

Workflow and Examples of Geographic Information Extraction with Large Language Models(LLMs) from Biodiversity Research Materials

Zheping Xu xuzp@mail.las.ac.cn

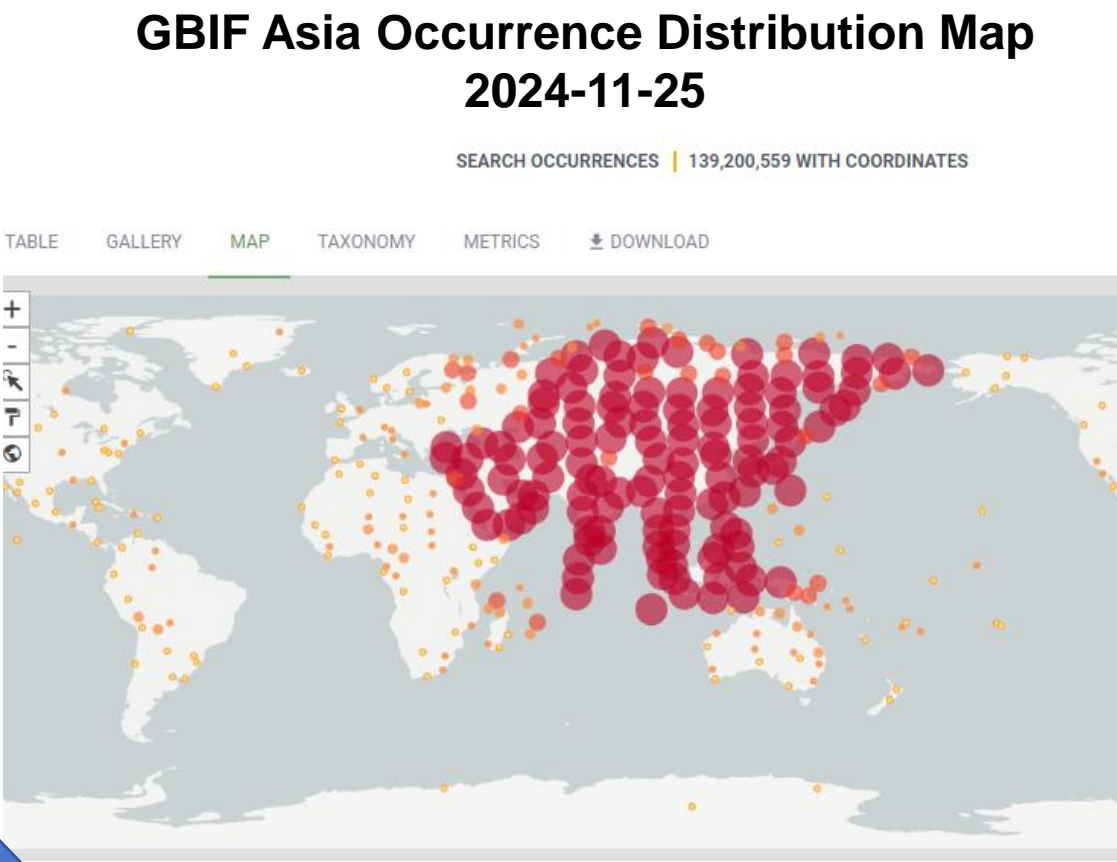
National Science Library, Chinese Academy of Sciences
The Third International Symposium of Mapping Asia Plants
Beijing, China, 2024-11-25

Outline

- **Background**
- Specimen Example
- Publication Example
- Discussion

Background

Statistics	All Data	Map Data	Percent(%)
Global	3,024,921,444	2,874,125,000	95.0
Global-Plant	497,634,292	428,991,179	86.2
Global-Animal	2,435,321,508	2,370,055,500	97.3
Asia	151,010,861	139,200,559	92.2
Asia-Plant	22,256,559	16,517,202	74.2
Asia-Animal	122,766,504	118,591,174	96.6



!!! Plant data has more gaps than others

Background: Specimen

OCCURRENCE | 14 JULY 2024

Sagina nodosa (L.) Fenzl

Knotted Pearlwort In English Collected in Russian Federation

Plantae > Tracheophyta > Magnoliopsida > Caryophyllales > Caryophyllaceae > *Sagi*

DETAILS

Species: *Sagina nodosa* (L.) Fenzl

Location: Russian Federation

Basis of record: Preserved specimen



Dataset: Moscow University Herbarium (MW)

Publisher: Lomonosov Moscow State Unive

Reference: <https://plant.depo.msu.ru/open/public/item/MW10925...>



Plant Species: *Sagina nodosa* (L.) Fenzl

Location:

- Latitude: 60°0'55" N
- Longitude: 29°43'3" E
- Area: Sankt-Peterburg, Kotlin District, Protective Dam D5 and Kronstadt Highway

Growth Environment: Wet Grassland

Growth Season: June to August

Seed Maturation Period: July

Collection Date: 14.07.2024

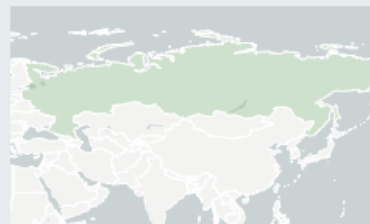
Collected by: D. Bochkov

Sample Number: 2433

Coordinates missing

This record is published without coordinates, but it includes a textual description of its location.

Location: Russian Federation



Sagina nodosa (L.) Fenzl

60°0'55" N 29°43'3" E

Санкт-Петербург, о. Котлин, защитная
дамба Д5, щебнистый участок близ
пересечения КАД и Кронштадтского шоссе

14.07.2024

собр. Бочков Д.

