

# Mining for Information in Texts from the Cultural Heritage

Marieke van Erp  
<http://ticc.uvt.nl/mitch>





Continuous  
Access  
to  
Cultural  
Heritage

Piroska Lendvai



Steve Hunt



Marieke van Erp



CONTINUOUS

Access

to

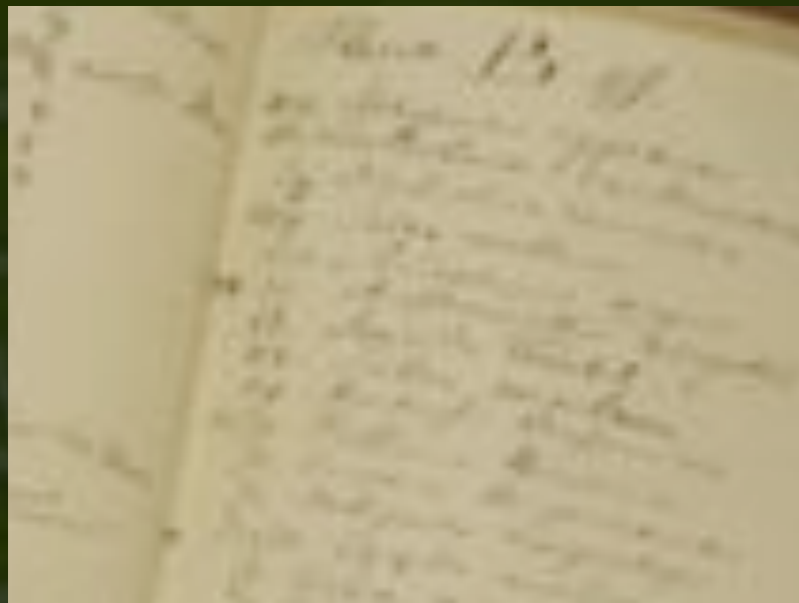
Cultu

terita

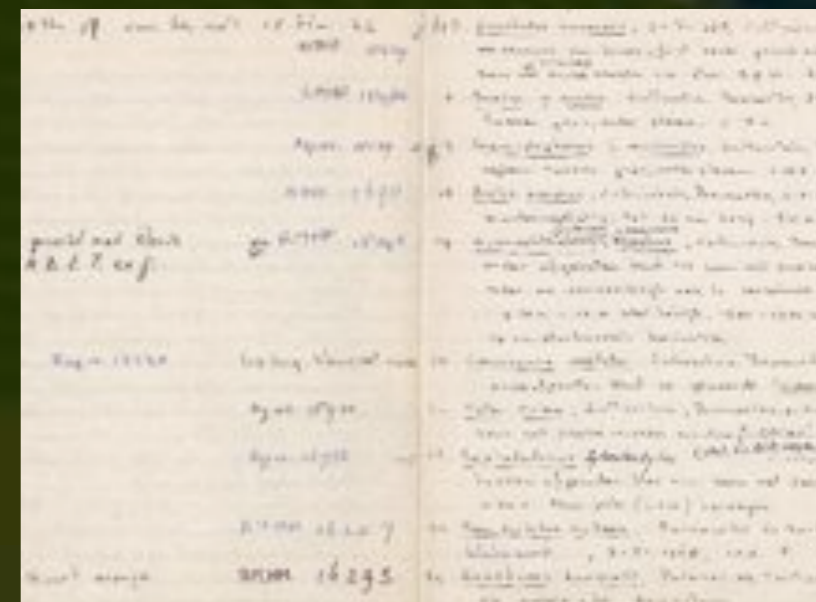
...

...





