

NATIONAL CENTER FOR SUPERCOMPUTING APPLICATIONS

Building the Data Infrastructure Solutions of Tomorrow



Research Data Challenges

- Storage
- Everything else!!!
 - The bytes are not enough on their own
00110100 00110010
 - Metadata, curation tools, indexes, storage abstraction, replication, data transfer, authentication, access control, transformation, analysis, tools, computation, ...

Cyberinfrastructure for the 21st Century Vision (CIF21) - 2012

- Develop a deep symbiotic **relationship between science and engineering users and developers of cyberinfrastructure** to simultaneously advance new research practices and open transformative opportunities across all science and engineering fields
- Provide an **integrated and scalable cyberinfrastructure** that leverages existing and new components across all areas of CIF21 and establishes a national data infrastructure and services capability
- Ensure long-term **sustainability** for cyberinfrastructure, via community development, learning and workforce development in CDS&E and transformation of practice

<http://www.nsf.gov/cif21/>





Architectural Vision for Research Cyberinfrastructure

Discipline Specific
Environments

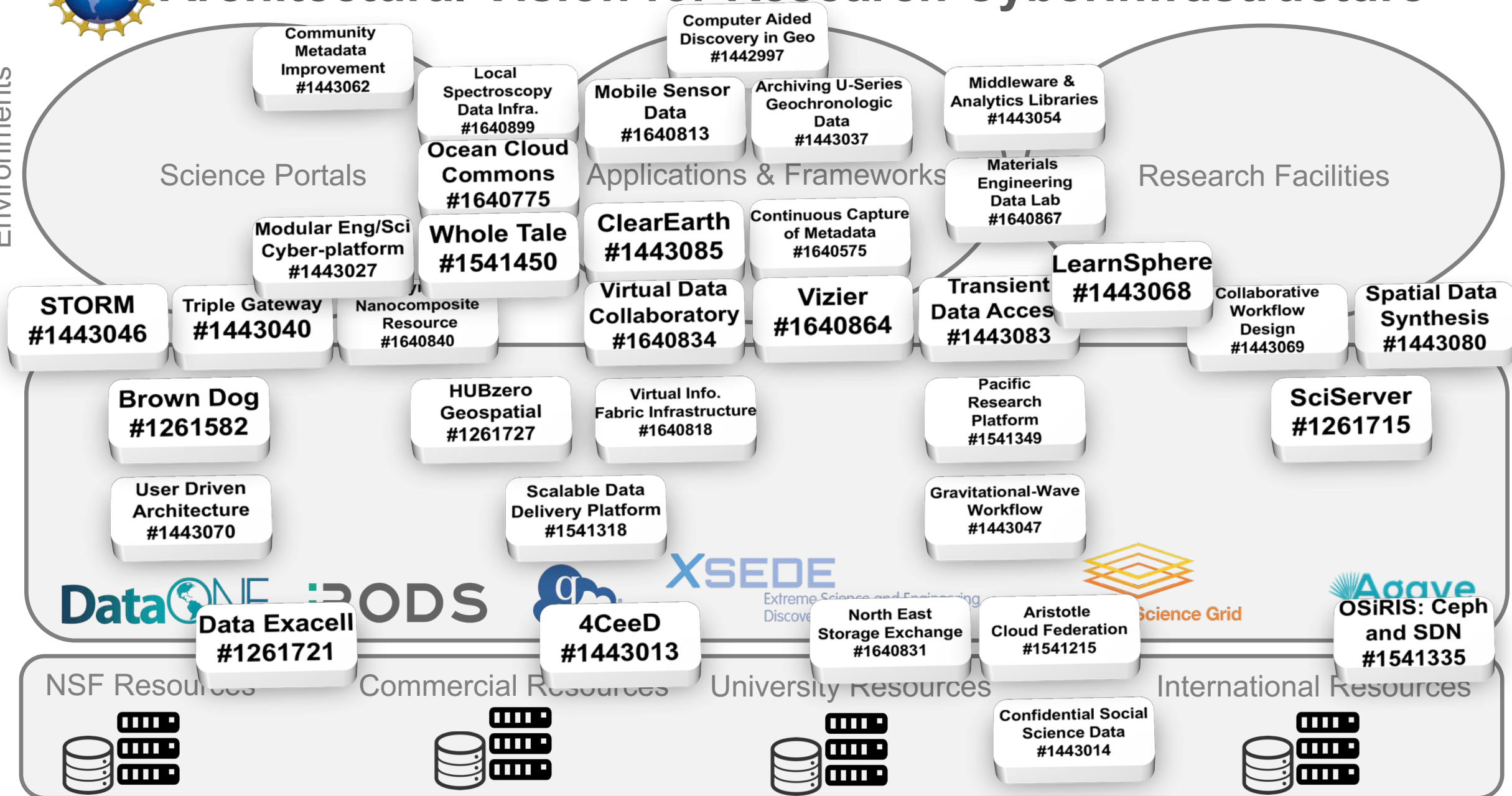
Science Portals

Applications & Frameworks

Research Facilities

Integrative Services

Resources

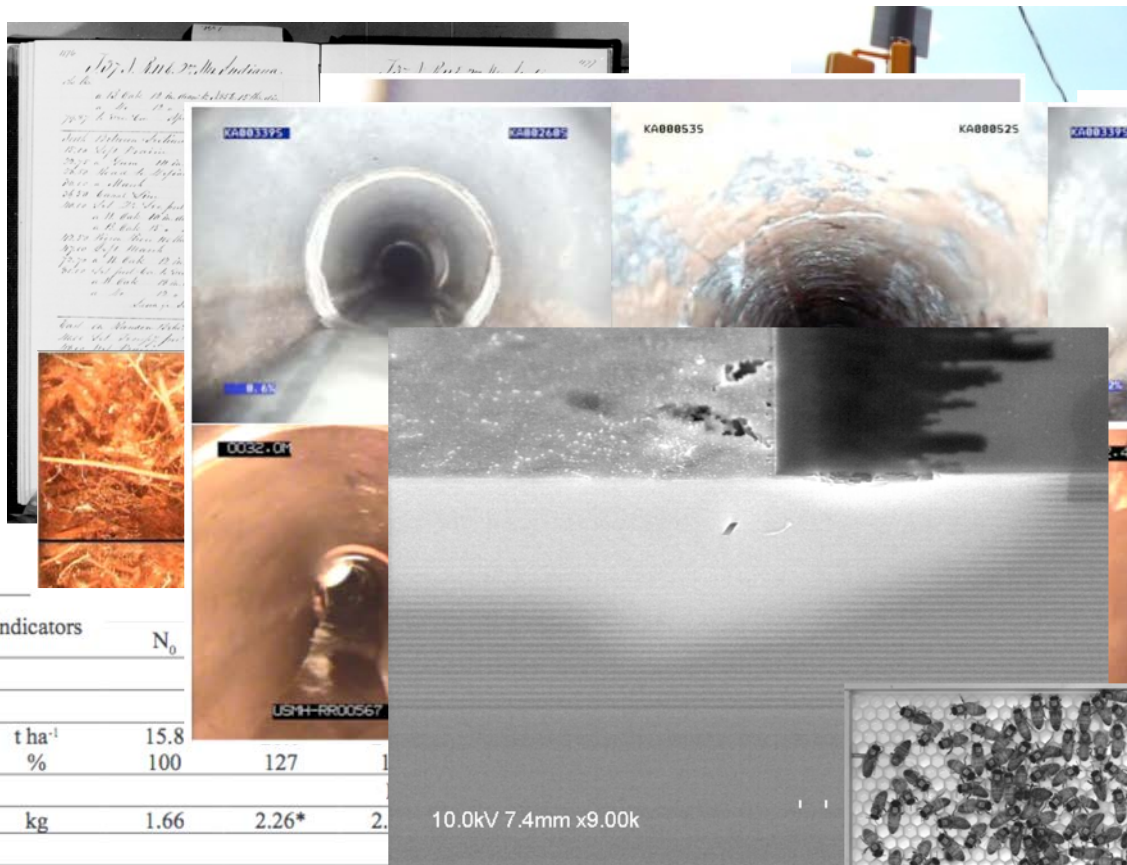


Smarr Taxonomy of Research CI Components


- **Data Applications**
 - Particles, Materials, Astro, Geo, GIS, Bio, Social, Environ, Ag, Medical, Sensors, etc