2433

Background: Literature

ع من الذباب الابيض لأول مره قى بغداد

ایمان محمد المانو
قسم وقایة النبات - كلية الزراعة
جامعة بغداد
بغداد / العراق

صور على نباتات الرازقي *Jasminum samba* خلال الموسم 2006-2007. وينتشر الذباب
بعه من الحشرات الى رتبة نصفيّة الاجنحه Hemiptera وتحت رتبة Sternorrhyncha
161جنس في العالم , تصيب هذه الحشرات مدى واسع من العوائل النباتية وبعض الانواع تكون
في جميع انحاء العالم , لها اربعة اطوار حورية الاطوار الثلاثة الاخيره غير متحركة وتشبه
بالعذراء pupa يعتمد في تصنيف الانواع والاجناس لعائلة الذباب الابيض على دراسة مايسمى
حوري الرابع , وان البحوث والدراسات جاريه ومستمره في تسجيل انواع جديده من هذه المجموعه
Bemisia التي تعود الى تحت عائلة Aleyrodinae تصيب الحمضيات ونباتات الزينة. سجلت في
اليه , سجلت الحشرة لأول مره في محافظة بغداد على نباتات الزينة (الرازقي) *Jasminum*
تخدام المفاتيح التصنيفيه ورسمت بواسطة كاميرا لوسيدا بالاعتماد على غلاف العذراء لهذه
من مقدمة الجسم ويصل الى بداية الفتحه الانبوبيه .

The Iraqi Journal of Agricultural Sciences 40 (6):17-20 (2009) NEW RECORD OF WHITEFLY SPECIES

Iman M. AL-Malo
Department of Plant Protection/ College
University of Baghdad

Новые местонахождения

O. A. Попова^{1*}, Н. А. Чашина
¹Забайкальский государственный университет
E-mails: olga.popova-54@yandex.ru
²Национальный парк «Чикойн»

³Бурятский государственный университет

Ключевые слова: Забайкальский край, флора, редкие растения.

Аннотация. В статье приводятся новые находки растений, впервые найденных в Забайкальском крае для флоры Забайкальского края и включены в Красную книгу Забайкальского края. Включены в Красную книгу Забайкальского края (2017), а *Artemisia furcata* – в Кра для территории края будет рекомендована для сохранения видов сосудистых растений.

New localities of

O. A. Popova

Flora of the vascular plants of the Baekdudaegan

Seung Hyun HWANG*, Yuji Asano

Department of Life Science, Jeonju National University
¹Natural Environment
²Department of Life Science, Jeonju National University
(Received 12 February 2015)

ABSTRACT: Baekdudaegan, the large mountainous area in the north of Korea, has served as a glacial refugium. This study was conducted for a total of twelve days in 2015. A list of vascular plants in the area was compiled, and a total of 441 taxa, consisting of 100 endemic taxa, 74 floristic regional indicator taxa, and 267 naturalized plants were distributed. The result of the survey is expected to be a basis for the development of conservation plans for the Baekdudaegan conservation area.

Keywords: Baekdudaegan conservation area, floristic regional indicator taxa, naturalized plants.

백두대간은 백두산에서 시작하여 금강산, 소백산을 거쳐 지리산으로 이어지는 총 길이 1,400km에 달해 동아시아 지역에서 가장 긴 큰 산줄기이다 (Kang et al., 2018). 백두대간은 한대성 식물과 온대성 식물이 공존하는 한반도 생물다양성의 핵심 지역이다.

園学雑誌 (J. Japan. Soc. Hort. Sci.) 73 (4): 364–373. 2004.

日本に自生する食用ヤマノイモ類の葉の形態と交雑親和性による分類

浅野裕司*

愛知県立農業大学校 444-0802 愛知県岡崎市美合町

Classification of Edible, Indigenous *Dioscorea* Species in Japan by Leaf Morphology and Cross-compatibility

Yuji Asano

Agricultural Academy of Aichi Prefecture, Okazaki, Aichi 444-0802

Summary

To speciate the edible, indigenous *Dioscorea* in Japan, 425 plants collected from 24 regions were classified as *Dioscorea japonica* Thunb., although some plants have leaf morphology characteristic of *Dioscorea opposita* Thunb. Typical *D. japonica* specimen were collected from 21 regions, whereas those with characteristics of *D. opposita* were collected from 17 regions. The former were collected exclusively from seven regions with mild climate, while the latter were collected from three regions with cold climate, indicating that the two species were adapted to different climatic zones. When *D. japonica* and cultivated *D. opposita* were hand-pollinated to observe the hybrid seed formation, few normal seeds were produced, indicating cross-incompatibility. Similarly, *D. japonica* and plants with characteristics of *D. opposita* were also incompatible, producing no normal seeds. Many crosses within *D. japonica* and those between *D. japonica* with characteristics of *D. opposita* and locally cultivated *D. opposita* produced viable seeds. These findings indicate that cross-incompatible plants with characteristics of *D. opposita* should not be classified as *D. japonica* but rather as *D. opposita* or its close relative.

Key Words: classification, compatibility, *Dioscorea japonica*, *Dioscorea opposita*, leaf shapes.

