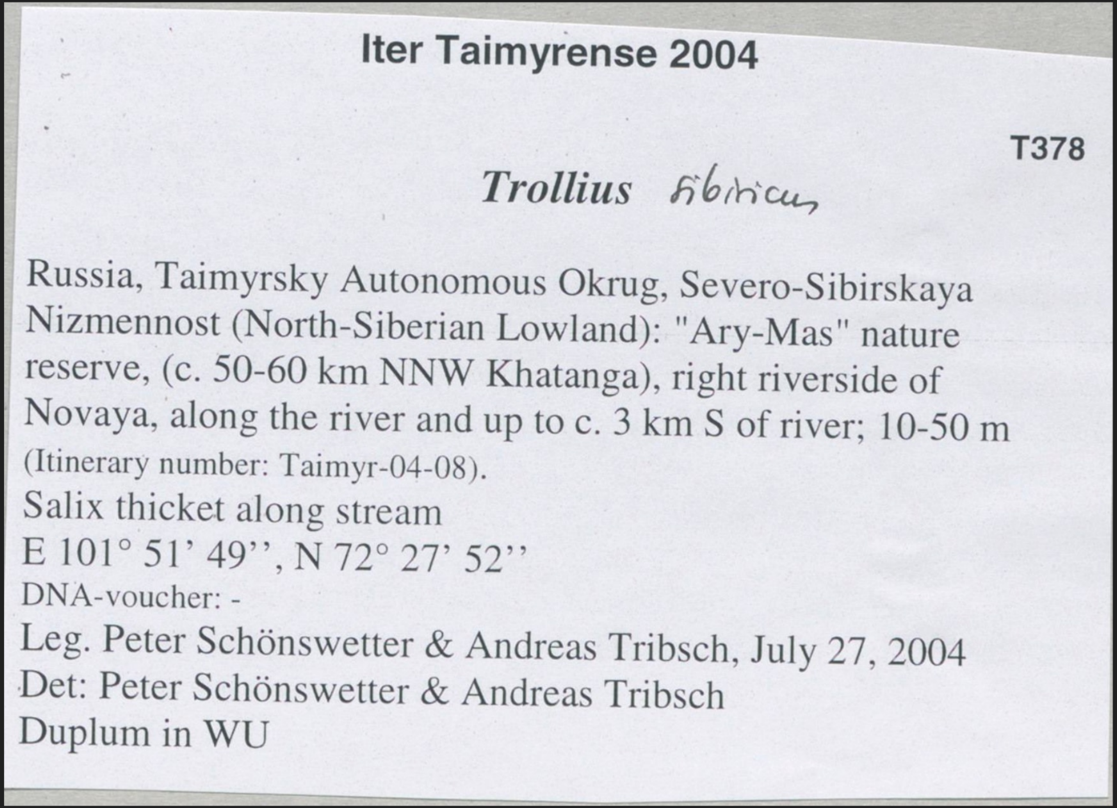
E 101° 51’ 49’’, N 72° 27’ 52’’

DNA-voucher: -

Leg. Peter Schönswetter & Andreas Tribsch, July 27, 2004

Det: Peter Schönswetter & Andreas Tribsch

Duplum in WU



Iter Taimyrense 2004

T378

Trollius sibiricu

Russia, Taimyrsky Autonomous Okrug, Severo-Sibirskaya Nizmennost (North-Siberian Lowland): "Ary-Mas" nature reserve, (c. 50-60 km NNW Khatanga), right riverside of Novaya, along the river and up to c. 3 km S of river; 10-50 m (Itinerary number: Taimyr-04-08).

Salix thicket along stream

E 101° 51’ 49’’, N 72° 27’ 52’’

DNA-voucher: -

Leg. Peter Schönswetter & Andreas Tribsch, July 27, 2004

Det: Peter Schönswetter & Andreas Tribsch

Duplum in WU

Iter Taimyrense 2004 T378

Trollius sibiricu

Russia, Taimyrsky Autonomous Okrug, Severo-Sibirskaya Nizmennost (North-Siberian Lowland): "Ary-Mas" nature reserve, (c. 50-60 km NNW Khatanga), right riverside of Novaya, along the river and up to c. 3 km S of river; 10-50 m (Itinerary number: Taimyr-04-08).

