EPA

Clowder

A Scalable Data Repository to share, organize and "analyze" data

Scientific Computing and Data Curation Division (SCDCD)

Enabling Scientists and Translation of Science

Outline

- Introduction to Clowder
- Data Model
- Metadata and Extractors
- Architecture
- Product Demonstration
- Q&A

Why Research Data Storage?

- Legislation
  - Evidence-based policymaking Act of 2018: Open Data Access and Management
  - Reproducibility
    - Data means “recorded information, regardless of form or the media on which the data is recorded.” (44 U.S.C. § 3502(16)).
    - Data Asset means “a collection of data elements or data sets that may be grouped together.” (44 112 U.S.C. § 3502(17)). A data asset refers to data systematically arranged such as in a table, relational or non-relational database, or more complex structure that contains content such as, but not limited to, statistical, geospatial, scientific, administrative, or operational information.
    - Machine-readable “when used with respect to data, means data in a format that can be easily processed by a computer without human intervention while ensuring no semantic meaning is lost.” (44 U.S.C. § 3502(18)).
    - Metadata means “structural or descriptive information about data such as content, format, source, rights, accuracy, provenance, frequency, periodicity, granularity, publisher or responsible party, contact information, method of collection, and other descriptions.” (44 U.S.C. § 3502(19)).
    - Open Format means a non-proprietary file format that has a fully-documented, published specification that is maintained by a standards body or a community driven process and has no restrictions placed upon its use.
- Data Provenance
- Document Catalog
- Data Automation

Clowder

Clowder

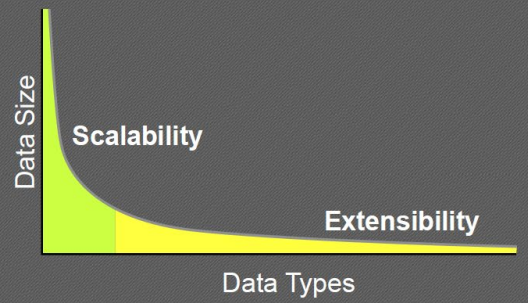
OPEN SOURCE DATA MANAGEMENT FOR RESEARCH

Build customized *cat*alogs in the clouds to help you manage your research data.

https://clowder.ncsa.illinois.edu/

clowder@lists.illinois.edu

Long (Full) Tail of Research Data



- Many domains
- Many data formats
- Domain specific metadata
- Custom processing
- Moving data is expensive

P Bryan Heidorn. 2008. Shedding light on the dark data in the long tail of science. *Library trends* 57, 2 (2008), 280–299.

What is Clowder?

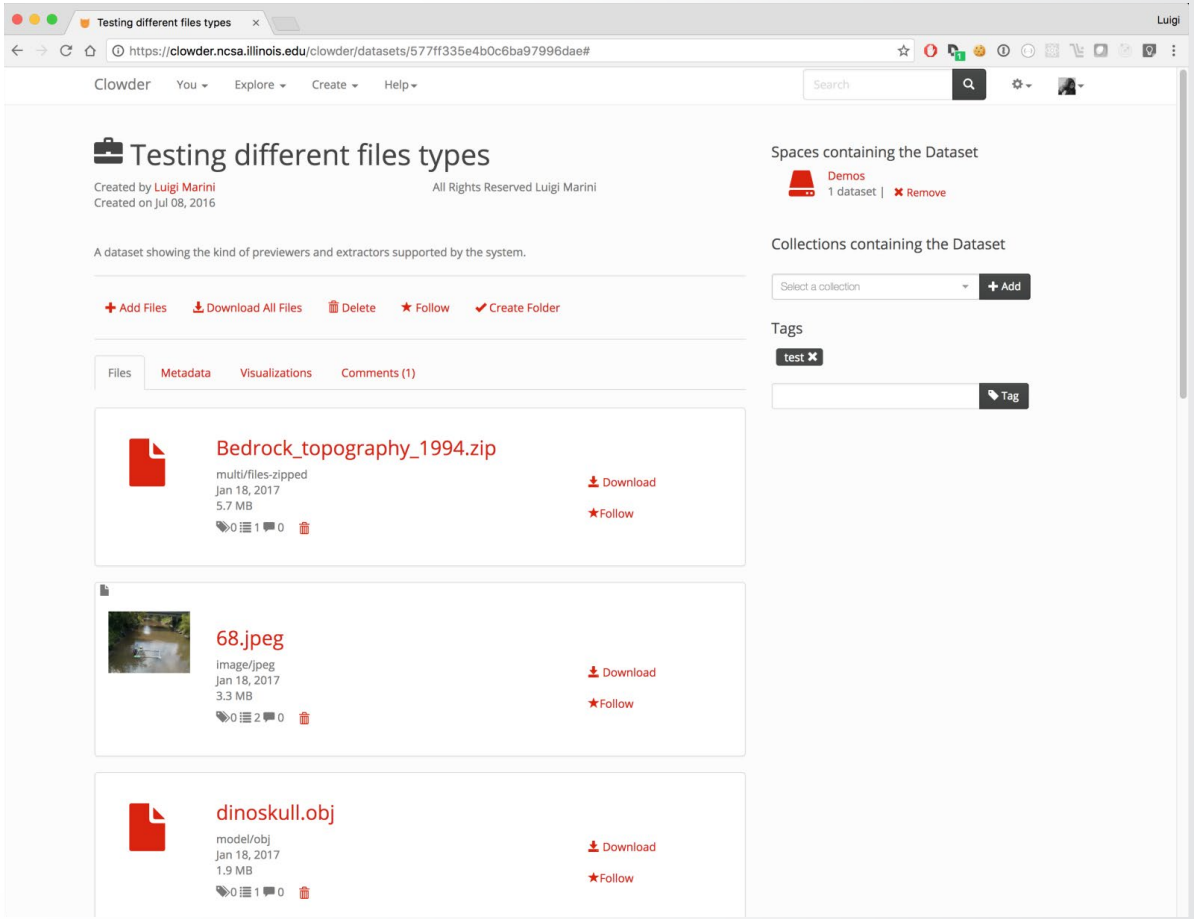
- A data management system designed to support any data format and multiple research domains
- You can install it and manage it on your own hardware resources
- It’s scalable
  - Petabytes of files, hundred of millions of resources
- Its extensible Consists of both Web interface (human readable) and Web Service API (machine readable)
- It’s open source

Funding for Clowder

- NARA/NSF OCI –Understanding Data Intensive and CPU Intensive Services to Support Preservation and Reconstruction of Electronic Records
- NSF CDI –Group Scope: Instrumenting Research on Interaction Networks in Complex Social Contexts
- **NSF EAR –Critical Zone Observatory Network for Intensively Managed Landscapes (IML-CZO)**
- NIH –Immunomodulatory and Regenerative Effects of Mesenchymal Stem Cells on Allografts
- **Illinois-Indiana Sea Grant –Great Lakes Monitoring**
- *European Commission –Linking Scientific Computing in Europe and the Eastern Mediterranean*
- NSF XSEDE –Various scientific gateways (Large Scale Video Analytics)
- **NSF ACI –CIF21 DIBBs: Brown Dog**
- NSF ACI -Sustainable Environment through Actionable Data (SEAD)
- **ARPA-E -TERRA-REF**
- **NSF ACI -CIF21 DIBBs: T2-C2: Timely and Trusted Curator and Coordinator Data Building Blocks**
- **NCSA Industry Program**

Scalibility

- A data management system designed to support any data format and multiple research domains
- You can install it and manage it on your own hardware resources
- It's scalable
  - Petabytes of files, hundred of millions of resources
- Its extensible
- Consists of both Web interface (human readable) and Web Service API (machine readable)
- It's open source



Security : Space and Control

IMLCZO You Explore Create Selections Help Search

Flux Tower Data

Flux Tower located near Monticello-includes atmospheric and soil moisture data. This data is currently NOT being shared with "Atmospheric" and "Soil" spaces.

