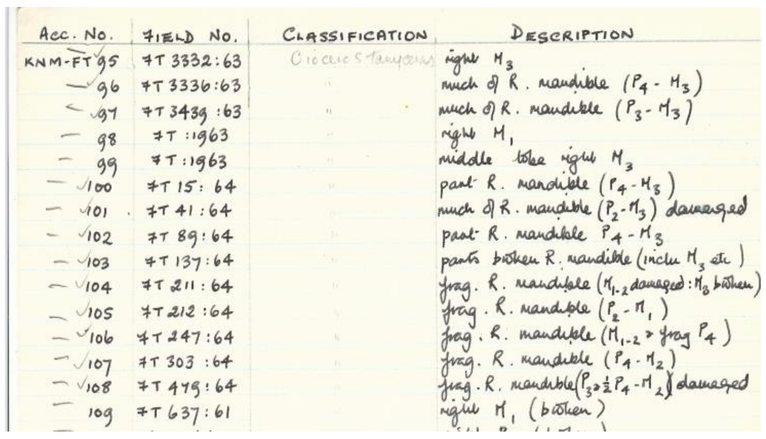
Digitization of handwritten fossil catalogues of the National Museums of Kenya

Riikka Korolainen

Overall goal: build a digital structured database for the fossil samples stored in the National Museums of Kenya.
MSc thesis goal: build an automated system for correcting and standardizing dental element notation. The data is corrected by checking each word and re-reading those that are suspected to be dental markings with a set of specialist classifiers.

1. Extract words with optical character recognition

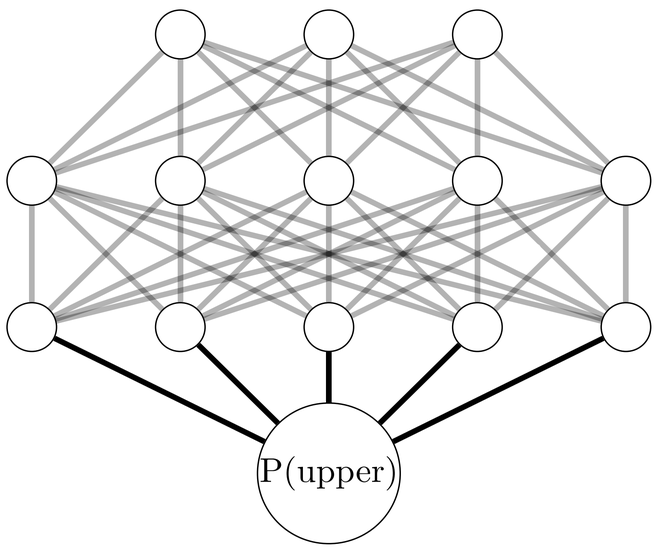
(Azure Vision API)



Acc. No.	Field No.	Classification	Description
KNM-FT 95	7T 3332:63	Ció ceros Tamyoua	right M_3
96	7T 3336:63		much of R. mandible (P_4 - H_3)
97	7T 3439 :63		much of R. mandible (P_3 - M_3)
98	7T :1963		right M_1
99	#T :1963		middle tobe right M_2
100	++ 15: 64		pant R. mandible (P_4 - M_3)
101	7T 41 : 64		much of R. mandible (P_2 - M_3) damaged
102	7T 89 : 64		pant R. mandible P_4 - M_3
103	7T 137 : 64		parts broken R. mandible (inclu M_3 etc)
104	7T 211 : 64		frag. R. mandible (M_{1-2} damaged: M_3 bother)
105	7T 212 :64		frag . R. mandible (P_2 - M_1)
106	7T 247 :64		frag. R. mandible (M_{1-2} > frag P_4)
107	7T 303 :64		frag. R. mandible (P_4 - M_2)
108	7T 419 : 64		frag. R. mandible(P_3>½P_4 - M_2) damaged
109	7T 637 : 61		right M_1 (broken)


```
"lines": [
    },
    {
        "text": "much of R. mandible (P4- H3)",
        "boundingPolygon": [
            {
                "x": 1385,
                "y": 284
            },
            {
                "x": 2233,
                "y": 282
            },
            {
                "x": 2233,
                "y": 358
            },
            {
                "x": 1385,
                "y": 359
            }
    ],
    "words": [
        {
            "text": "much",
            "boundingPolygon": [
                {
                    "x": 1388,
                    "y": 295
                },
                {
                    "x": 1507
```