Salix thicket along stream

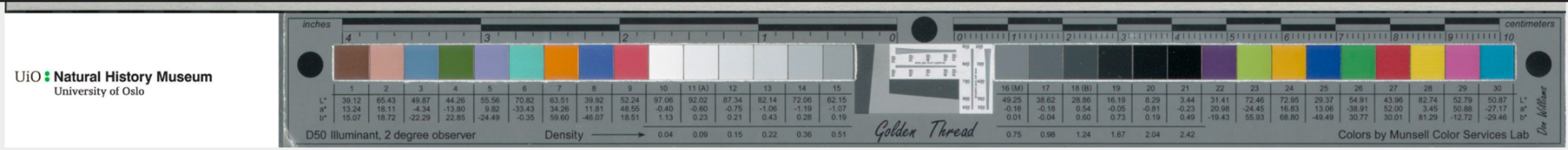
E 101° 51' 49'', N 72° 27' 52''

DNA-voucher: -

Leg. Peter Schönswetter & Andreas Tribsch, July 27, 2004

Unfortunately, OCR text also includes other non-label text

41230 1 2 3 4 5 6 7 8 9 10 L* 39.12 65.43 49.87 44.26 55.56 70.82 63.51 39.92 52.24 97.06 13.24 18.11 -4.34 -33.43 34.26 11.81 48.55 -0.40 18.72 -22.29 -0.35 59.60 -46.07 18.51 1.13 a* -13.80 9.82 22.85 -24.49 b\" 15.07 D50 Illuminant, 2 degree observer Density 0.04 11 (A) 92.02 -0.60 0.23 0.09 12 13 87.34 82.14 -0.75 -1.06 0.21 0.43 0.15 0.22 14 15 72.06 62.15 -1.19 -1.07 0.19 0.28 0.51 TeH'll! W 0.36 O 84899-8 # # # # 41230 Leucobrym juniperorden 600 200 300 B-66378 700400 800 500 850 550 Golden Thread Reg. 29.04. 2016 16 (M) 17 49.25 38.62 -0.16 -0.18 0.01 -0.04 0.75 0.98 18 (B) 28.86 0.54 0.60 1.24 19 16.19 -0.05 0.73 1.67 20 8.29 -0.81 0.19 2.04 21 3.44 -0.23 0.49 2.42 22 23 72.46 24 72.95 25 29.37 31.41 20.98 -24.45 16.83 13.06 -19.43 55.93 68.80 -49.49 SN: OL0222 centimeters 29 52.79 50.88 10 30 26 27 28 54.91 43.96 82.74 -38.91 52.00 3.45 50.87 -27.17 30.77 30.01 81.29 -12.72 -29.46 Colors by Munsell Color Services Lab 4:25 Don Williams"



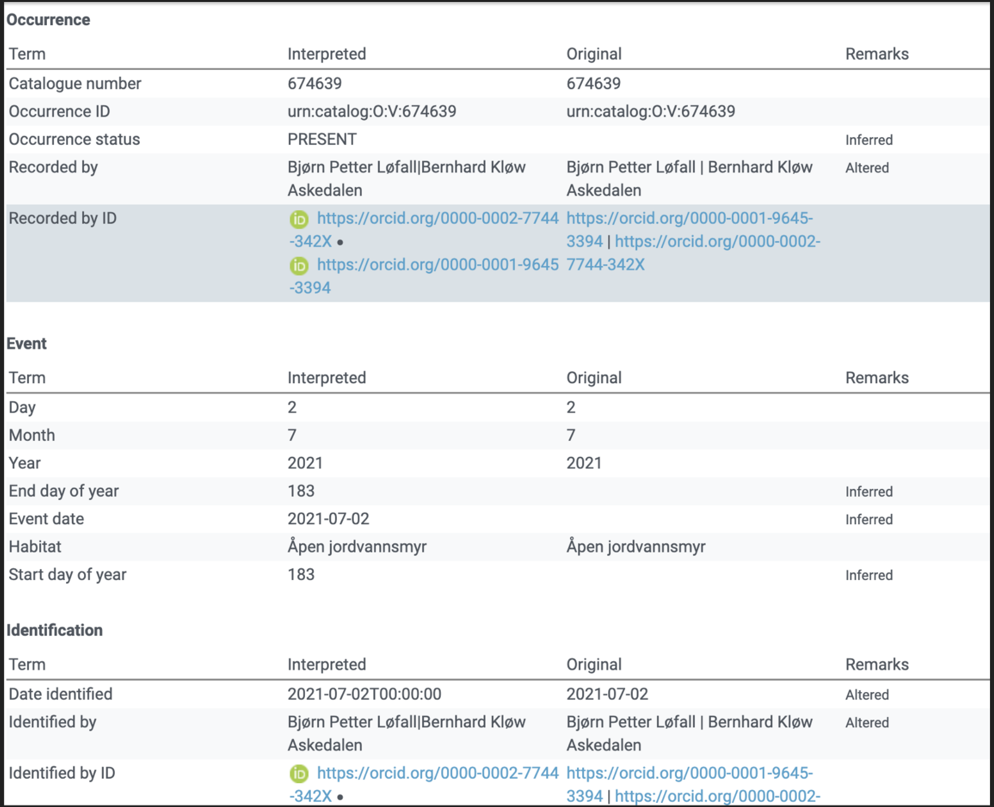
	1	2	3	4	5	6	7	8	9	10	11 (A)	12	13	14	15
L*	39.12	65.43	49.87	44.26	55.56	70.82	63.51	39.92	52.24	97.06	92.02	87.34	82.14	72.06	62.15
a*	13.24	18.11	-4.34	-13.80	9.82	-33.43	34.26	11.81	48.55	-0.40	-0.60	-0.75	-1.06	-1.19	-1.07
b*	15.07	18.72	-22.29	22.85	-24.49	-0.35	59.60	-46.07	18.51	1.13	0.23	0.21	0.43	0.28	0.19
Density →										0.04	0.09	0.15	0.22	0.36	0.51

