Clowder Open Source Data Management for Long Tail Data

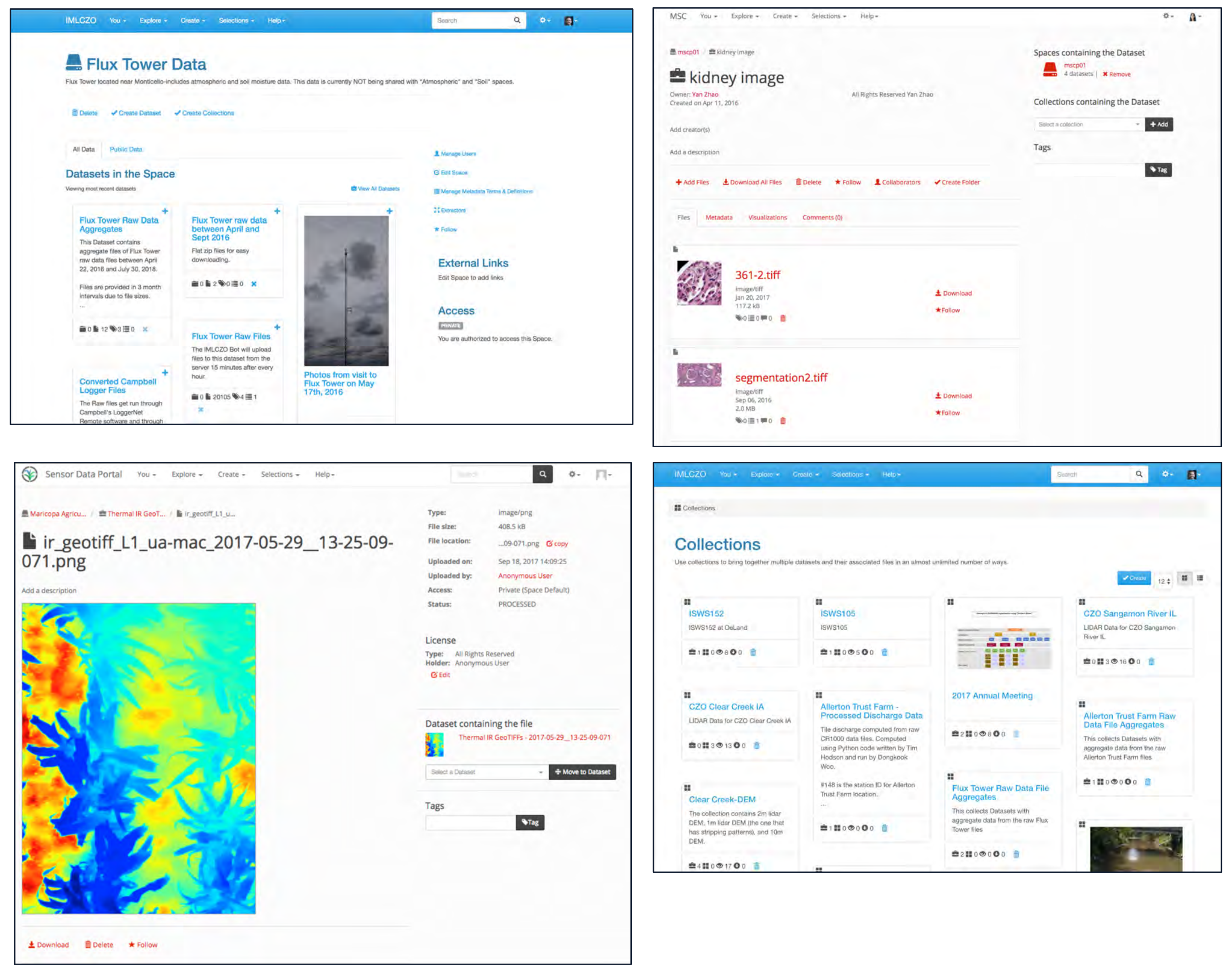
Luigi Marini, Rob Kooper, Indira Gutierrez, Max Burnette, Sandeep Puthanveetil Satheesan, Bing Zhang, Mike Lambert, Todd Nicholson, Yan Zhao, Jong Lee, Kenton McHenry
National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign

https://clowder.ncsa.illinois.edu

Introduction

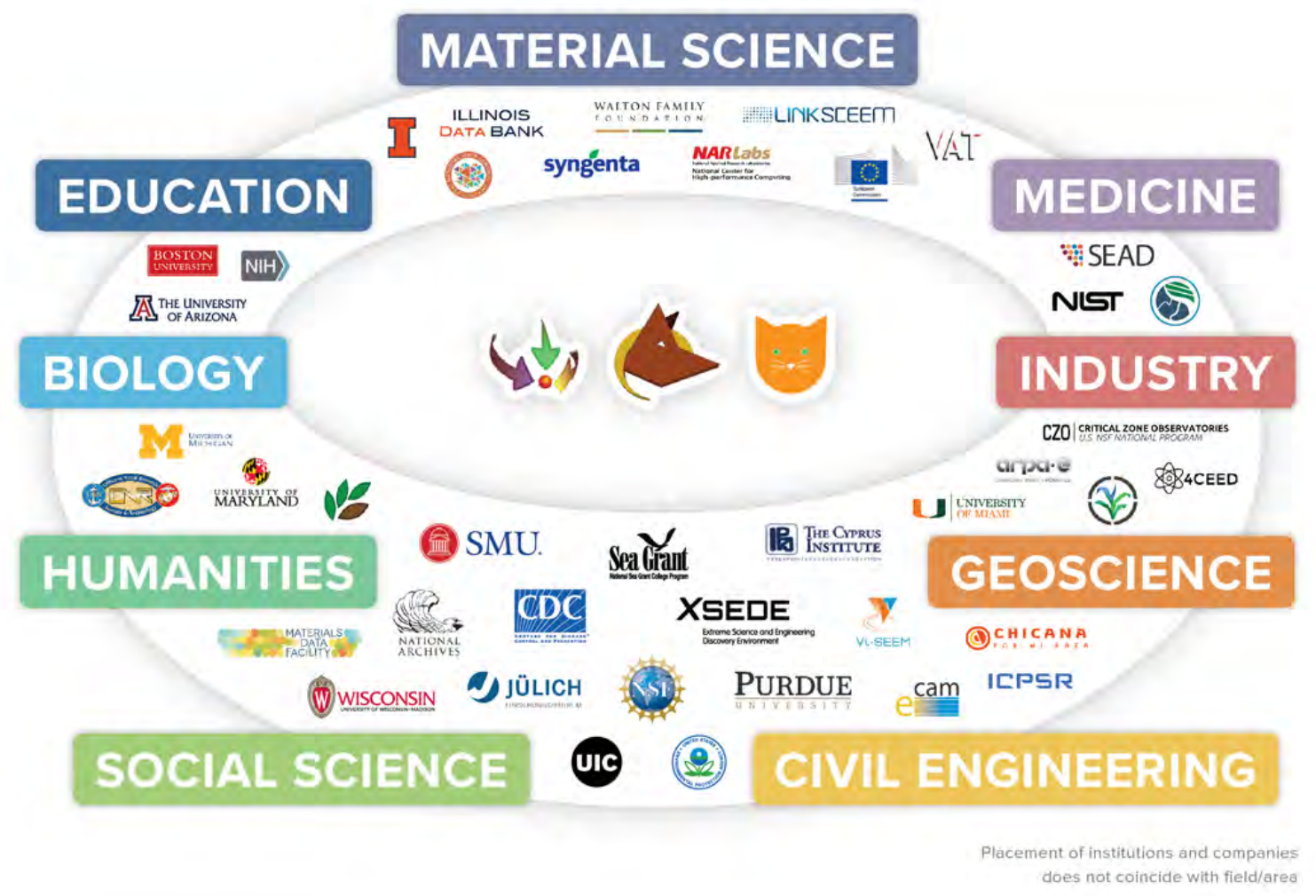
- Clowder is a customizable and scalable data management system you can install anywhere.
- You can install Clowder in the cloud, on your hardware, or you can partner with NCSA for a custom instance.
- You can contribute to the core software or by creating new metadata extractors and data visualizations.

Organize Data



Spaces (top-left), Datasets (top-right), Files (bottom-left), Collections (bottom-right).