- **Data Cyberinfrastructure**
 - Computing, Storage, Federation, Clouds, Networking, SDN
- **Data Trust, Security, and Privacy**
- **Data Curation**
 - Capture, Annotation, Documentation, Archiving, Libraries, Management, Publishing
- **Data Discovery and Exploration**
 - Semantics, Ontology, Metadata, Data Mining, Web, Search
- **Data Sharing Middleware**
 - Accessibility, Collaboration, Hubs, Repositories
- **Data Workflows**
- **Data Analytics and Analysis**
 - Data-Intensive Computing, Machine Learning, NLP, Statistics

Photos Structural Defects Image Stitching
Materials Development Gap Filling Green Infrastructure
Tabular Data Root Tip Tracking Loss of Organ Function
Evolution Underwater Photos Mapping NLP Databases
Video Lidar Hyperspectral Climate Modeling
Historic Maps Color Correction Microscopy Images
Flood Plain Analysis Documents Publications Radar
Cell Tracking Web Sites River Depth Distribution Hazard Modeling
Bee Colony Behavior Feedlot Tracking People Detection/Tracking
Phenomics River Maturity Paleoclimate Stream Detection and Sinuosity
Pollen Detection/Classification Human Preference Modeling Bee Detection/Tracking
Land Cover/Usage Regions in Conflict Satellite/Aerial Photos Tissue Classification
Coastline Changes Species Detection/Counting Disease Detection
Large Dynamic Group Behavior Water Detection (e.g. Lakes, Retaining Ponds) Reef Changes
Handwritten Documents River Meander Sentiment Analysis
3D Reconstruction Pre-Digital Datasets Food Supply
Event Detection Simulations Renal Failure