Delete Create Dataset Create Collections

All Data Public Data

Datasets in the Space

Viewing most recent datasets

View All Datasets

Flux Tower Raw Data Aggregates

This Dataset contains aggregate files of Flux Tower raw data files between April 22, 2016 and March 8, 2018.

Files are provided in 3 month intervals due to file sizes.

...

0 9 3 0

Flux Tower raw data between April and Sept 2016

Flat zip files for easy downloading.

0 2 0 0

Flux Tower Raw Files

The IMLCZO Bot will upload

Manage Users

Edit Space

Manage Metadata Terms & Definitions

Extractors

Follow

External Links

Edit Space to add links

Access

PRIVATE

You are authorized to access this Space.

Flux Tower Data / Manage Users

Manage Users of Space Flux Tower Data

Users Invites Requests (0)

Owner: Steve Sargent

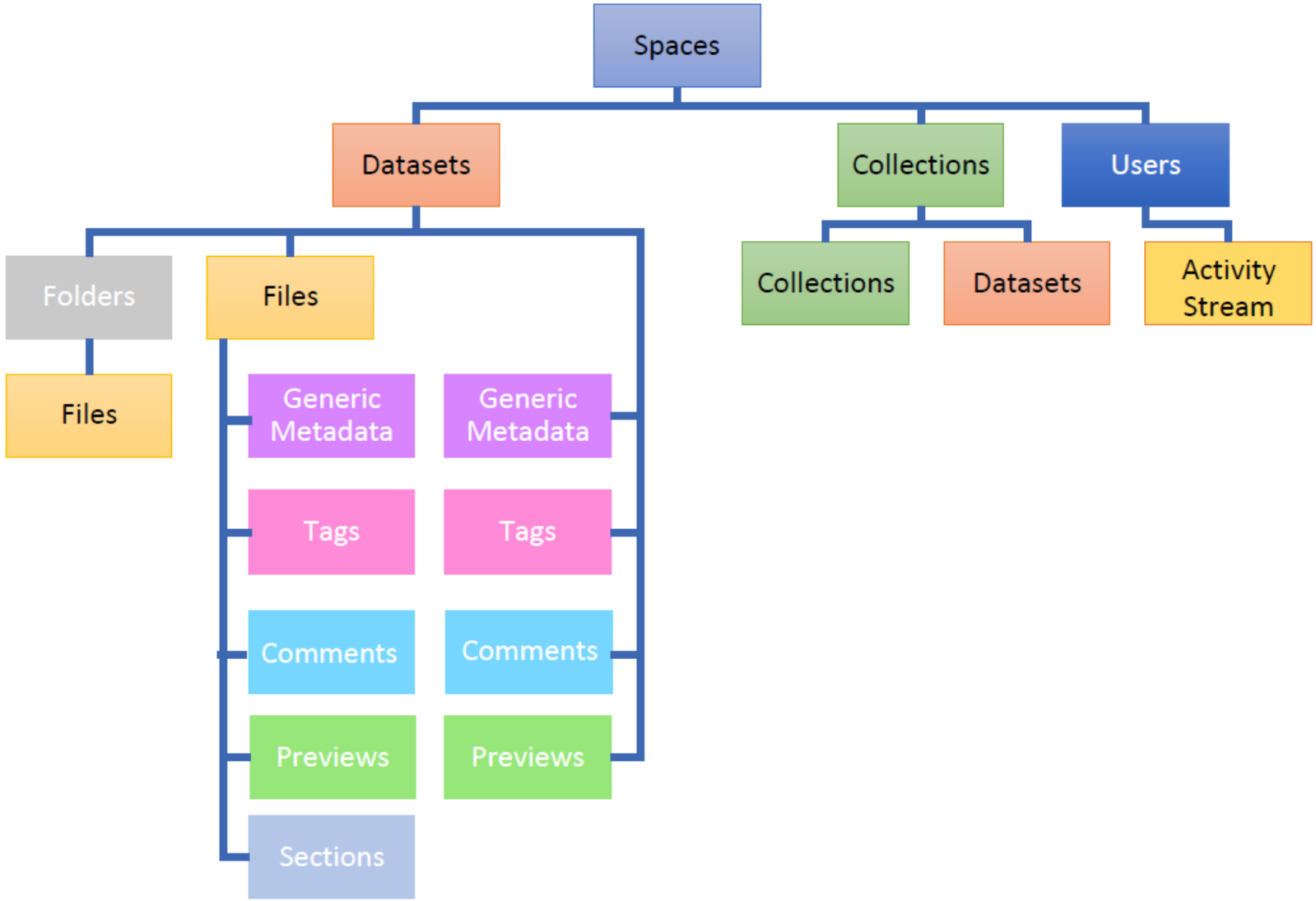
Admin

Admin Role

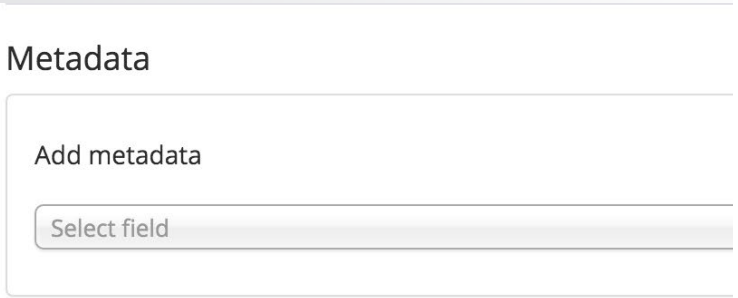
- Brock Angelo (jba@illinois.edu) [Local Account]
- Allison Goodwell (allison@goodwell.net) [Local Account]
- Steve Sargent (slsargen@illinois.edu) [Local Account]
- Michelle Pitcel (mpitcel2@illinois.edu) [Local Account]
- Luigi Marini (lmarini@illinois.edu) [Local Account]
- Praveen Kumar (kumar1@illinois.edu) [Local Account]

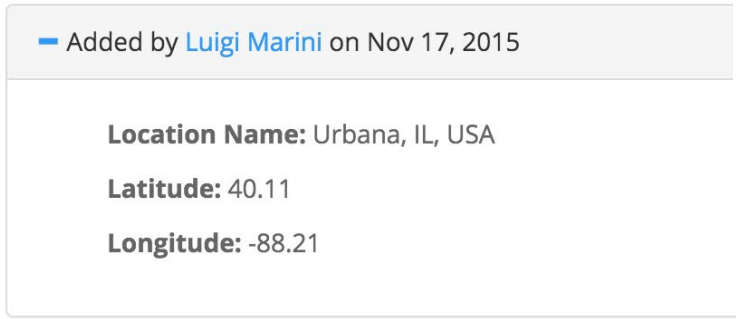
Select Users for this Level...