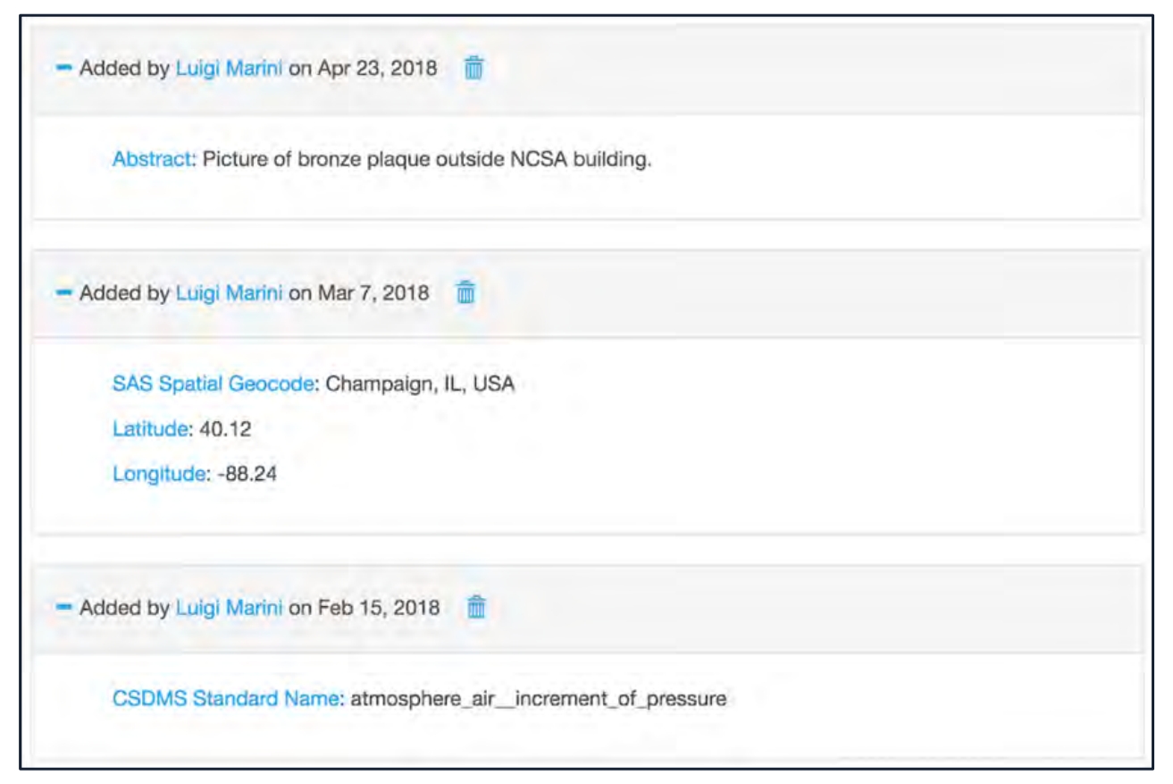
Open Source Community



Used across many different research areas.

Flexible Metadata Representation

Support for both user-defined and machine-defined metadata. System accepts metadata in a flexible representation based on JSON Linked Data.



- Added by Luigi Marini on Apr 23, 2018

Abstract: Picture of bronze plaque outside NCSA building.

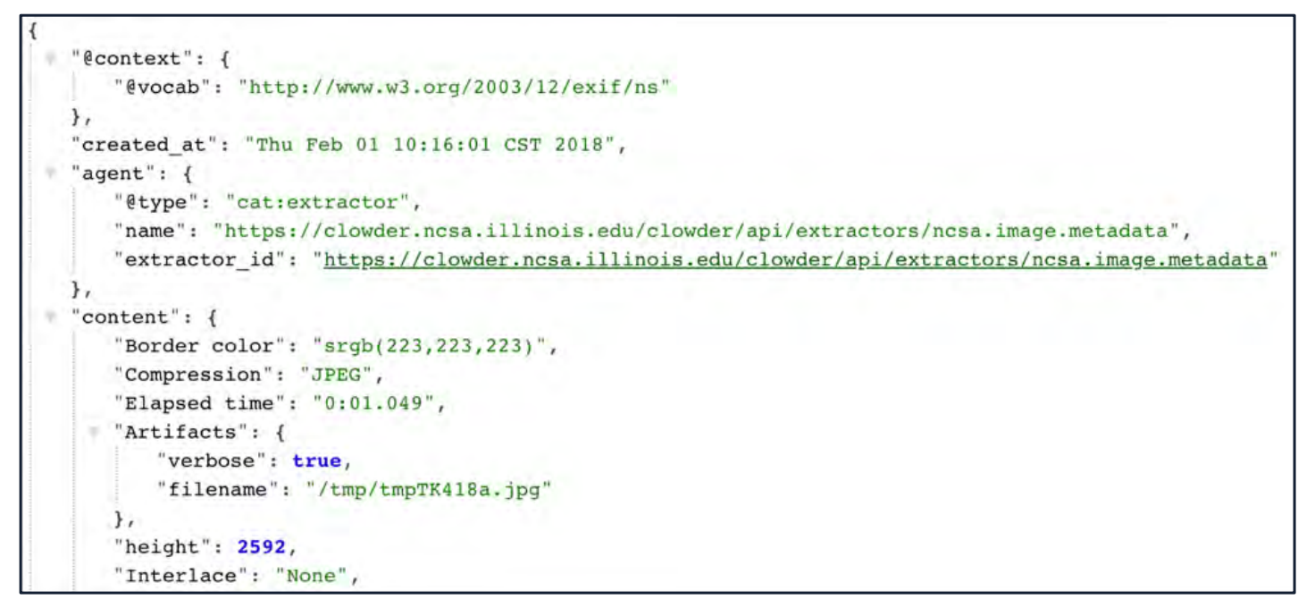
- Added by Luigi Marini on Mar 7, 2018

SAS Spatial Geocode: Champaign, IL, USA
Latitude: 40.12
Longitude: -88.24

- Added by Luigi Marini on Feb 15, 2018

CSDMS Standard Name: atmosphere_air__increment_of_pressure

Users can add metadata entries directly from the user interface.