D50 Illuminant, 2 degree observer

16 (M)	17	18 (B)	19	20	21	22	23	24	25	26	27	28	29	30	
49.25	38.62	28.86	16.19	8.29	3.44	31.41	72.46	72.95	29.37	54.91	43.96	82.74	52.79	50.87	L*
-0.16	-0.18	0.54	-0.05	-0.81	-0.23	20.98	-24.45	16.83	13.06	-38.91	52.00	3.45	50.88	-27.17	a*
0.01	-0.04	0.60	0.73	0.19	0.49	-19.43	55.93	68.80	-49.49	30.77	30.01	81.29	-12.72	-29.46	b*
0.75	0.98	1.24	1.67	2.04	2.42										

Colors by Munsell Color Services Lab

2. Send OCR text to a LLM to extract structured data



Occurrence

Term	Interpreted	Original	Remarks
Catalogue number	674639	674639	
Occurrence ID	urn:catalog:O:V:674639	urn:catalog:O:V:674639	
Occurrence status	PRESENT		Inferred
Recorded by	Bjørn Petter Løfall|Bernhard Kløw Askedalen	Bjørn Petter Løfall | Bernhard Kløw Askedalen	Altered
Recorded by ID	https://orcid.org/0000-0002-7744-342X • https://orcid.org/0000-0001-9645-3394	https://orcid.org/0000-0001-9645-3394 | https://orcid.org/0000-0002-7744-342X	

