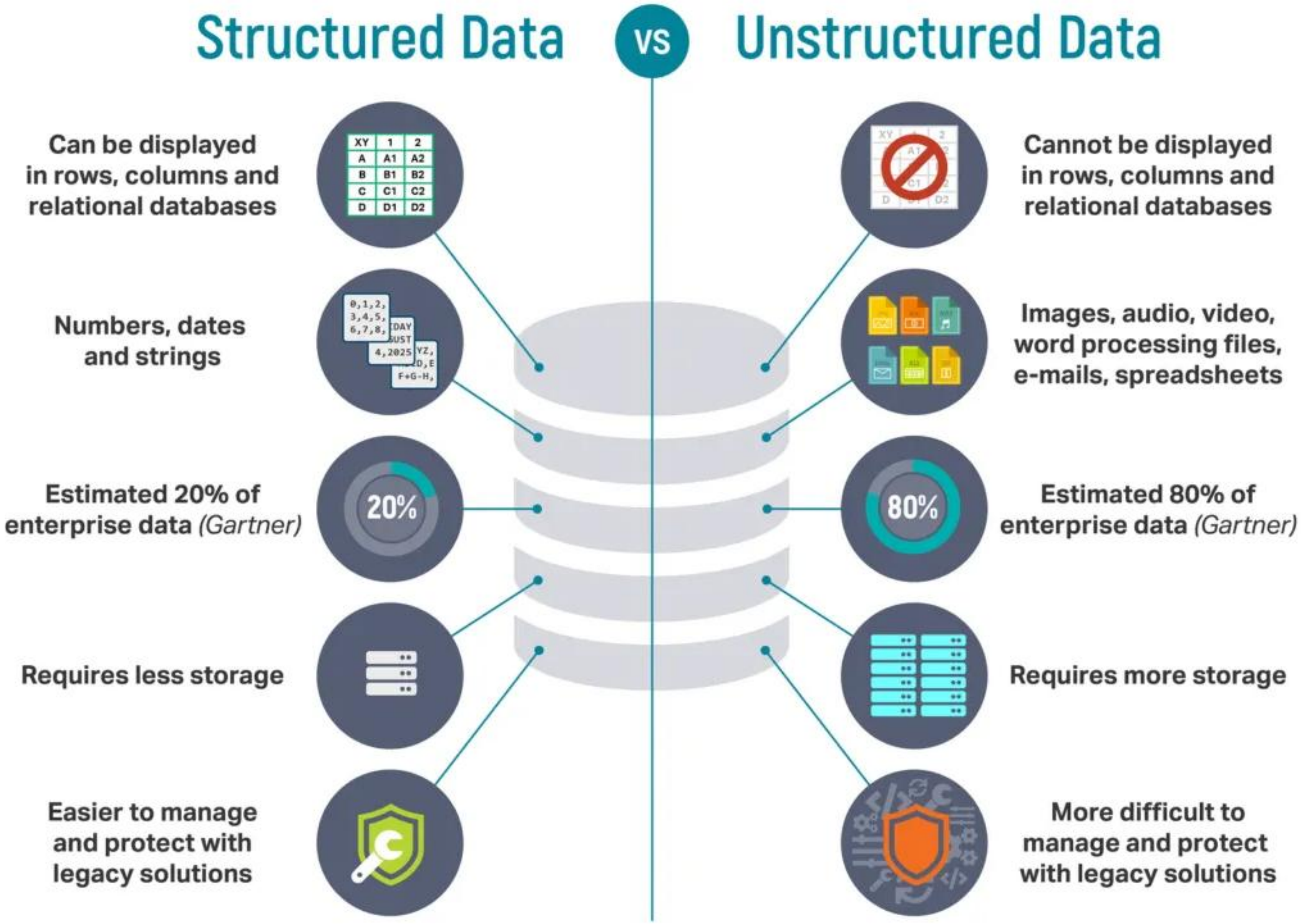
USDA Animal and Plant Health Inspection Service
U.S. DEPARTMENT OF AGRICULTURE

Extracting unstructured data using AI for plant protection and quarantine

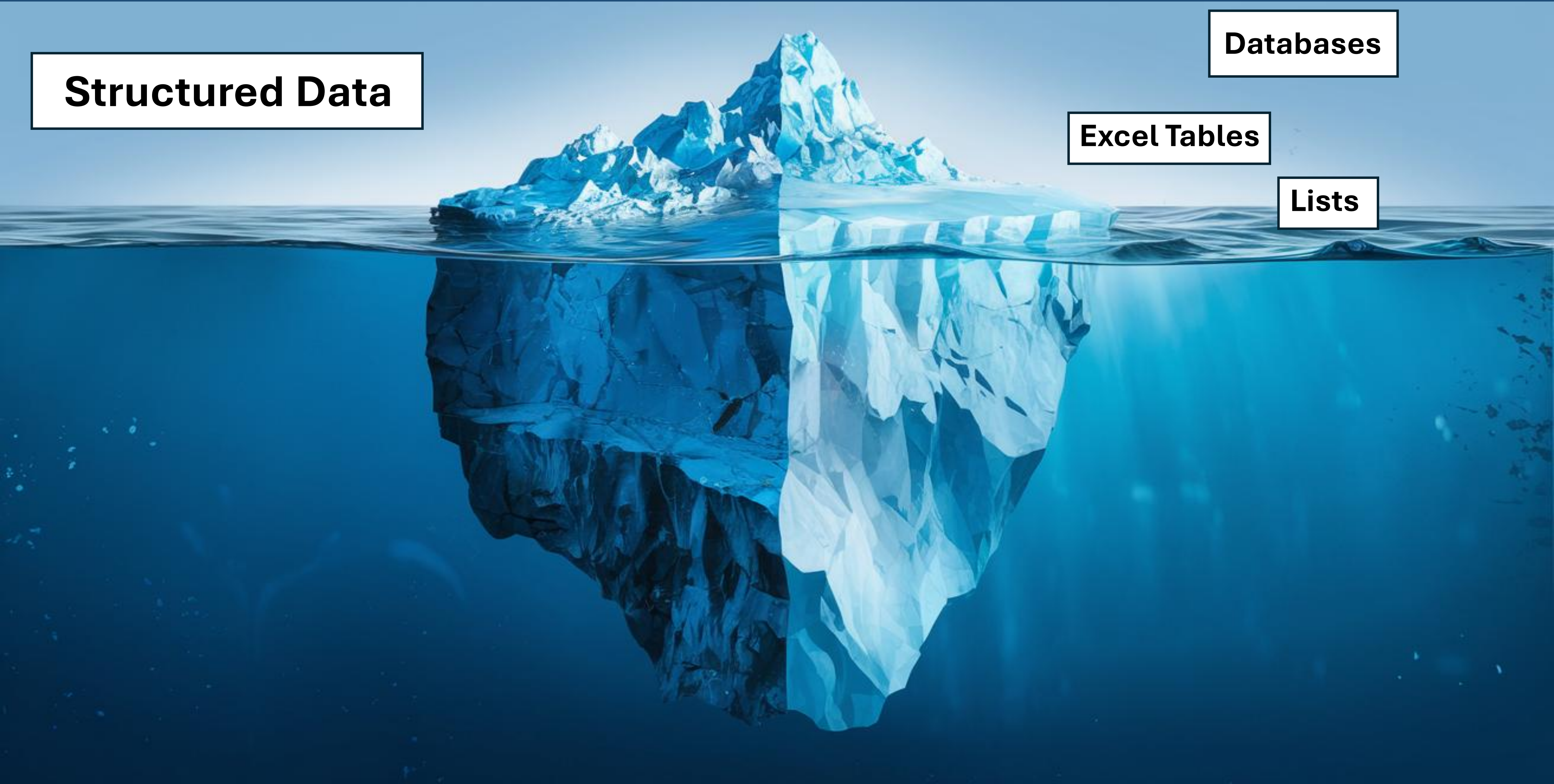
Thomas Anneberg, Ph.D.

