CLOWDER WEBINAR

SEPT *FRIDAY 13th*, 2019

AGENDA

- Schedule (10 mins)
- Search Updates (20 mins)
- Metadata Extractors Updates (20 mins)

GOALS

- A way to keep the community updated on new developments and plans
- A way for existing community members to learn more about what Clowder does and how
- A way for interested potential adopters to learn about Clowder

MONTHLY WEBINAR

- **When**: First Friday of the Month 11am CST
- **Who**: Users and Developers
- **What**: Discuss new and old features
  - 2 presentations by community members
  - Please suggest topics and volunteer to present
- **Where**: https://illinois.zoom.us/j/856788350

WEEKLY DEVELOPER MEETINGS

- **When**: 2nd, 3rd, 4th Friday of the Month 11am CST
- **Who**: Software Developers
- **What**: Technical discussions
- **Where**: https://illinois.zoom.us/j/856788350

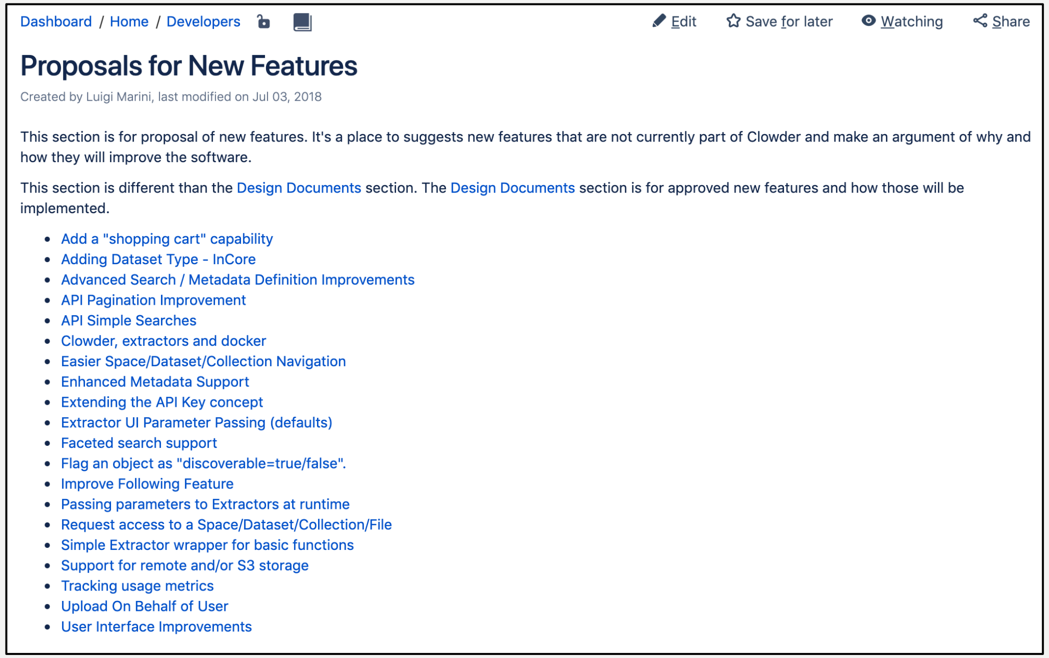
HOW TO CONTRIBUTE

- Contribute to CORE
- Develop new EXTRACTORS
- Develop new PREVIEWERS
- Develop new CLIENTS
- Contribute to DOCUMENTATION
- Contribute DESIGNS to improve INTERFACE
- Contribute IDEAS for NEW FEATURES

Not just software developers!

NEW FEATURES PROPOSALS

- Ask in Slack if anyone is working on it or worked on something similar
- Get feedback on your ideas and provide them in the Wiki



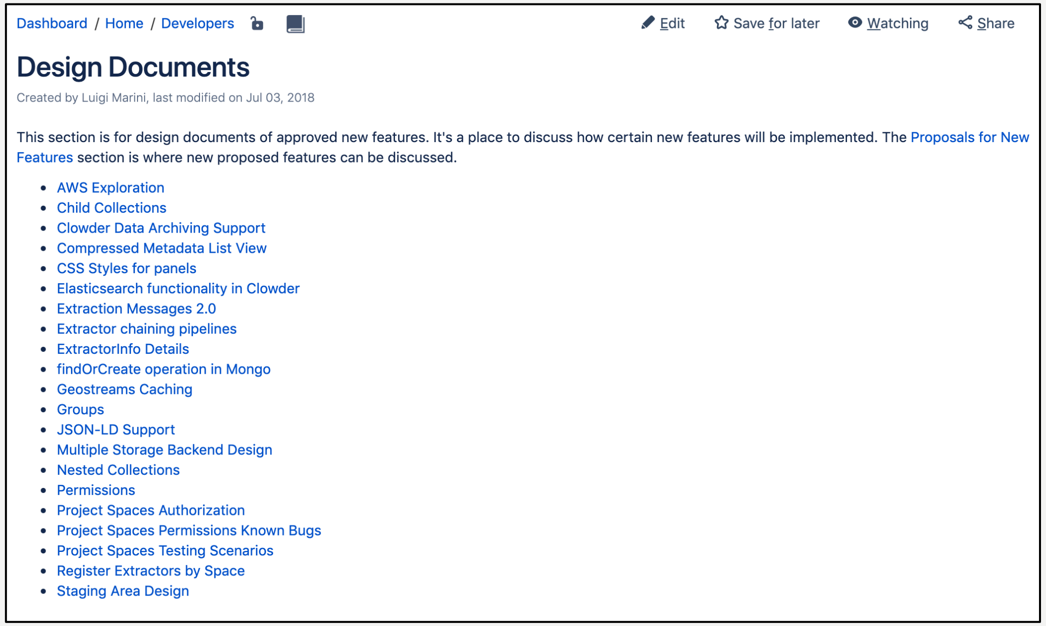
Proposals for New Features

Created by Luigi Marini, last modified on Jul 03, 2018

This section is for proposal of new features. It's a place to suggests new features that are not currently part of Clowder and make an argument of why and how they will improve the software.

This section is different than the Design Documents section. The Design Documents section is for approved new features and how those will be implemented.

- Add a "shopping cart" capability
- Adding Dataset Type - InCore
- Advanced Search / Metadata Definition Improvements
- API Pagination Improvement
- API Simple Searches
- Clowder, extractors and docker
- Easier Space/Dataset/Collection Navigation
- Enhanced Metadata Support
- Extending the API Key concept
- Extractor UI Parameter Passing (defaults)
- Faceted search support
- Flag an object as "discoverable=true/false".
- Improve Following Feature
- Passing parameters to Extractors at runtime
- Request access to a Space/Dataset/Collection/File
- Simple Extractor wrapper for basic functions
- Support for remote and/or S3 storage
- Tracking usage metrics
- Upload On Behalf of User
- User Interface Improvements



Design Documents

Created by Luigi Marini, last modified on Jul 03, 2018

This section is for design documents of approved new features. It's a place to discuss how certain new features will be implemented. The Proposals for New Features section is where new proposed features can be discussed.

- AWS Exploration
- Child Collections
- Clowder Data Archiving Support
- Compressed Metadata List View
- CSS Styles for panels
- Elasticsearch functionality in Clowder
- Extraction Messages 2.0
- Extractor chaining pipelines
- ExtractorInfo Details
- findOrCreate operation in Mongo
- Geostreams Caching
- Groups
- JSON-LD Support
- Multiple Storage Backend Design
- Nested Collections
- Permissions
- Project Spaces Authorization
- Project Spaces Permissions Known Bugs
- Project Spaces Testing Scenarios
- Register Extractors by Space
- Staging Area Design

SLACK & MAILING LIST

- All questions should be redirected to Slack
  - Link to join on https://clowderframework.org
- Or mailing list **clowder@lists.ncsa.illinois.edu**
  - Subscribe here https://lists.illinois.edu/lists/subscribe/clowder

CLOWDER SEARCH FEATURES

MAX BURNETTE
CLOWDER ALL-PAWS
SEPTEMBER 12, 2019

OVERVIEW

- **What is searchable?**
  - Supported syntax
- **Searching via GUI**
  - Search Box
  - Metadata Search
- **Searching via API**
  - What is the API?
  - Pagination
- **Technical Snippets**
  - Permissions
  - Elasticsearch & MongoDB

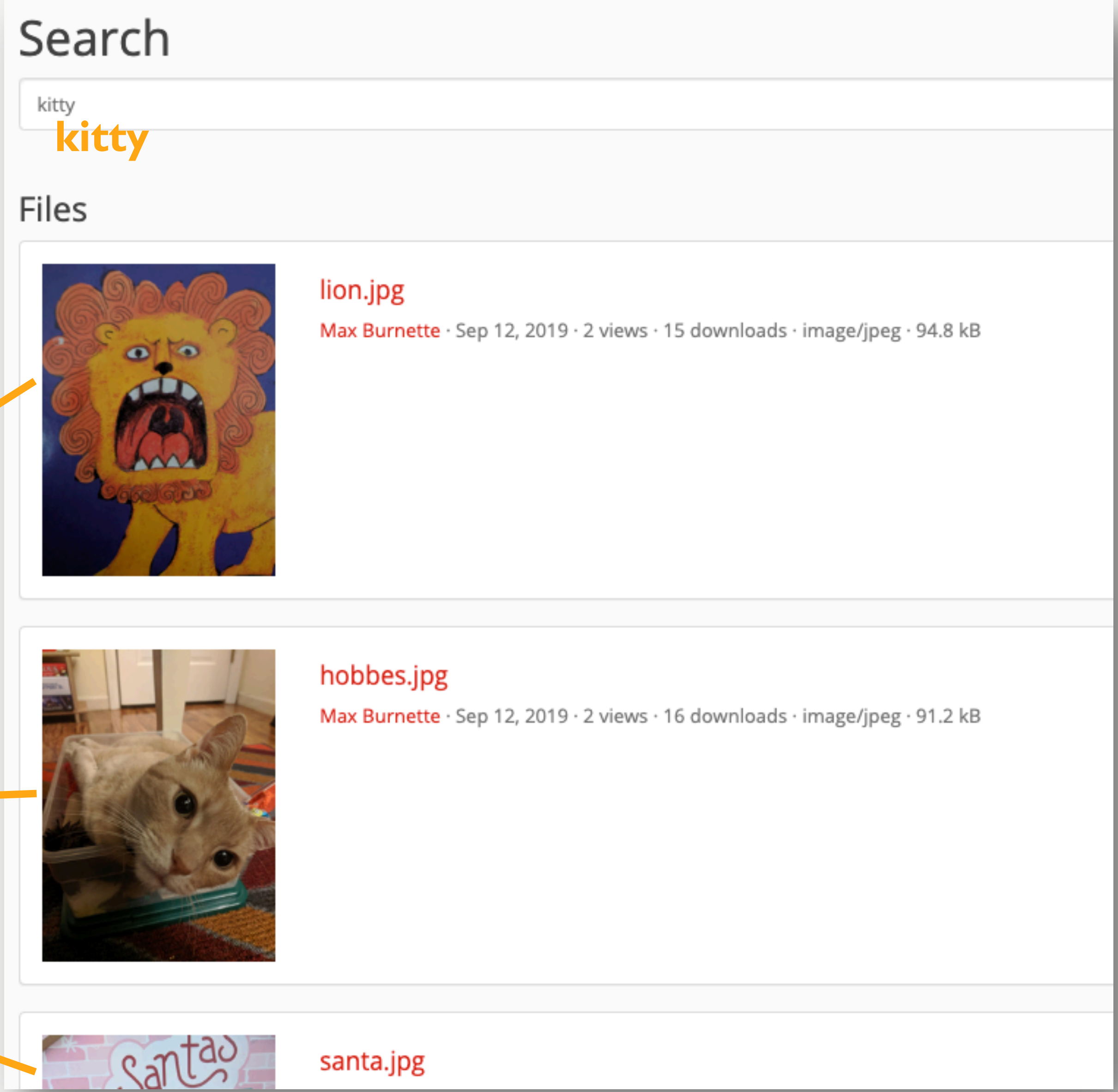
WHAT IS SEARCHABLE?