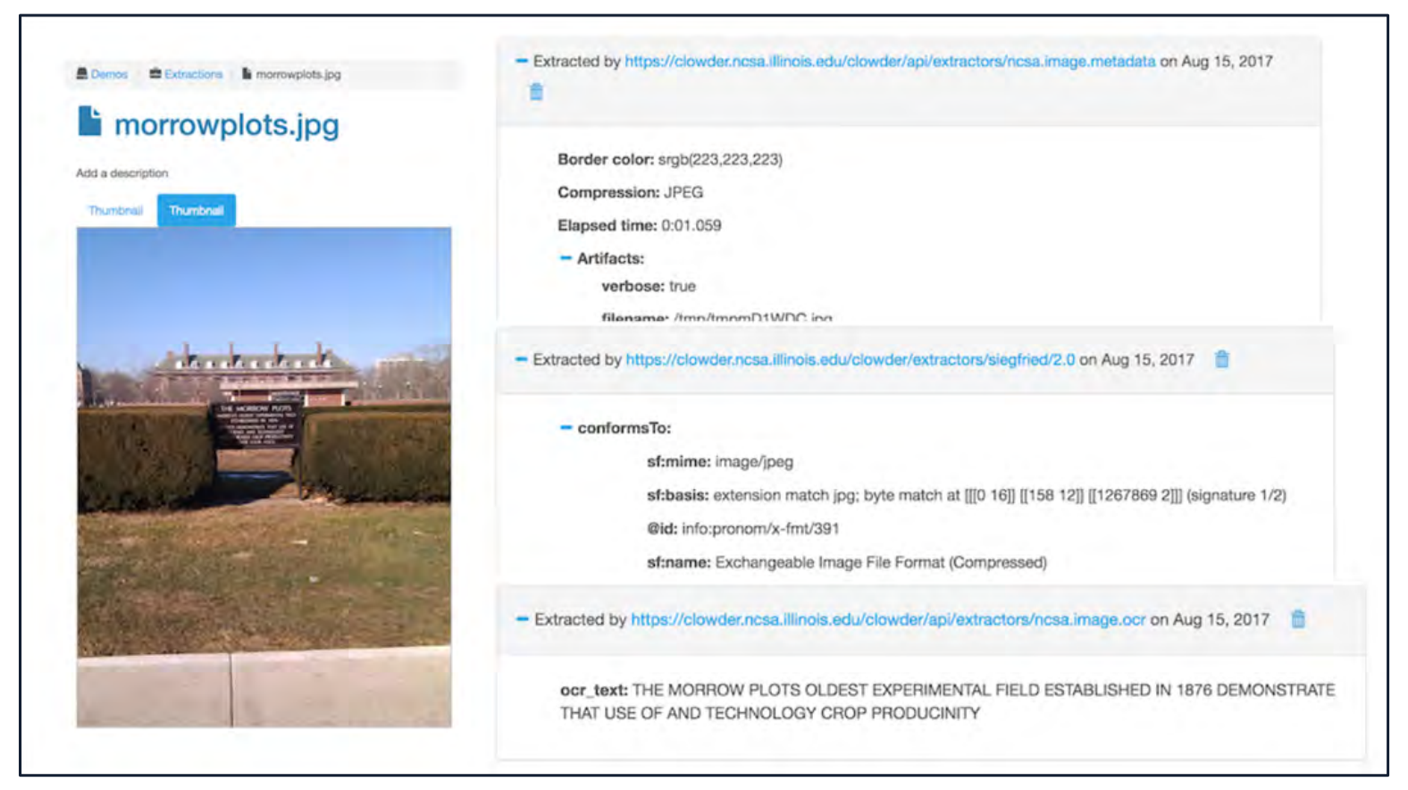
```
{
  "@context": {
    "@vocab": "http://www.w3.org/2003/12/exif/ns"
  },
  "created_at": "Thu Feb 01 10:16:01 CST 2018",
  "agent": {
    "@type": "cat:extractor",
    "name": "https://clowder.ncsa.illinois.edu/clowder/api/extractors/ncsa.image.metadata",
    "extractor_id": "https://clowder.ncsa.illinois.edu/clowder/api/extractors/ncsa.image.metadata"
  },
  "content": {
    "Border color": "srgb(223,223,223)",
    "Compression": "JPEG",
    "Elapsed time": "0:01.049",
    "Artifacts": {
      "verbose": true,
      "filename": "/tmp/tmpTK418a.jpg"
    },
    "height": 2592,
    "Interlace": "None",
```

Extractors and external clients can attach metadata to files and datasets using the Web service API.