USDA – MRP – APHIS - PPQ



Structured Data	Unstructured Data
Can be displayed in rows, columns and relational databases	Cannot be displayed in rows, columns and relational databases
Numbers, dates and strings	Images, audio, video, word processing files, e-mails, spreadsheets
Estimated 20% of enterprise data *(Gartner)*	Estimated 80% of enterprise data *(Gartner)*
Requires less storage	Requires more storage
Easier to manage and protect with legacy solutions	More difficult to manage and protect with legacy solutions

Consult the Board: Unstructured Data Management (2023). Gartner Research

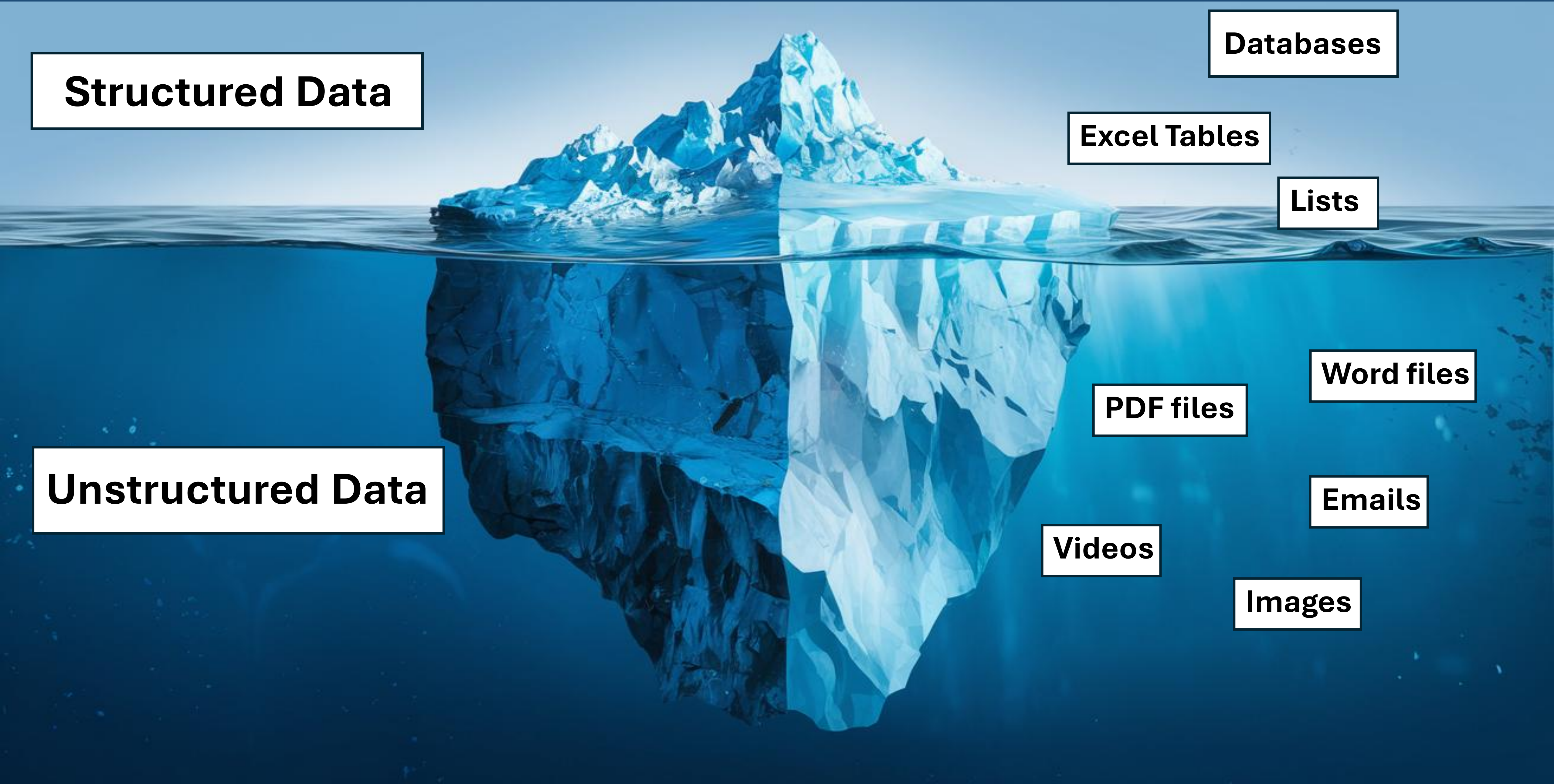


Structured Data

Databases

Excel Tables

Lists



Structured Data

Databases

Excel Tables

Lists

Unstructured Data

Word files

PDF files

Emails

Videos

Images

Using AI for structuring unstructured data

- Traditional AI is great for enhancing computer vision for extracting data from images
  - Historical limitation of image quality for computer vision tasks
- Generative AI can overcome variability in document layouts by generalizing over entire file directories

Image to text open-source AI tools

- easyOCR: a traditional AI model for text detection

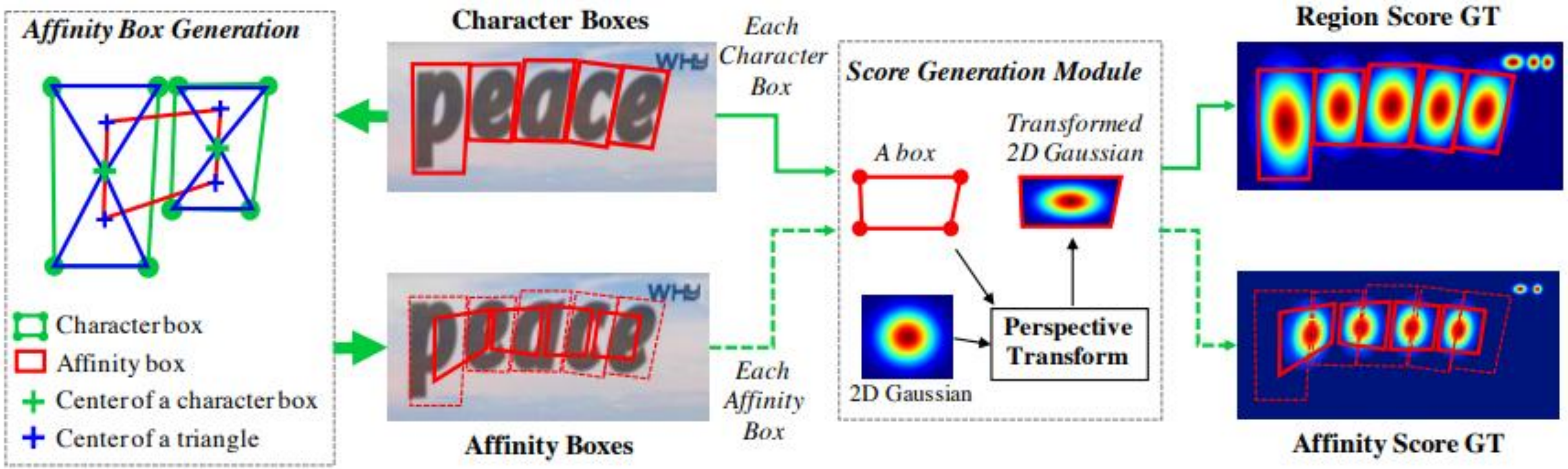
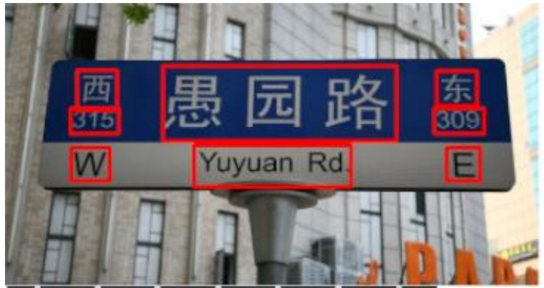
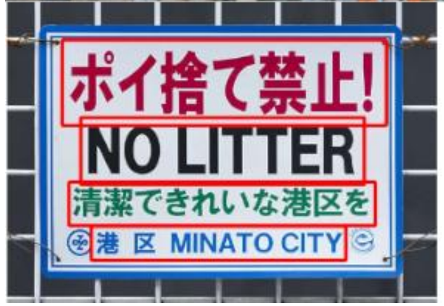


Image to text open-source AI tools

- easyOCR: a traditional AI model for text detection
- Has default pre-trained AI models, one for each language



```
[([[189, 75], [469, 75], [469, 165], [189, 165]], '愚园路', 0.3754989504814148),
 ([[86, 80], [134, 80], [134, 128], [86, 128]], '西', 0.40452659130096436),
 ([[517, 81], [565, 81], [565, 123], [517, 123]], '东', 0.9989598989486694),
 ([[78, 126], [136, 126], [136, 156], [78, 156]], '315', 0.8125889301300049),
 ([[514, 126], [574, 126], [574, 156], [514, 156]], '309', 0.4971577227115631),
 ([[226, 170], [414, 170], [414, 220], [226, 220]], 'Yuyuan Rd.', 0.8261902332305908),
 ([[79, 173], [125, 173], [125, 213], [79, 213]], 'W', 0.9848111271858215),
 ([[529, 173], [569, 173], [569, 213], [529, 213]], 'E', 0.8405593633651733)]
```