Resource Types:

- Files
- Datasets
- Collections

Fields by default:

- Name
- Description
- Creator name
- Tags



Search

kitty

kitty

Files

lion.jpg

Max Burnette · Sep 12, 2019 · 2 views · 15 downloads · image/jpeg · 94.8 kB

hobbes.jpg

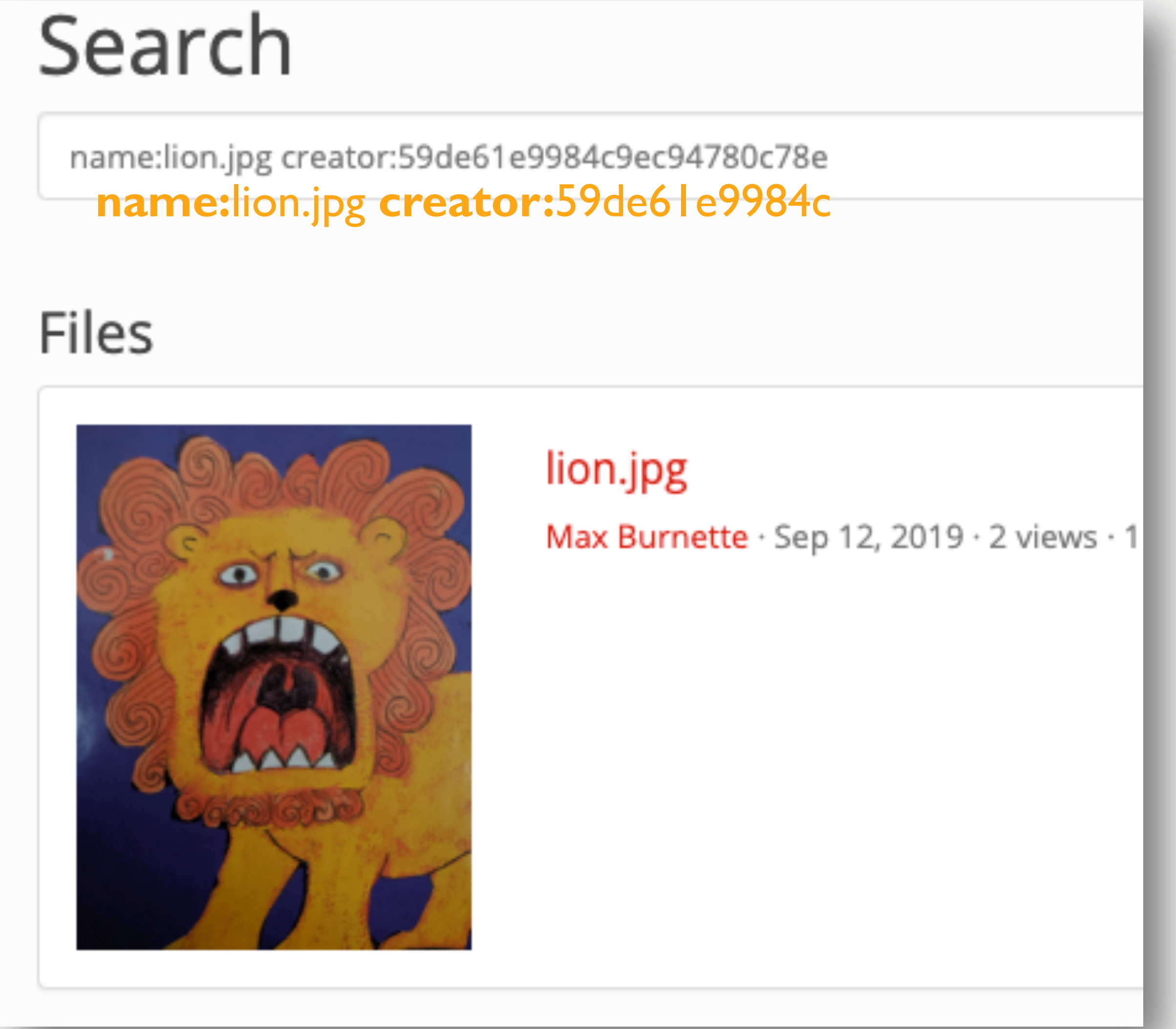
Max Burnette · Sep 12, 2019 · 2 views · 16 downloads · image/jpeg · 91.2 kB

santa.jpg

SUPPORTED SYNTAX

Several other fields can be specified in the query string (GUI or API) using **field:**value notation:

- **name:***resourceName*
- **creator:***creatorID*
- **resource_type:***type*
- **tag:***tagName*
- **in:***parentID*
- **contains:***childID*
- *metadata fields*
  - **"Alternative Title":**"*felis catus*"



Search

name:lion.jpg creator:59de61e9984c9ec94780c78e

name:lion.jpg **creator:**59de61e9984c

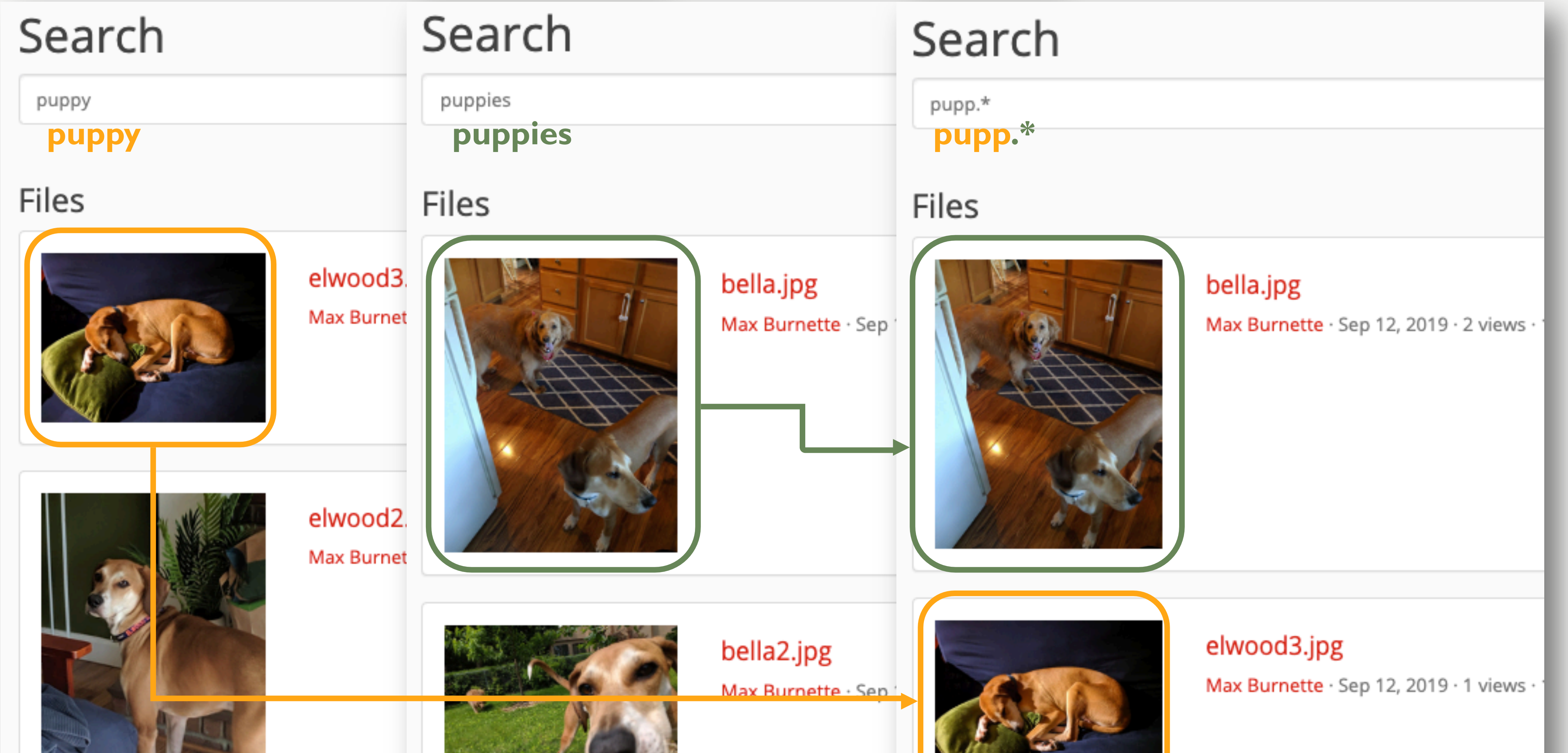
Files

lion.jpg

Max Burnette · Sep 12, 2019 · 2 views · 1

SUPPORTED SYNTAX

Regular expressions are also generally supported.



Search

puppy

Files

elwood3.

Max Burnet

elwood2.

Max Burnet

Search

puppies

Files

bella.jpg

Max Burnette · Sep

bella2.jpg

Max Burnette · Sep

Search

pupp.*

Files

bella.jpg

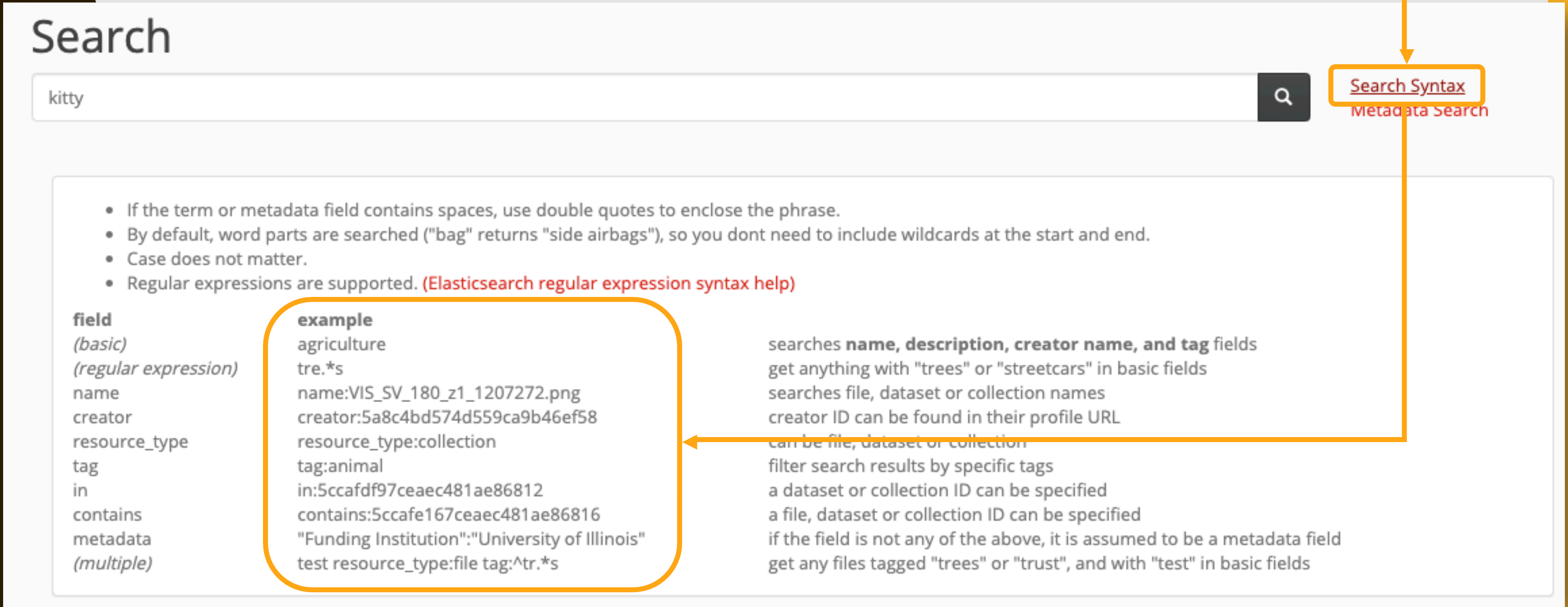
Max Burnette · Sep 12, 2019 · 2 views ·

elwood3.jpg

Max Burnette · Sep 12, 2019 · 1 views ·

SUPPORTED SYNTAX

There is a syntax help link next to the text box on the Search page.