Thomson N19 Ver6.00 Rev.00, Rev. N16 Exp. 00.000



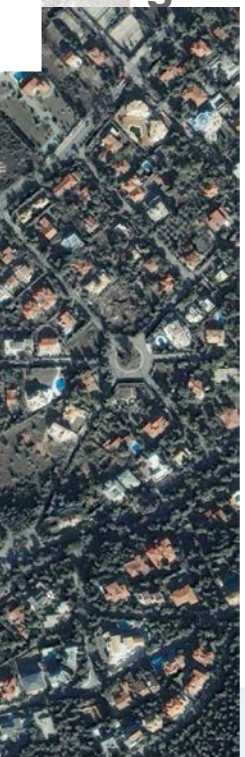
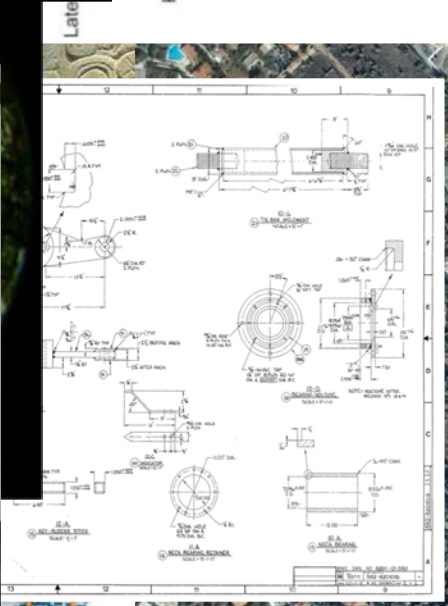
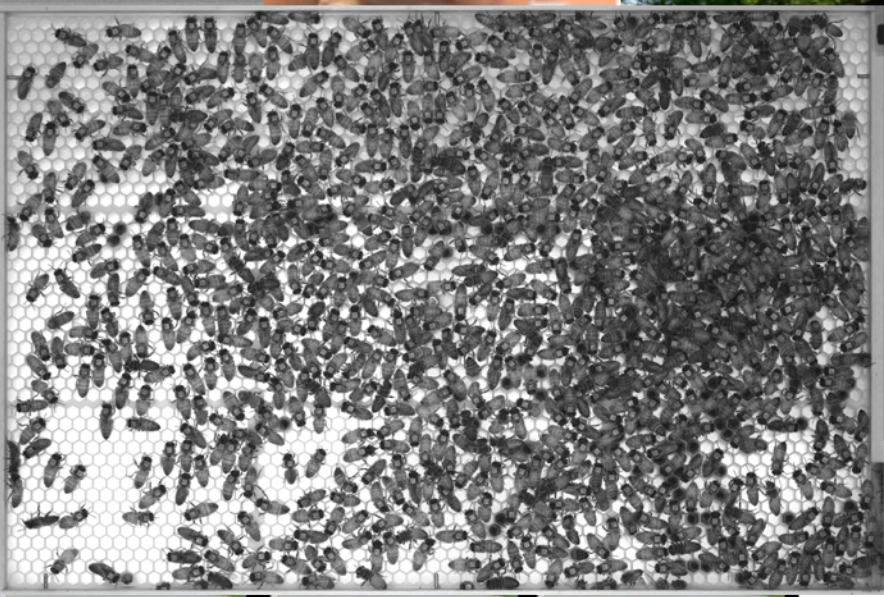
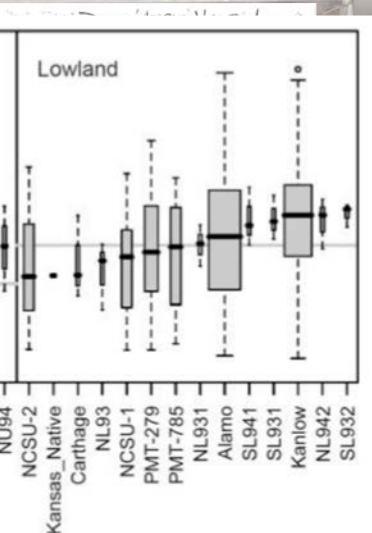
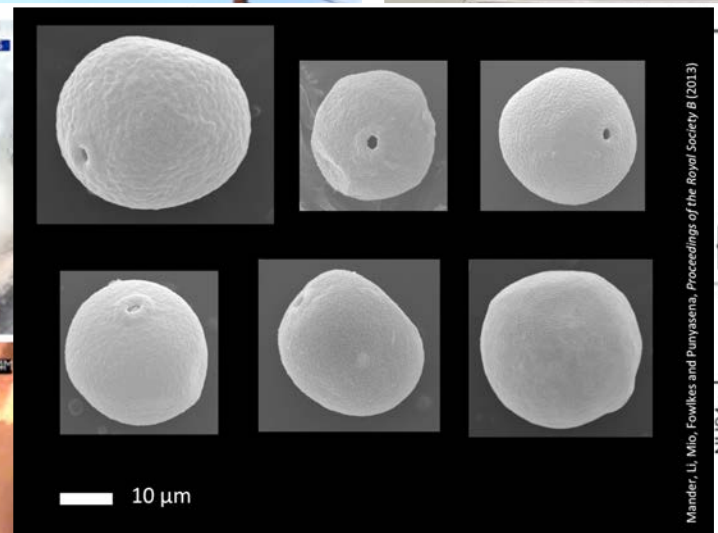
| Indicators | N ₀ | | | | | |
|--------------------|----------------|-------|----|--|--|--|
| t ha ⁻¹ | 15.8 | | | | | |
| % | 100 | 127 | 1 | | | |
| kg | 1.66 | 2.26* | 2. | | | |



USHH-RR00567

10.0kV 7.4mm x9.00k

| Annual biomass | | | | | | |
|--------------------------|------|-------|-------|-------|------|-------|
| t ha ⁻¹ | 27.0 | 28.5 | 29.7 | 5.31 | 10.5 | 10.7 |
| % | 100 | 105.6 | 110.1 | 18.68 | 100 | 102.3 |
| Biomass weight per plant | | | | | | |
| kg | 2.05 | 2.18 | 2.25 | 0.396 | 0.79 | 0.81 |



3D Reconstruction

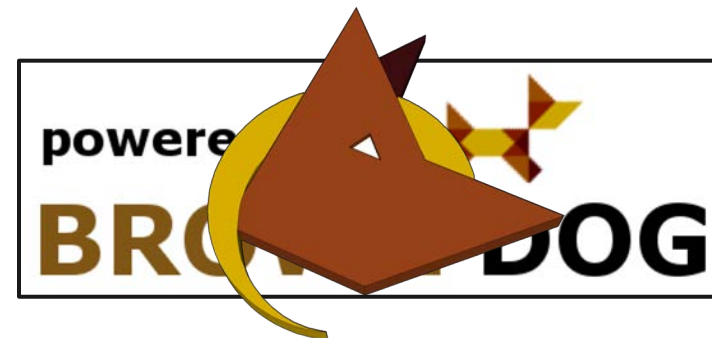
Event Detection

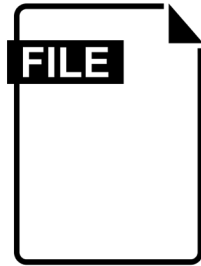
simulations

References

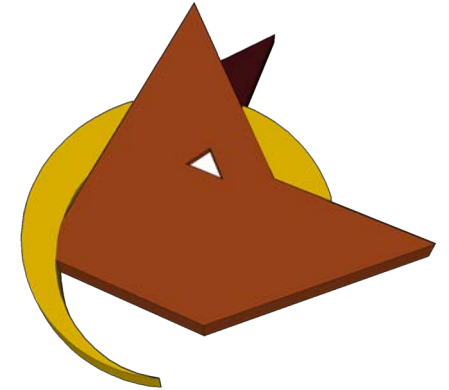
Brown Dog - A Science Driven Data Transformation Service