Automatic Metadata Extraction

Extend the system by creating new extractors to analyze data. For example, given the image on the right 3 separate extractors computed:

- EXIF from image bytes
- Siegfried file format identification
- OCR of text in image



Demos / Extractions / morrowplots.jpg

morrowplots.jpg

Add a description

Thumbnail Thumbnail

- Extracted by https://clowder.ncsa.illinois.edu/clowder/api/extractors/ncsa.image.metadata on Aug 15, 2017

Border color: srgb(223,223,223)
Compression: JPEG
Elapsed time: 0:01.059
- Artifacts:
  verbose: true
  filename: /tmp/tmpmD1WDC.jpg

- Extracted by https://clowder.ncsa.illinois.edu/clowder/extractors/siegfried/2.0 on Aug 15, 2017

- conformsTo:
  sf:mime: image/jpeg
  sf:basis: extension match jpg; byte match at [[[0 16]] [[158 12]] [[1267869 2]]] (signature 1/2)
  @id: info:pronom/x-fmt/391
  sf:name: Exchangeable Image File Format (Compressed)

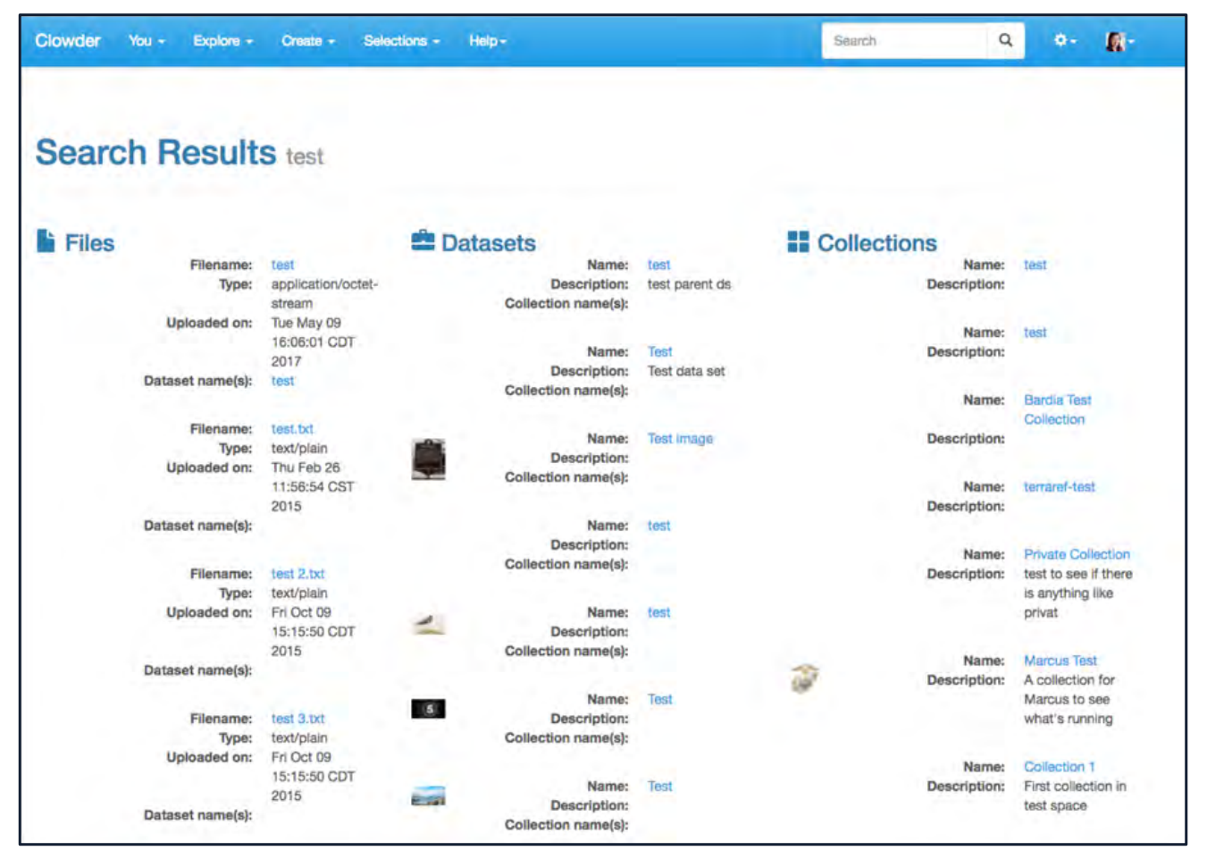
- Extracted by https://clowder.ncsa.illinois.edu/clowder/api/extractors/ncsa.image.ocr on Aug 15, 2017