Search

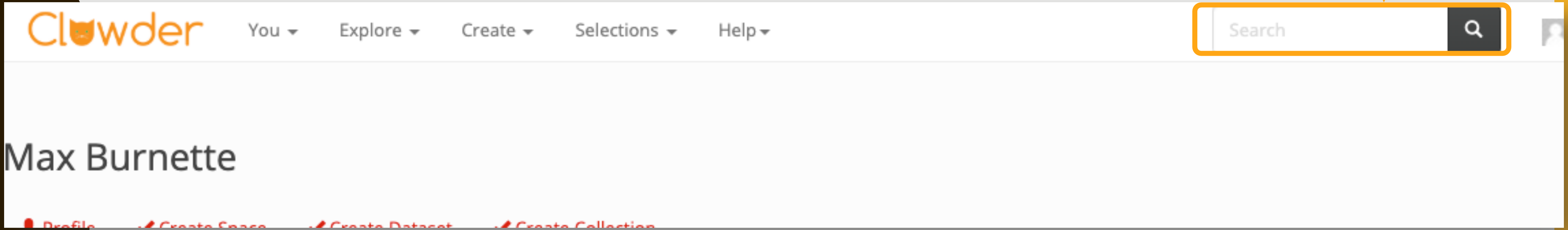
kitty

- If the term or metadata field contains spaces, use double quotes to enclose the phrase.
- By default, word parts are searched ("bag" returns "side airbags"), so you dont need to include wildcards at the start and end.
- Case does not matter.
- Regular expressions are supported. (Elasticsearch regular expression syntax help)

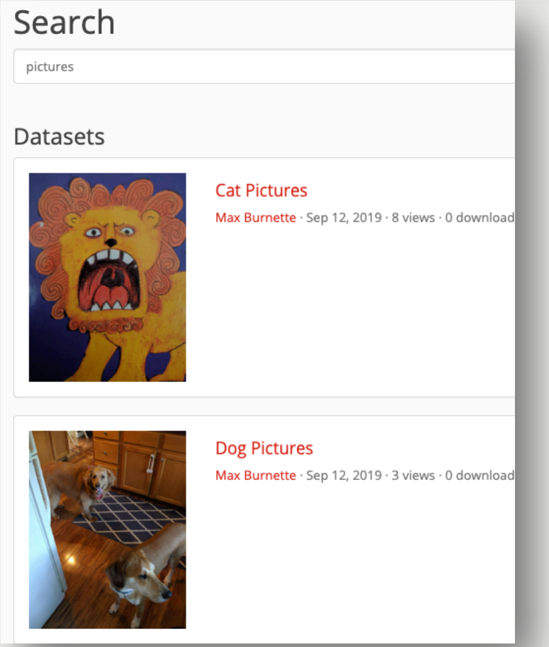
field	**example**	
(basic)	agriculture	searches **name, description, creator name, and tag** fields
(regular expression)	tre.*s	get anything with "trees" or "streetcars" in basic fields
name	name:VIS_SV_180_z1_1207272.png	searches file, dataset or collection names
creator	creator:5a8c4bd574d559ca9b46ef58	creator ID can be found in their profile URL
resource_type	resource_type:collection	can be file, dataset or collection
tag	tag:animal	filter search results by specific tags
in	in:5ccafdf97ceaec481ae86812	a dataset or collection ID can be specified
contains	contains:5ccafe167ceaec481ae86816	a file, dataset or collection ID can be specified
metadata	"Funding Institution":"University of Illinois"	if the field is not any of the above, it is assumed to be a metadata field
(multiple)	test resource_type:file tag:^tr.*s	get any files tagged "trees" or "trust", and with "test" in basic fields

SEARCHING VIA GUI – SEARCH BOX

- **Search Box** – perform basic queries as described on previous slides

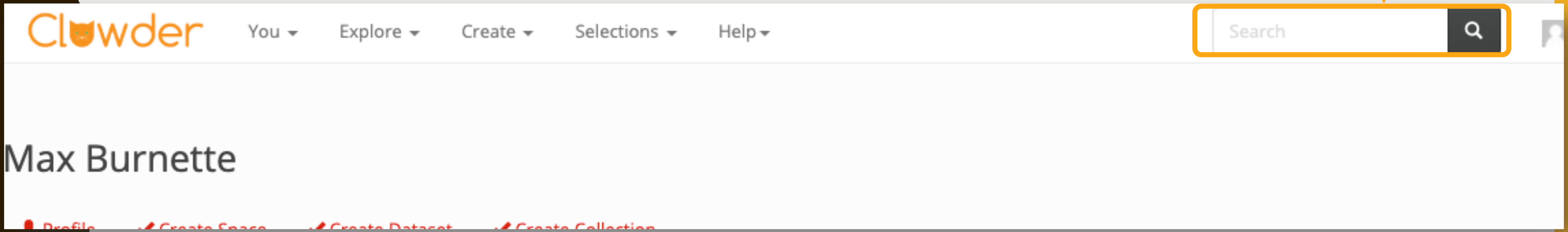


- Spaces between terms are treated as “AND”
- Use quotes to delineate phrases
- Will show results on the **/search** page
  - Currently capped at 60 results displayed
  - Infinite scrolling feature is in review
  - Will include total results count

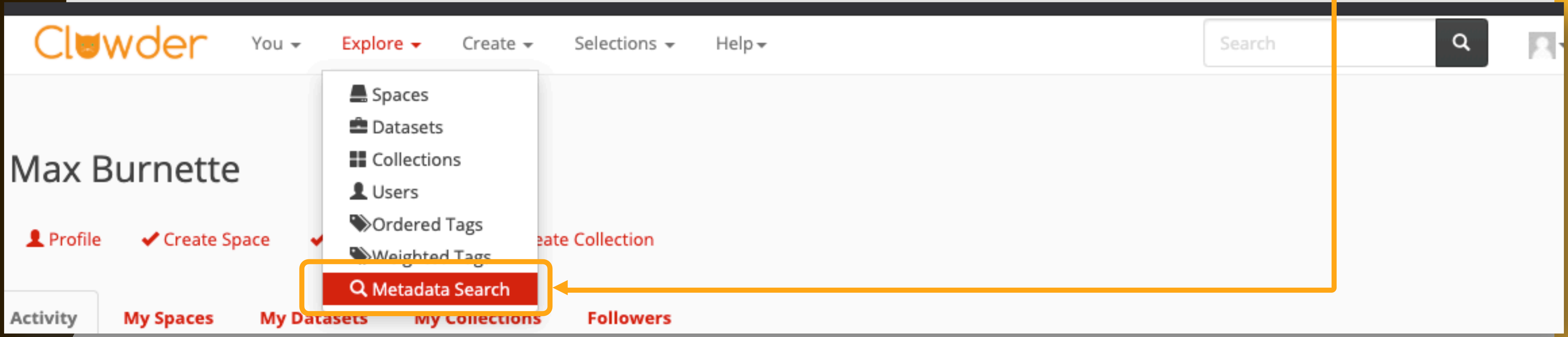

Search
pictures
Datasets
Cat Pictures
Max Burnette · Sep 12, 2019 · 8 views · 0 download
Dog Pictures
Max Burnette · Sep 12, 2019 · 3 views · 0 download

SEARCHING VIA GUI

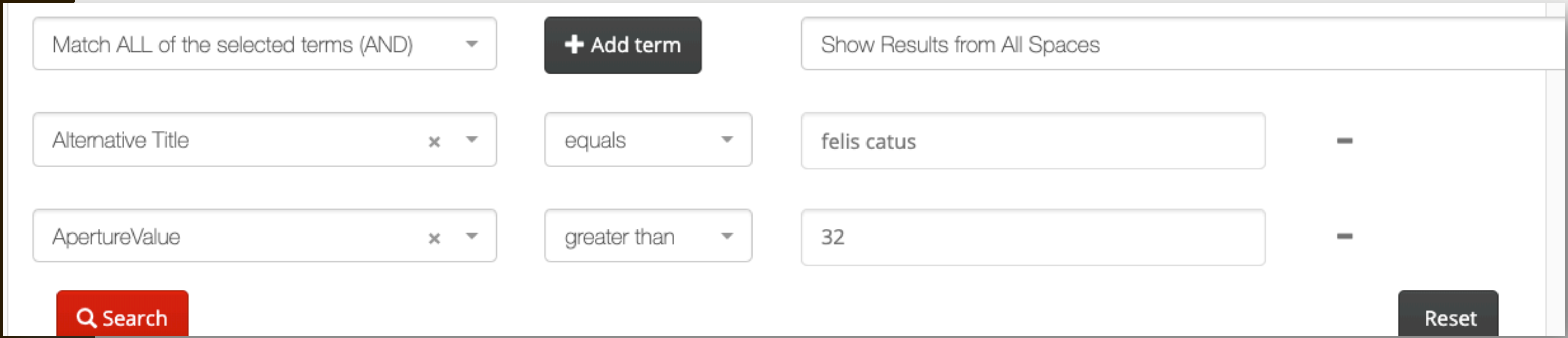
- **Search Box** – perform basic queries as described on previous slides



- **Metadata Search** – build complex query using GUI interface



SEARCHING VIA GUI – METADATA



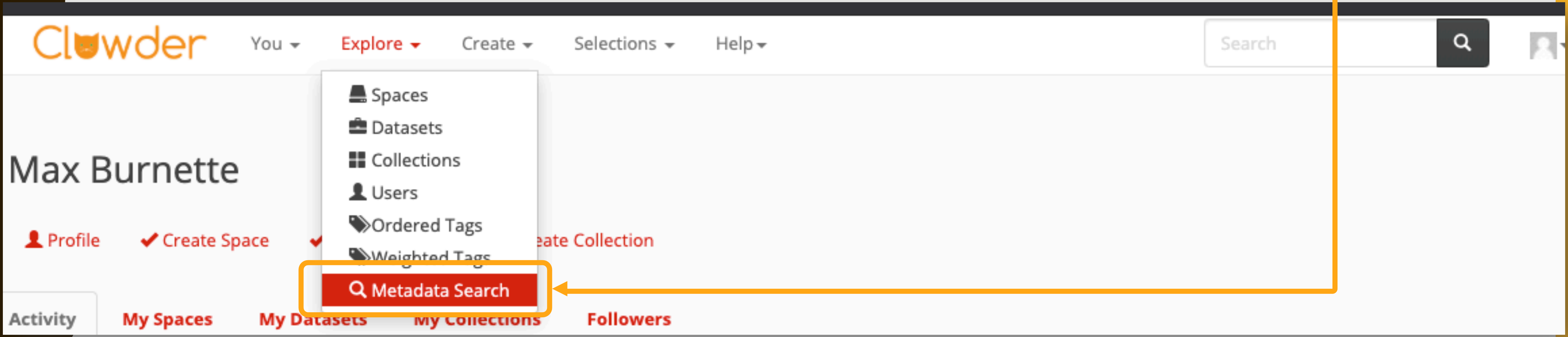
Match ALL of the selected terms (AND)