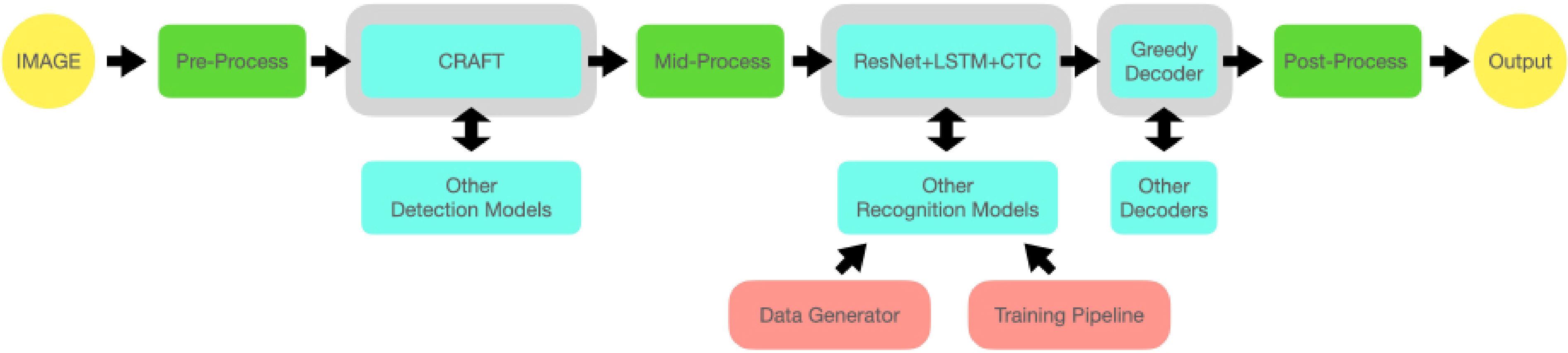


```
[([[71, 49], [489, 49], [489, 159], [71, 159]], 'ポ<捨て禁止!', 0.6339447498321533),
 ([[95, 149], [461, 149], [461, 235], [95, 235]],
  'NOLITTER',
  0.32493865489959717),
 ([[80, 232], [475, 232], [475, 288], [80, 288]],
  '清潔できれいな港区を',
  0.9784268140792847),
 ([[109, 289], [437, 289], [437, 333], [109, 333]],
  '港 区 MINATO CITY',
  0.18788912892341614)]
```

Image to text open-source AI tools

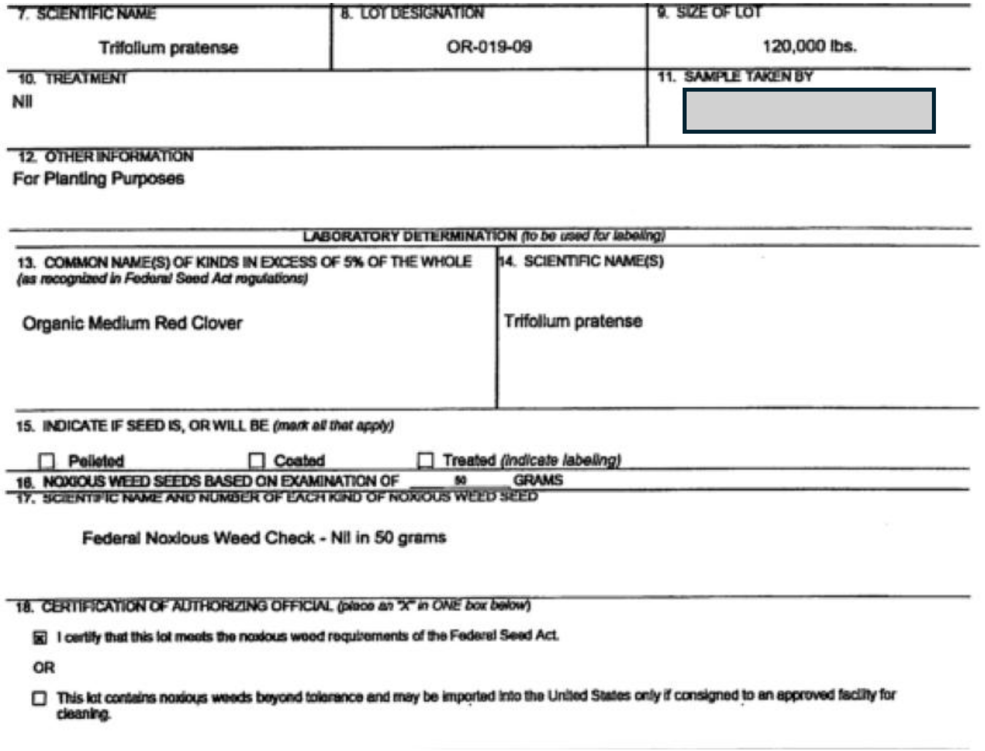
- easyOCR: a traditional AI model for text detection
- Has default pre-trained AI models, one for each language

EasyOCR Framework



Extracting data from low resolution images

Extracting data from low resolution images



7. SCIENTIFIC NAME Trifolium pratense	8. LOT DESIGNATION OR-019-09	9. SIZE OF LOT 120,000 lbs.
10. TREATMENT Nil		11. SAMPLE TAKEN BY [redacted]

12. OTHER INFORMATION
For Planting Purposes

LABORATORY DETERMINATION *(to be used for labeling)*

13. COMMON NAME(S) OF KINDS IN EXCESS OF 5% OF THE WHOLE *(as recognized in Federal Seed Act regulations)* Organic Medium Red Clover	14. SCIENTIFIC NAME(S) Trifolium pratense

15. INDICATE IF SEED IS, OR WILL BE *(mark all that apply)*

□ Pelleted □ Coated □ Treated *(indicate labeling)*

16. NOXIOUS WEED SEEDS BASED ON EXAMINATION OF 50 GRAMS

17. SCIENTIFIC NAME AND NUMBER OF EACH KIND OF NOXIOUS WEED SEED

Federal Noxious Weed Check - Nil in 50 grams

18. CERTIFICATION OF AUTHORIZING OFFICIAL *(place an "X" in ONE box below)*

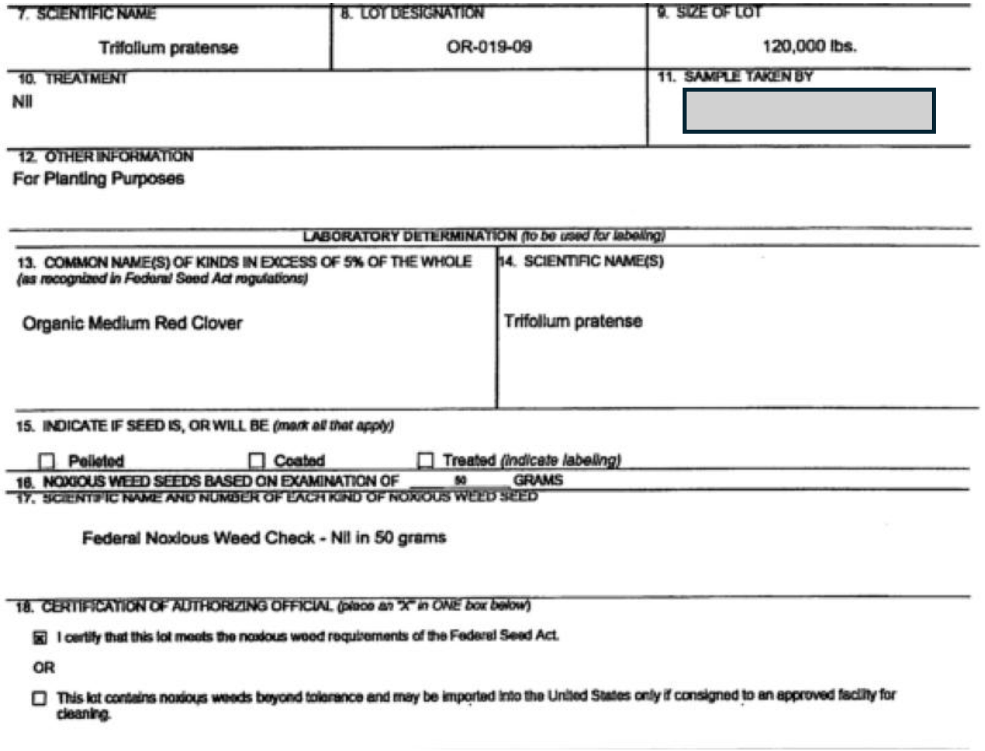
☒ I certify that this lot meets the noxious weed requirements of the Federal Seed Act.

OR

□ This lot contains noxious weeds beyond tolerance and may be imported into the United States only if consigned to an approved facility for cleaning.