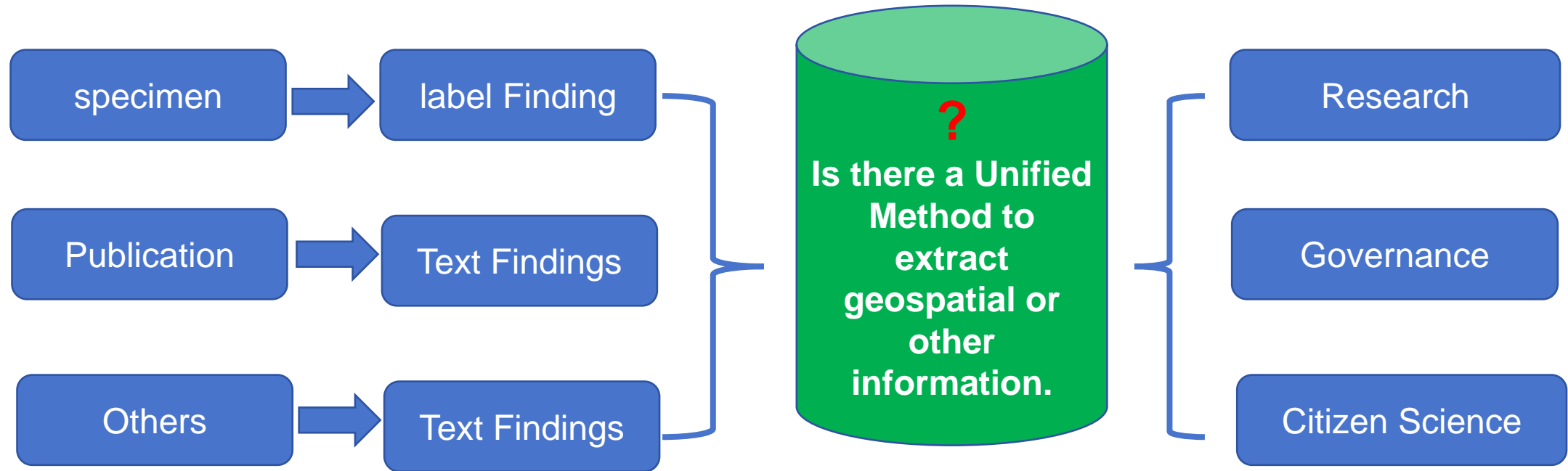
結 言

日本で食用にされるヤマノイモ類は、主にジネンジョ (*Dioscorea japonica* Thunb.) とナガイモ (*Dioscorea opposita* Thunb.) であり、これらは Enantiophyllum 節に属する。中国原産のナガイモは日本各地で栽培され、イモの形状により、長形のナガイモ群、扁形のイチョウイモ群および塊状のツクネイモ群に分類される。これに対し、ジネンジョは東北以南の山野に自生するとされ、古来より食用に供されてきた。

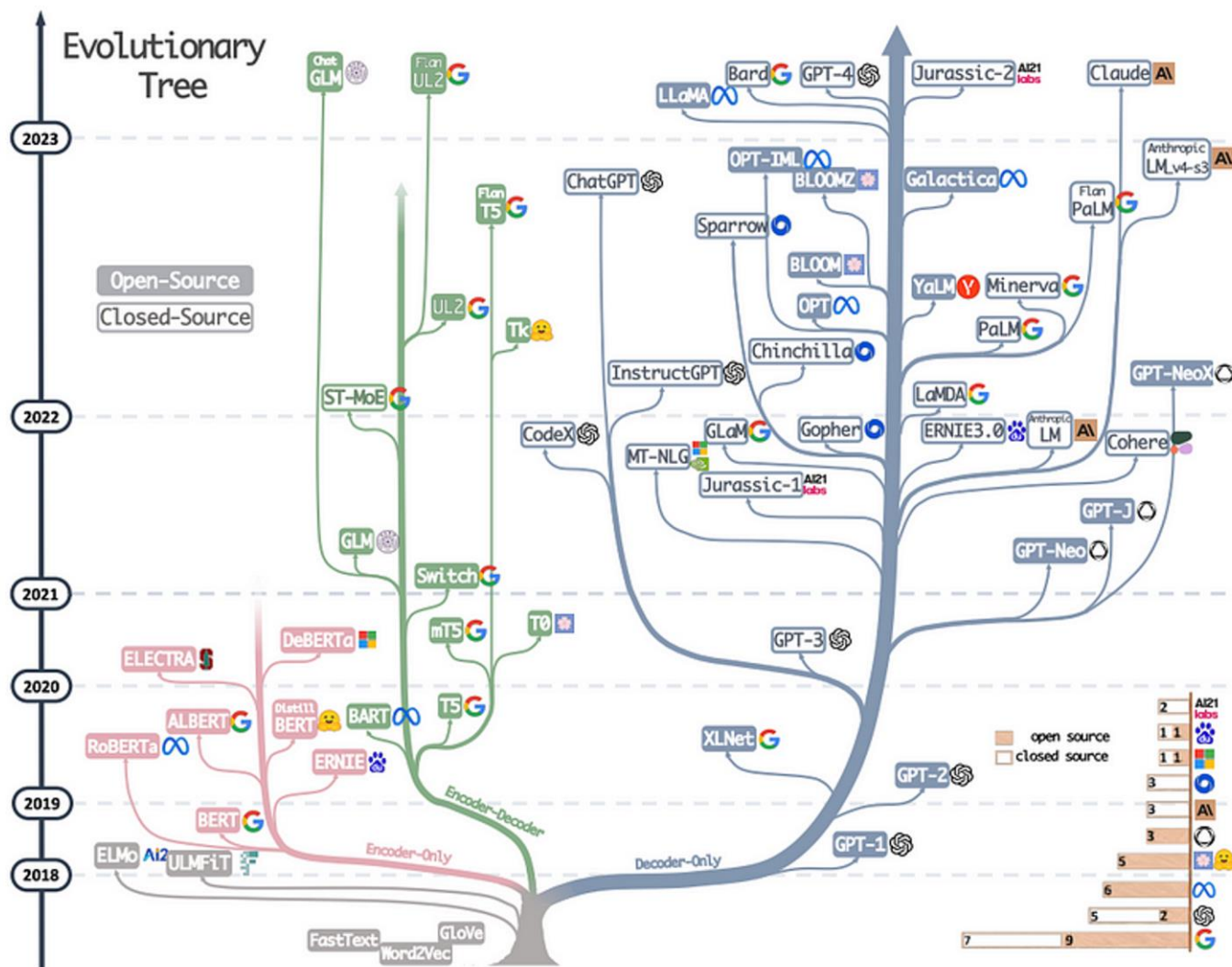
しかし、山野に自生する *D. opposita* の特性を調査し、栽培される *D. opposita* との類縁関係の報告なく、*D. opposita* の特性を有する株の分布状況は不明である。