Show Results from All Spaces

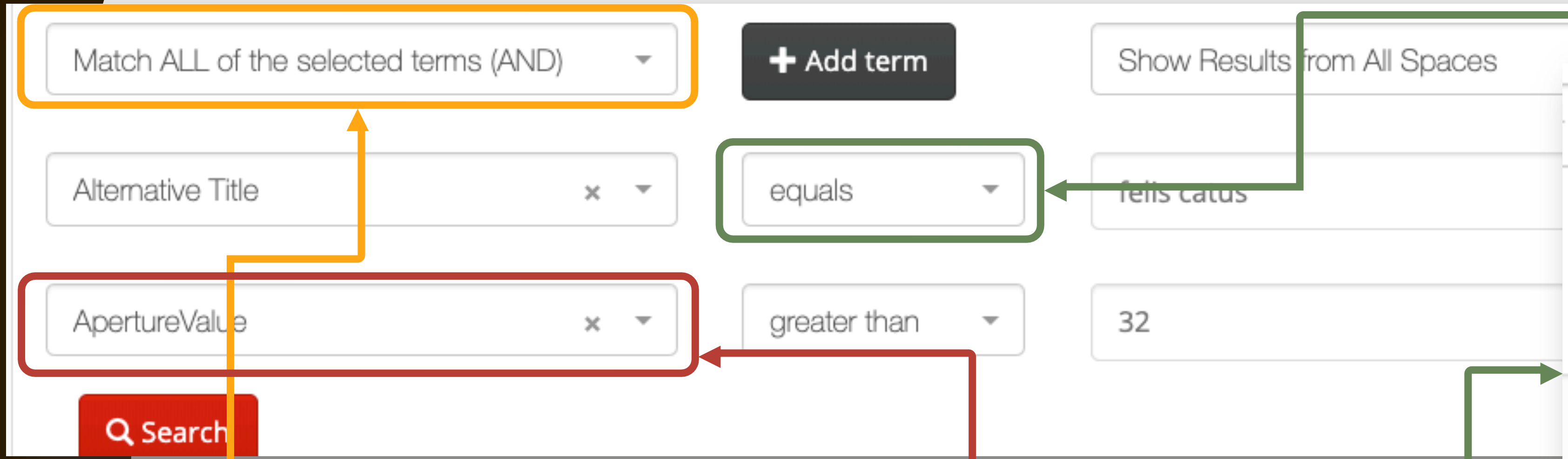
Alternative Title equals felis catus

ApertureValue greater than 32

- **Metadata Search** – build complex query using GUI interface



SEARCHING VIA GUI – METADATA

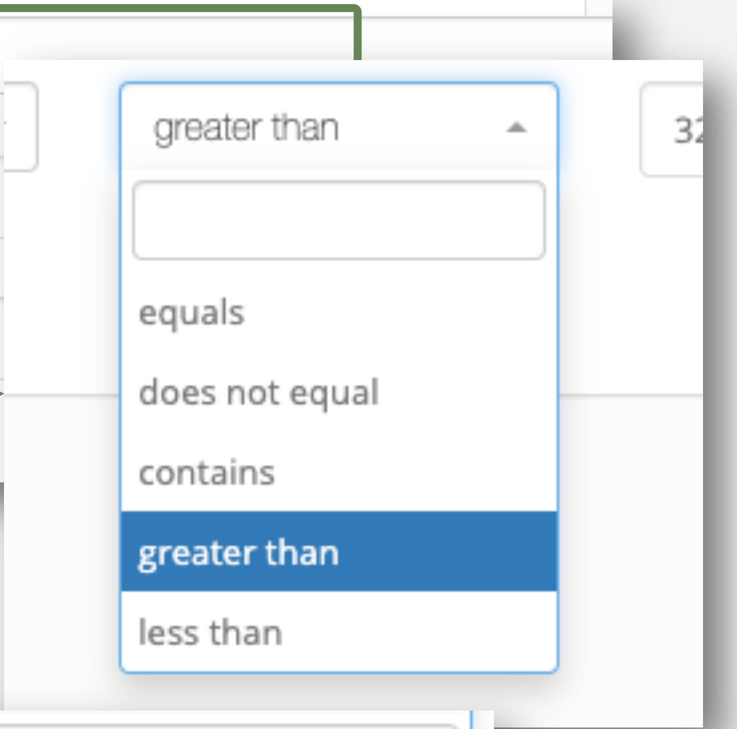


Match ALL of the selected terms (AND)

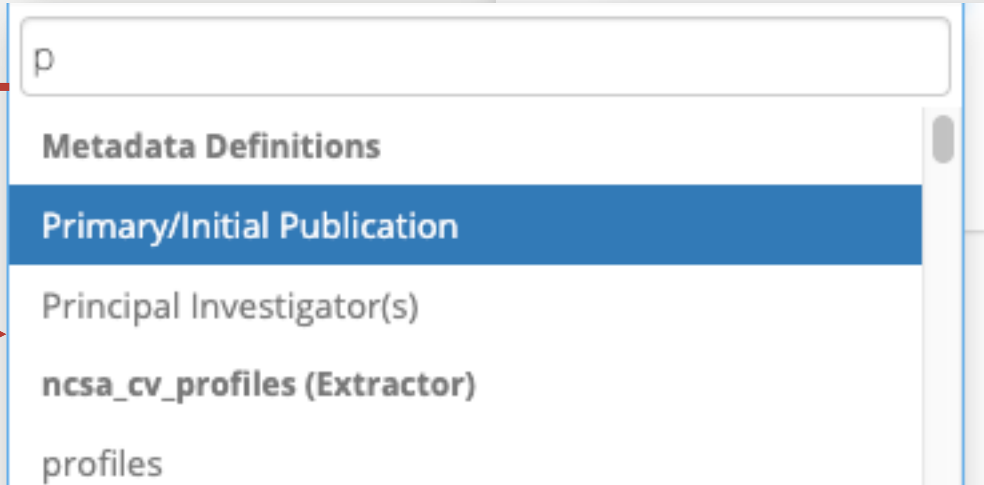
Show Results from All Spaces

Alternative Title | equals | felis catus

ApertureValue | greater than | 32



greater than | 32



p

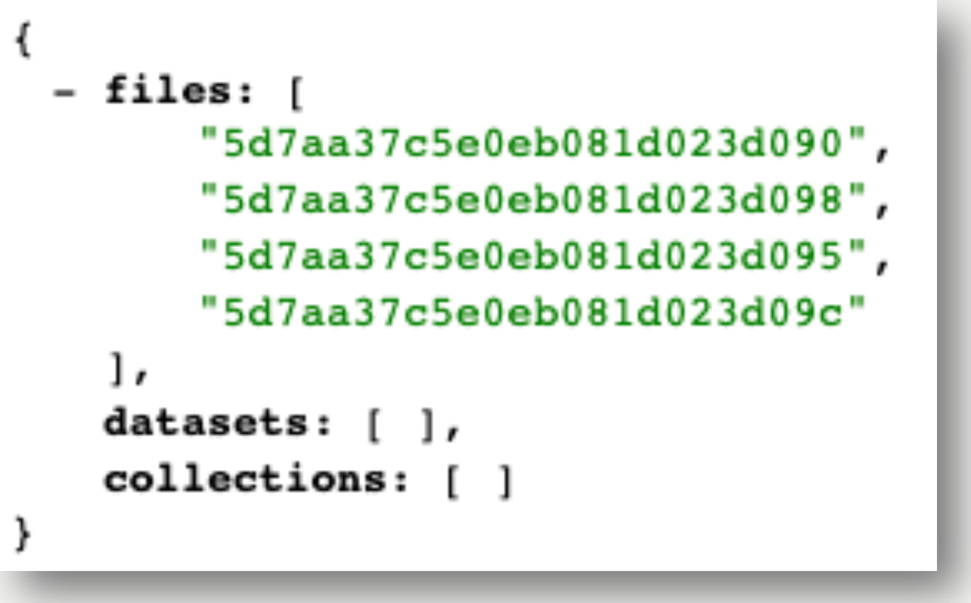
Metadata Definitions
Primary/Initial Publication
Principal Investigator(s)
ncsa_cv_profiles (Extractor)
profiles

- Supports OR grouping in addition to AND
- Allows multiple operators
- Auto-completion of fields in Metadata index, including extracted fields

SEARCHING VIA API – WHAT IS IT?

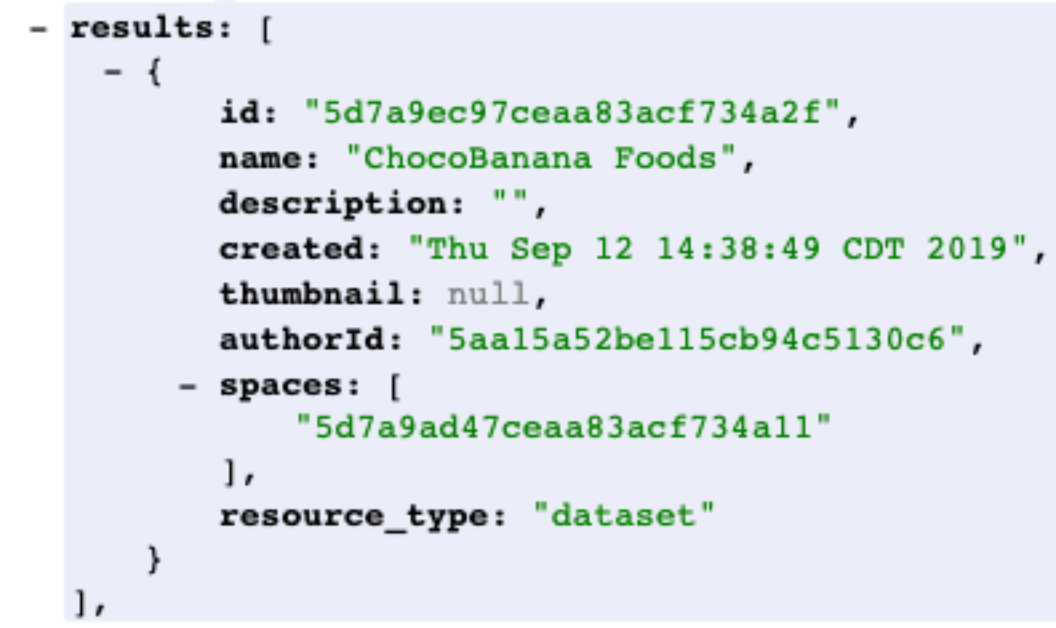
- Clowder’s API allows programs to perform searches
- Data is returned in JSON format, readable by code (and humans!)
- Query syntax is identical to Search Box

Current format



```
{
  - files: [
        "5d7aa37c5e0eb081d023d090",
        "5d7aa37c5e0eb081d023d098",
        "5d7aa37c5e0eb081d023d095",
        "5d7aa37c5e0eb081d023d09c"
    ],
    datasets: [ ],
    collections: [ ]
}
```

Revised (In review)

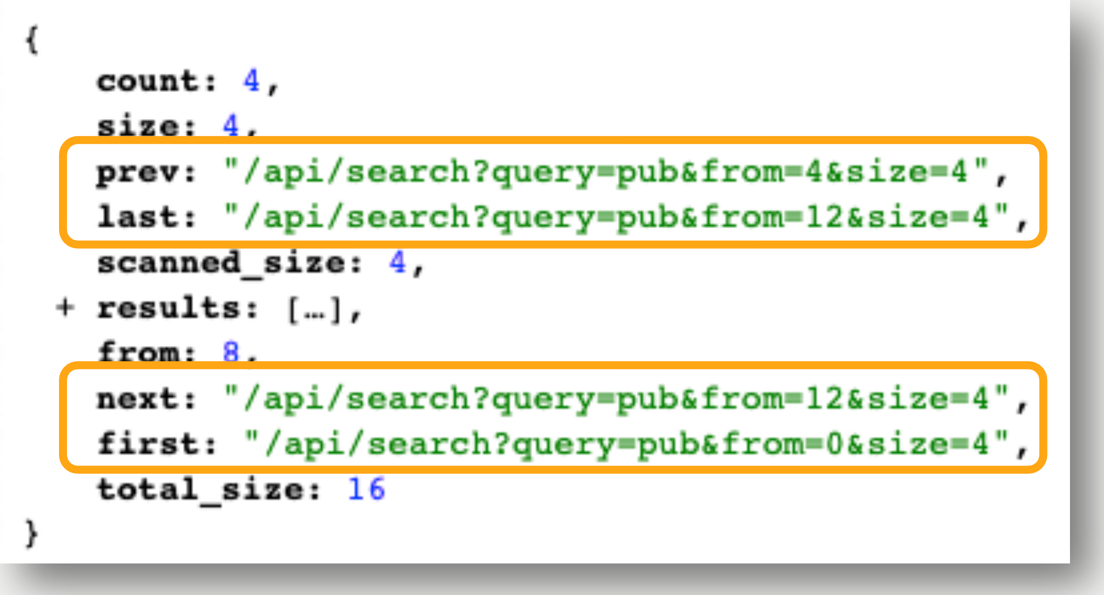


```
- results: [
   - {
          id: "5d7a9ec97ceaa83acf734a2f",
          name: "ChocoBanana Foods",
          description: "",
          created: "Thu Sep 12 14:38:49 CDT 2019",
          thumbnail: null,
          authorId: "5aa15a52be115cb94c5130c6",
        - spaces: [
              "5d7a9ad47ceaa83acf734a11"
          ],
          resource_type: "dataset"
     }
  ],
```

SEARCHING VIA API – PAGINATION

- Infinite scrolling is not possible when using API
- Instead, Clowder can return information on getting the next page of results
  - If user specifies a page number or page size, this information is included
  - This feature is currently in review