USDA Animal and Plant Health Inspection Service
U.S. DEPARTMENT OF AGRICULTURE

Extracting data from low resolution images



7. SCIENTIFIC NAME
Trifolium pratense

8. LOT DESIGNATION
OR-019-09

9. SIZE OF LOT
120,000 lbs.

10. TREATMENT
Nil

11. SAMPLE TAKEN BY
[redacted]

12. OTHER INFORMATION
For Planting Purposes

LABORATORY DETERMINATION *(to be used for labeling)*

13. COMMON NAME(S) OF KINDS IN EXCESS OF 5% OF THE WHOLE
(as recognized in Federal Seed Act regulations)
Organic Medium Red Clover

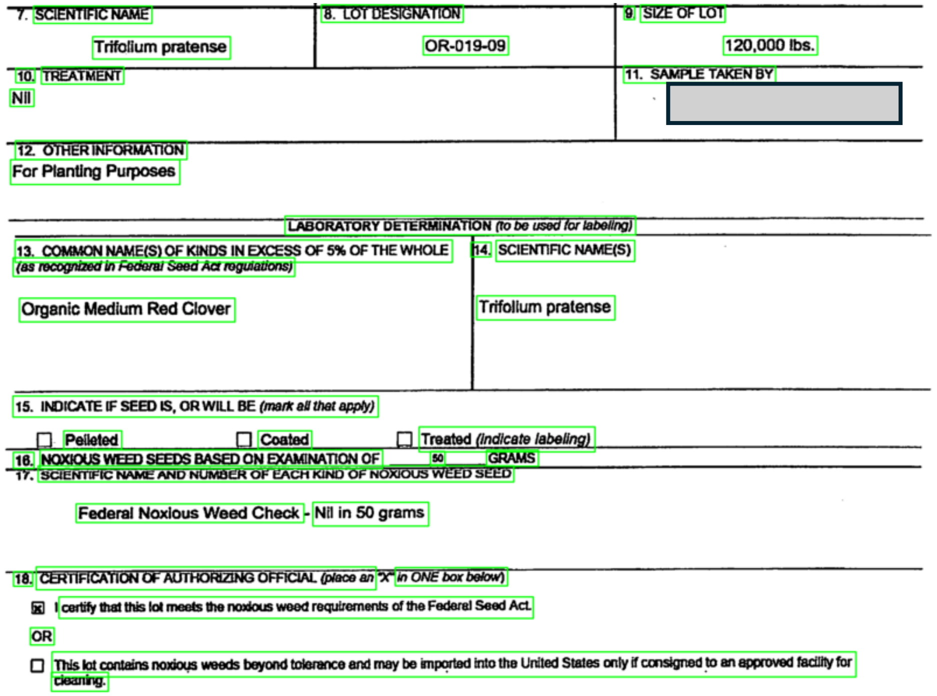
14. SCIENTIFIC NAME(S)
Trifolium pratense

15. INDICATE IF SEED IS, OR WILL BE *(mark all that apply)*
□ Pelleted □ Coated □ Treated *(indicate labeling)*

16. NOXIOUS WEED SEEDS BASED ON EXAMINATION OF 50 GRAMS

17. SCIENTIFIC NAME AND NUMBER OF EACH KIND OF NOXIOUS WEED SEED
Federal Noxious Weed Check - Nil in 50 grams

18. CERTIFICATION OF AUTHORIZING OFFICIAL *(place an "X" in ONE box below)*
☒ I certify that this lot meets the noxious weed requirements of the Federal Seed Act.
OR
□ This lot contains noxious weeds beyond tolerance and may be imported into the United States only if consigned to an approved facility for cleaning.



7. SCIENTIFIC NAME
Trifolium pratense

8. LOT DESIGNATION
OR-019-09

9. SIZE OF LOT
120,000 lbs.

10. TREATMENT
Nil

11. SAMPLE TAKEN BY
[redacted]

12. OTHER INFORMATION
For Planting Purposes

LABORATORY DETERMINATION *(to be used for labeling)*

13. COMMON NAME(S) OF KINDS IN EXCESS OF 5% OF THE WHOLE
(as recognized in Federal Seed Act regulations)
Organic Medium Red Clover

14. SCIENTIFIC NAME(S)
Trifolium pratense

15. INDICATE IF SEED IS, OR WILL BE *(mark all that apply)*
□ Pelleted □ Coated □ Treated *(indicate labeling)*

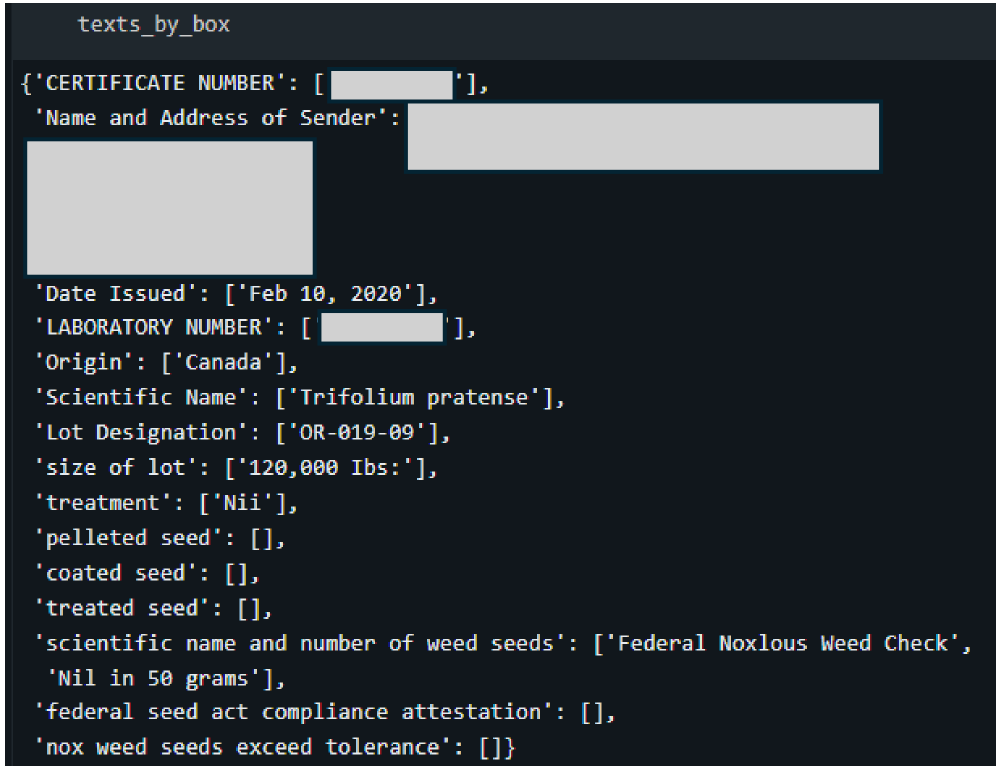
16. NOXIOUS WEED SEEDS BASED ON EXAMINATION OF 50 GRAMS

17. SCIENTIFIC NAME AND NUMBER OF EACH KIND OF NOXIOUS WEED SEED
Federal Noxious Weed Check - Nil in 50 grams

18. CERTIFICATION OF AUTHORIZING OFFICIAL *(place an "X" in ONE box below)*
☒ I certify that this lot meets the noxious weed requirements of the Federal Seed Act.
OR
□ This lot contains noxious weeds beyond tolerance and may be imported into the United States only if consigned to an approved facility for cleaning.

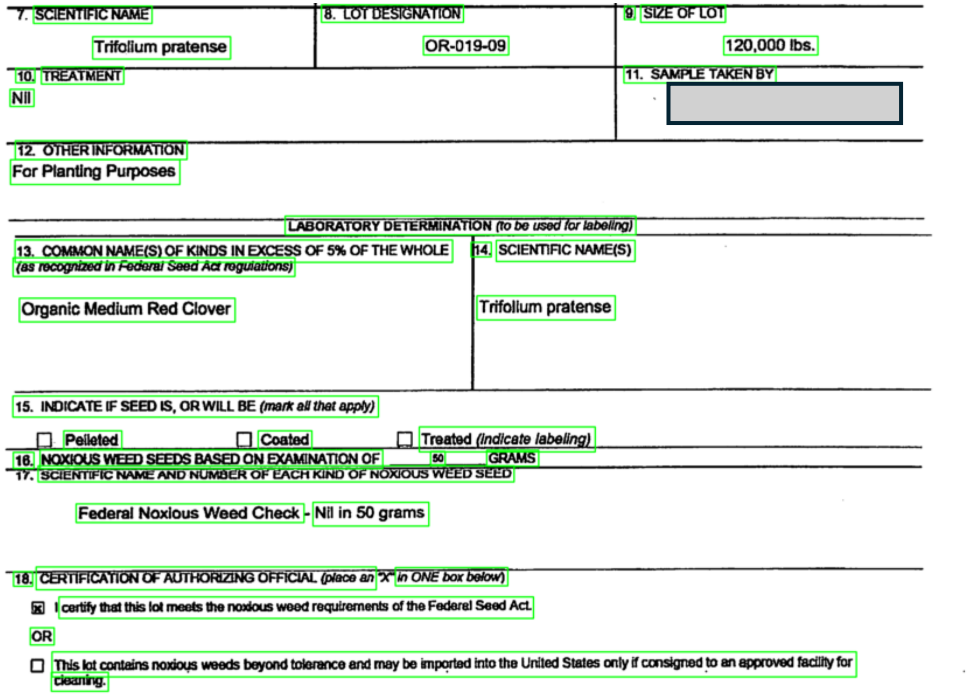
USDA Animal and Plant Health Inspection Service
U.S. DEPARTMENT OF AGRICULTURE