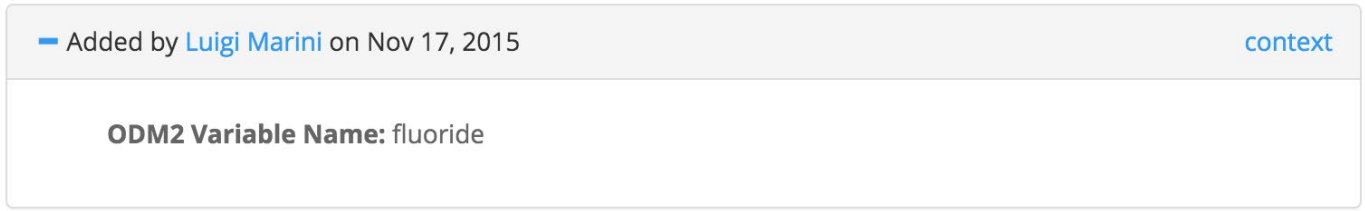
Data Model

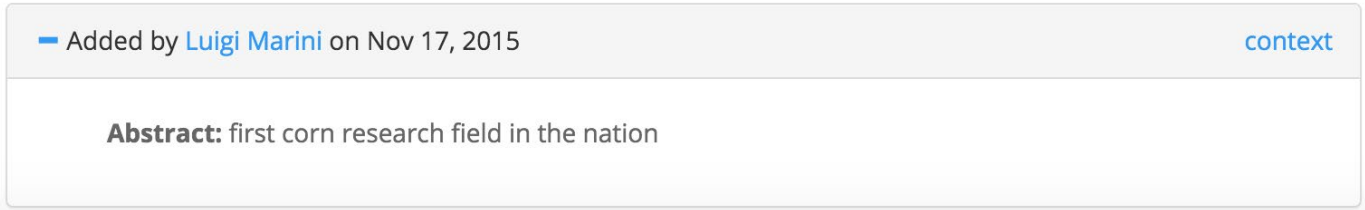

Spaces
Datasets
Collections
Users
Folders
Files
Collections
Datasets
Activity Stream
Files
Generic Metadata
Generic Metadata
Tags
Tags
Comments
Comments
Previews
Previews
Sections