- Extensibility
 - Easy to add new transformations (i.e. converters and extractors)
 - Encapsulated transformation software & dependencies
- API
 - Supporting other applications/frameworks to build on top of
- Support for diverse usage (i.e. clients, languages, community tools & applications)
- Scalability, Distributed, Data Movement, Provenance with File Validation & Information Loss, Tool Preservation & Publication, **Open Source**





Conversion



Extraction

```
{
  {
    "extractor_id": "ncsa.image.exif",
    "Image": "558c3d84e4b00c3a039d5ac5",
    "Format": "JPEG (Joint Photographic Experts Group JFIF format)",
    "Class": "DirectClass",
    "Geometry": "2592x1936+0+0",
    "Resolution": "72x72",
    "Print size": "36x26.8889",
    "Units": "PixelsPerInch",
    "Type": "TrueColor",
    "Endianness": "Undefined",
    "Colorspace": "sRGB",
    "Depth": "8-bit",
    "Channel depth": {
      "red": "8-bit",
      "green": "8-bit",
      "blue": "8-bit"
    },
    "Channel statistics": {
      "Red": {
        "min": "0 (0)",
```

```
{
  {
    "id": "558c3d84e4b00c3a039d5ac5",
    "filename": "IMG_0997.JPG",
    "tags": [
      "Human Face Automatically Detected",
      "Person Automatically Detected",
      "Human Eyes Automatically Detected"
    ]
  }
}
```

```
{
  {
    "extractor_id": "ncsa.image.ocr",
    "ocr_simple": [
      "EB BROWSER MOSAIC THE FIRST POPULAR BROWSER FOR THE WORLD WIDE",
      "BY MARC ANDREESSEN BINA THE NATIONAL CENTER COMPUTING APPLICATIONS",
      "1993 RELEASE TO THE PUBLIC INTERNET USERS EASY ACCESS TO SOURCES OF",
      "INFORMATION win HAVE TRANSFORMED THE INFORMATION UNIVERSITY OF "
    ],
    "Human Preference Extractor": {
      "Definitions": {
        "Human Preference": "A Computer Vision model that uses the",
        "Green Index": "The green index is the estimated percentage of green pixels w"
      },
      "Human Preference": "4",
      "Green Index": "53.8"
    },
    "tags": [
      "ts-dev.ncsa.illinois.edu:9000/files/558c3d84e4b00c3a039d5ac5"
    ],
    "datasets": [
      "ts-dev.ncsa.illinois.edu:9000/datasets/558c3dd6e4b00c3a039d5b77"
    ]
  }
}
```

Brown Dog

```
curl -s -F "File=@IMG0008.PCD" https://bd-api.ncsa.illinois.edu/v1/conversions/pgm/ -H "Transfer-Encoding: chunked" -H "Accept: text/plain" -H "Authorization: e6dab924-04c8-45c0-94aa-f0608c3c1a45"
```

```
response = requests.post('https://bd-api.ncsa.illinois.edu/v1/conversions/ed.zip/', files={'file': open("US-Dk3-2001-2003.xml", 'rb')}, headers={'Accept': 'text/plain', 'Authorization': 'e6dab924-04c8-45c0-94aa-f0608c3c1a45'})
```

```
curl -s https://bd-api.ncsa.illinois.edu/v1/extractions/url/ -X POST -d '{"fileurl":"http://browndog.ncsa.illinois.edu/examples/IMG_0997.jpg"}' -H "Content-Type: application/json" -H "Authorization: e6dab924-04c8-45c0-94aa-f0608c3c1a45" | jq -r ".id"
```


Selected Site

Set parameters for the run.

PFT*

populus
temperate.coniferous
temperate.deciduous

Start Date*

2004/01/01

End Date*

2004/12/31

Sipnet.climna*

Use Ameriflux

Email

Use [BrownDog](#)

Edit pecan.xml

Edit model config

Advanced setup

pecan-dev.ncsa.illinois.edu/pecan/03-inputs.php

Map Satellite

Missouri Ozark Site/B
Ashland, MO, US

Google

The [PEcAn project](#) is supported by the National Science Foundation

Generate xml file

pecan-dev.ncsa.illinois.edu/shiny/BrownDog/

Type

AmeriFlux

☒ I agree to AmeriFlux license.