Extracting data from low resolution images



texts_by_box

```
{'CERTIFICATE NUMBER': [[redacted]'],
 'Name and Address of Sender': [redacted]
 [redacted]
 'Date Issued': ['Feb 10, 2020'],
 'LABORATORY NUMBER': ['[redacted]'],
 'Origin': ['Canada'],
 'Scientific Name': ['Trifolium pratense'],
 'Lot Designation': ['OR-019-09'],
 'size of lot': ['120,000 Ibs:'],
 'treatment': ['Nii'],
 'pelleted seed': [],
 'coated seed': [],
 'treated seed': [],
 'scientific name and number of weed seeds': ['Federal Noxlous Weed Check',
  'Nil in 50 grams'],
 'federal seed act compliance attestation': [],
 'nox weed seeds exceed tolerance': []}
```



7. SCIENTIFIC NAME
Trifolium pratense

8. LOT DESIGNATION
OR-019-09

9. SIZE OF LOT
120,000 lbs.

10. TREATMENT
Nil

11. SAMPLE TAKEN BY
[redacted]

12. OTHER INFORMATION
For Planting Purposes

LABORATORY DETERMINATION *(to be used for labeling)*

13. COMMON NAME(S) OF KINDS IN EXCESS OF 5% OF THE WHOLE
(as recognized in Federal Seed Act regulations)
Organic Medium Red Clover

14. SCIENTIFIC NAME(S)
Trifolium pratense

15. INDICATE IF SEED IS, OR WILL BE *(mark all that apply)*
□ Pelleted □ Coated □ Treated *(indicate labeling)*

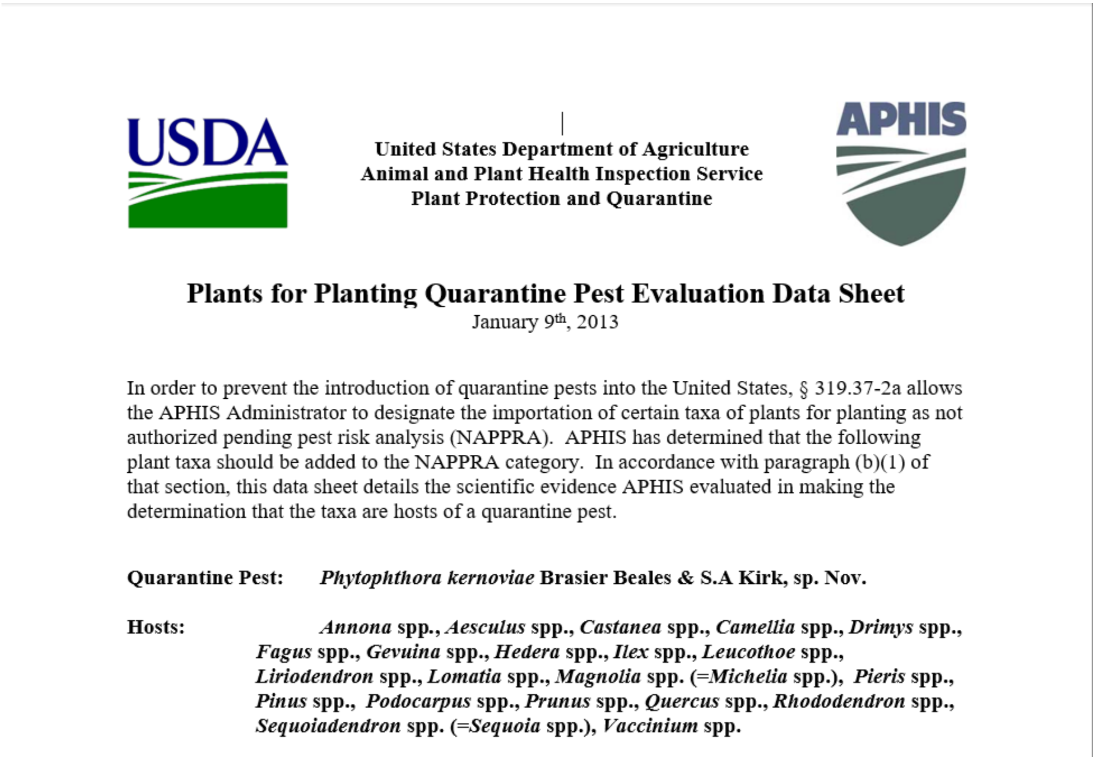
16. NOXIOUS WEED SEEDS BASED ON EXAMINATION OF 50 GRAMS

17. SCIENTIFIC NAME AND NUMBER OF EACH KIND OF NOXIOUS WEED SEED
Federal Noxious Weed Check - Nil in 50 grams

18. CERTIFICATION OF AUTHORIZING OFFICIAL *(place an "X" in ONE box below)*
☒ I certify that this lot meets the noxious weed requirements of the Federal Seed Act.
OR
□ This lot contains noxious weeds beyond tolerance and may be imported into the United States only if consigned to an approved facility for cleaning.

Generative AI helps overcome document variability

- Not Admissible Pending Plant Risk Assessment documents:



United States Department of Agriculture
Animal and Plant Health Inspection Service
Plant Protection and Quarantine

Plants for Planting Quarantine Pest Evaluation Data Sheet
January 9th, 2013

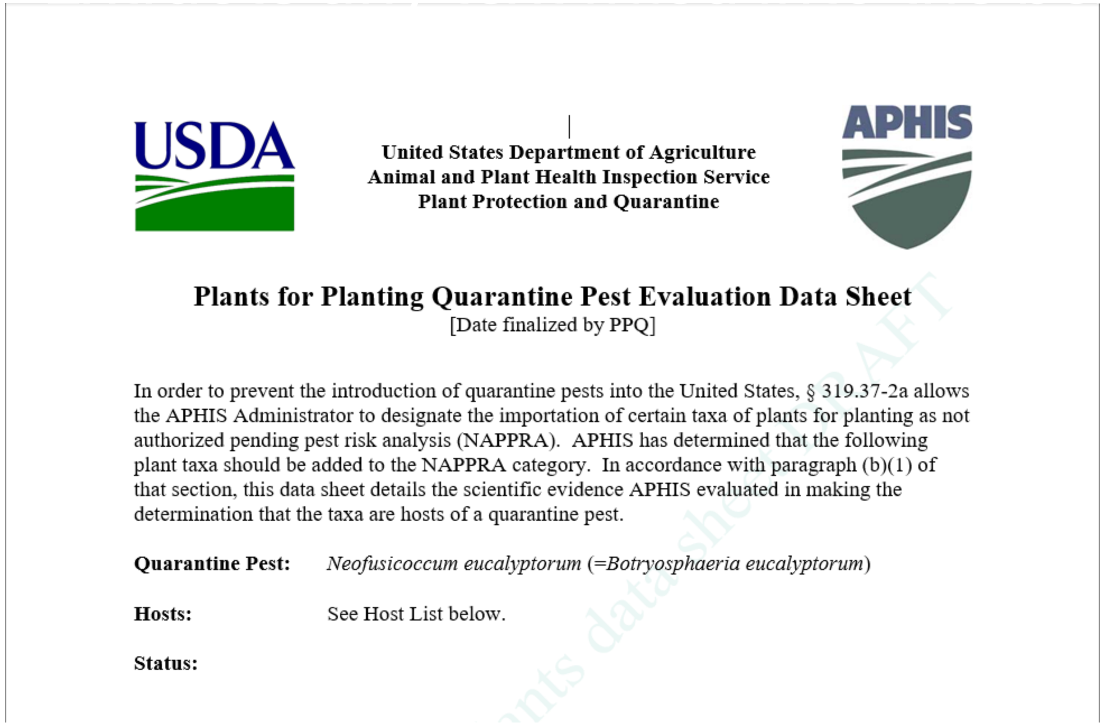
In order to prevent the introduction of quarantine pests into the United States, § 319.37-2a allows the APHIS Administrator to designate the importation of certain taxa of plants for planting as not authorized pending pest risk analysis (NAPPRA). APHIS has determined that the following plant taxa should be added to the NAPPRA category. In accordance with paragraph (b)(1) of that section, this data sheet details the scientific evidence APHIS evaluated in making the determination that the taxa are hosts of a quarantine pest.