Convolutional neural network, eg. EfficientNet [1]
Transfer learning: Freeze all but last layer(s), train to classify images of dental markings



Output: class probability

2. Table Inference

Data Science Project, spring 2024

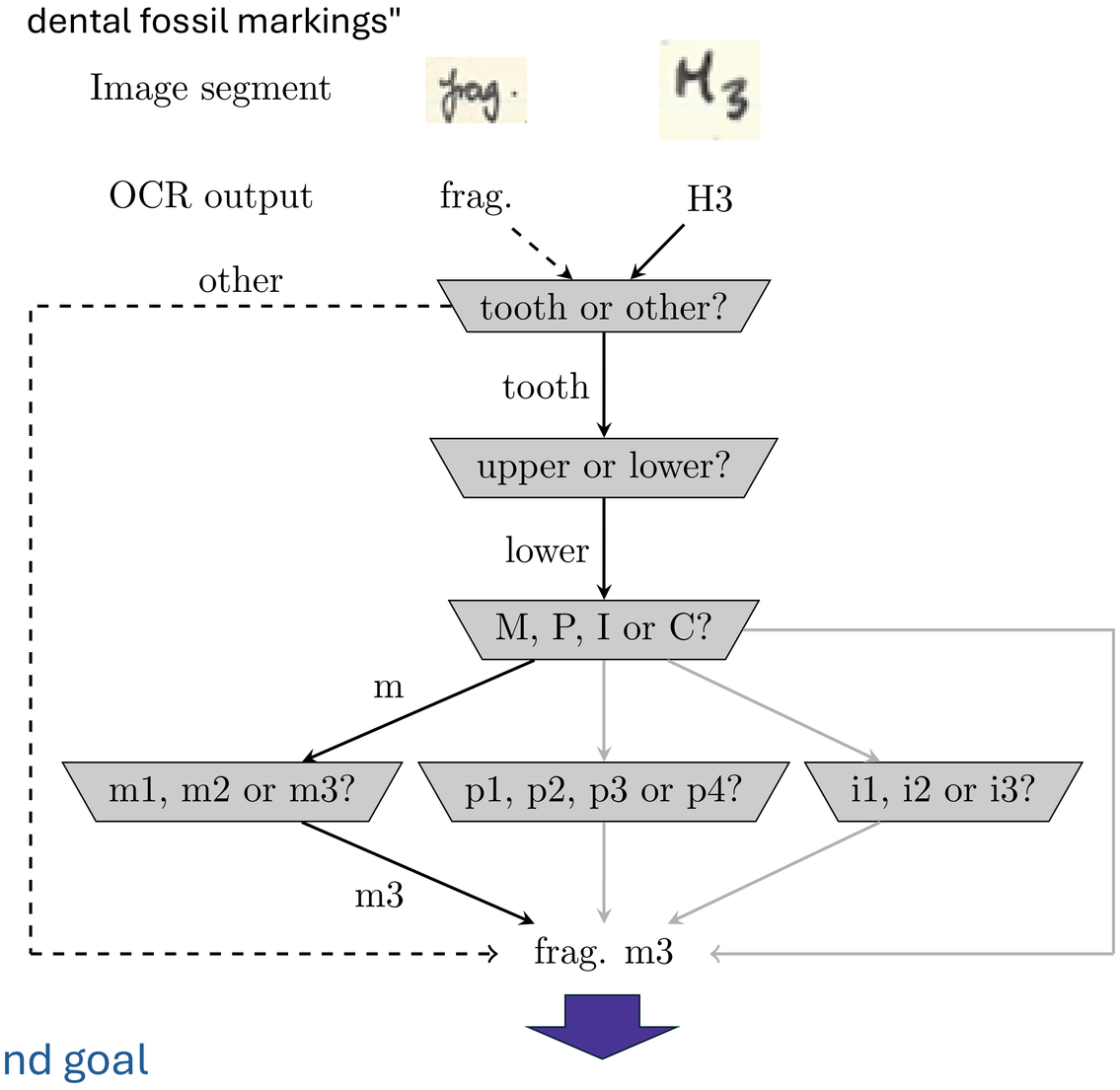
Rows: DBSCAN clustering
Columns: Hard-code common header names
Find headers on page
Assign each word under the header with x-direction middle point closest to word's x-direction middle point