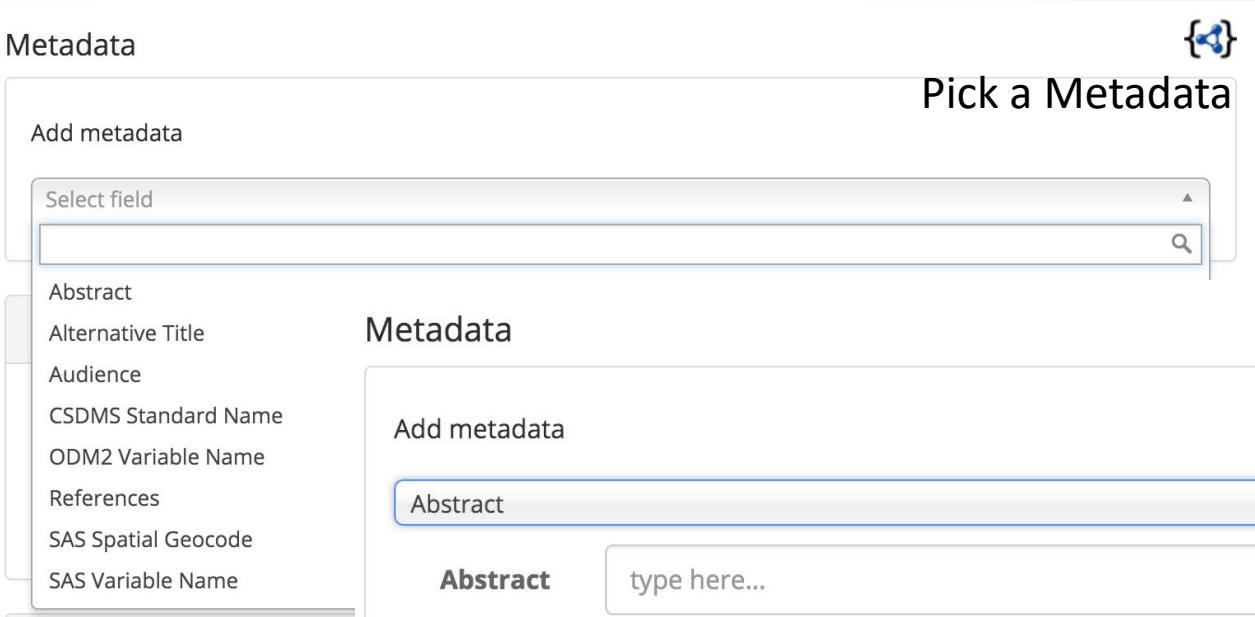
Metadata





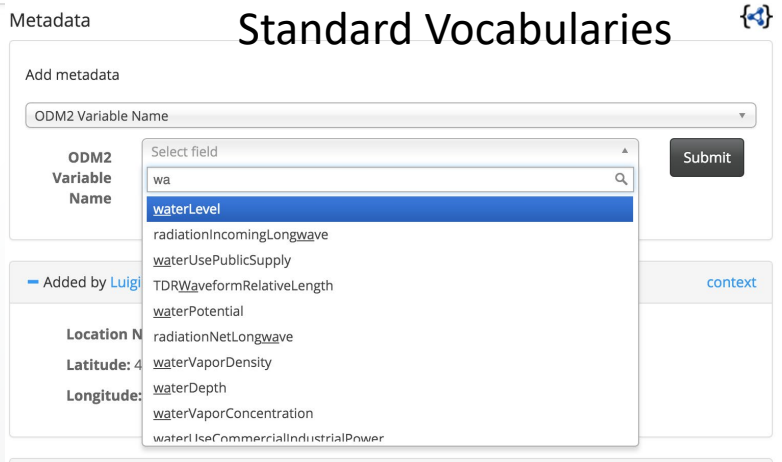






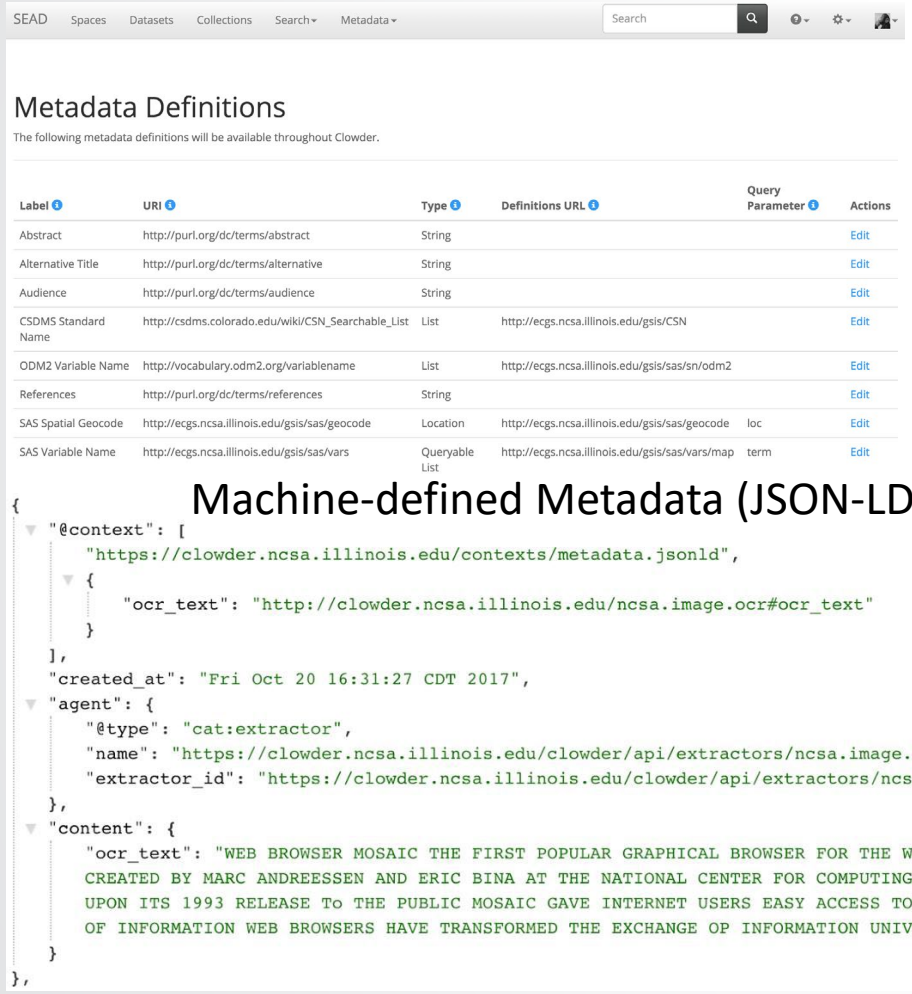
Pick a Metadata

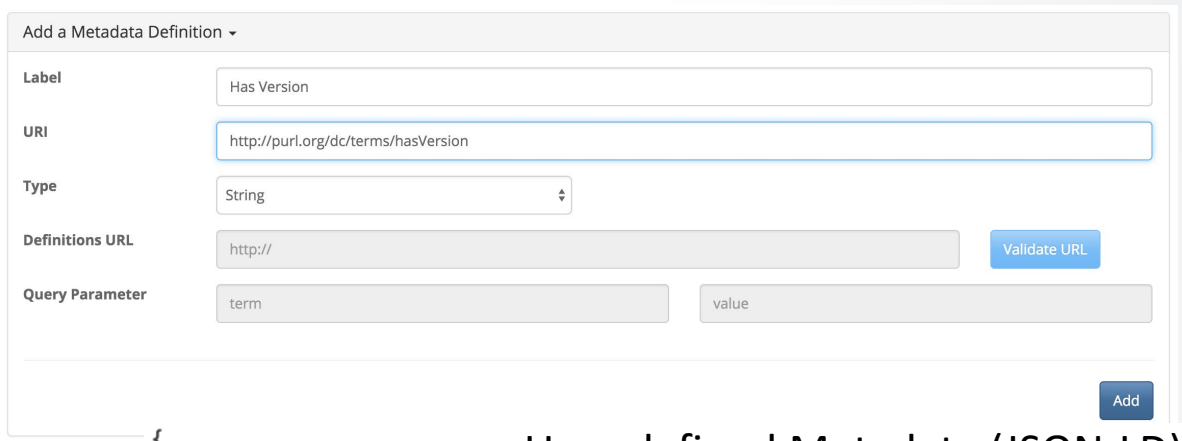
Free-Text Metadata



Standard Vocabularies

Metadata Definitions





Machine-defined Metadata (JSON-LD)


```
{
  "@context": [
    "https://clowder.ncsa.illinois.edu/contexts/metadata.jsonld",
    {
      "ocr_text": "http://clowder.ncsa.illinois.edu/ncsa.image.ocr#ocr_text"
    }
  ],
  "created_at": "Fri Oct 20 16:31:27 CDT 2017",
  "agent": {
    "@type": "cat:extractor",
    "name": "https://clowder.ncsa.illinois.edu/clowder/api/extractors/ncsa.image.ocr",
    "extractor_id": "https://clowder.ncsa.illinois.edu/clowder/api/extractors/ncsa.image.ocr"
  },
  "content": {
    "ocr_text": "WEB BROWSER MOSAIC THE FIRST POPULAR GRAPHICAL BROWSER FOR THE WORLD WIDE WEB
    CREATED BY MARC ANDREESSEN AND ERIC BINA AT THE NATIONAL CENTER FOR COMPUTING APPLICATIONS
    UPON ITS 1993 RELEASE To THE PUBLIC MOSAIC GAVE INTERNET USERS EASY ACCESS TO MULTIMEDIA S
    OF INFORMATION WEB BROWSERS HAVE TRANSFORMED THE EXCHANGE OP INFORMATION UNIVERSITY OF ILL
  }
},
```


User-defined Metadata (JSON-LD)


```
{
  "@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"
  }
}
```

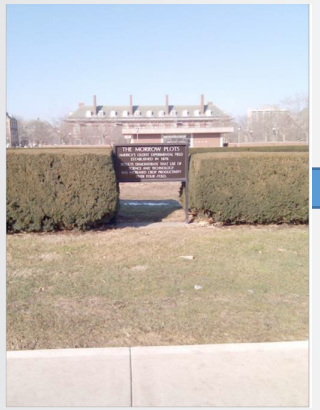

Exporting Metadata

OAI ORE endpoint: direct communication between long term preservation systems and Clowder

BagIT : downloading dataset as zip archive of files and metadata

ata Extraction using Simple Extractors

Object Character Recognition



Extractors

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

EXIF

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:

height: 2592

Interlace: None

Compose: Over

Tainted: false

Filesize: 1.268MB

Pixels per second: 83.98MB

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

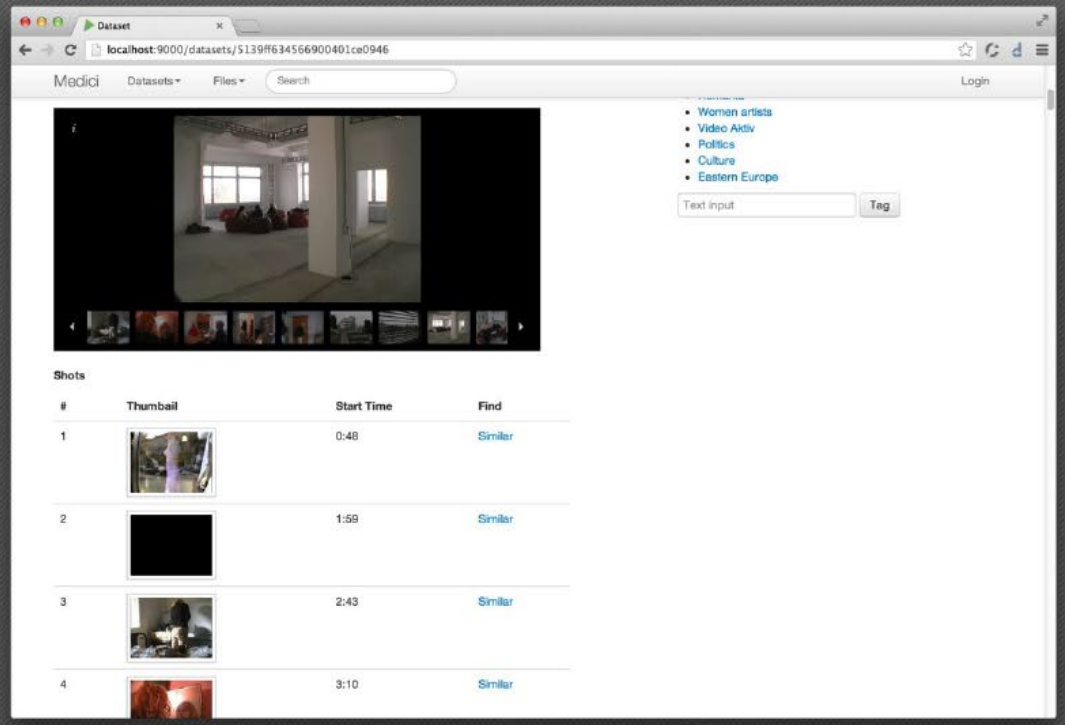
sf:name: Exchangeable Image File Format (Compressed)

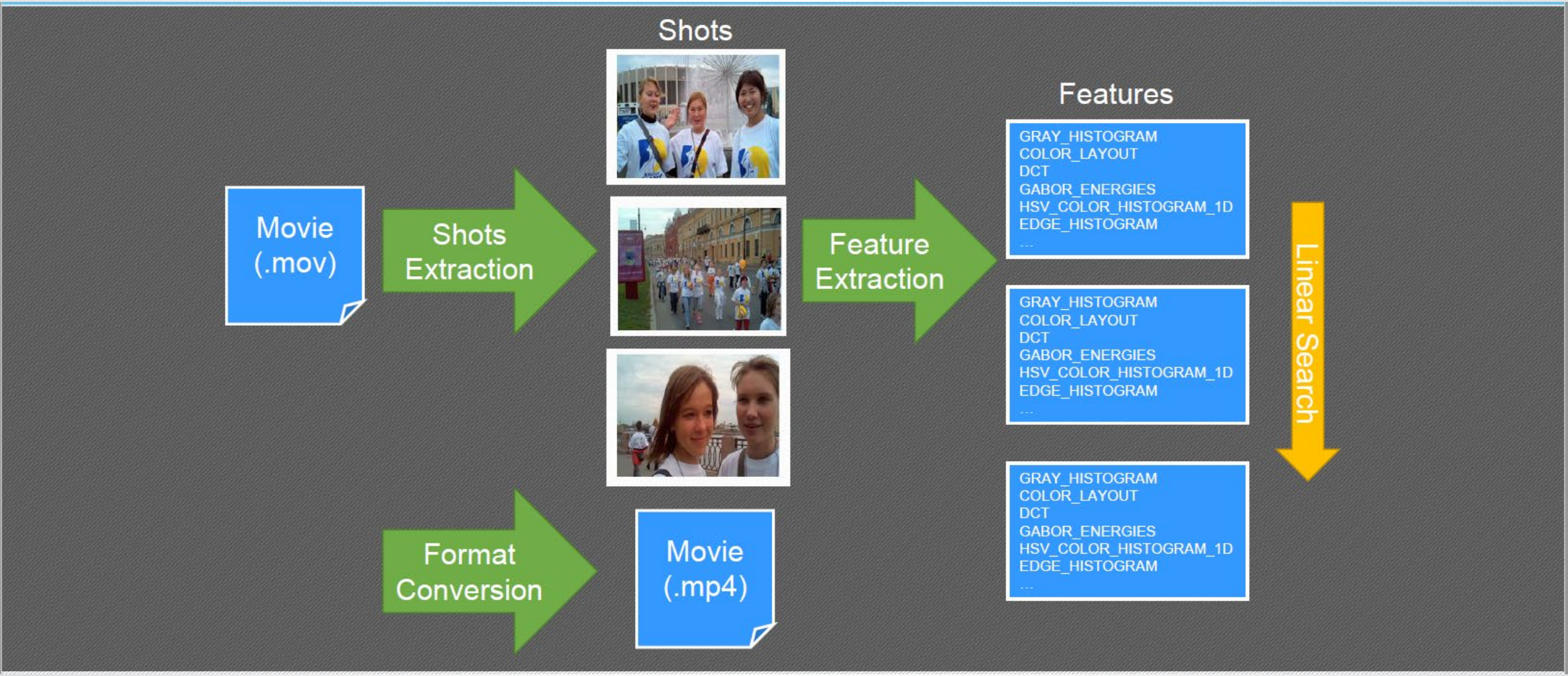
Extracted by https://clowder.ncsa.illinois.edu/clowder/api/extractors/ncsa.image.ocr on Aug 15, 2017

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

Video Analysis Tableau (1)

- Input: Movie file
- Output: Sections representing “shots,” feature vectors for Image based retrieval