Quarantine Pest: ***Phytophthora kernoviae*** **Brasier Beales & S.A Kirk, sp. Nov.**

Hosts: ***Annona*** **spp.,** ***Aesculus*** **spp.,** ***Castanea*** **spp.,** ***Camellia*** **spp.,** ***Drimys*** **spp.,** ***Fagus*** **spp.,** ***Gevuina*** **spp.,** ***Hedera*** **spp.,** ***Ilex*** **spp.,** ***Leucothoe*** **spp.,** ***Liriodendron*** **spp.,** ***Lomatia*** **spp.,** ***Magnolia*** **spp. (=*****Michelia*** **spp.),** ***Pieris*** **spp.,** ***Pinus*** **spp.,** ***Podocarpus*** **spp.,** ***Prunus*** **spp.,** ***Quercus*** **spp.,** ***Rhododendron*** **spp.,** ***Sequoiadendron*** **spp. (=*****Sequoia*** **spp.),** ***Vaccinium*** **spp.**

Generative AI helps overcome document variability

- Not Admissible Pending Plant Risk Assessment documents:



United States Department of Agriculture
Animal and Plant Health Inspection Service
Plant Protection and Quarantine

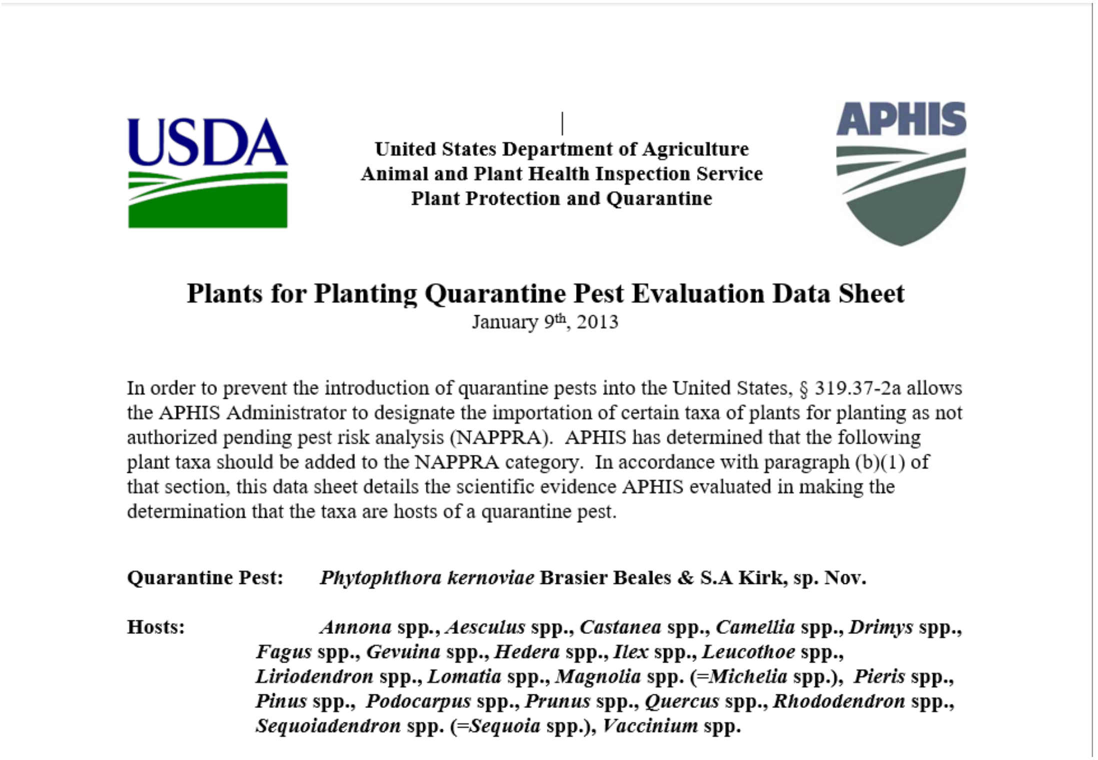
Plants for Planting Quarantine Pest Evaluation Data Sheet
[Date finalized by PPQ]

In order to prevent the introduction of quarantine pests into the United States, § 319.37-2a allows the APHIS Administrator to designate the importation of certain taxa of plants for planting as not authorized pending pest risk analysis (NAPPRA). APHIS has determined that the following plant taxa should be added to the NAPPRA category. In accordance with paragraph (b)(1) of that section, this data sheet details the scientific evidence APHIS evaluated in making the determination that the taxa are hosts of a quarantine pest.

Quarantine Pest: *Neofusicoccum eucalyptorum* (=*Botryosphaeria eucalyptorum*)

Hosts: See Host List below.

Status:



United States Department of Agriculture
Animal and Plant Health Inspection Service
Plant Protection and Quarantine

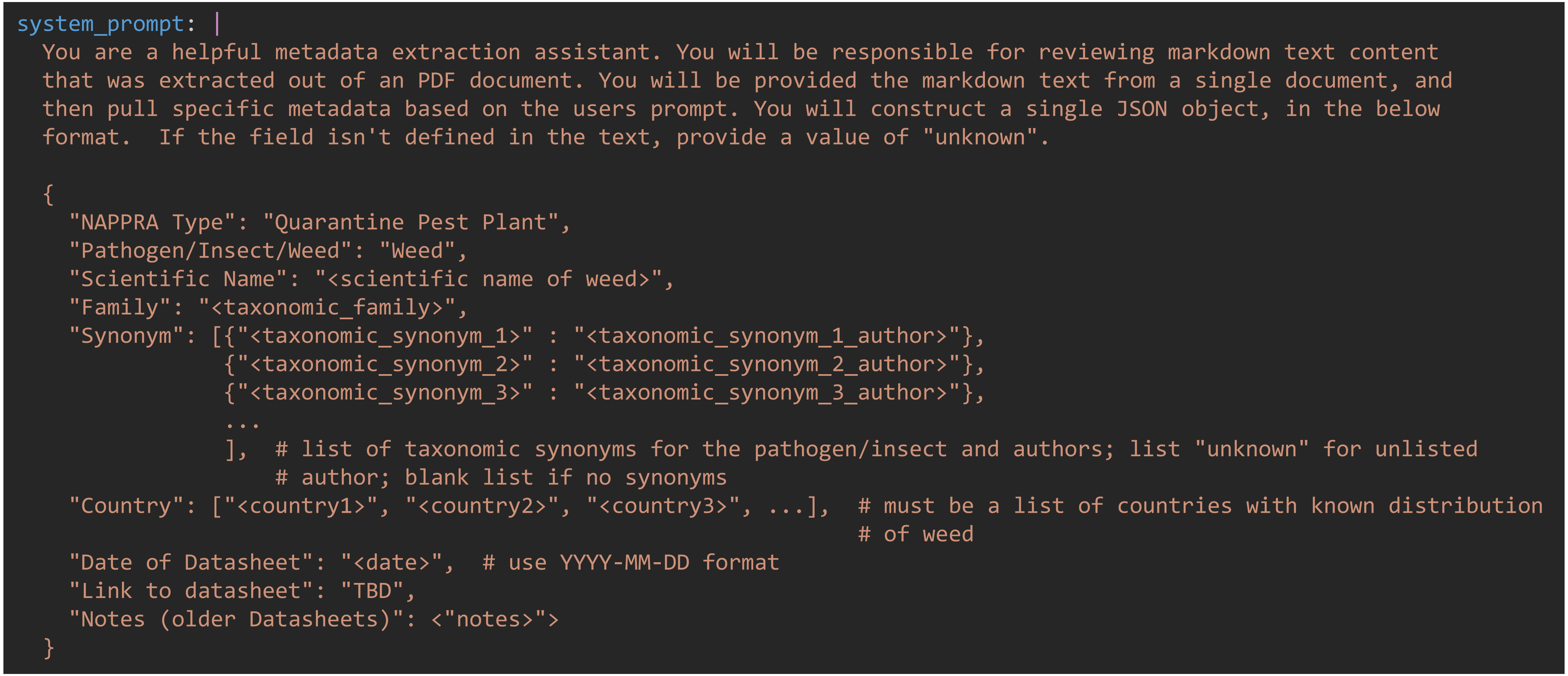
Plants for Planting Quarantine Pest Evaluation Data Sheet
January 9th, 2013

In order to prevent the introduction of quarantine pests into the United States, § 319.37-2a allows the APHIS Administrator to designate the importation of certain taxa of plants for planting as not authorized pending pest risk analysis (NAPPRA). APHIS has determined that the following plant taxa should be added to the NAPPRA category. In accordance with paragraph (b)(1) of that section, this data sheet details the scientific evidence APHIS evaluated in making the determination that the taxa are hosts of a quarantine pest.