Start year

2001

End year

2001

BrownDog Token

c8ac077a-0cd8-4486-8be6-d2bb70c14abb

Model

ed.zip

ameriflux.zip
clim
dalec
ed.zip
linkages
pecan.nc
pecan.zip

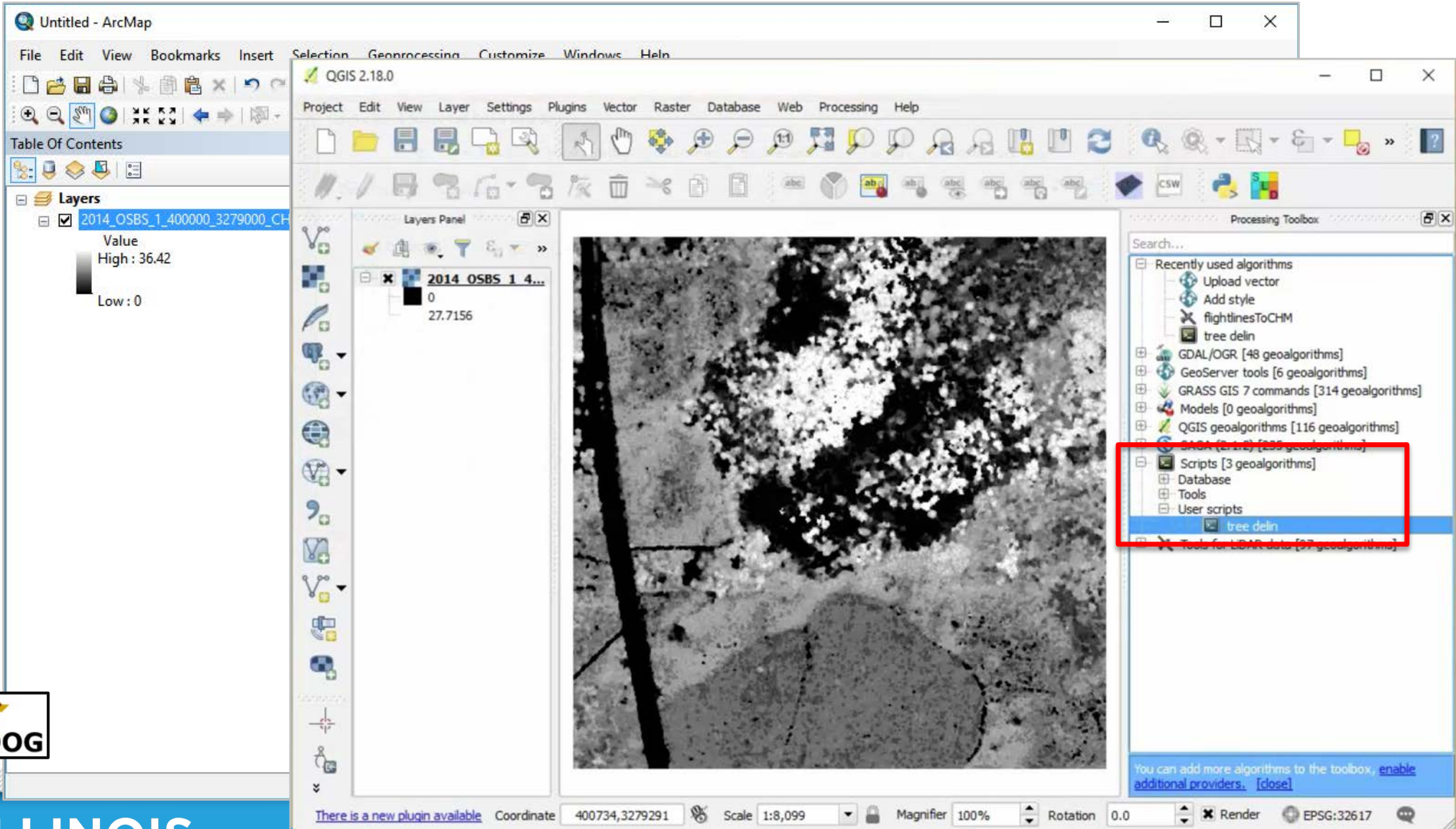
Freeman Ranch- Woodland (US-FR3)

```
<input>
<type>AmeriFlux</type>
<site>Freeman Ranch- Woodland (US-FR3)</site>
<lat>29.94</lat>
<lon>-97.99</lon>
<start_date>2001-01-01 00:00:00</start_date>
<end_date>2001-12-31 23:59:59</end_date>
</input>
```

Download XML

Download Data

Geospatial Software



General Software

example-path

Search Sheet

Home Insert Page Layout Formulas Data Review View

Paste

Calibri (Body) 12 A A

B I U

Alignment

Number

Conditional Formatting

Format as Table

Cell Styles

Cells

Editing

A1 Route

| | A | B | C | D | E |
|----|-------|----------|-----------|-------------|-------------|
| 1 | Route | Lat | Long | Green Index | Orientation |
| 2 | 1 | 40.11123 | -88.17377 | 30.8 | 0 |
| 3 | 1 | 40.11123 | -88.17377 | 37.7 | 90 |
| 4 | 1 | 40.11123 | -88.17377 | 38.9 | 180 |
| 5 | 1 | 40.11123 | -88.17377 | 27.6 | 270 |
| 6 | 1 | 40.11202 | -88.17488 | 32.1 | 270 |
| 7 | 1 | 40.11202 | -88.17488 | 20.9 | 180 |
| 8 | 1 | 40.11202 | -88.17488 | 32.2 | 0 |
| 9 | 1 | 40.11202 | -88.17488 | 47.1 | 90 |
| 10 | 1 | 40.11289 | -88.17517 | 31.6 | 0 |
| 11 | 1 | 40.11289 | -88.17517 | 16.9 | 90 |
| 12 | 1 | 40.11289 | -88.17517 | 19.9 | 270 |
| 13 | 1 | 40.11289 | -88.17517 | 24.1 | 180 |

Brown Dog

Green Route Index

1. Provide an Access Token:

8fde62b9-436f-4d69-a64...

