Applications and Roadmap of the Clowder Open-Source Data Management Framework

Chen Wang[a], Luigi Marini[a], Maxwell Burnette[a], Todd Nicholson[a], Mike Lambert[a], Rob Kooper[a], Hoang Nguyen[b], Kenton McHenry[a]

[a] National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign
[b] Research Computing Centre, The University of Queensland

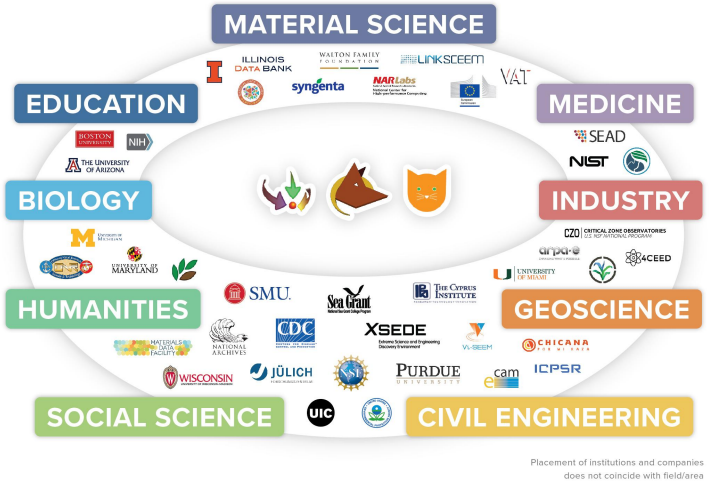
Introduction

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

Objectives

- Provide a framework for scientific data management
- Bootstrap the creation of data pipelines for specific use cases
- Deploy data infrastructure next to computation resources and sensors

Open Source Community

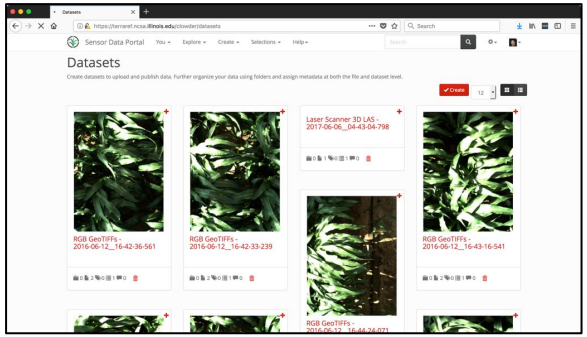


https://clowderframework.org/

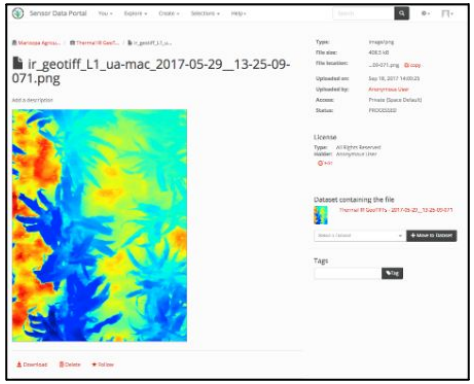
2022 Australian Computer Science Week

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.

Rich Web Interface

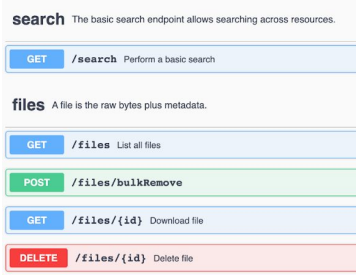


TERRA-REF phenotyping reference platform



Comprehensive Web API

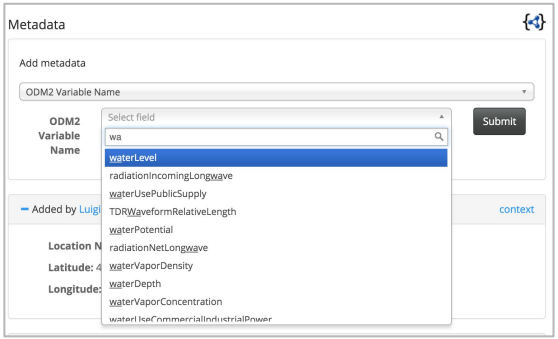
- Simple JSON based RESTful API with user API keys and OpenAPI spec
- Almost all functionality available in the Web UI is available in the Web API
  - ✔ Extractors use the Web API to write back results
  - ✔ Projects can create custom clients specifically tailored to their use cases
  - ✔ Data pipelines can leverage the API for custom workflows