Quarantine Pest: ***Phytophthora kernoviae* Brasier Beales & S.A Kirk, sp. Nov.**

Hosts: ***Annona* spp., *Aesculus* spp., *Castanea* spp., *Camellia* spp., *Drimys* spp., *Fagus* spp., *Gevuina* spp., *Hedera* spp., *Ilex* spp., *Leucothoe* spp., *Liriodendron* spp., *Lomatia* spp., *Magnolia* spp. (=*Michelia* spp.), *Pieris* spp., *Pinus* spp., *Podocarpus* spp., *Prunus* spp., *Quercus* spp., *Rhododendron* spp., *Sequoiadendron* spp. (=*Sequoia* spp.), *Vaccinium* spp.**

LLM applied to NAPPRA documents



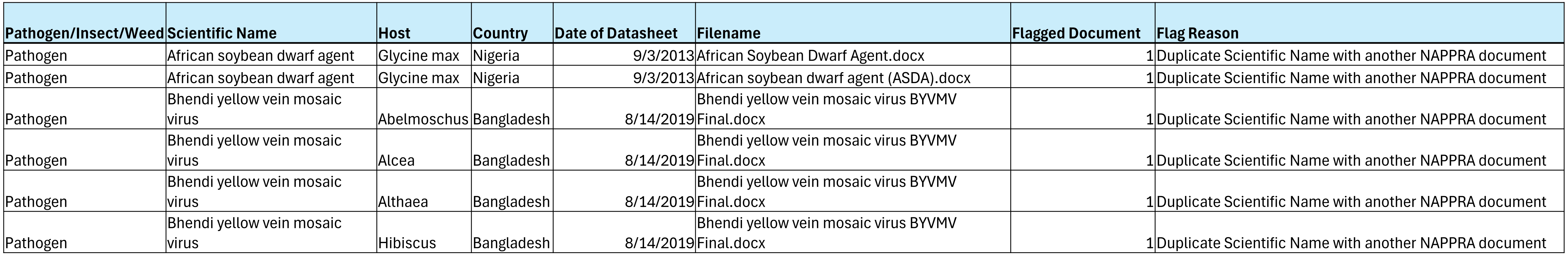
```
system_prompt: |
  You are a helpful metadata extraction assistant. You will be responsible for reviewing markdown text content
  that was extracted out of an PDF document. You will be provided the markdown text from a single document, and
  then pull specific metadata based on the users prompt. You will construct a single JSON object, in the below
  format.  If the field isn't defined in the text, provide a value of "unknown".

  {
    "NAPPRA Type": "Quarantine Pest Plant",
    "Pathogen/Insect/Weed": "Weed",
    "Scientific Name": "<scientific name of weed>",
    "Family": "<taxonomic_family>",
    "Synonym": [{"<taxonomic_synonym_1>" : "<taxonomic_synonym_1_author>"},
                {"<taxonomic_synonym_2>" : "<taxonomic_synonym_2_author>"},
                {"<taxonomic_synonym_3>" : "<taxonomic_synonym_3_author>"},
                ...
                ],  # list of taxonomic synonyms for the pathogen/insect and authors; list "unknown" for unlisted
                    # author; blank list if no synonyms
    "Country": ["<country1>", "<country2>", "<country3>", ...],  # must be a list of countries with known distribution
                                                                 # of weed
    "Date of Datasheet": "<date>",  # use YYYY-MM-DD format
    "Link to datasheet": "TBD",
    "Notes (older Datasheets)": <"notes>">
  }
```

LLM applied to NAPPRA documents

- 881 NAPPRA forms were extracted with the LLM
  - Data table composed of 117,0000 rows was produced

- Agency savings of ~50 hours for five employees



Pathogen/Insect/Weed	Scientific Name	Host	Country	Date of Datasheet	Filename	Flagged Document	Flag Reason
Pathogen	African soybean dwarf agent	Glycine max	Nigeria	9/3/2013	African Soybean Dwarf Agent.docx	1	Duplicate Scientific Name with another NAPPRA document
Pathogen	African soybean dwarf agent	Glycine max	Nigeria	9/3/2013	African soybean dwarf agent (ASDA).docx	1	Duplicate Scientific Name with another NAPPRA document
Pathogen	Bhendi yellow vein mosaic virus	Abelmoschus	Bangladesh	8/14/2019	Bhendi yellow vein mosaic virus BYVMV Final.docx	1	Duplicate Scientific Name with another NAPPRA document
Pathogen	Bhendi yellow vein mosaic virus	Alcea	Bangladesh	8/14/2019	Bhendi yellow vein mosaic virus BYVMV Final.docx	1	Duplicate Scientific Name with another NAPPRA document
Pathogen	Bhendi yellow vein mosaic virus	Althaea	Bangladesh	8/14/2019	Bhendi yellow vein mosaic virus BYVMV Final.docx	1	Duplicate Scientific Name with another NAPPRA document
Pathogen	Bhendi yellow vein mosaic virus	Hibiscus	Bangladesh	8/14/2019	Bhendi yellow vein mosaic virus BYVMV Final.docx	1	Duplicate Scientific Name with another NAPPRA document

Document Harvest Toolkit

- Partnership with Johns Hopkins University Applied Physics Lab
- Developing a general toolkit to extract text from documents, use LLM prompts to create structured metadata from documents, and validate outputs
  - Generative AI
  - Using LLMs hosted on secure server
  - Python package, with graphical user interface
- The AI model works well for extracting and categorizing unstructured data sources (word files, pdfs, images…etc…)

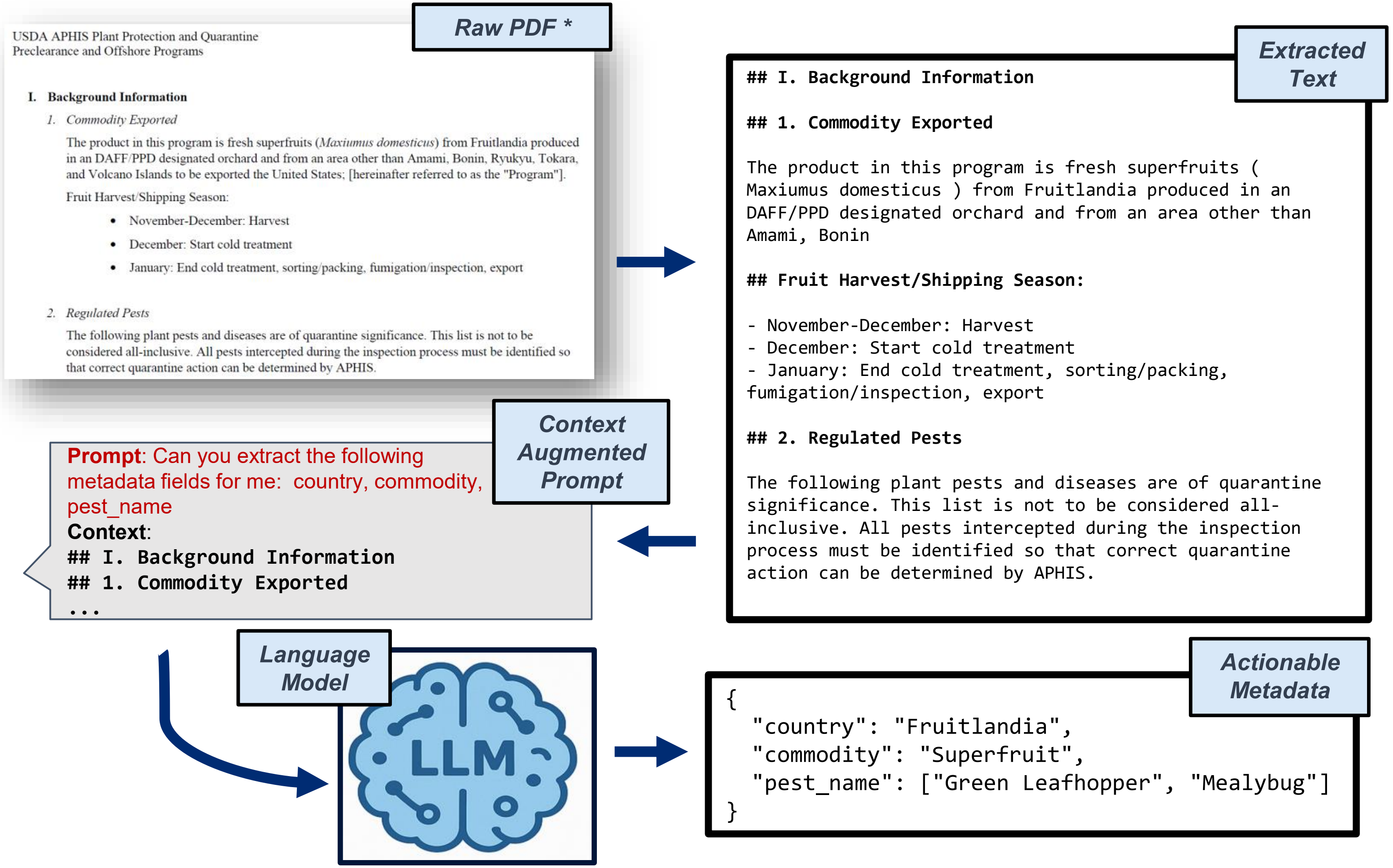
Doc-Harvest-Toolkit Example Workflow

To the right is an example workflow using the Doc-Harvest Toolkit.