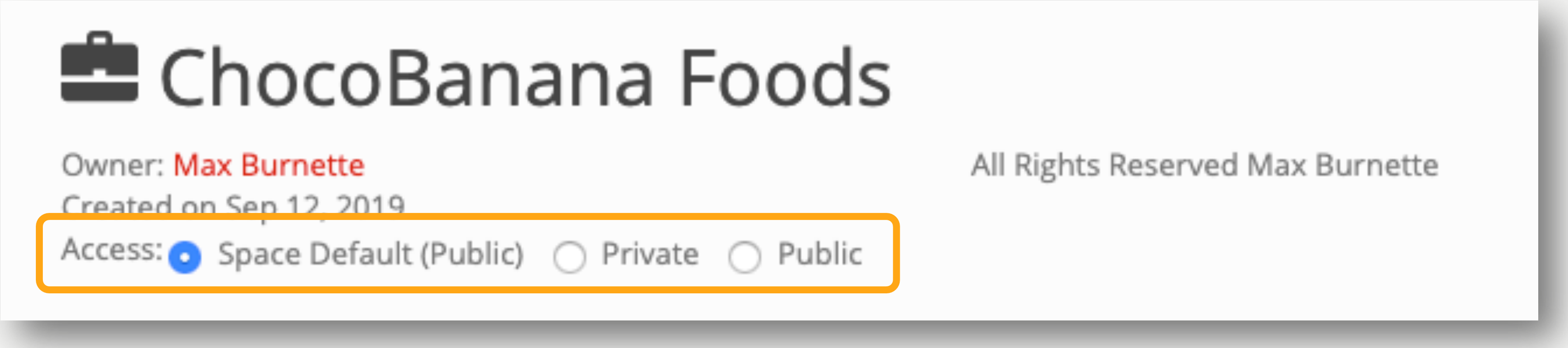
clowder/**api**/search?**query=pub**&**size=4**&**from=8**



```
{
    count: 4,
    size: 4,
    prev: "/api/search?query=pub&from=4&size=4",
    last: "/api/search?query=pub&from=12&size=4",
    scanned_size: 4,
  + results: […],
    from: 8,
    next: "/api/search?query=pub&from=12&size=4",
    first: "/api/search?query=pub&from=0&size=4",
    total_size: 16
}
```

TECH SNIPPETS I - PERMISSIONS

- Across all search methods, you only see results you have permission to see!
- As a reminder:
  - Spaces can be made private or public
  - Datasets can be made private, *even in public spaces*
- You may see different results from another user for the same search



ChocoBanana Foods

Owner: Max Burnette

All Rights Reserved Max Burnette

Created on Sep 12, 2019

Access: ● Space Default (Public) ○ Private ○ Public

TECH SNIPPETS II - DATABASES

- Clowder has two ways of storing searchable data:
  - MongoDB (***required***, *always enabled)*
  - Elasticsearch (***optional***, *supports additional features)*

- Some of the search features in this presentation require Elasticsearch to be enabled
  - in addition to MongoDB, not instead of!
- Many features are also supported by MongoDB alone, but may be slightly slower
- Some features such as the Search Box are not currently enabled with MongoDB alone

QUESTIONS?

- **What is searchable?**
  - Supported syntax
- **Searching via GUI**
  - Search Box
  - Metadata Search
- **Searching via API**
  - What is the API?
  - Pagination
- **Technical Snippets**
  - Permissions
  - Elasticsearch & MongoDB

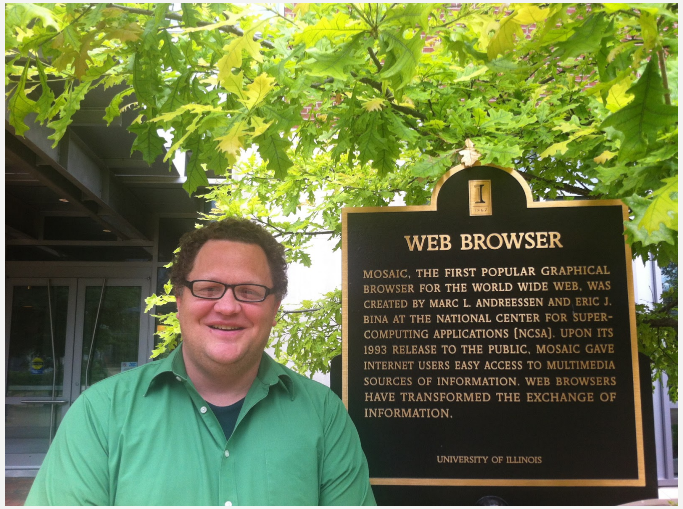
Max Burnette, NCSA

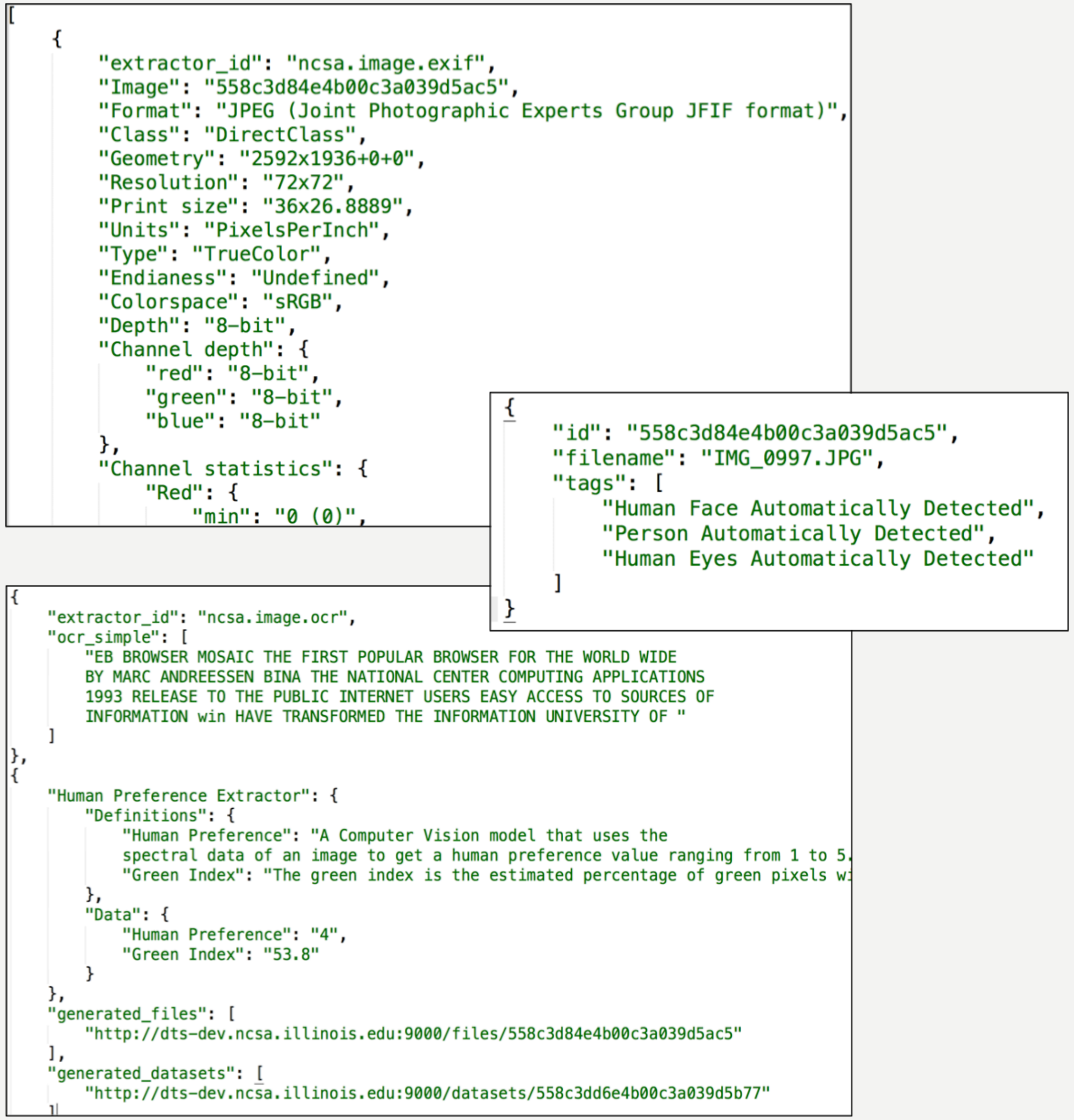
mburnet2@Illinois.edu

METADATA EXTRACTORS

LUIGI MARINI

AUTOMATIC METADATA EXTRACTION

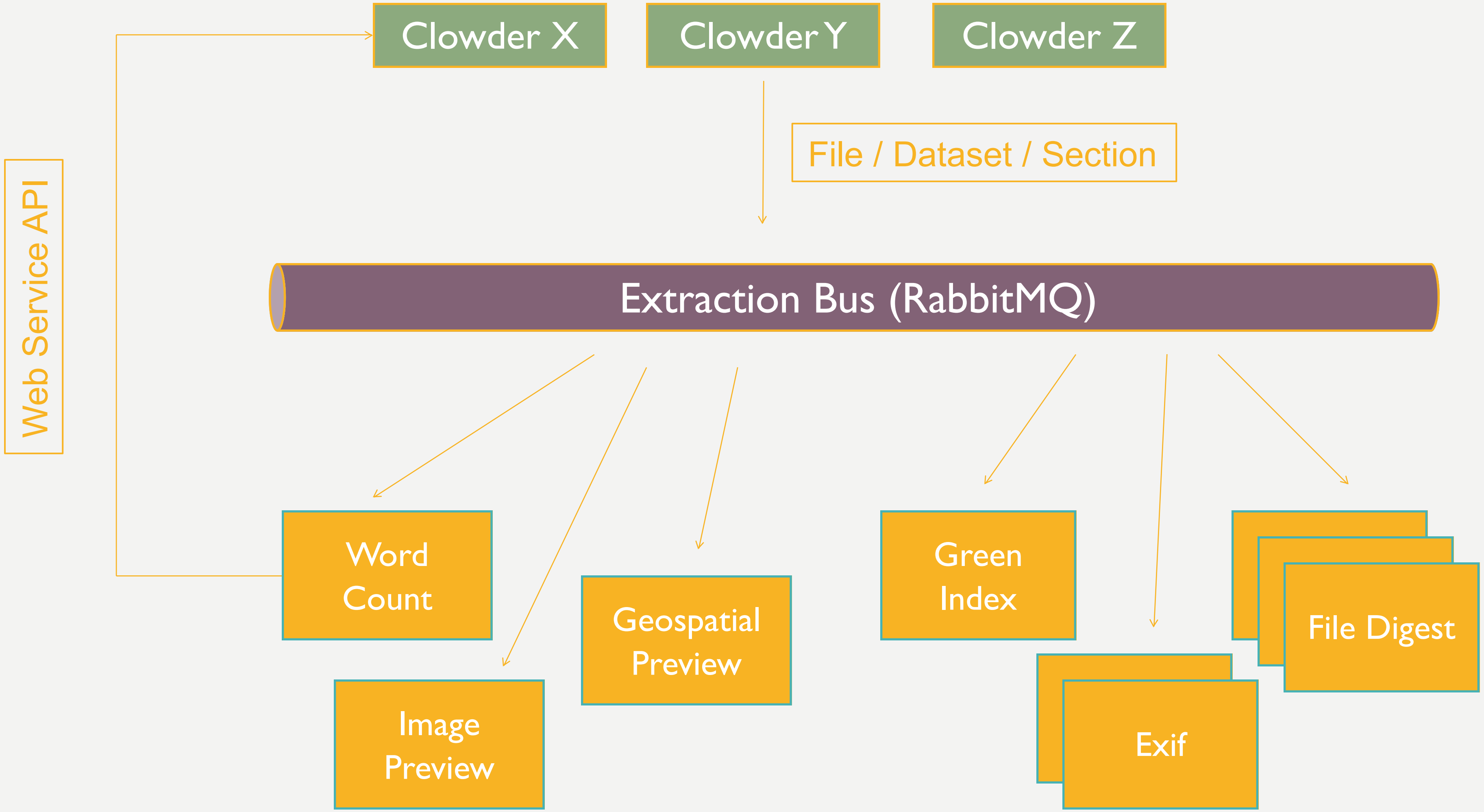

WEB BROWSER
MOSAIC, THE FIRST POPULAR GRAPHICAL BROWSER FOR THE WORLD WIDE WEB, WAS CREATED BY MARC L. ANDREESSEN AND ERIC J. BINA AT THE NATIONAL CENTER FOR SUPER-COMPUTING APPLICATIONS [NCSA]. UPON ITS 1993 RELEASE TO THE PUBLIC, MOSAIC GAVE INTERNET USERS EASY ACCESS TO MULTIMEDIA SOURCES OF INFORMATION. WEB BROWSERS HAVE TRANSFORMED THE EXCHANGE OF INFORMATION.
UNIVERSITY OF ILLINOIS