ACC_NO	FIELD_NO	CLASSIFICATION	DESCRIPTION	TOOTH_RECORDS
KNM-FT 95	7T 3332:63	Ció ceros Tamyoua	M3	
96	7T 3336:63	much	of R. mandible (P4- H3)	
	7T 3439 : 63	much	of R. mandible (P3-M3)	
98	7T : 1963	right	M.	
99	#T : 1963		middle tobe right M2	
100	++ 15:64	pant	R. mandible (P4-M3)	
101	7T 41:64	much	of R. mandible (P2-M3) damaged	
102	7T 89:64	paul-	R. Mandible P4-M3	
−103	7T137:64	parts	broken R. mandible (inclu M2 etc)	
−104	7T 211: 64	frag.	R. mandible (M1, damaged: M3 bother)	
105	7T 211: :64		frag. R. Mandible (P. M.)	
~106	7T247:64	Jag.	R. mandible (M1-2 > frag P4)	
/107	#T 303 :64		Jurag. R. Mandible (PA-M2)	
V108	7T 419:64	Jag	. R. mandible (P == P4-M 2) damaged	
9	IT 637:61	right	M (broken)	

Unsolved: how to get teeth in description cleaned to tooth_records?

3. Element description cleaning

MSc thesis in Data Science, "Fine-tuned optical character recognition for dental fossil markings"



End goal

ACC_NO	FIELD_NO	CLASSIFICATION	DESCRIPTION	TOOTH_RECORDS
KNM-FT 95	7T 3332:63	Ció ceros Tamyoua	m3	(m3)
96	7T 3336:63	much	of R. Mandible (p4- m3)	(p4, m3)
	7T 3439 : 63	much	of R. Mandible (p3-m3)	(p3, m3)
98	7T : 1963	right	m1	(m1)
99	#T : 1963		middle tobe right m3	(m3)
100	++ 15:64	pant	R. Mandible (p4-m3)	(p4, m3)
101	7T 41:64	much	of R. Mandible (p2-m3) damaged	(p2, m3)
102	7T 89:64	paul-	R. Mandible p4-m3	(p4, m3)
−103	7T137:64	parts	broken R. Mandible (inclu m3 etc)	(m3)
−104	7T 211: 64	frag.	R. Mandible (m1-2, damaged: m3 bothe	(m1, m2, m3)
105	7T 211: :64		frag. R. Mandible (p4. M2.)	(p2, m1)
~106	7T247:64	Jag.	R. Mandible (m1-2 > frag p4)	(m1, m2, p4)
/107	#T 303 :64		Jurag. R. Mandible (p4-m2)	(p4,m2)

Challenges

- Errors in column division
- Errors in line splitting to words by Azure
  - imperfect inputs to downstream models

Suggestion for future work: word bounding box detection model trained on fossil catalogues

Institutional address

University of Helsinki, Faculty of Science
P. O. Box 68 (Pietari Kalmin katu 5)
00014 University of Helsinki

References

[1] Tan M, Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning 2019 May 24 (pp. 6105-6114). PMLR.

Acknowledgements

Data Science project supervision:
Kari Lintulaakso (LUOMUS),
Stephen Maikweki (ICT officer, National Museums of Kenya)
Data Science project group:
Max Väistö, Riikka Korolainen, Janne Tuukkanen, Yinong Li and Axel Wester.
Thesis supervisor:
Indrė Žliobaitė (Professor, University of Helsinki, Department of Computer Science)