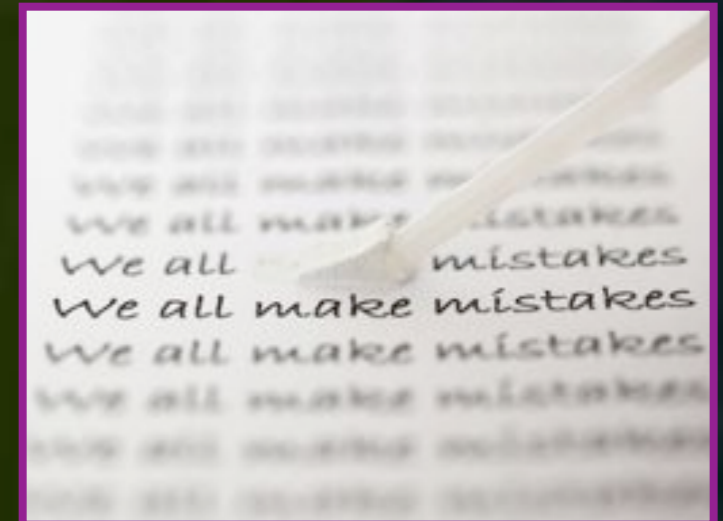
data	
<i>Anolis cristatellus</i>	DUMERIL & BIBRON
<b>Polychrotidae, Sauria (lizards)</b>	
<i>Anolis cristatellus</i> DUMERIL & BIBRON 1837: 143	
<i>Ptychocheilus (Istiocercus) cristatellus</i> — FITZINGER 1843: 6	
<i>Xiphosurus cristatellus</i> — O'SHAUGHNESSY 1875: 271	
<i>Anolis cristatellus</i> — BOULENGER 1885: 26	
<i>Anolis linderi</i> RUTHVEN 1912	
<i>Anolis cozumelae</i> SMITH 1939	
<i>Anolis cozumelae</i> — SMITH & TAYLOR 1950: 59	
<i>Otenonotus cristatellus</i> - GUYER & SAVAGE 1986	
<i>Anolis cristatellus cristatellus</i> DUMERIL & BIBRON 1837	
<i>Anolis cristatellus wiloyae</i> GRANT 1931	
Common Puerto Rican anole, Crested anole	
Puerto Rico (including many offshore islands), Isla Vieques, Isla Culebra, Isla Gorda, Isla Santa Rosa, Isla Santa Catalina, Isla Santa Rosa, Isla Santa Rosa	



- 16,870 records describing characteristics and history of animal specimens in a natural history database
- 39 columns
- Dutch, English, German and Portuguese
- numeric and textual values (both atomic and elaborate)

column Name	value
order	Anura
genus	Megophrys
country	Indonesia
biotope	in rain near road
collection date	01.02.1888
type	holotype
determinator	A. Dubois
defined by	(Linnaeus, 1758)
special remarks	in bad condition, was eaten by <i>Leptodactylus rugosus</i> (3023) at night and thrown up again the next morning when killed, partly digested

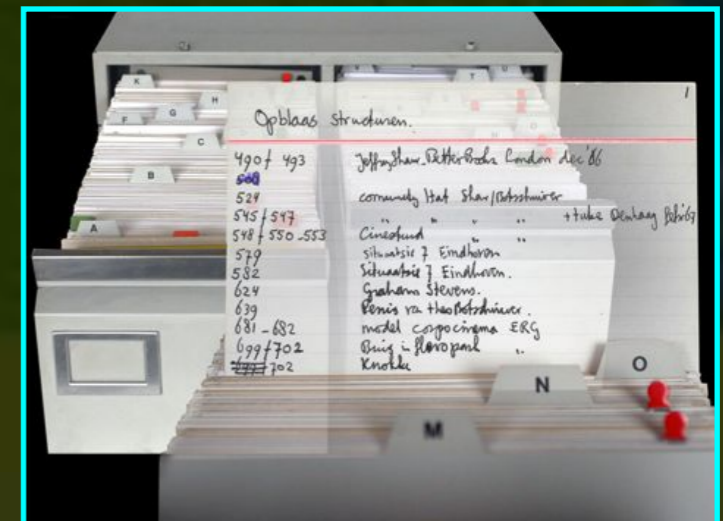
data cleaning



data structuring



data retrieval



data cleaning



author	determinator	family	genus	country	preservation method
(Daudin, 1802)	-----	Bataguridae	Anolis	Cambodja	(shield, dry)
(Schlegel)	G. vd. Boog	Colubridae	Geophis	Indonesia	-----
Schneider	M. S. Hoogmoed	-----	Bufo	Suriname	-----
(Horst, 1883)	Tyler, M J	Hylidae	Litoria	-----	alcohol

author	determinator	family	genus	country	preservation method
(Daudin, 1802)	-----	Bataguridae	Anolis	Cambodja	(shield, dry)
(Schlegel)	G. vd. Boog	Colubridae	<b>Geophis</b>	Indonesia	-----
Schneider	M. S. Hoogmoed	-----	Bufo	Suriname	-----
(Horst, 1883)	Tyler, M J	Hylidae	Litoria	-----	alcohol