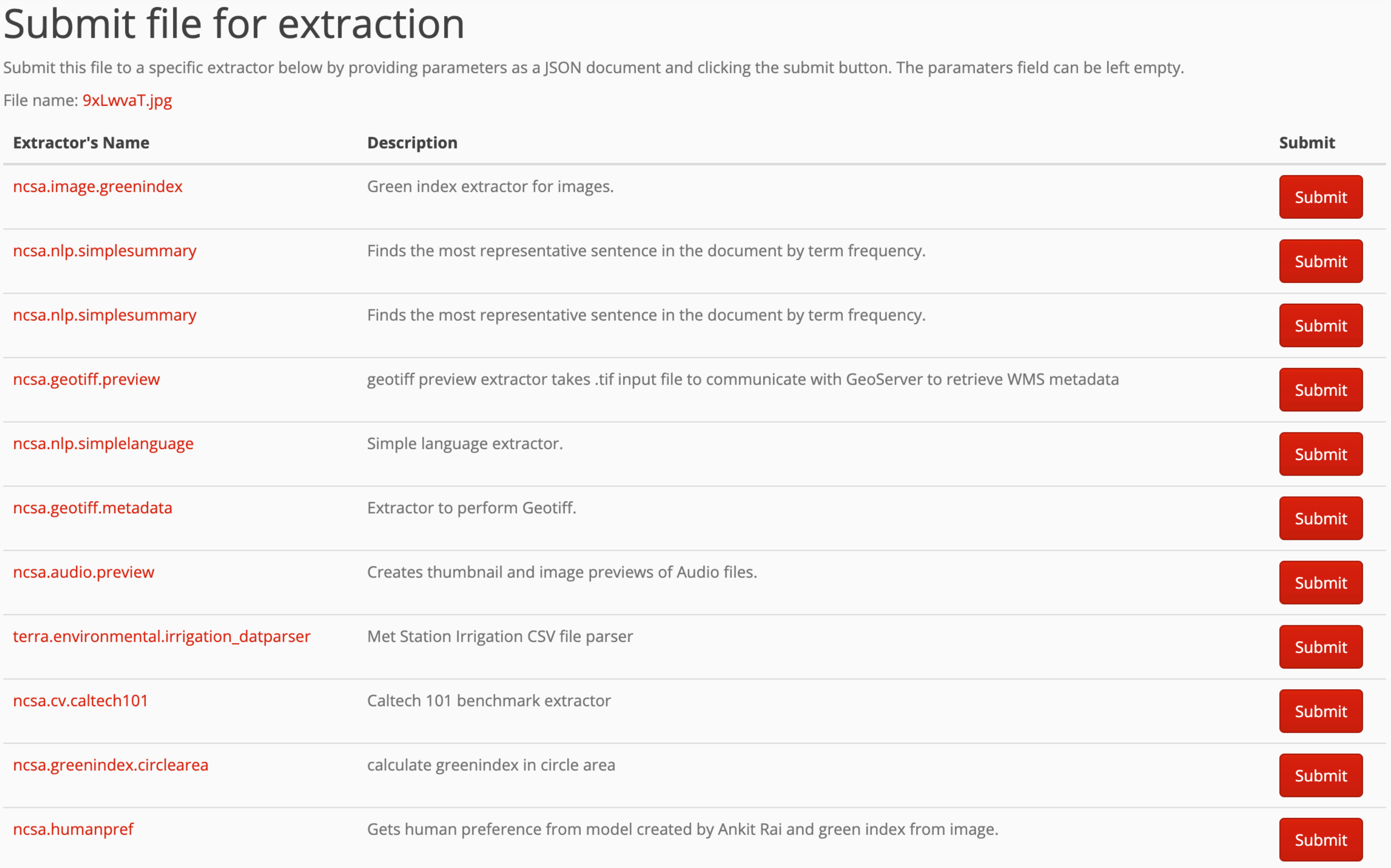
```
[
    {
        "extractor_id": "ncsa.image.exif",
        "Image": "558c3d84e4b00c3a039d5ac5",
        "Format": "JPEG (Joint Photographic Experts Group JFIF format)",
        "Class": "DirectClass",
        "Geometry": "2592x1936+0+0",
        "Resolution": "72x72",
        "Print size": "36x26.8889",
        "Units": "PixelsPerInch",
        "Type": "TrueColor",
        "Endianess": "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
            spectral data of an image to get a human preference value ranging from 1 to 5.
            "Green Index": "The green index is the estimated percentage of green pixels wi
        },
        "Data": {
            "Human Preference": "4",
            "Green Index": "53.8"
        }
    },
    "generated_files": [
        "http://dts-dev.ncsa.illinois.edu:9000/files/558c3d84e4b00c3a039d5ac5"
    ],
    "generated_datasets": [
        "http://dts-dev.ncsa.illinois.edu:9000/datasets/558c3dd6e4b00c3a039d5b77"
    ]
```


Clowder X
Clowder Y
Clowder Z
File / Dataset / Section
Web Service API
Extraction Bus (RabbitMQ)
Word Count
Image Preview
Geospatial Preview
Green Index
Exif
File Digest

FILE MANUAL SUBMISSION



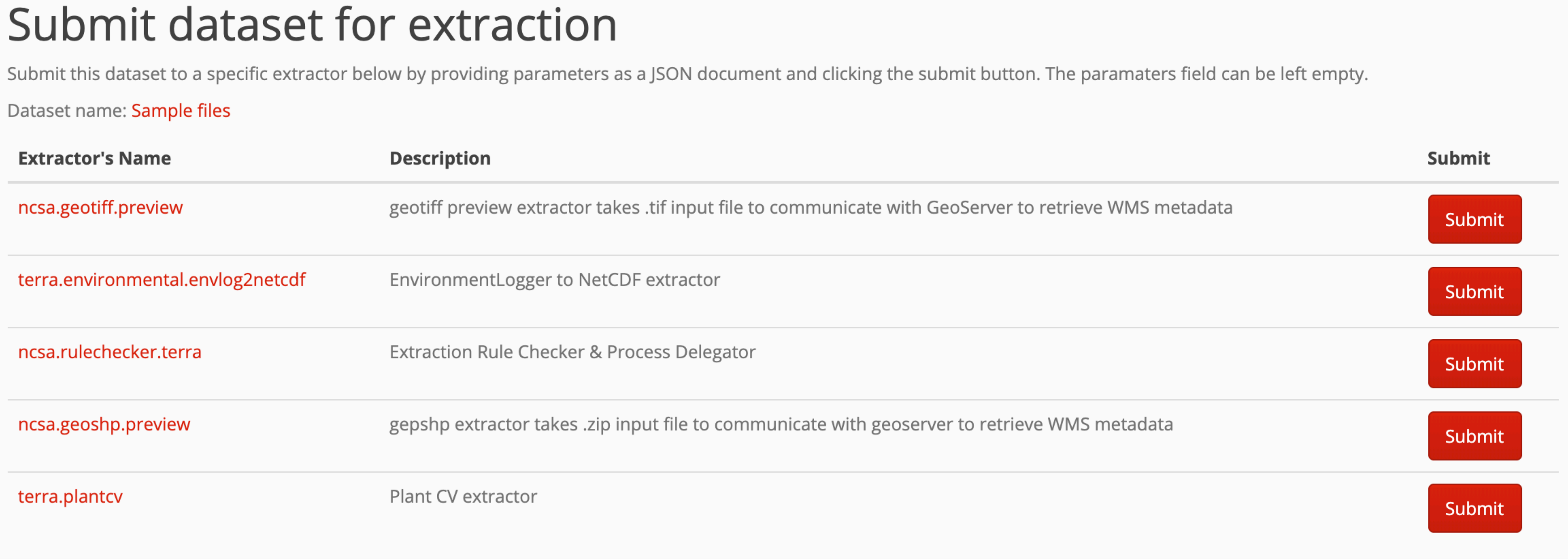
Submit file for extraction

Submit this file to a specific extractor below by providing parameters as a JSON document and clicking the submit button. The paramaters field can be left empty.

File name: 9xLwvaT.jpg

Extractor's Name	Description	Submit
ncsa.image.greenindex	Green index extractor for images.	Submit
ncsa.nlp.simplesummary	Finds the most representative sentence in the document by term frequency.	Submit
ncsa.nlp.simplesummary	Finds the most representative sentence in the document by term frequency.	Submit
ncsa.geotiff.preview	geotiff preview extractor takes .tif input file to communicate with GeoServer to retrieve WMS metadata	Submit
ncsa.nlp.simplelanguage	Simple language extractor.	Submit
ncsa.geotiff.metadata	Extractor to perform Geotiff.	Submit
ncsa.audio.preview	Creates thumbnail and image previews of Audio files.	Submit
terra.environmental.irrigation_datparser	Met Station Irrigation CSV file parser	Submit
ncsa.cv.caltech101	Caltech 101 benchmark extractor	Submit
ncsa.greenindex.circlearea	calculate greenindex in circle area	Submit
ncsa.humanpref	Gets human preference from model created by Ankit Rai and green index from image.	Submit

DATASET MANUAL SUBMISSION



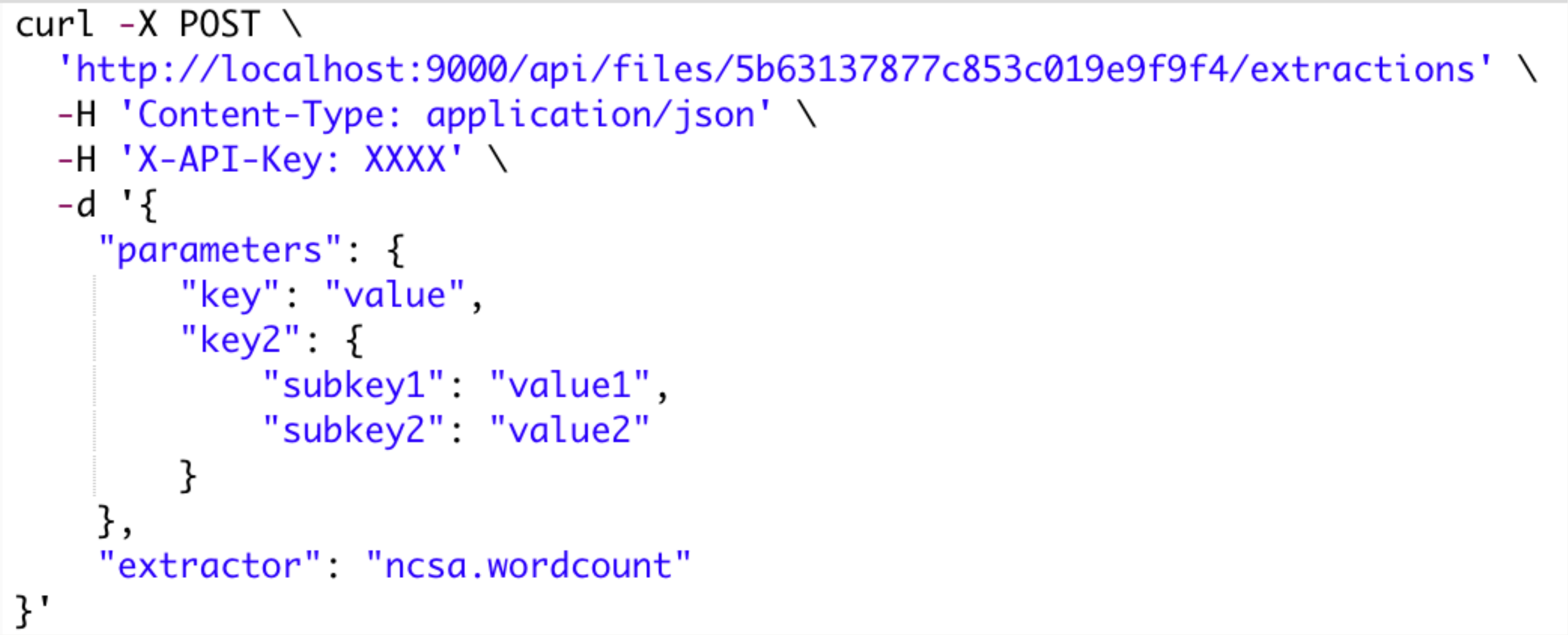
Submit dataset for extraction

Submit this dataset to a specific extractor below by providing parameters as a JSON document and clicking the submit button. The paramaters field can be left empty.