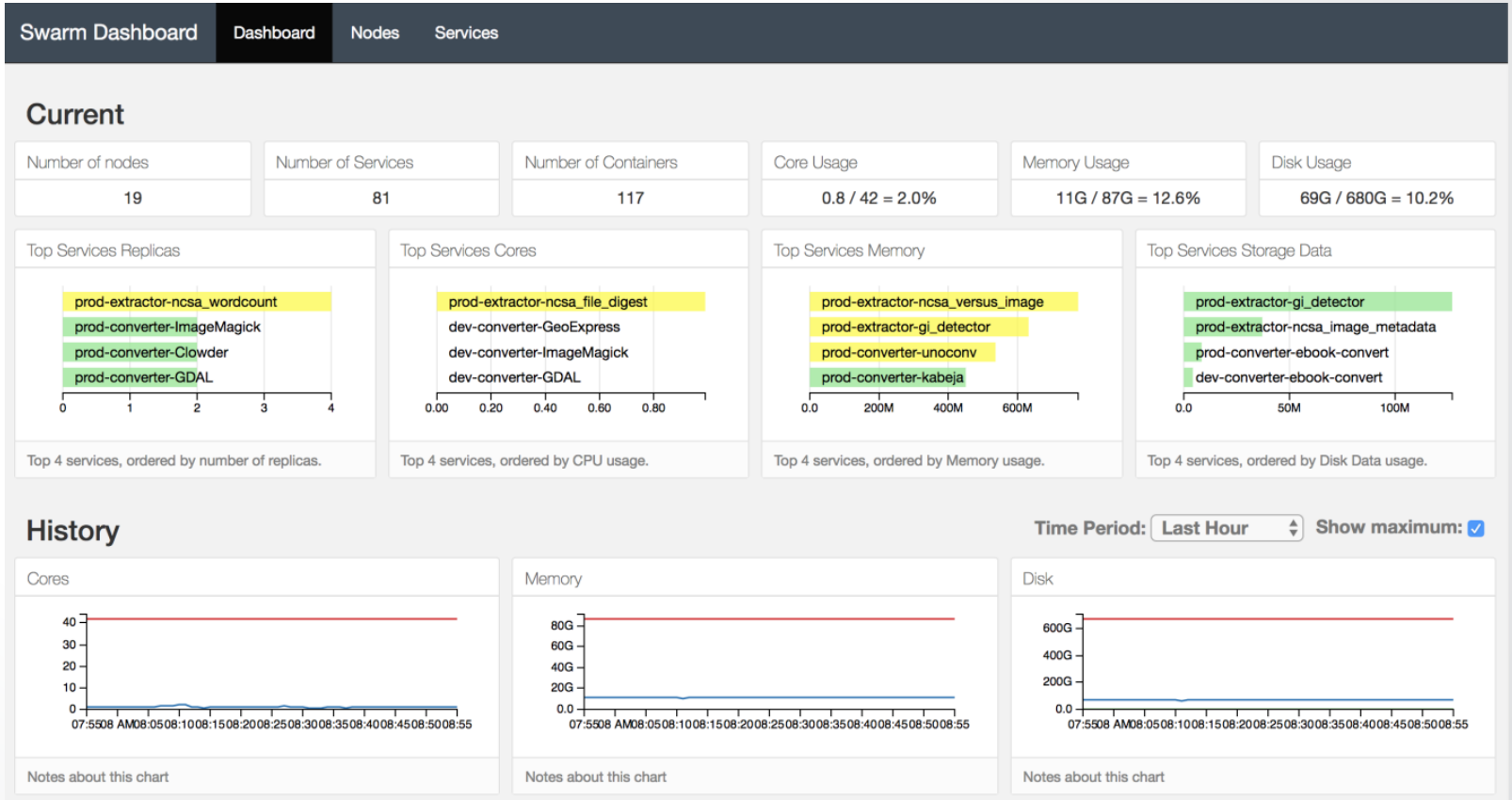

Shots
Features
Movie (.mov)
Shots Extraction
Feature Extraction
GRAY_HISTOGRAM
COLOR_LAYOUT
DCT
GABOR_ENERGIES
HSV_COLOR_HISTOGRAM_1D
EDGE_HISTOGRAM
...
GRAY_HISTOGRAM
COLOR_LAYOUT
DCT
GABOR_ENERGIES
HSV_COLOR_HISTOGRAM_1D
EDGE_HISTOGRAM
...
GRAY_HISTOGRAM
COLOR_LAYOUT
DCT
GABOR_ENERGIES
HSV_COLOR_HISTOGRAM_1D
EDGE_HISTOGRAM
...
Linear Search
Format Conversion
Movie (.mp4)

Developing Extractors

- PyClowder
  - A Python library to simplify the process
  - Modules to call API
  - HPC Extractor
- JClowder
  - A Java library to simplify the process
  - Experimental, few functions
- From Scratch
  - RabbitMQ client library
  - HTTP/JSON client libraries

Elastic Scaling

- All extractors are dockerized and available on Docker Hub
- Scale number of instances of each extractors based on RabbitMQ queues
- Uses Docker Swarm