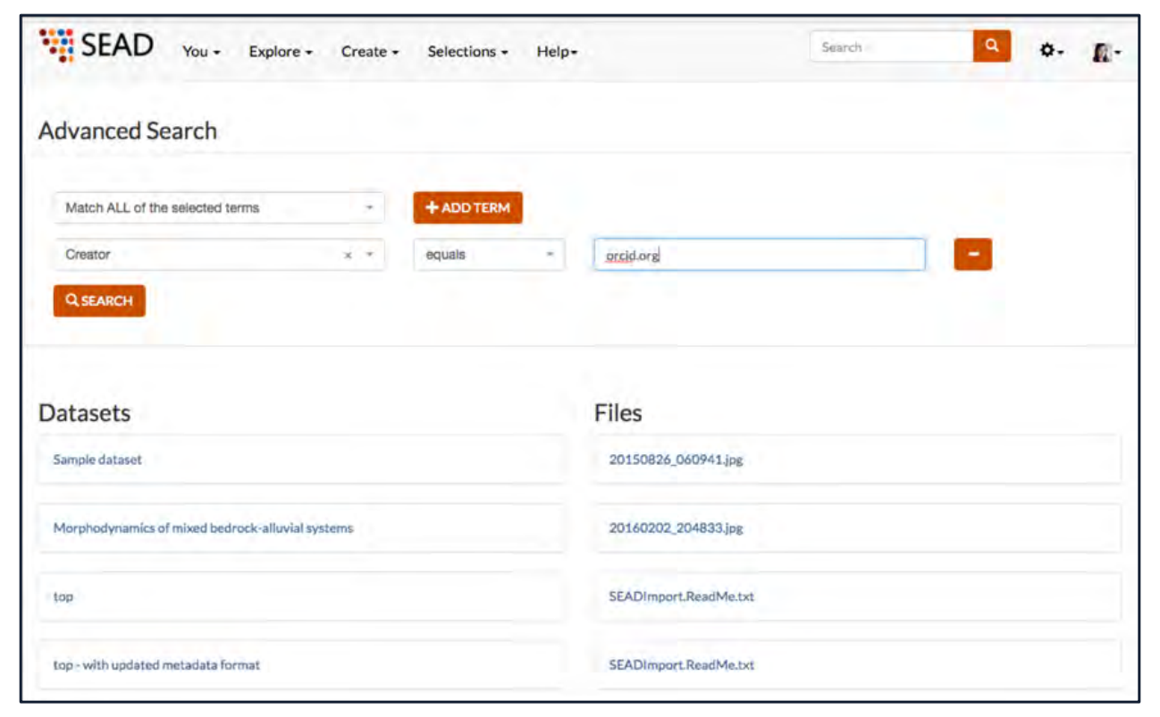
ocr_text: THE MORROW PLOTS OLDEST EXPERIMENTAL FIELD ESTABLISHED IN 1876 DEMONSTRATE THAT USE OF AND TECHNOLOGY CROP PRODUCINITY

Write your extractors using the Clowder Python SDK or in most other programming languages.

Search

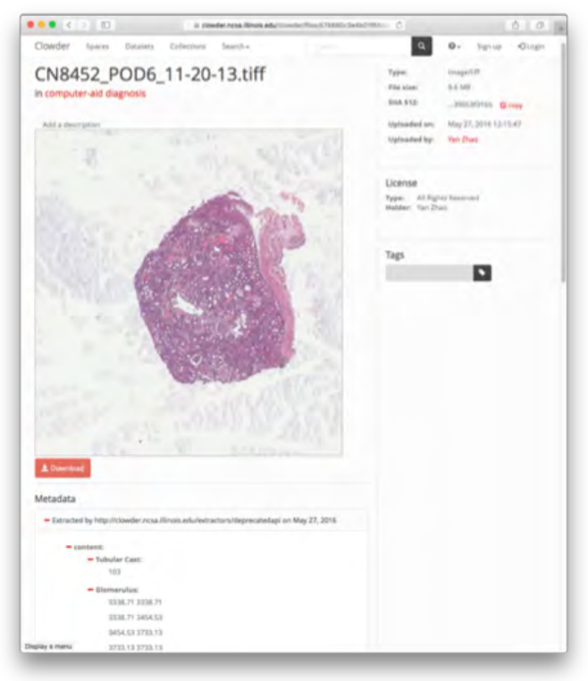
Flexible text search over basic attributes of files, datasets and collections. Advanced search over metadata fields.