API docs see: https://clowderframework.org/documentation.html

Flexible Metadata

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

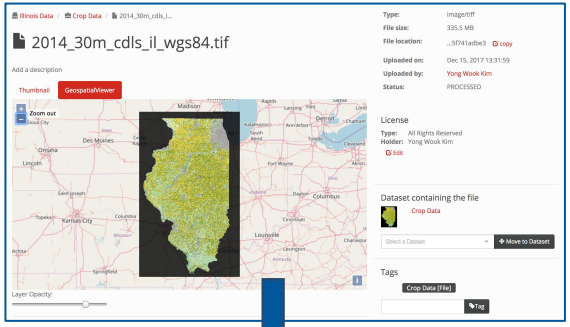


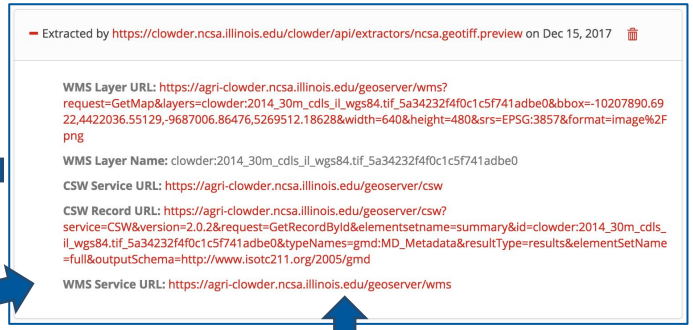

```
{
  "@context": [
    "https://clowder.ncsa.illinois.edu/contexts/metadata.jsonld",
    {
      "CSDMS Standard Name": "http://csdms.colorado.edu/wiki/CSN_Searchable_List"
    }
  ],
  "created_at": "Thu Feb 15 11:12:45 CST 2018",
  "agent": {
    "@type": "cat:user",
    "name": "Luigi Marini",
    "user_id": "http://clowder.ncsa.illinois.edu/clowder/api/users/54b84415621bb34a2f4bed3b"
  },
  "content": {
    "CSDMS Standard Name": "atmosphere_air__increment_of_pressure"
  }
},
```


Users can add metadata entries directly from the user interface.