そこで、日本各地に自生するヤマノイモ類 (Enantiophyllum 節) を収集し、葉の形態によりジネンジョと *D. opposita* の特性を有する株の2種類に分類し、これらの分類を調査した。そして、ジネンジョ、*D. opposita* の自生株および栽培される *D. opposita* の間の交雑親和性を調査し、自生するヤ

Background



Background: Multimodal LLMs



What's in this picture?

Looks like a duck.

That's not a duck. Then what's it?

Looks more like a bunny.

Why?

It has bunny ears.

Description of three toed woodpecker: It has black and white stripes throughout the body and a yellow crown.

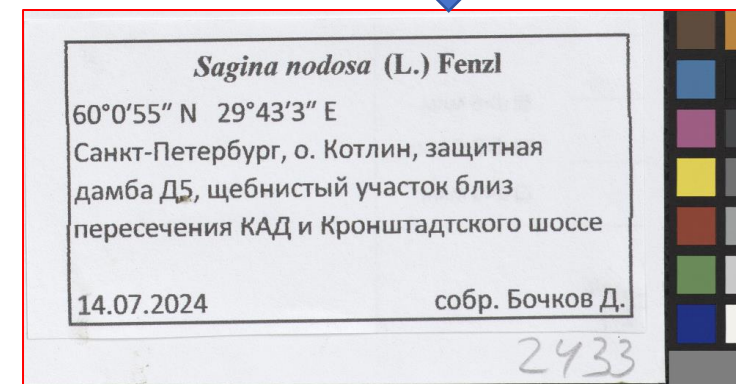
Description of downy woodpecker: It has white spots on its black wings and some red on its crown.

Question: what is the name of the woodpecker in the picture?

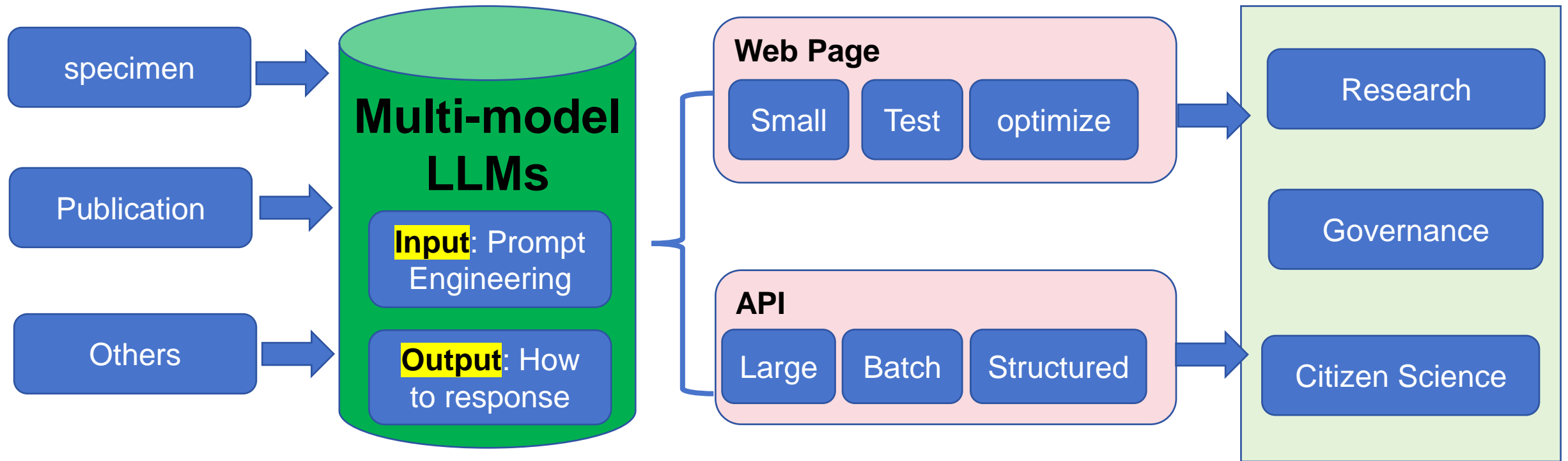
Downy

Background: Problems

- Multiple languages
- Different layout of the label
- Label may be rotate in different direction
- Free text style, not structured data
- Old geonames, may be not used now
- Want to get more description from the specimens
-



Architecture



Multi-model LLMs: Yiyan, Tongyi, ChatGLM...