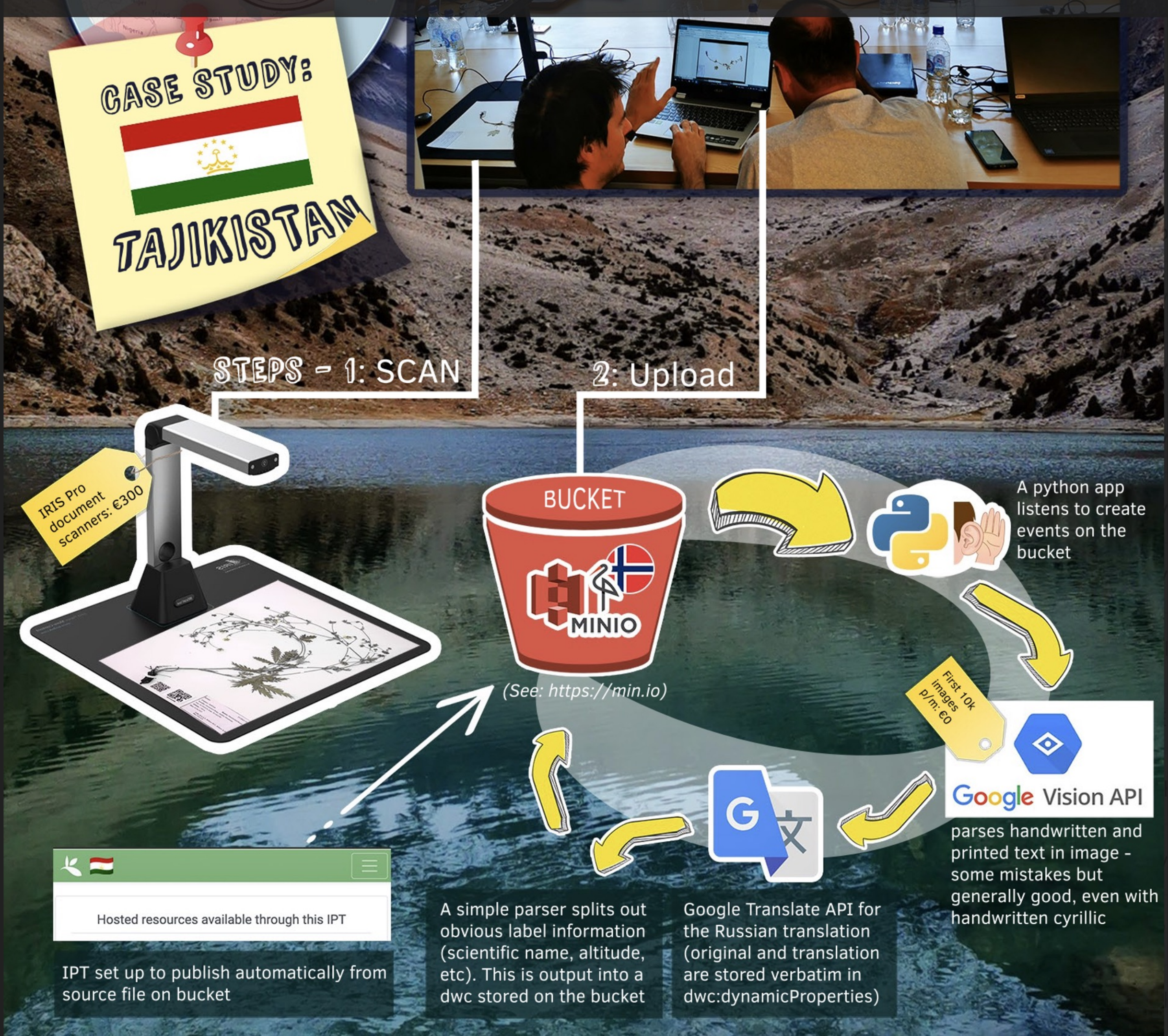
Event

Term	Interpreted	Original	Remarks
Day	2	2	
Month	7	7	
Year	2021	2021	
End day of year	183		Inferred
Event date	2021-07-02		Inferred
Habitat	Åpen jordvannsmyr	Åpen jordvannsmyr	
Start day of year	183		Inferred

Identification

Term	Interpreted	Original	Remarks
Date identified	2021-07-02T00:00:00	2021-07-02	Altered
Identified by	Bjørn Petter Løfall|Bernhard Kløw Askedalen	Bjørn Petter Løfall | Bernhard Kløw Askedalen	Altered
Identified by ID	https://orcid.org/0000-0002-7744-342X •	https://orcid.org/0000-0001-9645-3394 | https://orcid.org/0000-0002-	

CASE STUDY: TAJIKISTAN



STEPS - 1: SCAN

2: Upload

IRIS Pro document scanners: €300

(See: https://min.io)

A python app listens to create events on the bucket

First 10k images p/m: €0

Google Vision API parses handwritten and printed text in image - some mistakes but generally good, even with handwritten cyrillic

Google Translate API for the Russian translation (original and translation are stored verbatim in dwc:dynamicProperties)

A simple parser splits out obvious label information (scientific name, altitude, etc). This is output into a dwc stored on the bucket