Dataset name: Sample files

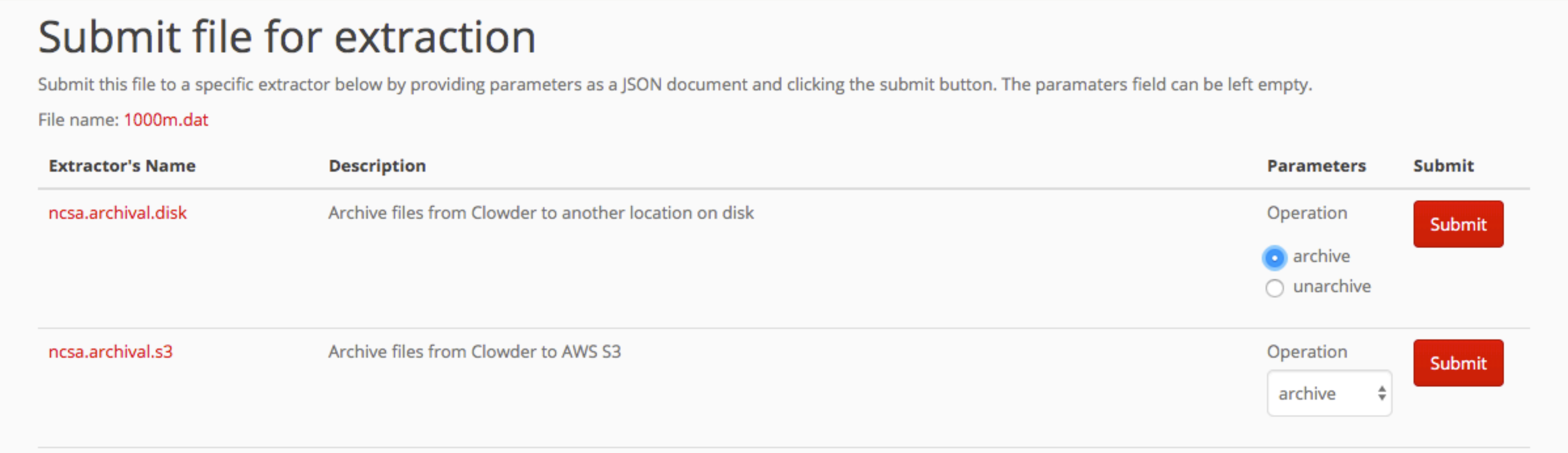
Extractor's Name	Description	Submit
ncsa.geotiff.preview	geotiff preview extractor takes .tif input file to communicate with GeoServer to retrieve WMS metadata	Submit
terra.environmental.envlog2netcdf	EnvironmentLogger to NetCDF extractor	Submit
ncsa.rulechecker.terra	Extraction Rule Checker & Process Delegator	Submit
ncsa.geoshp.preview	gepshp extractor takes .zip input file to communicate with geoserver to retrieve WMS metadata	Submit
terra.plantcv	Plant CV extractor	Submit

SUBMIT PARAMETERS THROUGH API



```
curl -X POST \
  'http://localhost:9000/api/files/5b63137877c853c019e9f9f4/extractions' \
  -H 'Content-Type: application/json' \
  -H 'X-API-Key: XXXX' \
  -d '{
    "parameters": {
        "key": "value",
        "key2": {
            "subkey1": "value1",
            "subkey2": "value2"
        }
    },
    "extractor": "ncsa.wordcount"
}'
```

SUBMISSION PARAMETERS
(IN DEVELOPMENT)



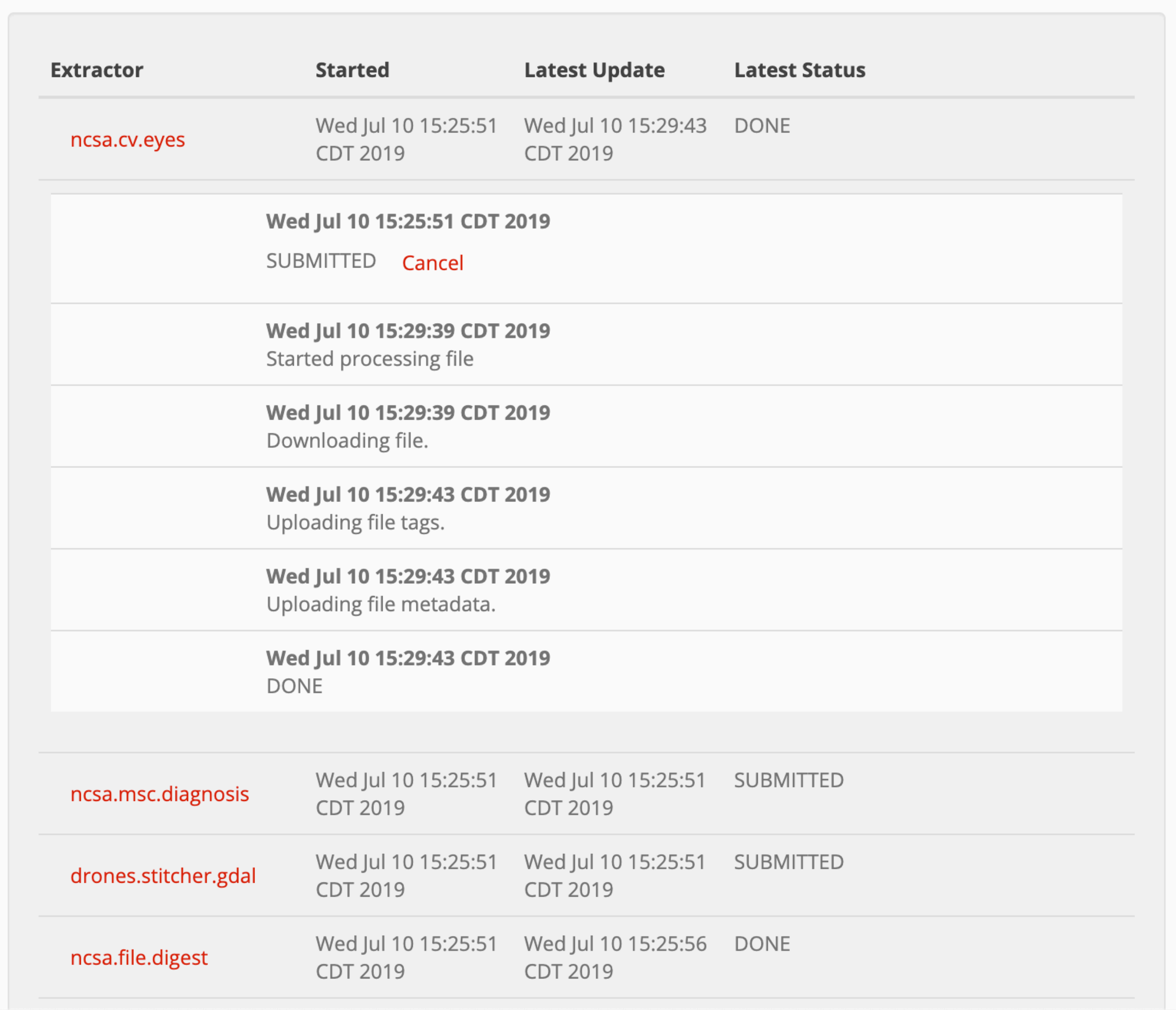
Submit file for extraction

Submit this file to a specific extractor below by providing parameters as a JSON document and clicking the submit button. The paramaters field can be left empty.

File name: 1000m.dat

Extractor's Name	Description	Parameters	Submit
ncsa.archival.disk	Archive files from Clowder to another location on disk	Operation ● archive ○ unarchive	Submit
ncsa.archival.s3	Archive files from Clowder to AWS S3	Operation archive	Submit

EXTRACTION EVENTS UI



Extractor	Started	Latest Update	Latest Status
ncsa.cv.eyes	Wed Jul 10 15:25:51 CDT 2019	Wed Jul 10 15:29:43 CDT 2019	DONE

Wed Jul 10 15:25:51 CDT 2019
SUBMITTED Cancel

Wed Jul 10 15:29:39 CDT 2019
Started processing file

Wed Jul 10 15:29:39 CDT 2019
Downloading file.

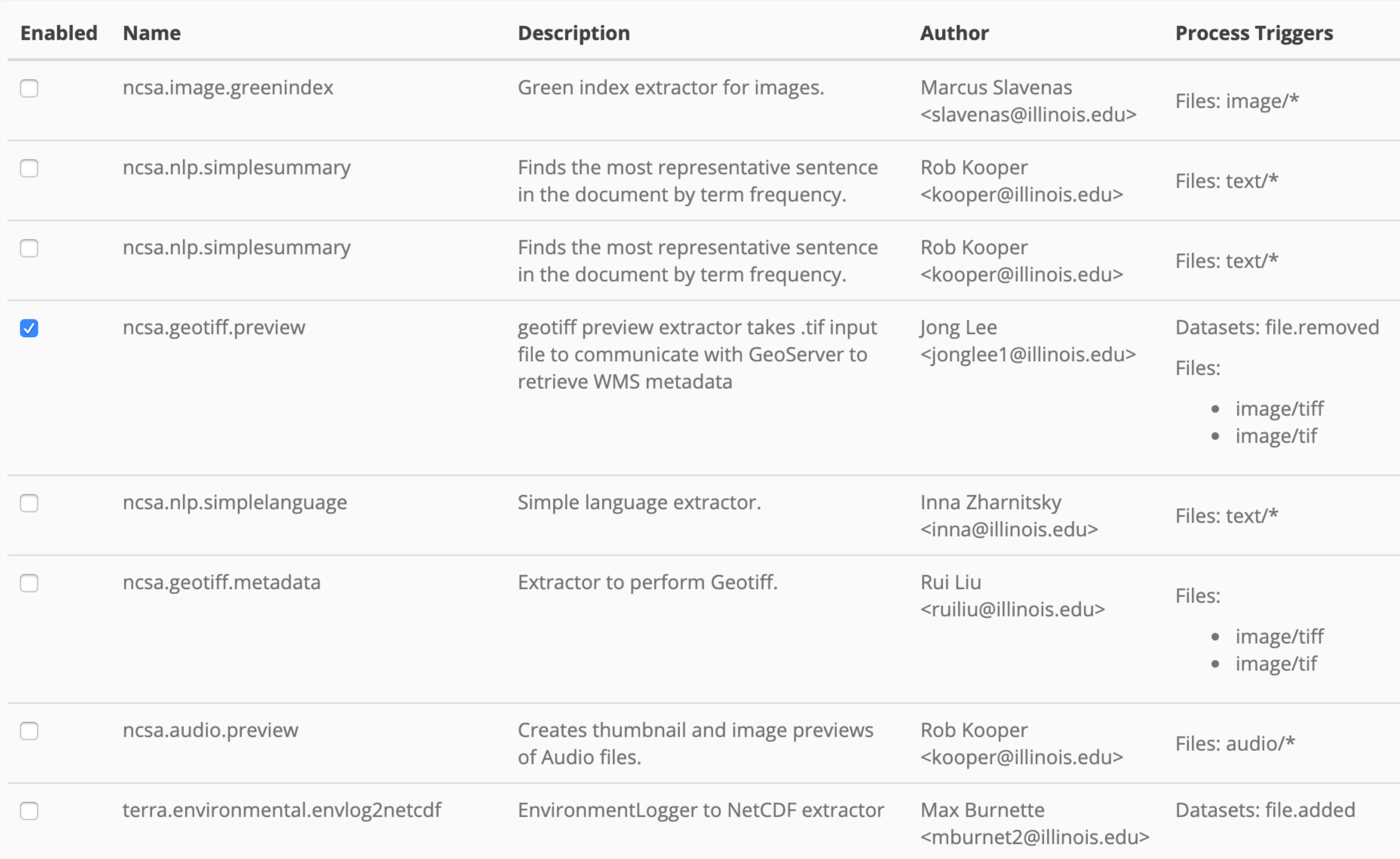
Wed Jul 10 15:29:43 CDT 2019
Uploading file tags.