2. Select cells in the sheet as a 2 column list of latitude, longitude pairs along a path.

3. Submit to Brown Dog service. A new sheet will be added to the workbook with the results once they are ready.

Submit Selection

Submitted 6 rows and 2 columns

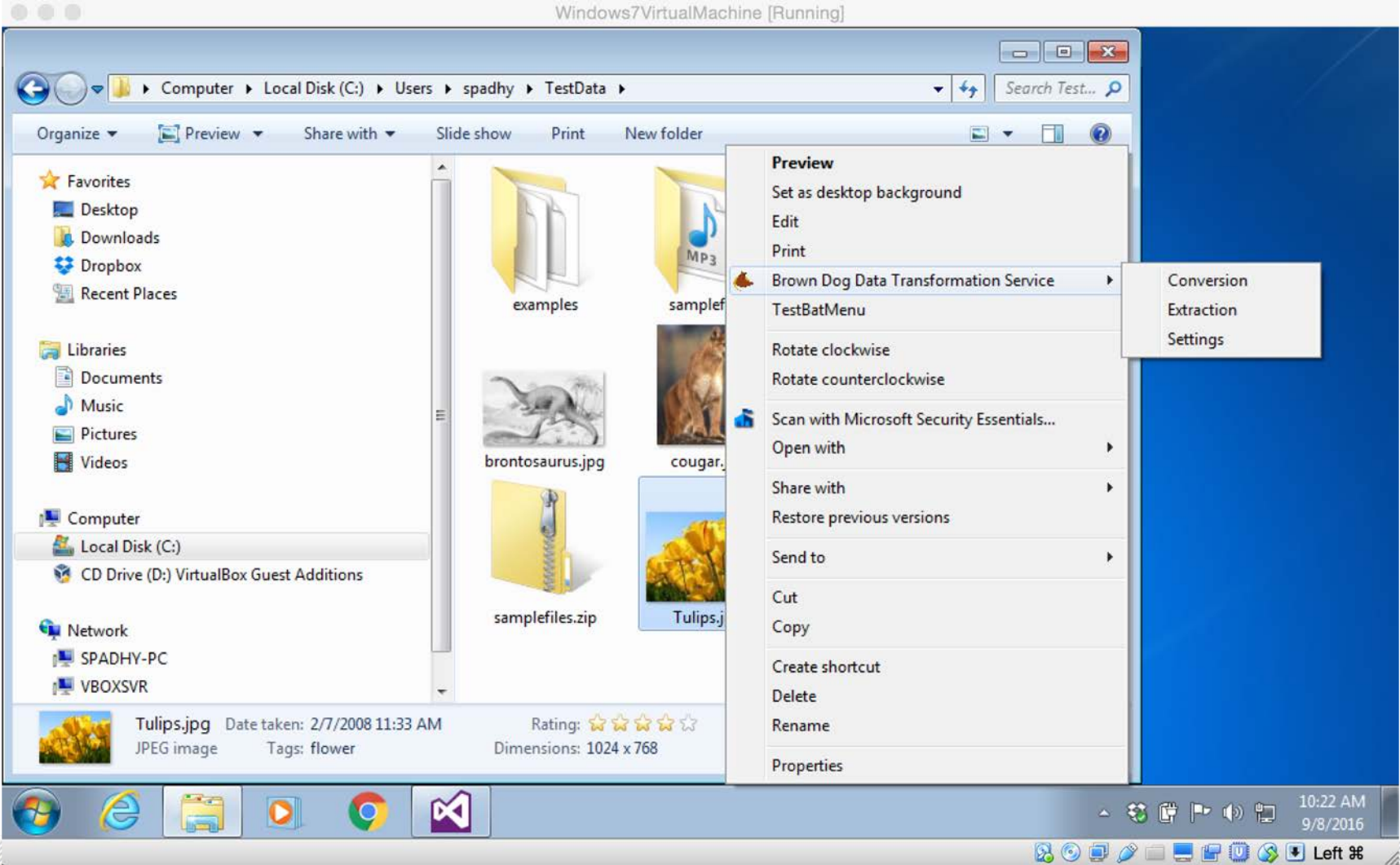
Adding metadata to sheet

[Original File Metadata](#)

- Derived File 0 metadata
- Derived File 1 metadata
- Derived File 2 metadata
- Derived File 3 metadata
- Derived File 4 metadata
- Derived File 5 metadata
- Derived File 6 metadata
- Derived File 7 metadata
- Derived File 8 metadata
- Derived File 9 metadata
- Derived File 10 metadata
- Derived File 11 metadata



Operating Systems



| | |
|---------------|----------------------------------------------------------------------------------|
| Lines of Code | 273 |
| Other files | Model Image Sample |
| Dependencies | numpy, argparse, glob, cv2, cPickle, random, h5py, skimage, sklearn, scipy |
| Difficulties | Install OpenCV (cv2) |



```

ap.add_argument("-p", "--path",
                help="path to gi_detector")
args = vars(ap.parse_args())

path_1 = '/Users/ankitrai/Dropbox/ppao_VM/gi_detector/'
def pyramid(image, scale=1.5, minSize=(55,55)):
    # yield the original image
    yield image

    # keep looping over the pyramid
    while True:
        # compute the new dimensions of the image and resize it
        w = int(image.shape[1] / scale)
        image = imutils.resize(image, width=w)

        # if the resized image does not meet the supplied minimum
        # size, then stop constructing the pyramid
        if image.shape[0] < minSize[1] or image.shape[1] < minSize[0]:
            break

        # yield the next image in the pyramid
        yield image
def sliding_window(image, stepSize, windowSize):
    # slide a window across the image
    for y in xrange(0, image.shape[0], stepSize):
        for x in xrange(0, image.shape[1], stepSize):
            # yield the current window
            yield (x, y, image[y:y + windowSize[1], x:x + windowSize[0]])
def detect(image, winDim, winStep=4, pyramidScale=1.5, minPr
    # initialize the list of bounding boxes and associat

    # loop over the image pyramid
    pyramid_layers = pyramid(image, scale=1.5, minSize=(100,100))
    for layer in pyramid_layers:
        # determine the current scale of the pyramid
        scale = image.shape[0] / float(layer.shape[0])
        # loop over the sliding windows for the current pyramid layer
        for (x, y, window) in sliding_window(layer, winStep, winDim):
            (winH, winW) = window.shape[:2]
            if winH == winDim[1] and winW == winDim[0]:
                # extract HOG features from the current window and classifi