Outline

- Background
- Specimen Example**
- Publication Example
- Discussion

Specimen Example



There is a label block on this picture, please extract the data on the block

The label on the picture contains the following data:

- Herbarium Name: Herbarium Universitatis Mosquensis (MW)
- Specimen Number: MW1092533
- Scientific Name: *Sagina nodosa* (L.) Fenzl
- Location:
 - 60°05'5"N, 29°43'3"E
 - Санкт-Петербург, о. Котлин, защитная дамба Д5, щебнистый участок близ пересечения КАД и Кронштадтского шоссе
- Date Collected: 14.07.2024
- Collector: Бочков Д.
- Scale: Indicates measurements up to 150 mm (as referenced by marks on the scale)



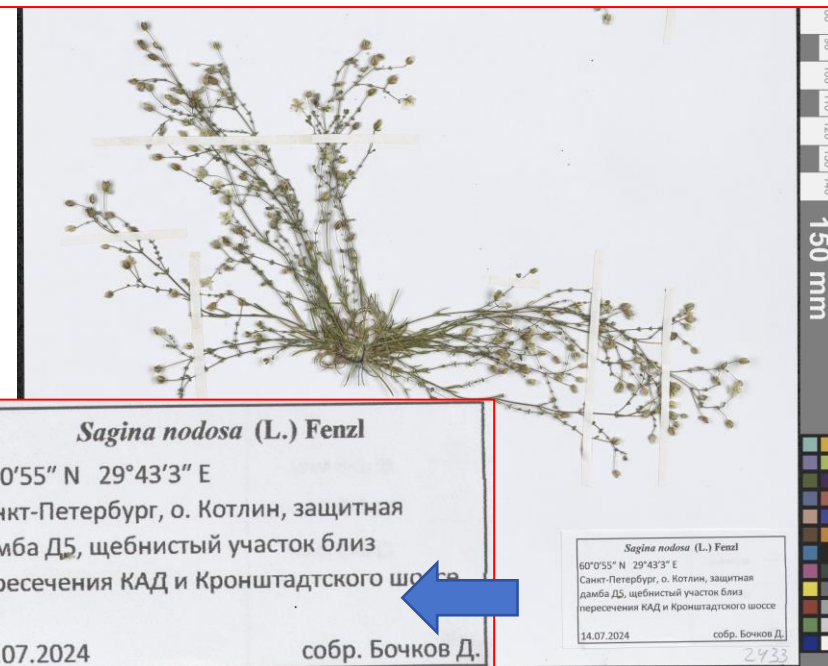
please translate location into Chinese and English

中文翻译:

圣彼得堡, 科特林岛, 防护堤D5, 碎石地段, 靠近KAD与克罗恩施塔德公路交汇处。

英文翻译:

St. Petersburg, Kotlin Island, protective dam D5, gravel area, near the intersection of KAD and Kronstadt Highway.



Specimen Example



Q: please extract locality information from the picture, and match the coordinate to this place and then return information in JSON format including following fields: country, state/province, locality, latitude, longitude

```
json
1 {
2   "country": "日本",
3   "state/province": "鹿児島県",
4   "locality": "奄美群島加計呂麻島秋徳-押角",
5   "latitude": "28.099333",
6   "longitude": "129.280833"
7 }
```

Q: return these information in English and Korea in JSON format

```
json
1 {
2   "English": {
3     "country": "Japan",
4     "state/province": "Kagoshima Prefecture",
5     "locality": "Kakeroma Island, Amami Islands, Kagoshima Prefecture",
6     "latitude": "28.099333",
7     "longitude": "129.2685" // Adjusted longitude based on the provided
8   },
9   "Korean": {
10    "country": "일본",
11    "state/province": "鹿児島県",
12    "locality": "카カロ마섬, 아마미諸島, 鹿児島県",
13    "latitude": "28.099333",
14    "longitude": "129.2685" // Adjusted longitude based on the provided
15  }
16 }
```

Specimen Example

Q: please extract the data from the label block and return information by JSON format

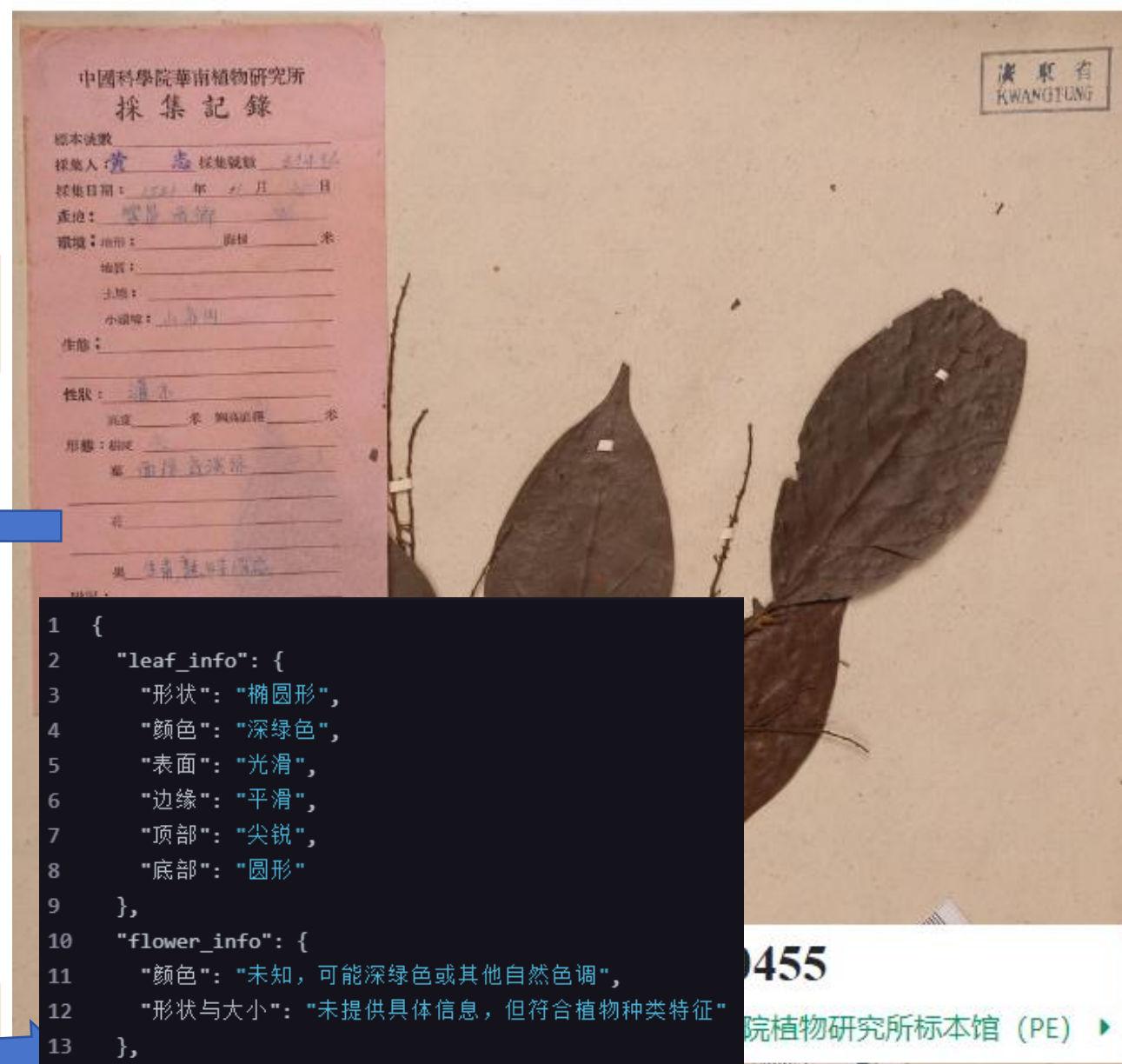
json OCR for hand-writing

```
1 {
2   "采集地": "乐昌西乡（广东）",
3   "采集人": "黄志",
4   "采集时间": "1531年（识别信息有误，实际采集日期可能为2017年7",
5   "标本编号": "31436",
6   "保存地点": "中国科学院华南植物研究所标本室",
7   "学名": "Euphorbiaceae Antidesma japonicum Sieb. & Zucc.",
8   "海拔": "信息未详",
9   "备注": "标本上附有蛱蝶"
10 }
```