actual value: Geophis

author	determinator	family	genus	country	preservation method
(Daudin, 1802)	-----	Bataguridae	Anolis	Cambodja	(shield, dry)
(Schlegel)	G. vd. Boog	Colubridae	?	Indonesia	-----
Schneider	M. S. Hoogmoed	-----	Bufo	Suriname	-----
(Horst, 1883)	Tyler, M J	Hylidae	Litoria	-----	alcohol

actual value: Geophis

author	determinator	family	genus	country	preservation method
(Daudin, 1802)	-----	Bataguridae	Anolis	Cambodja	(shield, dry)
(Schlegel)	G. vd. Boog	Colubridae	?	Indonesia	-----
Schneider	M. S. Hoogmoed	-----	Bufo	Suriname	-----
(Horst, 1883)	Tyler, M J	Hylidae	Litoria	-----	alcohol

actual value: Geophis

author	determinator	family	genus	country	preservation method
(Daudin, 1802)	-----	Bataguridae	Anolis	Cambodja	(shield, dry)
(Schlegel)	G. vd. Boog	Colubridae	?	Indonesia	-----
Schneider	M. S. Hoogmoed	-----	Bufo	Suriname	-----
(Horst, 1883)	Tyler, M J	Hylidae	Litoria	-----	alcohol

actual value: Geophis

author	determinator	family	genus	country	preservation method
(Daudin, 1802)	-----	Bataguridae	Anolis	Cambodja	(shield, dry)
(Schlegel)	G. vd. Boog	Colubridae	?	Indonesia	-----
Schneider	M. S. Hoogmoed	-----	Bufo	Suriname	-----
(Horst, 1883)	Tyler, M J	Hylidae	Litoria	-----	alcohol

actual value: Geophis

author	determinator	family	genus	country	preservation method
(Daudin, 1802)	-----	Bataguridae	Anolis	Cambodja	(shield, dry)
(Schlegel)	G. vd. Boog	Colubridae	?	Indonesia	-----
Schneider	M. S. Hoogmoed	-----	Bufo	Suriname	-----
(Horst, 1883)	Tyler, M J	Hylidae	Litoria	-----	alcohol

actual value: Geophis  
predicted value: Rhapsophis

author	determinator	family	genus	country	preservation method
(Daudin, 1802)	-----	Bataguridae	Anolis	Cambodja	(shield, dry)
(Schlegel)	G. vd. Boog	Colubridae	?	Indonesia	-----
Schneider	M. S. Hoogmoed	-----	Bufo	Suriname	-----
(Horst, 1883)	Tyler, M J	Hylidae	Litoria	-----	alcohol

- <100 cells to check for a column instead of 16,780
- recall (estimate): 90-100%
- one-size-fits-all

CONTINUOUS  
Access

subject	relation	object
specimen collection	occurs before	entry in museum
species	has broader term	genus
city	falls within	country

# CONTINUOUS Access

- detects inconsistencies database usage
- small scope
- high recall and precision within scope

CONTINUOUS  
ACCESS

data structuring



CONTINUOUS  
ACCESS



	number	reference	preservation method	country	location	collector	class	order	genus	coll. date
1	3	Daudin, 1802	alcohol	Suriname	Paramaribo	M. S. Hoogmoed	Reptilia	Sauria	Anolis	28.08.1968
2	1	Spix, 1825	alcohol	Surinam	Raleigh Cataracts, Coppename River	K.W.R. Zwart	Reptilia	Sauria	Kentropyx	24.12.1968
3	1	Linnaeus	alcohol		Sipaliwini, between Base Bivouac and Meyers' farm	M. S. Hoogmoed	Amphibia		Bufo	02.03.1951
4	1		alcohol	Suriname	Galibi	M. S. Hoogmoed		Sauria		15.04.1951
5	1	Linnaeus, 1758	alcohol	Surinam		F. G. Mees	Amphibia		Anura	

Amphibia

Anura

## Frog

From Wikipedia, the free encyclopedia

*For other uses, see Frog (disambiguation).*

The frog is an **amphibian** in the order **Anura** (meaning "tail-less", from Greek an-, without + oura, tail), formerly referred to as Salientia (Latin saltare, to jump). The name frog derives from Old English frogga,<sup>[1]</sup> (compare Old Norse frauki, German Frosch, older Dutch spelling kikvorsich), cognate with Sanskrit plava (frog), probably deriving from Proto-Indo-European praw = "to jump".<sup>[2]</sup>

Most frogs are characterized by long hind legs, a short body, webbed digits (fingers or toes), protruding eyes and the absence of a tail. Most frogs have a semi-aquatic lifestyle, but move easily on land by jumping or climbing. They typically lay their eggs in puddles, ponds or lakes, and their larvae, called tadpoles, have gills and develop in water. Adult frogs follow a carnivorous diet, mostly of arthropods, annelids and gastropods. Frogs are most noticeable by their call, which can be widely heard during the night or day, mainly in their mating season.