```



Total Code
from 2 Files

| | |
|---------------|---------------------------------------------------|
| Lines of Code | 47 |
| Other files | None |
| Dependencies | bd, requests, os, glob, argparse, time, json, PIL |
| Difficulties | |

```

import sys
import requests
import os
import glob
import argparse
import time
import json
from PIL import Image, ImageDraw
from bd import extract

# Construct the argument parser and parse the arguments
ap = argparse.ArgumentParser()
ap.add_argument("-t", "--token", required=True, help="Token")
ap.add_argument("-i", "--images", help="Path to image directory")
ap.add_argument("-b", "--bdapi", help="Brown Data API")
args = vars(ap.parse_args())

# Parse the arguments
token = args["token"]
if args["images"] != None:
    image_path = args["images"]
else:
    image_path = "./images"
if args["bdapi"] != None:
    bdapi = args["bdapi"]
    if bdapi[-1] != "/":
        bdapi = bdapi + "/"
else:
    bdapi = "https://bd-api.ncsa.illinois.edu/"

# Loop through images in image directory
for image_path in glob.glob(image_path + "/*.jpg"):
    metadata = extract(bdapi, image_path, token)['metadata.jsonld']

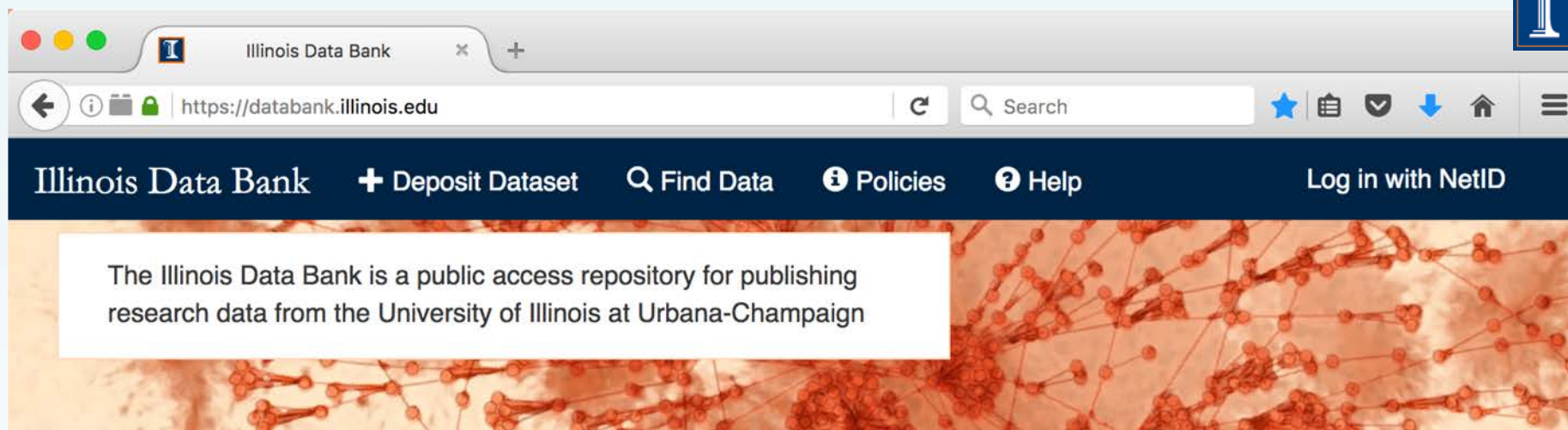
    for mdata in metadata:
        if "y0" in mdata["content"]:
            x0 = mdata["content"]["x0"]
            x1 = mdata["content"]["x1"]
            y0 = mdata["content"]["y0"]
            y1 = mdata["content"]["y1"]
            print "coordinates for bioswale bounding box in" + image_path
            print "[(x0,y0),(x1,y1)] = [{" + str(x0) + "," + str(y0) + "}, {" + str(x1) + "," + str(y1) + "}]"]

    img = Image.open(image_path)
    draw = ImageDraw.Draw(img)
    draw.rectangle([(x0, y0), (x1, y1)])
    img.show()

```

Total Code
from 1 File





The screenshot shows a web browser window with the URL <https://databank.illinois.edu>. The page features a dark blue navigation bar with the following links: Illinois Data Bank, + Deposit Dataset, 🔍 Find Data, ⓘ Policies, ⓘ Help, and Log in with NetID. Below the navigation bar is a large banner with a network diagram background. A white text box on the banner contains the text: "The Illinois Data Bank is a public access repository for publishing research data from the University of Illinois at Urbana-Champaign".

You are ready to deposit data if:

- your data is in a final state and not expected to undergo revisions.
- you have removed any private, confidential, or other legally protected information from your data.
- you are a faculty member, staff member, or graduate student at the University of Illinois at Urbana-Champaign.
- you have permission to publicly distribute data from all creator(s) and/or copyright owner(s).

[Learn how to publish your data](#)

Published data:

- is open to anyone in the world.
- receives a stable identifier ([DOI](#)) for easy reference and citation.
- is readily available for anyone to access for a minimum of 5 years.
- is located in a stable environment that complies with many funder and publisher requirements.

[Review our policies](#)

The Illinois Data Bank is a product of the [Research Data Service](#) at the University Library. [See our Access and Use Policy](#). [Contact us](#) for questions and to provide feedback.



Clowder (2013-Present)

NSF Innovative Systems and Software: Applications to NARA Research Problems (OCI-0525308)



The screenshot shows the Clowder homepage in a web browser. The address bar displays `https://clowder.ncsa.illinois.edu/clowder/`. The page has a blue header with the Clowder logo, navigation links for 'Explore' and 'Help', a search bar, and 'Sign up' and 'Login' buttons. The main content area features a 'Welcome to Clowder' message, explaining that it is a scalable data repository for sharing, organizing, and analyzing data. A 'Resources' sidebar on the right lists 'Spaces', 'Collections', 'Datasets', 'Files', 'Bytes', and 'Users'. The footer indicates the system is powered by Clowder (1.3.2#16 branch:master sha1:b7a81e8).

Welcome to Clowder

Welcome to Clowder, a scalable data repository where you can share, organize and analyze data. This is a demo instance to try the system out. Please do not use this instance to store real data. We delete the content of this instance when we need to and it does not have very much disk space available. Thank you.

Powered by [Clowder](#) (1.3.2#16 branch:master sha1:b7a81e8).

This screenshot shows a file view in Clowder. The browser address bar shows a file path: `https://clowder.ncsa.illinois.edu/clowder/files/5a1f1481e4b0cfb1ad158e1f?dataset=5a1d7faae4b0cfb1ad158e1f`. The file is named `6011937799.jpg` and is an aerial image from Google Earth. The right sidebar displays file metadata: Type (image/jpeg), File size (494.8 kB), Upload date (Nov 29, 2017 14:11:45), Upload by (Bardia Heidari Haratmeh), and Status (PROCESSED). It also shows the license (All Rights Reserved) and the dataset it belongs to (GI identification dataset). A tag 'Bioswale Detected [Section]' is visible at the bottom.