Q: and re-extract the information of leaf, flower and skin

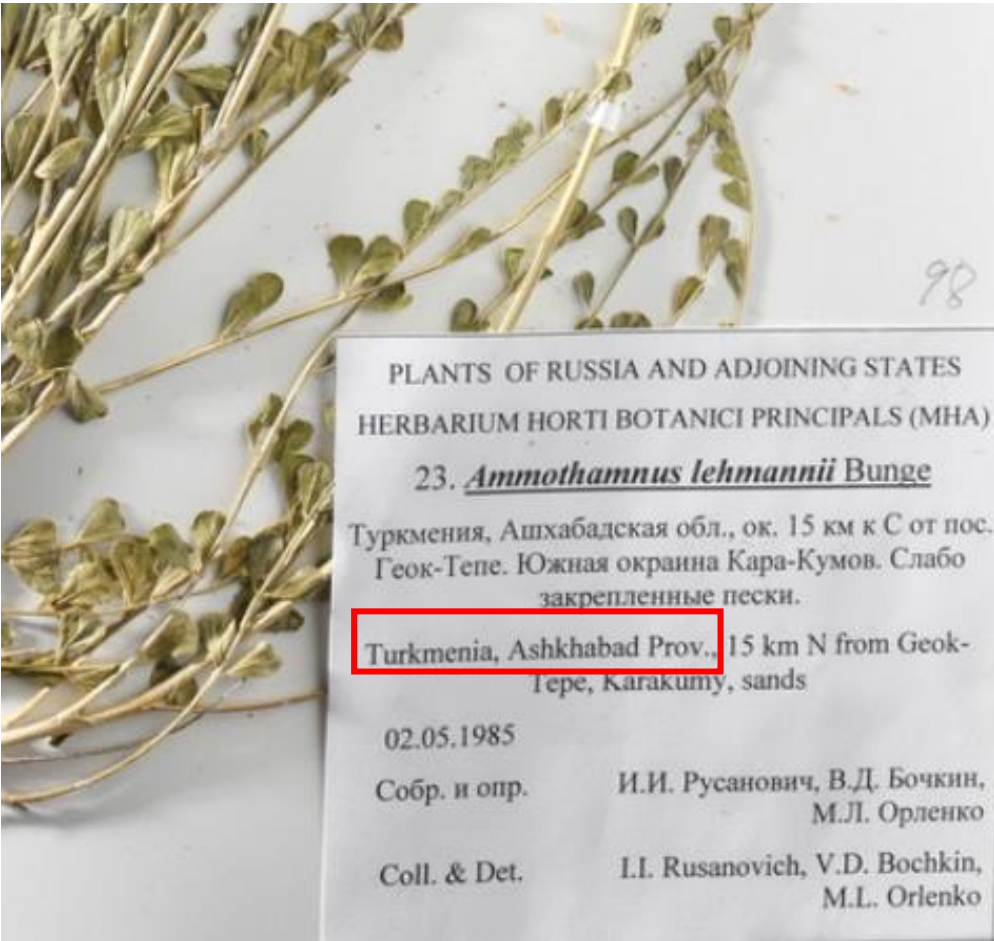
```
1 {
2   "leaf_info": {
3     "形状": "椭圆形",
4     "颜色": "深绿色",
5     "表面": "光滑",
6     "边缘": "平滑",
7     "顶部": "尖锐",
8     "底部": "圆形"
9   },
10  "flower_info": {
11    "颜色": "未知，可能深绿色或其他自然色调",
12    "形状与大小": "未提供具体信息，但符合植物种类特征"
13  },
14  "skin_info": {
15    "颜色": "灰色",
16    "质地与纹理": "未提供具体信息，但符合植物种类特征"
17  }
18 }
```