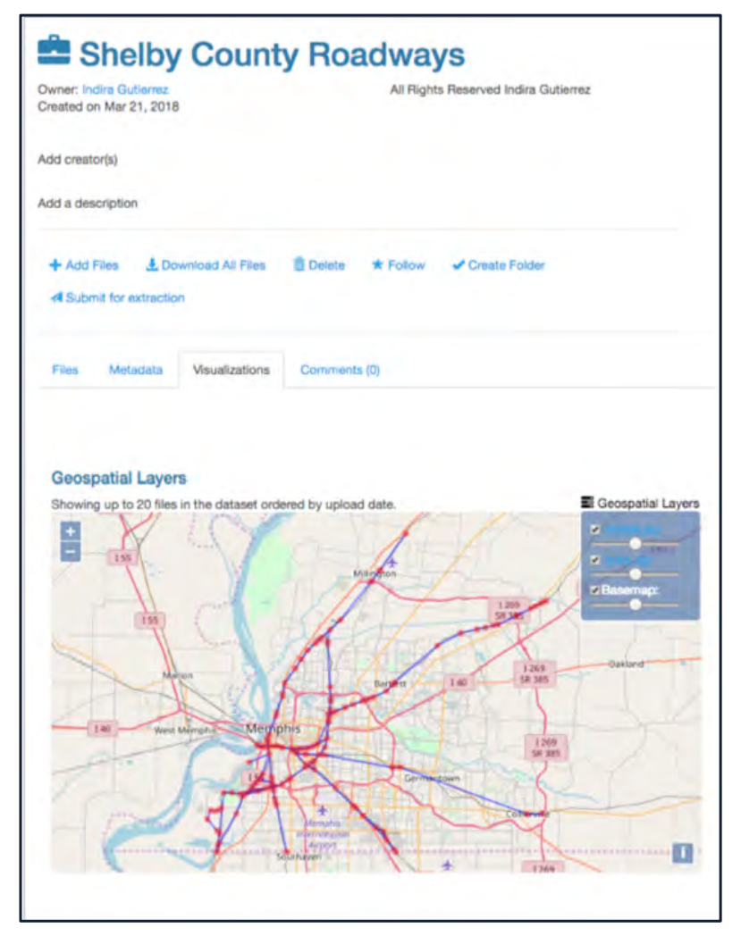


Data Visualizations

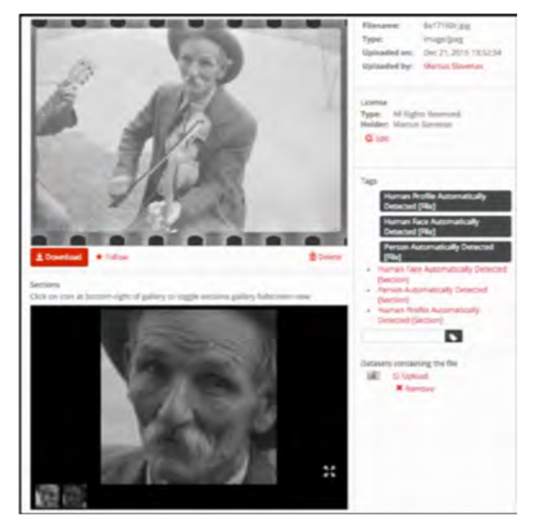
- To preview the content of large files and visualize the information contained in files and datasets in a meaningful way, Clowder provides ways to write Javascript-based visualizations.
- Often these data previews are added by automatic extractions.