Extractors and external clients can attach metadata to files and datasets

Automatic Data Extraction & Visualizations

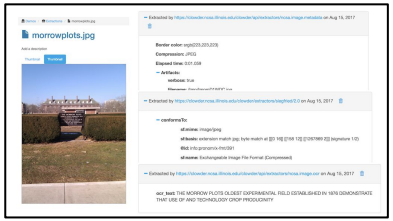


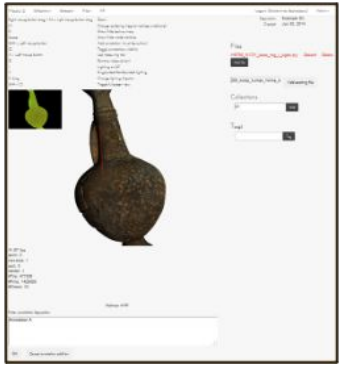


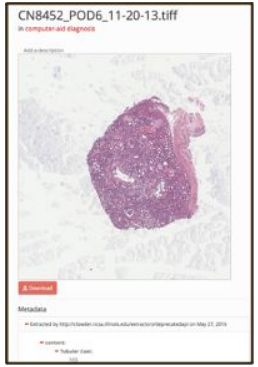
Shapefile / Geotiff Extractor

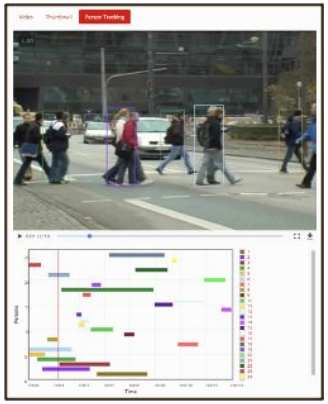
Geoserver

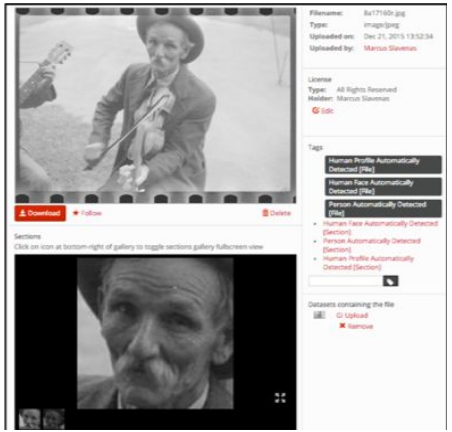
Example Extractors and Visualizations









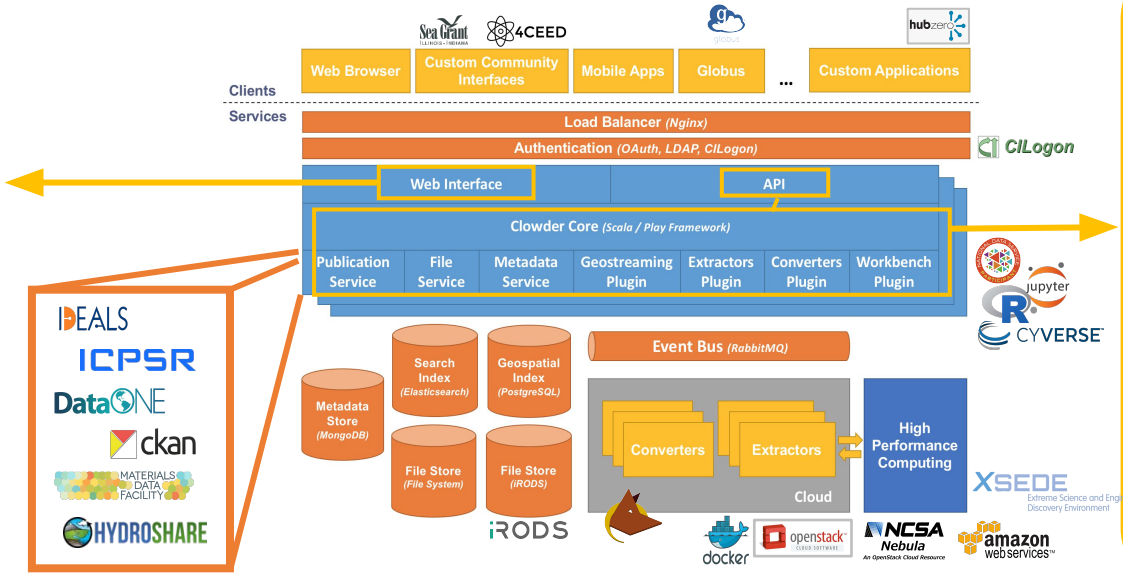


Architecture

- 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 requests and sits around 150 instances of these extractors.

React

- Rebuilding frontend in React.js
- Encourage wider adoption and ability for contributors to learn (vs Play Framework and JQuery)
- Further decoupling of frontend and backend is better for long term sustainability



FastAPI

- FastAPI is a modern, fast web framework for building async APIs with type hints
- In 2022 Scala is not as popular: Libraries are not always well documented or require a strong understanding of the functional aspects of the language. Tutorials, stack overflow, general support not great