<https://www.cvh.ac.cn/spms/detail.php?id=e708b634>



Specimen Example: Data checking

Q: please extract information from the label block, match the coordinate form the locality and return in JSON format, including scientific name, country, state/province, locality, collect date,latitude, longitude



Json Tongyi

```
{
  "scientific_name": "Ammothamnus lehmannii",
  "country": "Russia",
  "state_province": "Turkmenistan, Ashgabat",
  "locality": "15 km N from Geok-Tepe, Karakumy",
  "collect_date": "02.05.1985",
  "latitude": null,
  "longitude": null
}
```

采集信息 CVH

采集人 V.D.Bochkin,M.L.Orlenko

采集号

采集时间 19850502

采集地 俄罗斯

json Yiyan

```
1 {
2   "scientific_name": "Ammodaemus lehmannii Bunge",
3   "country": "Turkmenia",
4   "state_province": "Ashgabat Prov.",
5   "locality": "15 km N from Geok-Tepe, Karakumy",
6   "collect_date": "02.05.1985",
7   "latitude": "Unknown",
8   "longitude": "Unknown"
}
```

json Zhipu

```
1 {
2   "scientific_name": "Ammothamnus lehmannii",
3   "country": "Turkmenistan",
4   "state_province": "Ashkhabad Prov.",
5   "locality": "15 km N from Geok-Tepe, Karakumy",
6   "collect_date": "02.05.1985",
7   "latitude": "",
8   "longitude": ""
9 }
10 }
```

Outline

- Background
- Specimen Example
- Publication Example**
- Discussion

Publication Example Korea

Flora of the vascular plants of the Baekdudaegan conservation area: Deok-chi to Yuk-sim-nyeong

Seung Hyun HWANG*, Jin Woong LEE¹, Eun Hwa LA² and Jin Kap AHN²

Department of Biology, Daejeon University, Daejeon 34520, Korea

¹*Natural Environmental Restoration Institute, Daejeon 34428, Korea*

²*Department of Life Sciences, Jeonbuk National University, Jeonju 54896, Korea*

(Received 12 February 2019; Revised 12 March 2020; Accepted 19 March 2020)

ABSTRACT: Baekdudaegan, the largest mountain range in eastern Asia, is a biodiversity hotspot in Korea that may have served as a glacial refugium. This study presents the flora of vascular plants on Deok-chi upto the Yuk-sim-nyeong area of the Baekdudaegan conservation area. The survey area was divided into four subareas and fieldwork was conducted for a total of twelve days in 2015. Voucher specimens were collected during the survey and were deposited at Daejeon University. A list of vascular plants was prepared based on the voucher specimens. The results of the survey showed that a total of 441 taxa, consisting of 100 families, 265 genera, 398 species, 9 subspecies, 32 varieties, and 2 forms, were found in the survey area. There was one endangered species, *Aconitum coreanum*, in the Bonghwasan Mt. area. Sixteen endemic taxa, 74 floristic regional indicator plants, as designated by the Ministry of the Environment, and eleven naturalized plants were distributed. The results of this study can serve as basic information to establish conservation and management plans for the Baekdudaegan conservation area.

Keywords: Baekdudaegan conservation area, flora, endemic, endangered plants, naturalized plants

백두대간은 백두산에서 시작하여 금강산, 설악산, 태백산, 소백산을 거쳐 지리산으로 이어지는 총 연장 1,600 km에 달해 동아시아 지역에서 가장 긴 큰 산줄기다(Chung et al., 2018). 백두대간은 한대성 식물과 온대성 식물이 함께

의 핵심축으로 2005년 2,634.27 km²로 최초 지정된 이. 2013년 2,750.77 km²로 확대 지정되었다.

본 연구는 지리산국립공원과 덕유산 국립공원 사이의 치-옥십령구간을 조사구간으로 하여 구시봉(1,014 m),

Q: Please extract distribution information from this article and return following data in JSON format: page number, line sentence, locality information



```
json
1  [
2    {
3      "page": 57,
4      "line": "Table 1.",
5      "locality": "Gonamsan Mt. area",
6    },
7    {
8      "page": 57,
9      "line": "Table 1.",
10     "locality": "Bonghwasan Mt. area",
11   },
12   {
13     "page": 57,
14     "line": "Table 1.",
15     "locality": "Baegunsan Mt. area",
16   },
17   {
```

Q: Please extract distribution information in Korean:



```
json
1  [
2    {
3      "page": 57,
4      "line": "Table 1.",
5      "locality": "고남산 지역",
6    },
7    {
8      "page": 57,
9      "line": "Table 1.",
10     "locality": "봉화산 지역",
11   },
12   {
13     "page": 57,
14     "line": "Table 1.",
15     "locality": "백운산 지역",
16   },
17   {
```

Publication Example

Russian

УДК 582.5/9(1-751.2)(571.55)

Новые местонахождения редких видов растений в Забайкальском крае

О. А. Попова^{1*}, Н. А. Чашчина¹, А. П. Лесков¹, С. Н. Щеглова¹, Е. А. Андриевская², Б. Б. Намзалов³

¹ Забайкальский государственный университет, ул. Александрово-Заводская, 30, г. Чита, 30672039, Россия.
E-mails: olga.popova-54@yandex.ru*, bupleurum24@mail.ru, leskova-olga@inbox.ru, svetashcheglova@mail.ru

² Национальный парк «Чикой», ул. Первомайская, 1, с. Красный Чикой. Забайкальский край, 763060, Россия.
E-mail: npchikoi@yandex.ru

³ Бурятский государственный университет, ул. Смолина, 24а, г. Улан-Удэ, 670000, Россия. E-mail: namzalov@rambler.ru

* Автор для переписки

Ключевые слова: Забайкальский край, Национальный парк «Чикой», новые местонахождения, редкие растения.