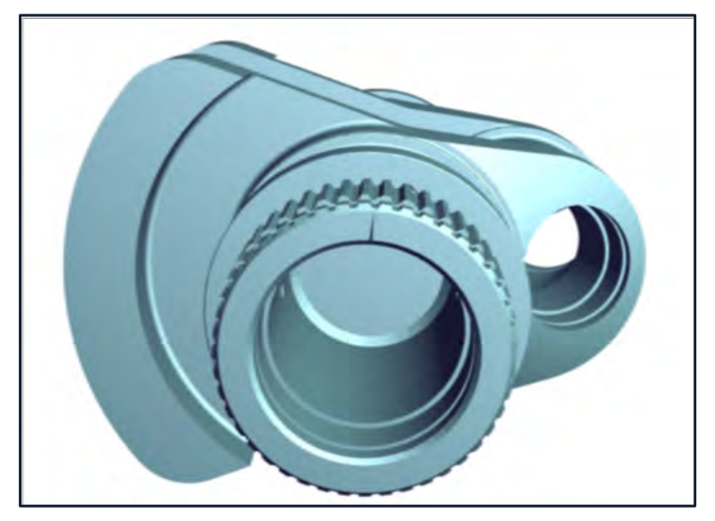
Cell identification



GIS Shapefile previewer



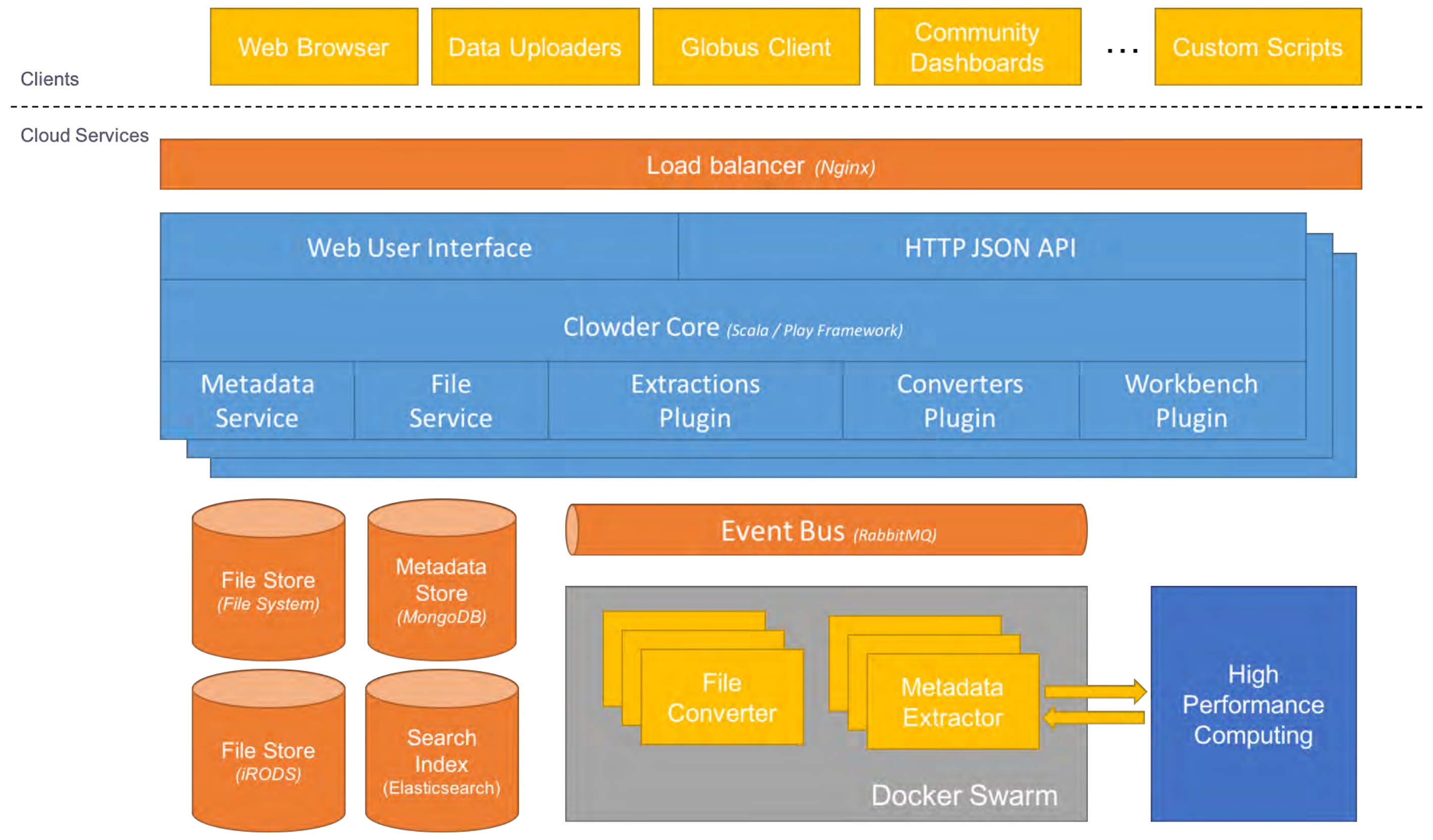
Face identification



Interactive 3D object viewer

Architecture

Clowder is built from the ground up to be extensible and support new use cases and domains.



- The scalability of the Clowder system has been proven with one instance (TERRA-REF) having 1 PB of data and close to 40 million files.
- The Brown Dog project runs 60 different extractors as services in a docker swarm, which consists of 30 machines. Extractors are scaled elastically based on the number of requests and the swarm runs around 150 instances of these extractors.

ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

NCSA

Clowder has been funded by: NSF #BCS-0941268, #EAR-331906, #ACI-1261582, #ACI-1443013, #OCI-0940824, #OCI-0525308, #OAC- 1835834; NIH #1P01AI089556-01A1. Illinois - Indiana SeaGrant #DW92329201, European Commission #RI-261600, XSEDE #OCI-1053575, ARPA-E #DE-AR0000594.