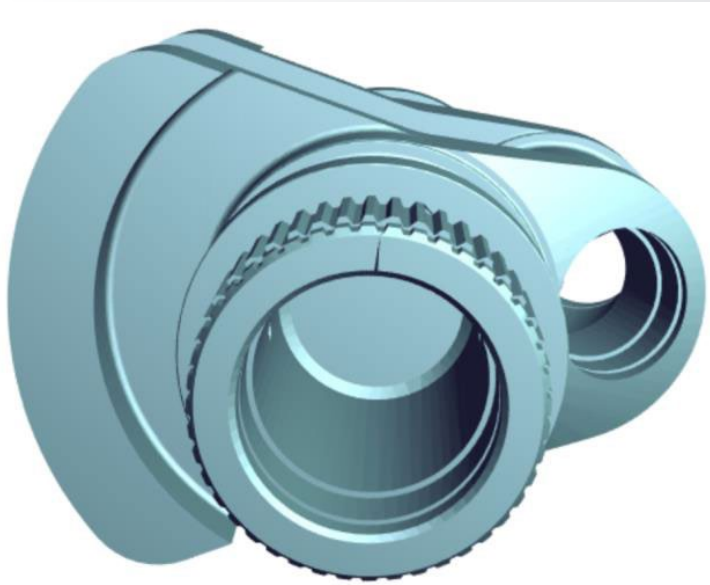


Simple Previewers

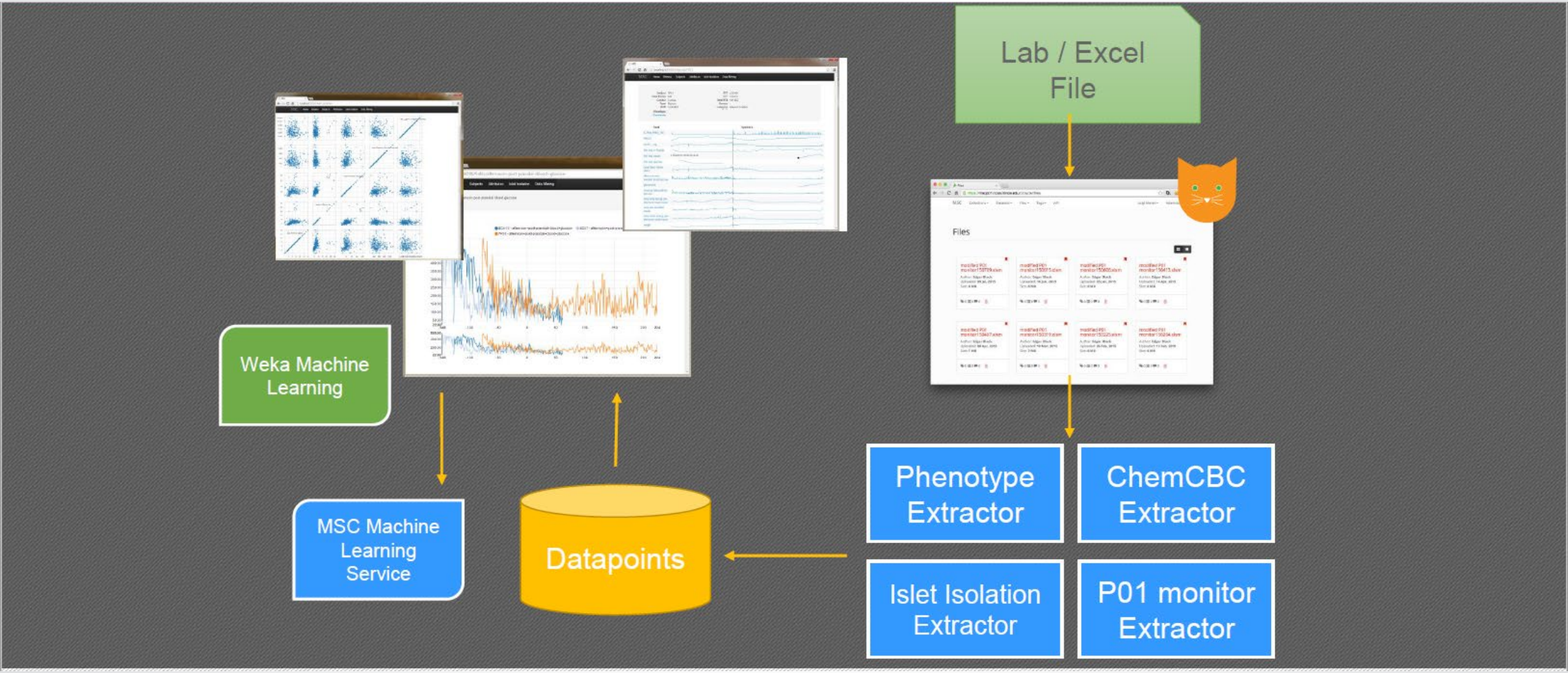
Video player

4.082000017166132011e+01
3.883999991416924047e+01
3.757999982833854347e+01
3.613999991416922342e+01
3.506000008583059468e+01
3.433999991416921205e+01
3.379999999999989768e+01
3.344000002145756412e+01
3.289999999999989200e+01
3.254000002145755843e+01
3.254000002145755843e+01
3.289999999999989200e+01
3.308000004291523766e+01
3.325999997854222556e+01
3.344000002145756412e+01
3.379999999999989768e+01
3.398000004291524334e+01
3.398000004291524334e+01
3.416000008583058900e+01

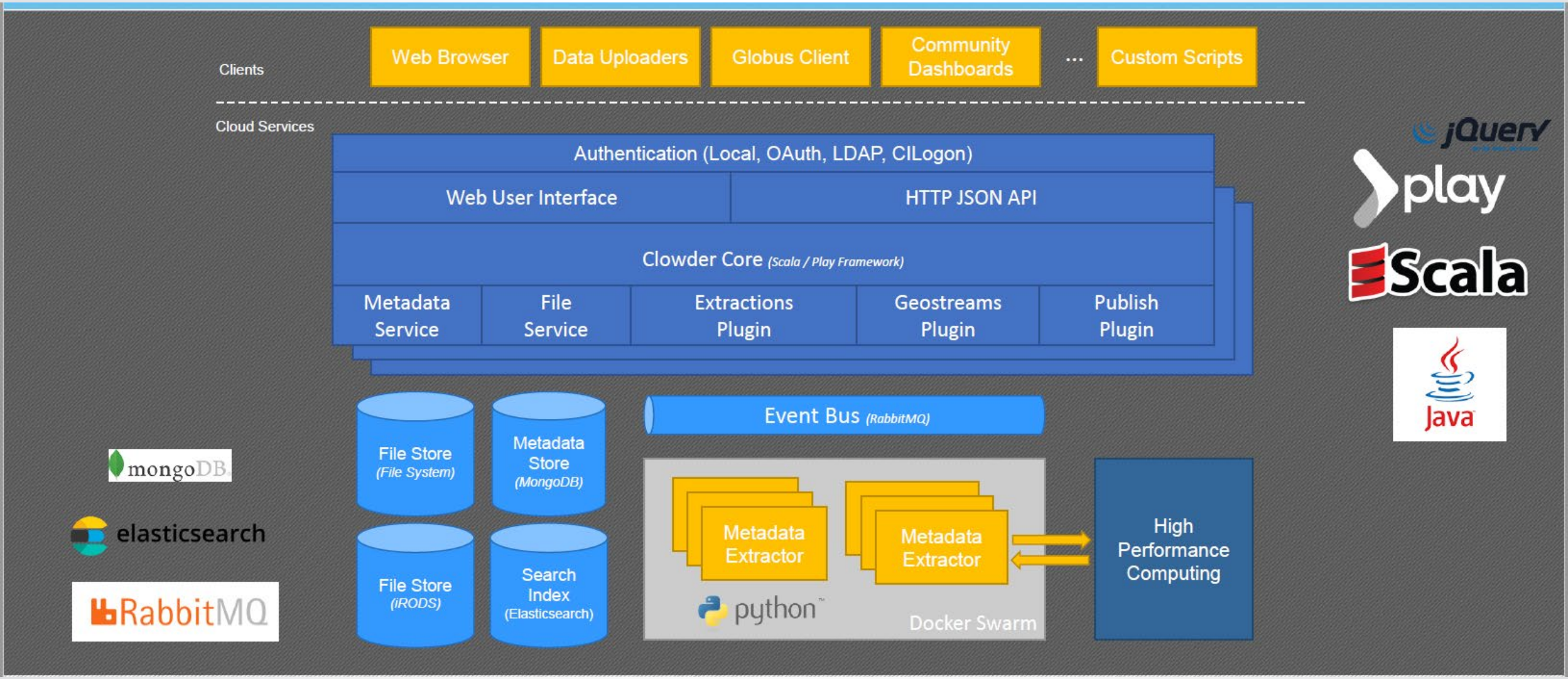
3D Objects (.obj)



Use Case – NIH Immunomodulatory and Regenerative Effects of Mesenchymal Stem Cells on Allografts