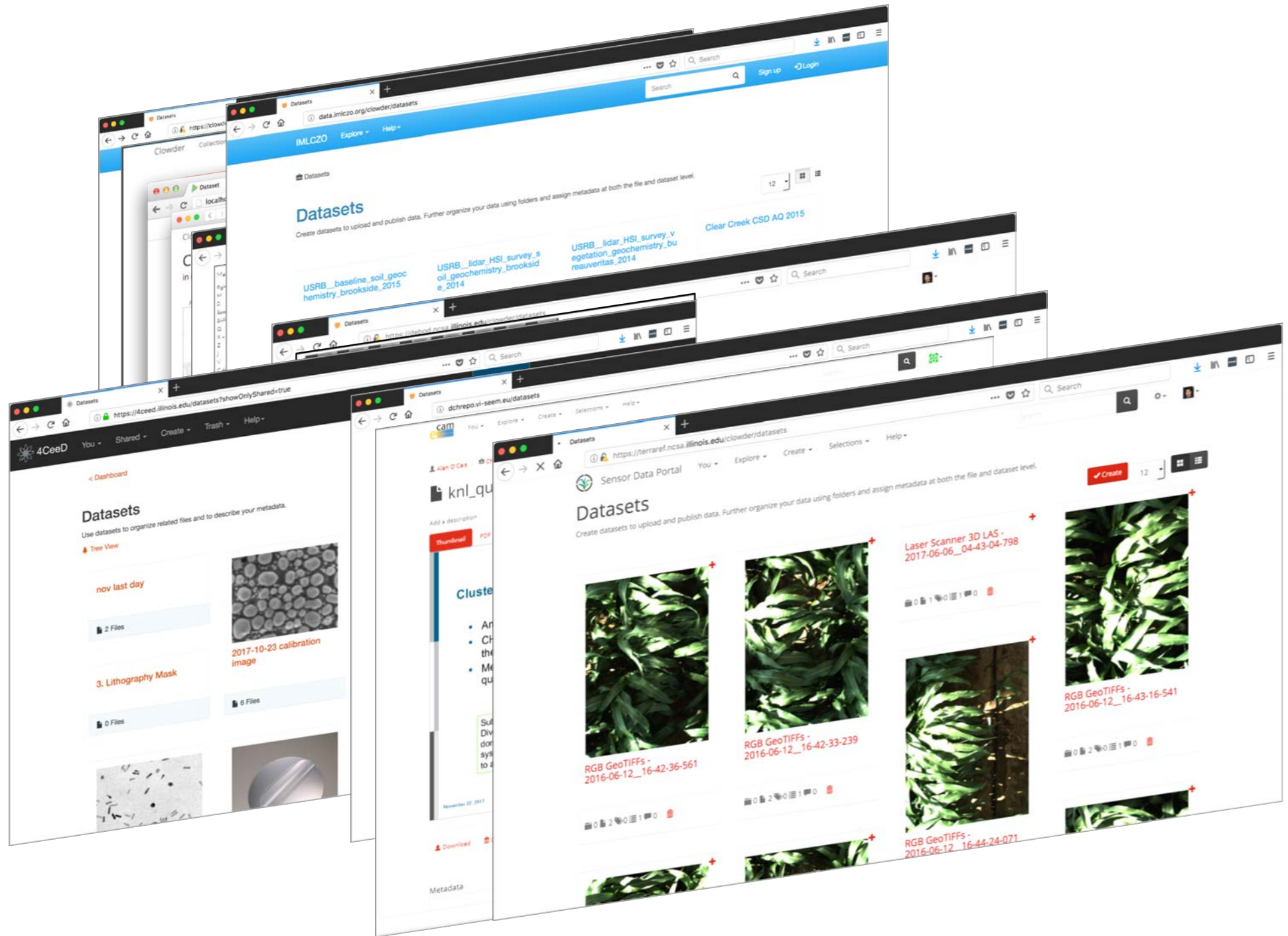
6011937799.jpg

Type: image/jpeg
File size: 494.8 kB
Uploaded on: Nov 29, 2017 14:11:45
Uploaded by: [Bardia Heidari Haratmeh](#)
Status: PROCESSED

License
Type: All Rights Reserved
Holder: Bardia Heidari Haratmeh

Dataset containing the file
[GI identification dataset](#)

Tags
• [Bioswale Detected \[Section\]](#)



MATERIAL SCIENCE

EDUCATION



BIOLOGY



HUMANITIES



SOCIAL SCIENCE



CIVIL ENGINEERING

MEDICINE

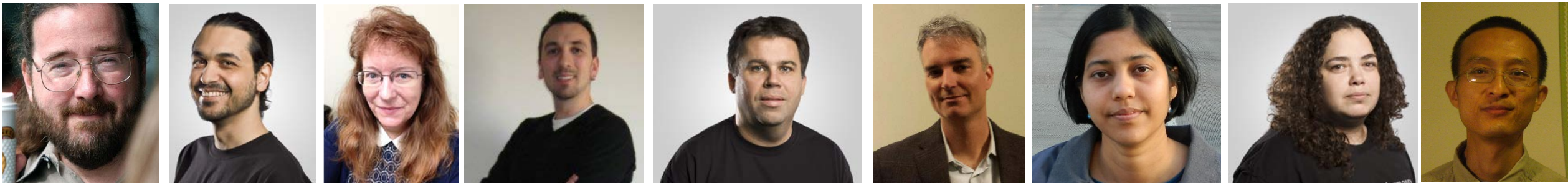


INDUSTRY



GEOSCIENCE





<http://browndog.ncsa.illinois.edu>

 **@NCSABrownDog**