The distribution of frogs ranges from tropic to subarctic regions, but most species are found in tropical rainforests. Consisting of more than 5 000

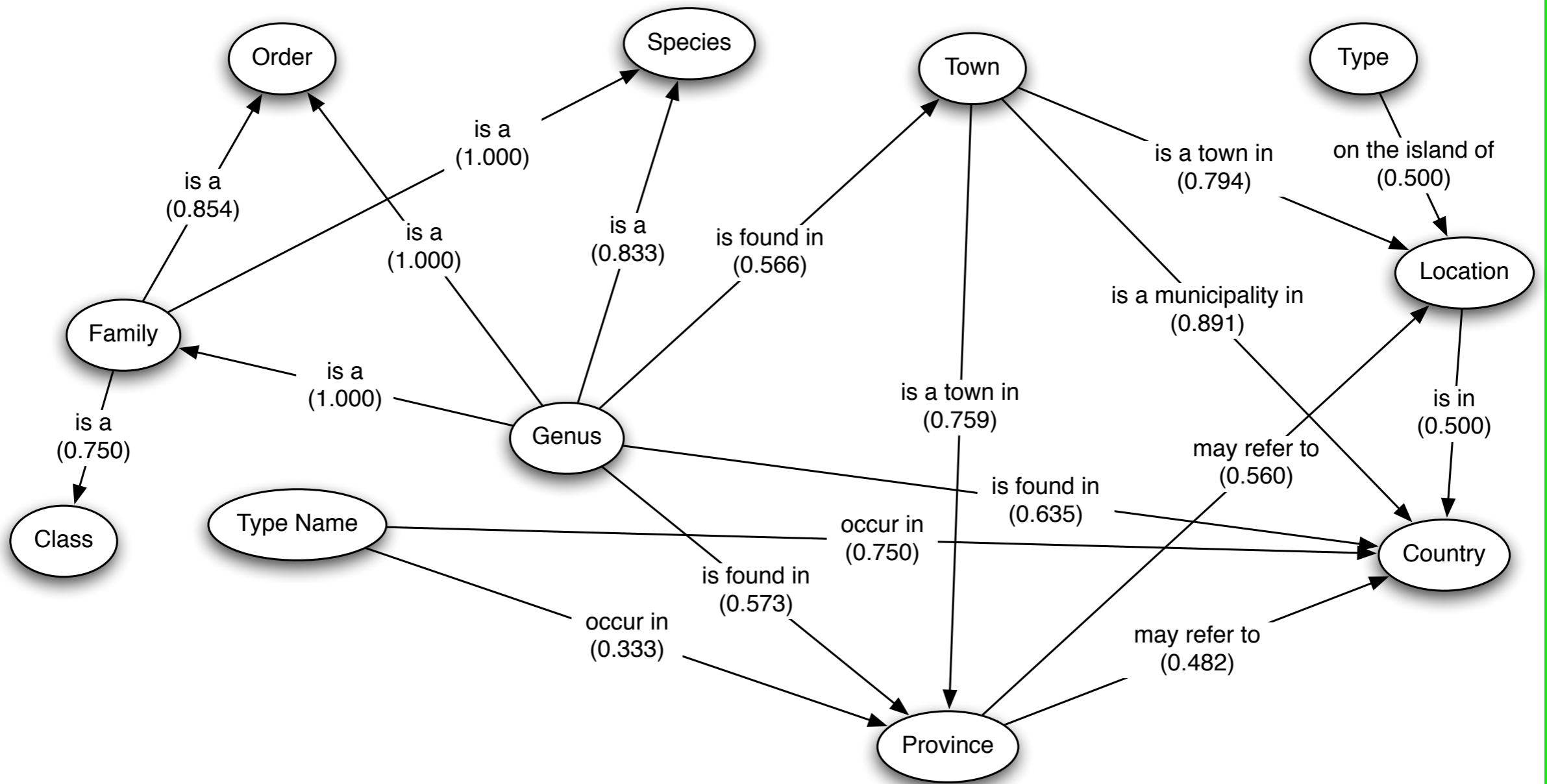
Frogs are most noticeable by their call, which can be widely heard during the night or day, mainly in their mating season.