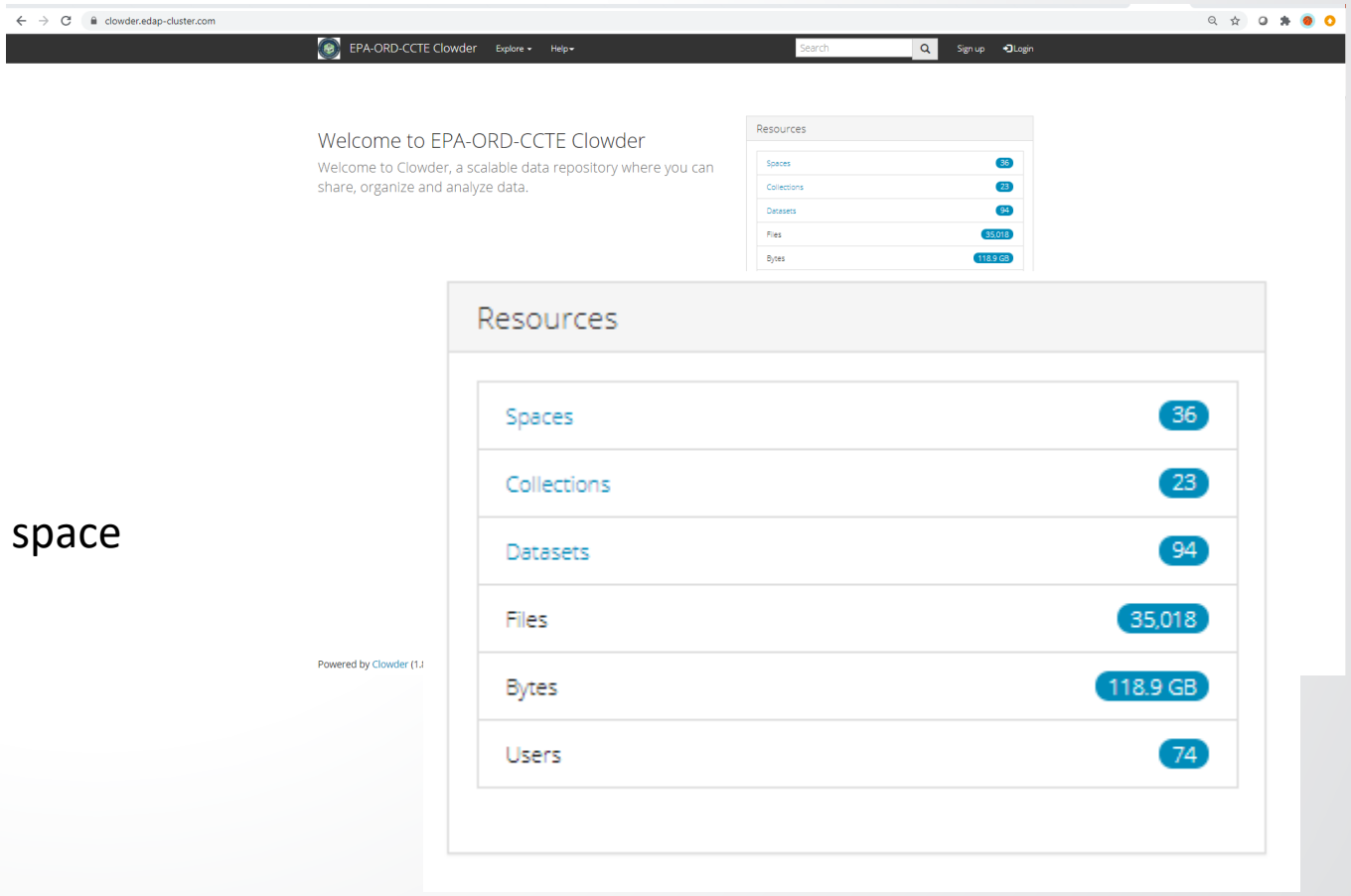


Architecture



CCTE Clowder

- https://clowder.edap-cluster.com/
- https://clowder.edap-cluster.com/signup
- EPA AWS
  - Kubernetes



<u>Simple</u> Process

- Send email to ccte_scdcd with desired space name
- Ask the space members to signup (if they have not)
- Once you get the space you can add them into your space

Future (To-be and in-progress)

- Get the ATO for Clowder Prototype and move to Production
- Upgrade to newer version (2.0) to include S3 as storage (in progress)
- Propose to use to store all research data with metatags for data provenance
- Expand the use case from Just file storage to Video analysis and development of extractors
- Adding more extractors from NCSA and other sources
- Process for approving data for public

DEMO

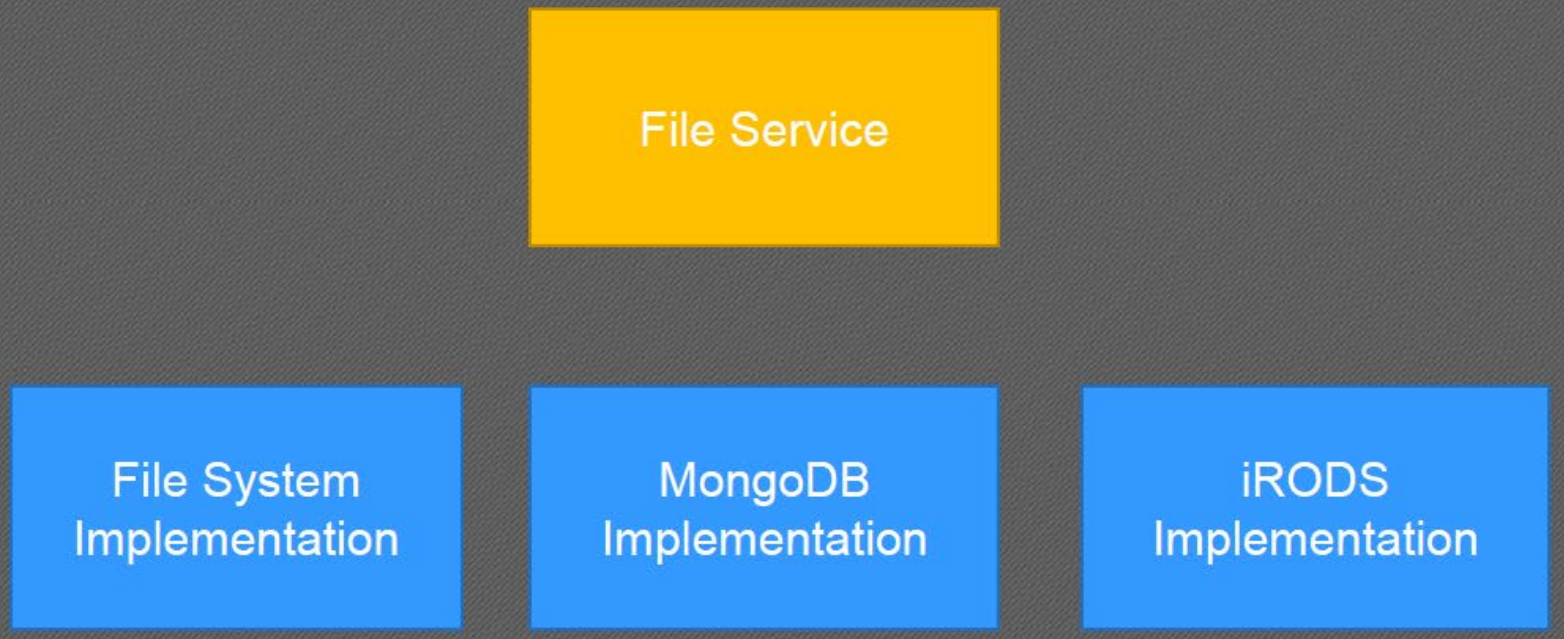
Security – Norman Adkins

DERs & ToxRefDB visualization– Jason Brown

Use case DCT – Amar Singh

Services

- Ability to swap underlying implementation



Plugin and Configs

- Example config flag: datasets belong to multiple spaces
- Entire feature sets for very specific use cases:

Geostreams API

POST	/api/geostreams/datapoints
POST	/api/geostreams/datapoints/bulk
DELETE	/api/geostreams/datapoints/:id
GET	/api/geostreams/datapoints
GET	/api/geostreams/datapoints/averages
GET	/api/geostreams/datapoints/trends
GET	/api/geostreams/datapoints/bin/:time/:depth
GET	/api/geostreams/datapoints/:id
GET	/api/geostreams/cache
GET	/api/geostreams/cache/invalidate
GET	/api/geostreams/cache/:id
POST	/api/geostreams/sensors
GET	/api/geostreams/sensors/update
GET	/api/geostreams/sensors/:id
PUT	/api/geostreams/sensors/:id
GET	/api/geostreams/sensors/:id/stats
GET	/api/geostreams/sensors/:id/streams
GET	/api/geostreams/sensors/:id/update
GET	/api/geostreams/sensors

SEAD Publish