Stages:

1. Obtain Raw PDF
2. Extract Text and Cleanly Format
3. Augment Prompt with Document Context
4. Submit Prompt to Language Model (LLM)
5. Parse LLM response for actionable, structured metadata

Approach: Raw PDF to Actionable Metadata



Raw PDF *

USDA APHIS Plant Protection and Quarantine
Preclearance and Offshore Programs

I. Background Information

1. *Commodity Exported*

The product in this program is fresh superfruits (*Maxiumus domesticus*) from Fruitlandia produced in an DAFF/PPD designated orchard and from an area other than Amami, Bonin, Ryukyu, Tokara, and Volcano Islands to be exported the United States; [hereinafter referred to as the "Program"].

Fruit Harvest/Shipping Season:

- November-December: Harvest
- December: Start cold treatment
- January: End cold treatment, sorting/packing, fumigation/inspection, export

2. *Regulated Pests*

The following plant pests and diseases are of quarantine significance. This list is not to be considered all-inclusive. All pests intercepted during the inspection process must be identified so that correct quarantine action can be determined by APHIS.

Extracted Text

```
## I. Background Information

## 1. Commodity Exported

The product in this program is fresh superfruits (
Maxiumus domesticus ) from Fruitlandia produced in an
DAFF/PPD designated orchard and from an area other than
Amami, Bonin

## Fruit Harvest/Shipping Season:

- November-December: Harvest
- December: Start cold treatment
- January: End cold treatment, sorting/packing,
fumigation/inspection, export

## 2. Regulated Pests

The following plant pests and diseases are of quarantine
significance. This list is not to be considered all-
inclusive. All pests intercepted during the inspection
process must be identified so that correct quarantine
action can be determined by APHIS.
```

Context Augmented Prompt

Prompt: Can you extract the following metadata fields for me: country, commodity, pest_name

Context:

```
## I. Background Information
## 1. Commodity Exported
...
```

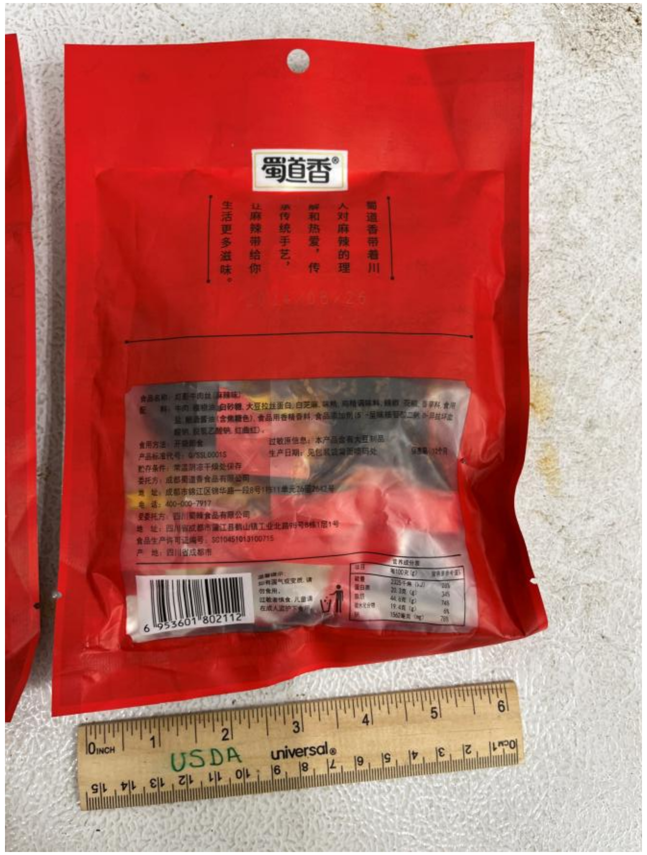
Language Model

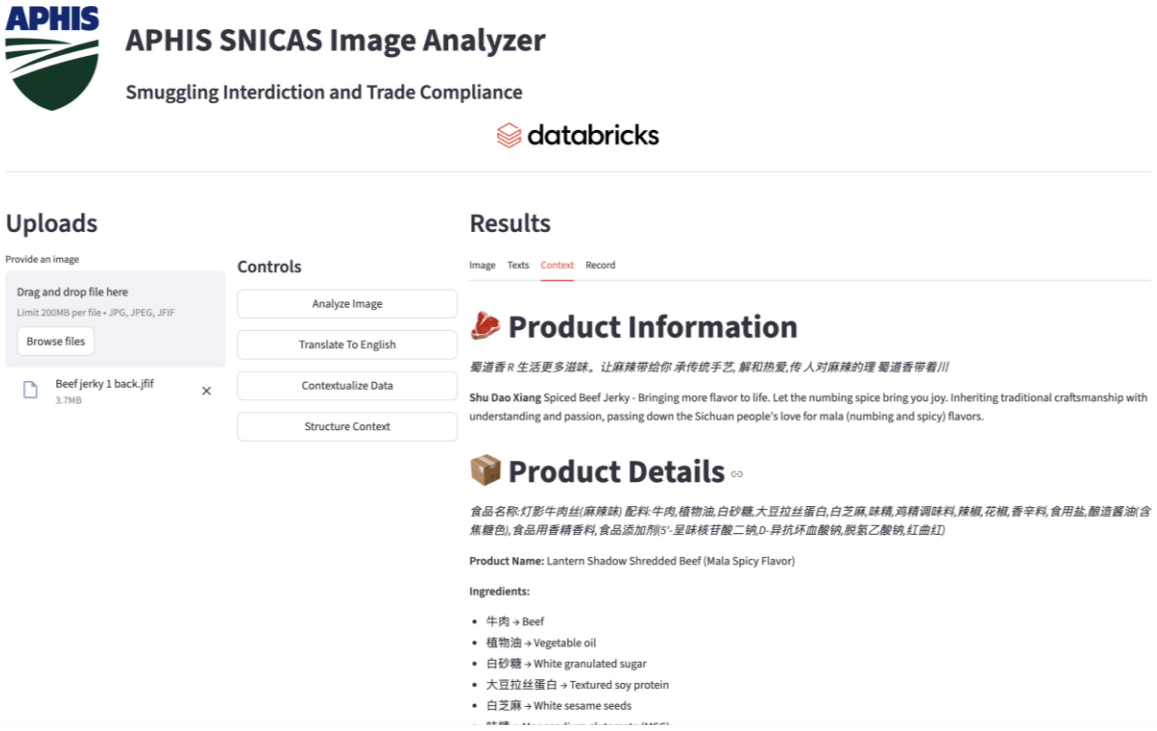
Actionable Metadata

```
{
  "country": "Fruitlandia",
  "commodity": "Superfruit",
  "pest_name": ["Green Leafhopper", "Mealybug"]
}
```

* Note: Documents in this example are notional for illustrative purposes – no real documents are represented

Smuggling Interdiction and Trade Compliance translator tool





APHIS SNICAS Image Analyzer

Smuggling Interdiction and Trade Compliance

databricks

Uploads

Provide an image

Drag and drop file here
Limit 200MB per file • JPG, JPEG, JFIF

Browse files

Beef jerky 1 back.jfif
3.7MB

Controls

Results

Product Information

蜀道香R 生活更多滋味。让麻辣带给你 承传统手艺, 解和热爱,传 人对麻辣的理 蜀道香带着川

Shu Dao Xiang Spiced Beef Jerky - Bringing more flavor to life. Let the numbing spice bring you joy. Inheriting traditional craftsmanship with understanding and passion, passing down the Sichuan people's love for mala (numbing and spicy) flavors.

Product Details

食品名称:灯影牛肉丝(麻辣味) 配料:牛肉,植物油,白砂糖,大豆拉丝蛋白,白芝麻,味精,鸡精调味料,辣椒,花椒,香辛料,食用盐,酿造酱油(含焦糖色),食品用香精香料,食品添加剂(5'-呈味核苷酸二钠,D-异抗坏血酸钠,脱氢乙酸钠,红曲红)

Product Name: Lantern Shadow Shredded Beef (Mala Spicy Flavor)

Ingredients:

- 牛肉 → Beef
- 植物油 → Vegetable oil
- 白砂糖 → White granulated sugar
- 大豆拉丝蛋白 → Textured soy protein
- 白芝麻 → White sesame seeds
- 味精 → Monosodium glutamate (MSG)

Next steps for using AI for structured data creation in PPQ

- Implement human-in-the-loop quality control steps in toolkit
  - Human as a grader of AI-generated products
  - Allow analysts to refine queries and provide reference lists for improving AI model accuracy

Next steps for using AI for structured data creation in PPQ

- Implement human-in-the-loop quality control steps in toolkit
  - Human as a grader of AI-generated products
  - Allow analysts to refine queries and provide reference lists for improving AI model accuracy
- Work with Department to develop approved access and use of LLMs in PPQ's cloud computing environment
- Make the Document Harvest Toolkit widely available to analysts across PPQ

Collaborators:

APL
JOHNS HOPKINS
APPLIED PHYSICS LABORATORY

databricks

Questions?