Larvae, called tadpoles, have gills and develop in water. Adult frogs follow a carnivorous diet, mostly of arthropods, annelids and gastropods. Frogs have a semi-aquatic lifestyle, but move easily on land by jumping or climbing. They typically lay their eggs in puddles, ponds or lakes, and their



# relation candidates for town and country

direction	relation candidate	frequency	rating
	is a municipality and a town in	45	+
	is a municipality and a city in	19	+
	is a municipality in	10	+
	is one of the five districts of	5	-
	is the name of two provinces in	5	-



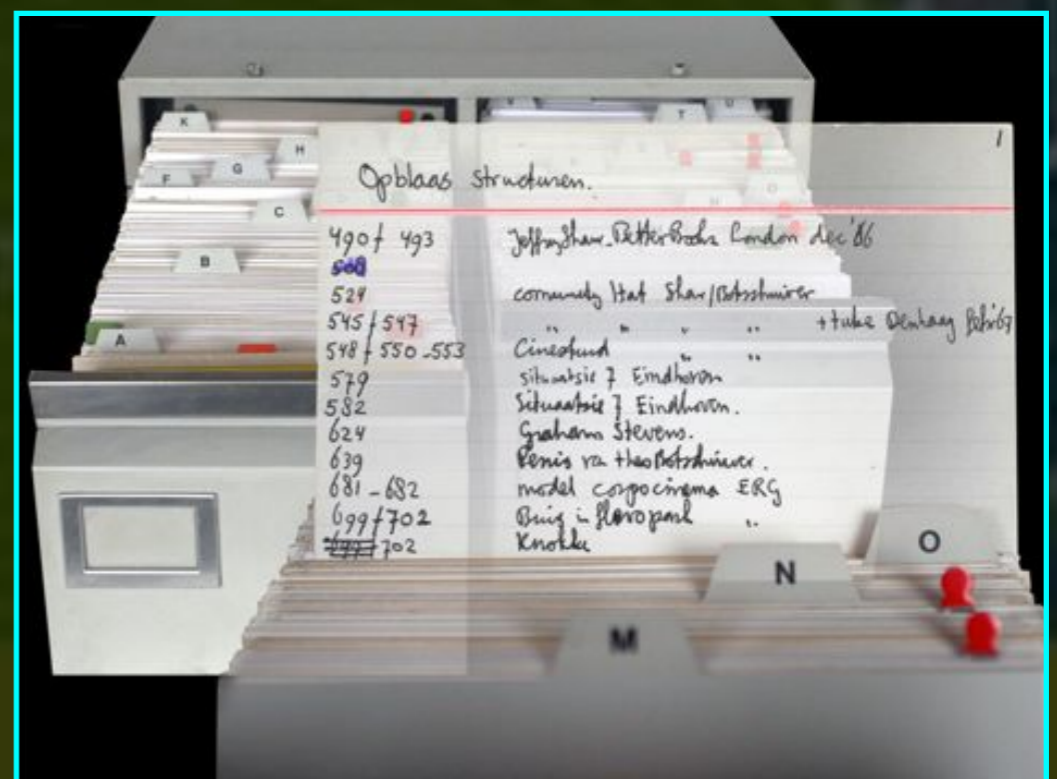
PROVINCE

- Some time in captivity. Collected on 24-09-1995 and died on 24-12-1995. Homopus juv. was born on 26-01-1997 and died 02-02-1997. The egg was laid on 10-10-1996. Other info same as RMNH 27497
- Slides MSH 1975-XVIII-27/29, 1975-XIX-20/25; tape recording 1975 II B 297-304. Acquired as gift from the British Museum (Nat. Hist.), BMNH 1975. 1348

born	found as egg, hatched 15-04-1972
died	killed in April 1998
formerly	formerly determined as <i>Lygosoma temmincki</i>
length	length approx. 1.75 m
loan	loaned to Dr. X on 28-09-1979
museum	gift from the British Museum
slide	slides MSH 1975-xviii-27/29, 1975-xix-20/25
tank	in jar with 27946

# CONTINUOUS ACCESS TO Cultural Heritage

data retrieval



pollen

○ Data  
● Field

Results 1-13 of 13 for 'pollen'

[Start New Entry From Fieldbook \(23523\)](#)

*Dia* s *Physshadens* 9 *mpten* 2 van *kyong* E . R , , *poetje* *lange* *weg* *Tema* - *Kyong* , *kwakend* *in* *ondergelopen* *stake* *land* *water* *water* [ *plus* *minus* ] *5* *cm* *hoog* *stond* , *tussen* *pollen* *gras* *in* *water* , *savannah* *15* - *VI* - *1972* , *21* . *00* - *23* . *30* *u* , *luchttemperatuur* *25* . *50* , *80* *m* *boven* *zeer* *niveau* [LEFT] *RMNH* *23523* *1* *1* *m* *2000* *Opgenomen* *op* *land* *Ghana* *1* .