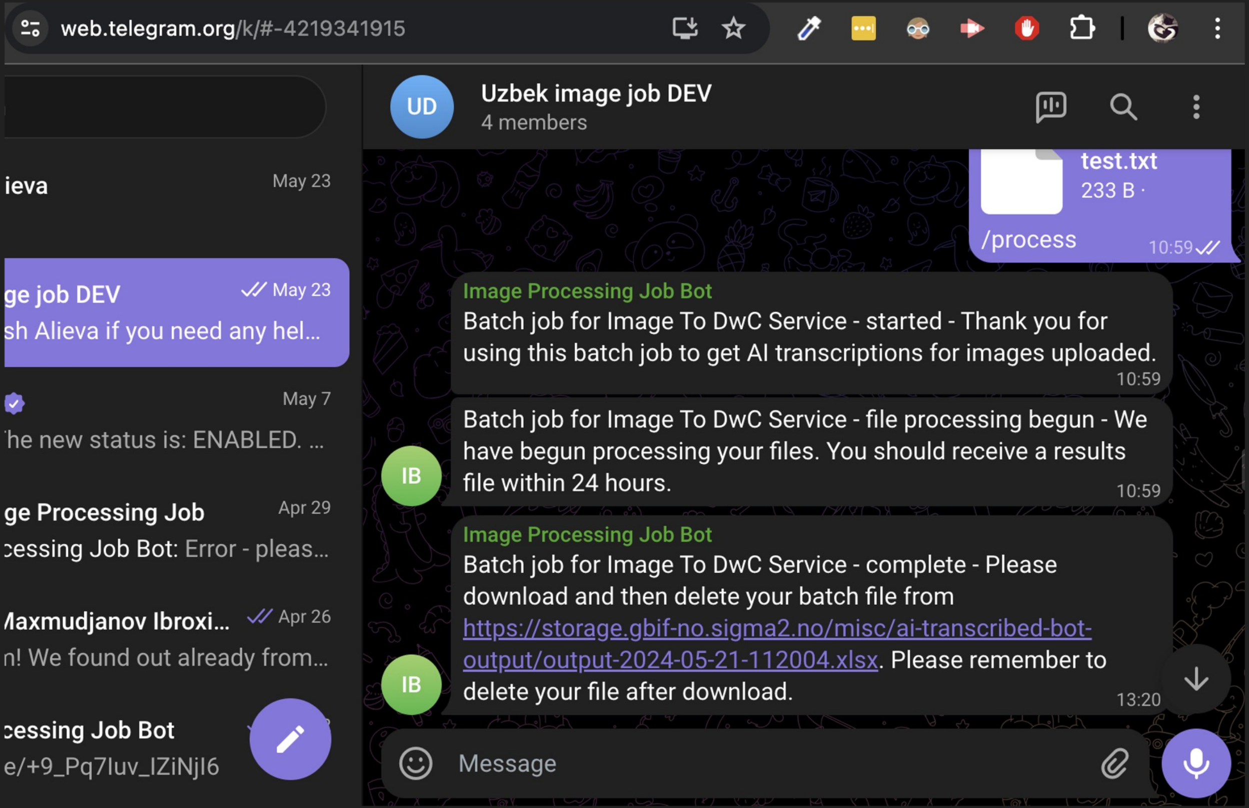
Hosted resources available through this IPT

IPT set up to publish automatically from source file on bucket

Uzbekistan

They publish and host their own images, publishing happens manually.

Step 1: Host images
Step 2: Generate list of images and send to telegram bot
Step 3: Receive back Darwin Core and check it
Step 4: Upload to IPT



/process 10:59

Image Processing Job Bot
Batch job for Image To DwC Service - started - Thank you for using this batch job to get AI transcriptions for images uploaded. 10:59

Batch job for Image To DwC Service - file processing begun - We have begun processing your files. You should receive a results file within 24 hours. 10:59

Image Processing Job Bot
Batch job for Image To DwC Service - complete - Please download and then delete your batch file from https://storage.gbif-no.sigma2.no/misc/ai-transcribed-bot-output/output-2024-05-21-112004.xlsx. Please remember to delete your file after download. 13:20

NHM @ UiO

Mainly used for quality checking or in an ad hoc way to speed up digitisation

https://gbif-norway.github.io/label-classification-gpt/python-interactive/code/github_pages/index-uio.html

E.g. Bjørn Petter sends a prepared dataset for QA or a series of catalog numbers

We run it through OCR + LLMs, and send it back to him

He checks it

Complications

- Image hosting
- Data publication via the IPT or another system
- Do we encourage manual verification of each record? How much should get checked?
- Rerunning with newer models

Workflows

Every workflow will differ, but common factors:

- A dedicated programmer
  - Optional pipeline 1: Method to extract list of images/catalogue numbers for processing
  - Pipeline 2: Images in, OCR out
  - Pipeline 3: OCR in, DwC out
  - Optional pipeline 4: DwC to publication platform
- Image hosting
- Some data publication platform (probably an IPT)

Ways of improving this

- Every time we run a specimen image through OCR, we are basically creating information which can be considered as an “annotation” to the specimen

- Every time we run that OCR through an LLM to extract structured data, we are creating another machine annotation

  Wouldn’t it be cool if we had a system where we could see each other’s annotations and add to them/suggest corrections?

DiSSCo annotation system (and over to Sam on Zoom)

- Specimens in Europe are covered by DiSSCo:
  - The DiSSCo RI aims to create a new business model for one European collection that digitally unifies all European natural science assets under common access, curation, policies and practices
- DiSSCo is building a framework - a data model - to capture annotations alongside specimen data and ensure FAIRness.
  - Aims to accommodates commenting, editing, and data improvement
- https://uio.zoom.us/j/4769565894?pwd=TWg5L05vZnJNbWl1R3lyZ3R2Zk13Zz09

https://dissco.tech/2024/01/14/the-data-model-behind-disscos-annotation-service/