Аннотация. В статье приводятся список и краткая характеристика распространения 15 видов цветковых растений, впервые найденных в южном Забайкалье на территории Национального парка «Чикой». Впервые для флоры Забайкальского края указывается местонахождение *Artemisia furcata*, собранного на Чикоконском хребте в районе озера Шебеты. Из числа найденных видов два вида рода *Rhodiola* (*R. quadrifida*, *R. pinnatifida*) включены в Красную книгу Забайкальского края (2017). По одному виду редких растений занесены в Красные книги сопредельных территорий: *Saxifraga davurica* включена в Красную книгу Республики Саха (Якутия) (2017), а *Artemisia furcata* – в Красную книгу Амурской области (2009). *Oxycoccus microcarpus* как редкий вид для территории края будет рекомендован для включения в список редких и находящихся под угрозой исчезновения видов сосудистых растений Забайкальского края.

New localities of rare plant species in Trans-Baikal Territory

O. A. Popova^{1*}, N. A. Chashchina¹, A. P. Leskov¹, S. N. Shcheglova¹,
E. A. Andrievskaya², B. B. Namzalov³

Q: Please extract distribution information from this article and return following data in JSON format: page number, line sentence, locality information



```
1  [
2    {
3      "page": "87",
4      "line": "4",
5      "locality": "Забайкальский край, Красночикойский р-н, Национальный парк
6    },
7    {
8      "page": "87",
9      "line": "9",
10     "locality": "Забайкальский край, Красночикойский р-н, Национальный парк
11   },
12   {
13     "page": "87",
14     "line": "14",
15     "locality": "Забайкальский край, Красночикойский р-н, Национальный парк
16   },
17   {
18     "page": "87",
19     "line": "19",
20     "locality": "Забайкальский край, Красночикойский р-н, Национальный парк
```

Outline

- Background
- Specimen Example
- Publication Example
- Discussion**

Discussion: Database+ API in batch

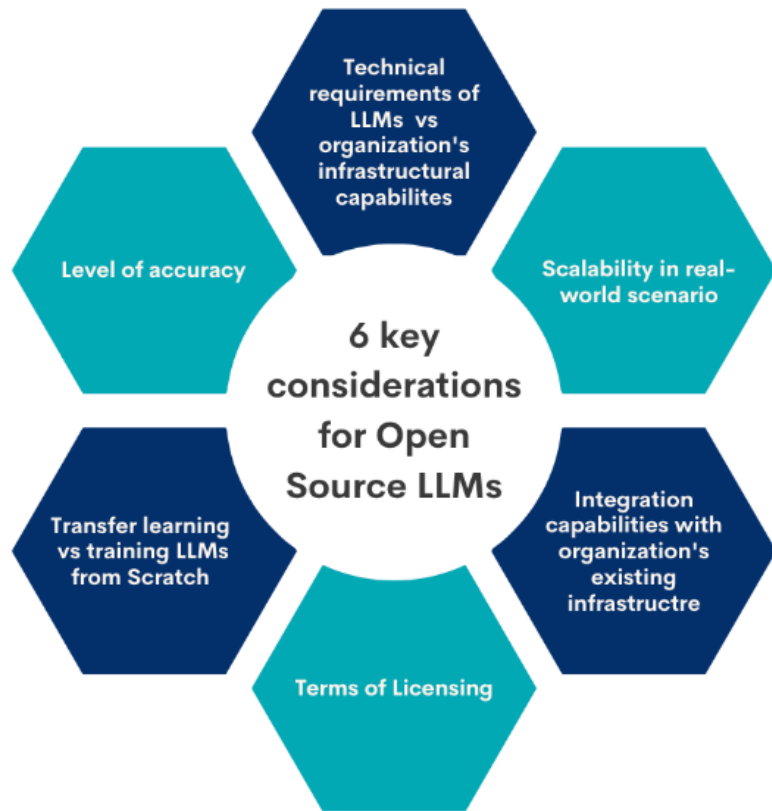
```
with open(img_path, 'rb') as img_file:
    img_base = base64.b64encode(img_file.read()).decode('utf-8')

client = ZhipuAI(api_key=" ") #
填写您自己的APIKey
response = client.chat.completions.create(
    model="glm-4v-plus", # 填写需要调用的模型名称
    messages=[
        {
            "role": "user",
            "content": [
                {
                    "type": "image_url",
                    "image_url": {
                        "url": img_base
                    }
                },
                {
                    "type": "text",
                    "text": "Please extract the place name information from this image, match
                           it with the decimal format of latitude and longitude, and return the
                           following information in JSON format: country, state/province, small
                           place name, latitude and longitude"
                }
            ]
        }
    ],
)
print(response.choices[0].message)
```

```
json
{
  "country": "俄罗斯",
  "state": "圣彼得堡",
  "small_locality": "舒瓦洛夫公园草地",
  "coordinates": {
    "latitude": 60.02,
    "longitude": 30.2099
  }
}
```



Discussion: Open source+ Local Deployment



Open source

Less Money

High talent

More models

More Control

Smarter

Self Adjust

Commercial Solution

More Money

Lower talent

High Accuracy

Less Control

Fixed

Provider Adjust

Discussion

- Specimen and publication include so many information for distribution and trait character and also is a good source for filling data gap.
- LLM is a good way to extract information in multiple language and multiple models
- LLM can be used in the whole data life cycle: collect, clean, check, gap filling
- LLM should be matched with some human work to get high-quality result.
- Open source and commercial solution are both considered for different application case as for your detailed need.