[Start New Entry From Fieldbook \(\)](#)

*Lacerta* *muralis* 2 [ *man* ] [ *man* ] *Frankrijk* , *6* - *VII* - *1976* , *16* . *00* *u* , *op* *muur* *boven* *droge* *kreekbedding* , *in* *pollen* *van* *vegetatie* , *in* *dors* , [ *plus* *10* *m* , *1* [ *stuk* ] *u* . *M. S. Hoogmoed* . *Gedroogd* [ *nembuta* ] *15* - *VII* - *1976* , *op* *13* - *VII* - *1976* *in* *vangzakje* *7* *eieren* *gevonden* *in* *een* *groepje* *van* *4* *aan* *elkaar* *gevonden* *gekweekt* *en* *in* *eer* *van* *3* .

[Start New Entry From Fieldbook \(\)](#)

*Leptodactylus* *Fussus* *1* *hgr* . *Canario* , *3* *km* *Z* . *O* . *van* *kamp* , *bij* *vlegveld* , *30* - *II* - *1978* , *12* . *00* *u* , *springend* *in* *vochtige* *wizard* *savanne* , *tussen* *pollen* *vegetatie* . *85* *m* . *M. S. Hoogmoed* . *Gedroogd* [ *nembuta* ] *15* - *VII* - *1976* , *op* *13* - *VII* - *1976* *in* *vangzakje* *7* *eieren* *gevonden* *in* *een* *groepje* *van* *4* *aan* *elkaar* *gevonden* *gekweekt* *en* *in* *eer* *van* *3* .

*Leptodactylus* *Fussus* *1* *hgr* . *Canario* , *3* *km* *Z* . *O* . *van* *kamp* , *bij* *vlegveld* , *30* - *II* - *1978* , *12* . *00* *u* , *springend* *in* *vochtige* *wizard* *savanne* , *tussen* *pollen* *vegetatie* . *85* *m* . *M. S. Hoogmoed* . *Gedroogd* [ *nembuta* ] *15* - *VII* - *1976* , *op* *13* - *VII* - *1976* *in* *vangzakje* *7* *eieren* *gevonden* *in* *een* *groepje* *van* *4* *aan* *elkaar* *gevonden* *gekweekt* *en* *in* *eer* *van* *3* .

bE  
bB

# CONTINUOUS ACCESS

- query interpretation
- query expansion
- result ranking



- *Are there any specimens of species Dipsas catesbeyi from Guyana and Venezuela in the collection?*
- ▶ `all(Dipsas,catesbeyi,any(Guyana,Venezuela))`

- *Dendrophis pictus*

- ▶ any(all(Dendrophis,pictus),all(Dendrelaphis,inornatus))

- *Spanje*

- ▶ any(Spanje,España,Spain,Espanha)

- rank matches from genus and species fields higher

- recall: from 31.67% to 83.30%
- unanswered queries: from 52 to 6
- MAP: from 28.28% to 42.57%

- data cleaning is essential
- “digitising” a heritage institution is complicated
- don't try to tame text

# CONTINUOUS Access

- roll out
- scale-up
- stay involved



- [IEEEIS09] Antal van den Bosch, Marieke van Erp, and Caroline Sporleder (2009) Making a Clean Sweep of Cultural Heritage. IEEE Intelligent Systems, Special Issue on AI and Cultural Heritage March/April 2009 (vol. 25 no. 2), pp. 54-63
- [LaTeCH09] Marieke van Erp, Antal van den Bosch, Sander Wubben and Steve Hunt (2009) Instance-Driven Discovery of Ontological Relation Labels. Proceedings of EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education (LaTeCH - SHELT&R 2009), Athens, Greece, March 30, 2009
- [CiCLing08] Piroska Lendvai (2008) Alignment-based Expansion of Textual Database Fields. In: A. Gelbukh (Ed.), Proceedings of the Computational Linguistics and Intelligent Text Processing 9th International Conference, CICLing 2008. Lecture Notes in Computer Science, Vol. 4919/2008, Berlin / Heidelberg: Springer, pp. 522-531.

<http://ticc.uvt.nl/mitch>