Wed Jul 10 15:29:43 CDT 2019
Uploading file metadata.

Wed Jul 10 15:29:43 CDT 2019
DONE

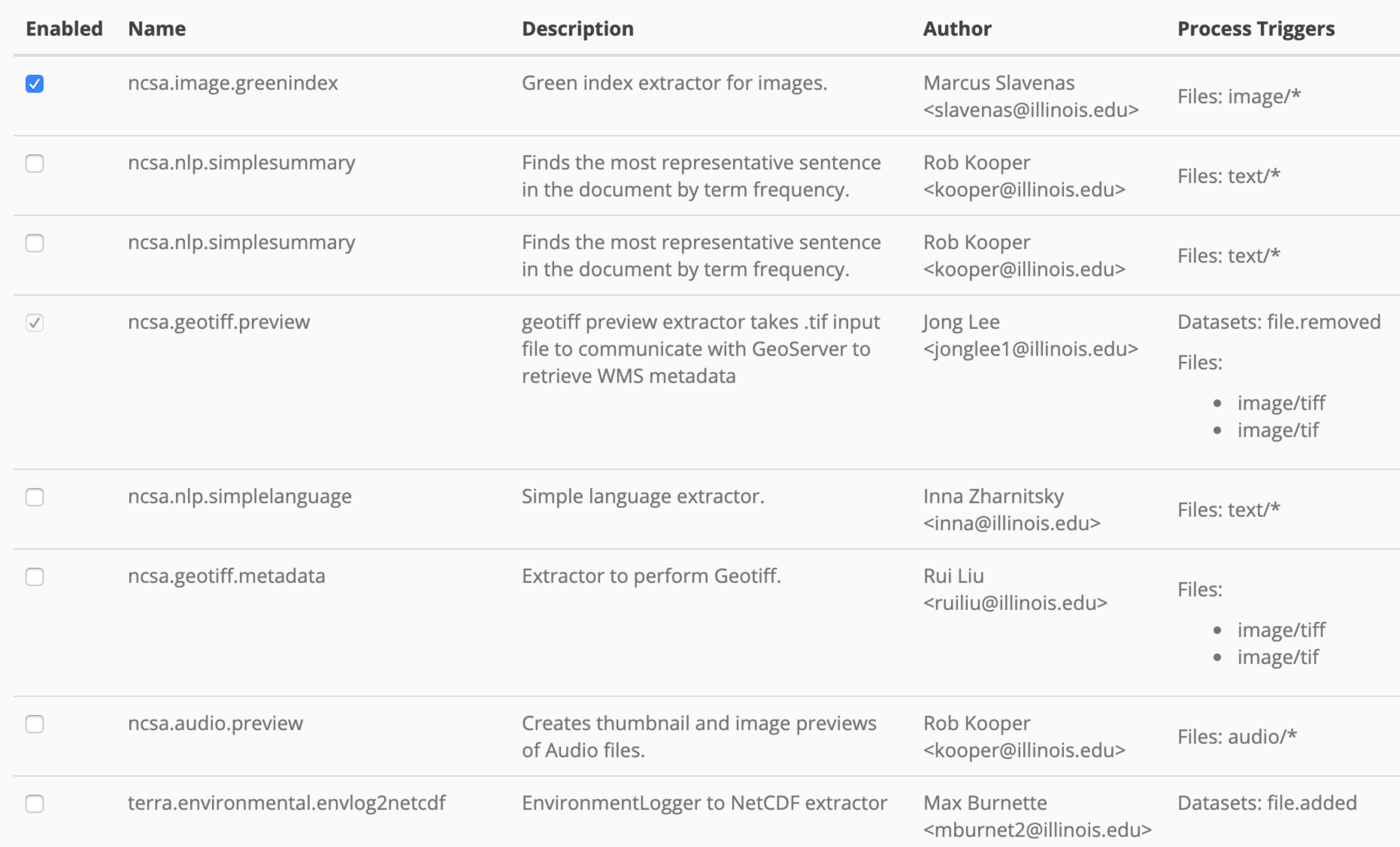
ncsa.msc.diagnosis	Wed Jul 10 15:25:51 CDT 2019	Wed Jul 10 15:25:51 CDT 2019	SUBMITTED
drones.stitcher.gdal	Wed Jul 10 15:25:51 CDT 2019	Wed Jul 10 15:25:51 CDT 2019	SUBMITTED
ncsa.file.digest	Wed Jul 10 15:25:51 CDT 2019	Wed Jul 10 15:25:56 CDT 2019	DONE

GLOBAL EXTRACTOR ENABLE



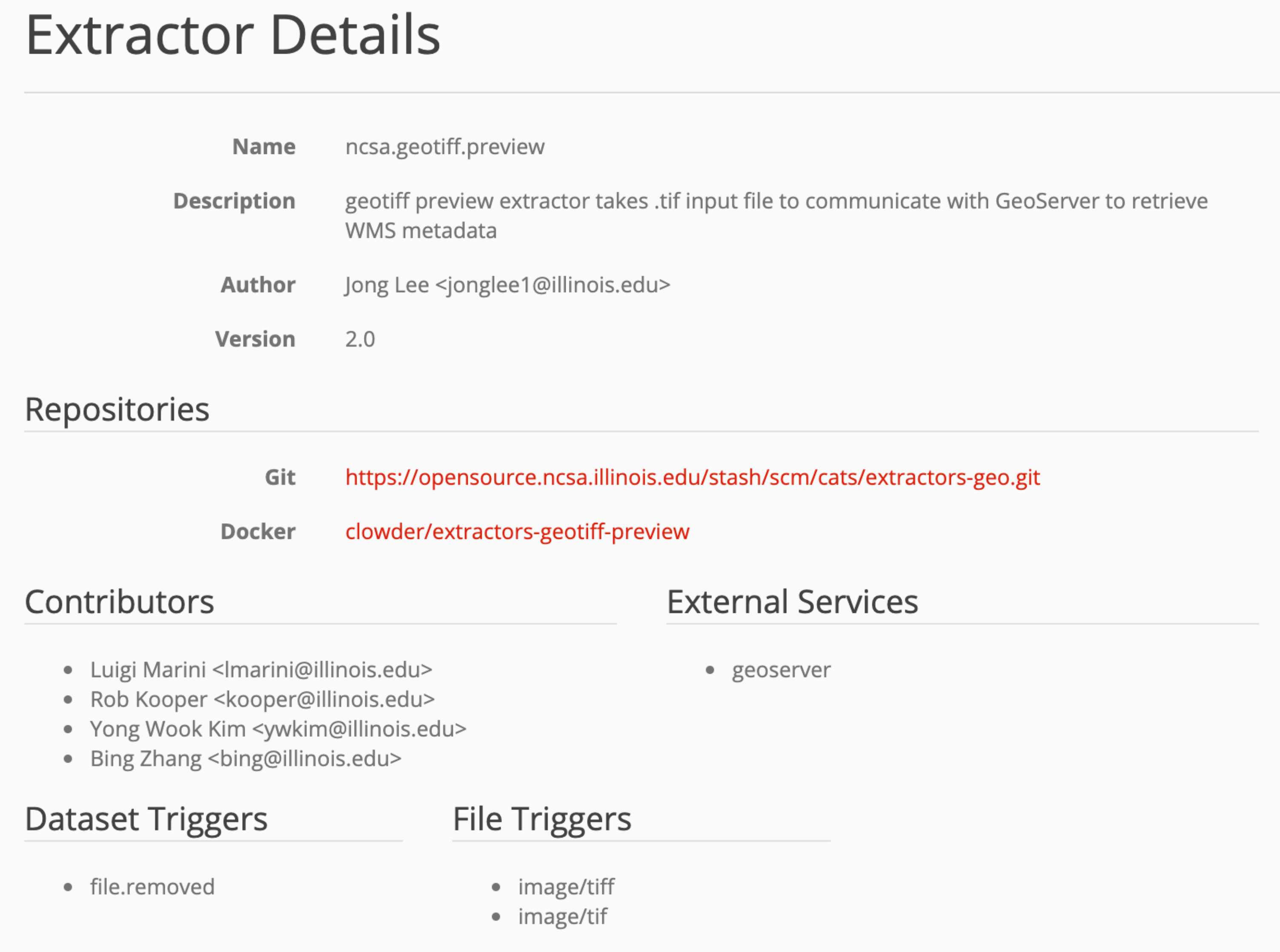
Enabled	Name	Description	Author	Process Triggers
☐	ncsa.image.greenindex	Green index extractor for images.	Marcus Slavenas <slavenas@illinois.edu>	Files: image/*
☐	ncsa.nlp.simplesummary	Finds the most representative sentence in the document by term frequency.	Rob Kooper <kooper@illinois.edu>	Files: text/*
☐	ncsa.nlp.simplesummary	Finds the most representative sentence in the document by term frequency.	Rob Kooper <kooper@illinois.edu>	Files: text/*
☑	ncsa.geotiff.preview	geotiff preview extractor takes .tif input file to communicate with GeoServer to retrieve WMS metadata	Jong Lee <jonglee1@illinois.edu>	Datasets: file.removed Files: • image/tiff • image/tif
☐	ncsa.nlp.simplelanguage	Simple language extractor.	Inna Zharnitsky <inna@illinois.edu>	Files: text/*
☐	ncsa.geotiff.metadata	Extractor to perform Geotiff.	Rui Liu <ruiliu@illinois.edu>	Files: • image/tiff • image/tif
☐	ncsa.audio.preview	Creates thumbnail and image previews of Audio files.	Rob Kooper <kooper@illinois.edu>	Files: audio/*
☐	terra.environmental.envlog2netcdf	EnvironmentLogger to NetCDF extractor	Max Burnette <mburnet2@illinois.edu>	Datasets: file.added

SPACE EXTRACTOR ENABLE



Enabled	Name	Description	Author	Process Triggers
☑	ncsa.image.greenindex	Green index extractor for images.	Marcus Slavenas <slavenas@illinois.edu>	Files: image/*
☐	ncsa.nlp.simplesummary	Finds the most representative sentence in the document by term frequency.	Rob Kooper <kooper@illinois.edu>	Files: text/*
☐	ncsa.nlp.simplesummary	Finds the most representative sentence in the document by term frequency.	Rob Kooper <kooper@illinois.edu>	Files: text/*
☑	ncsa.geotiff.preview	geotiff preview extractor takes .tif input file to communicate with GeoServer to retrieve WMS metadata	Jong Lee <jonglee1@illinois.edu>	Datasets: file.removed Files: • image/tiff • image/tif
☐	ncsa.nlp.simplelanguage	Simple language extractor.	Inna Zharnitsky <inna@illinois.edu>	Files: text/*
☐	ncsa.geotiff.metadata	Extractor to perform Geotiff.	Rui Liu <ruiliu@illinois.edu>	Files: • image/tiff • image/tif
☐	ncsa.audio.preview	Creates thumbnail and image previews of Audio files.	Rob Kooper <kooper@illinois.edu>	Files: audio/*
☐	terra.environmental.envlog2netcdf	EnvironmentLogger to NetCDF extractor	Max Burnette <mburnet2@illinois.edu>	Datasets: file.added

EXTRATORS INFO GUI



Extractor Details

Name	ncsa.geotiff.preview
Description	geotiff preview extractor takes .tif input file to communicate with GeoServer to retrieve WMS metadata
Author	Jong Lee <jonglee1@illinois.edu>
Version	2.0

Repositories

Git	https://opensource.ncsa.illinois.edu/stash/scm/cats/extractors-geo.git
Docker	clowder/extractors-geotiff-preview

Contributors

- Luigi Marini <lmarini@illinois.edu>
- Rob Kooper <kooper@illinois.edu>
- Yong Wook Kim <ywkim@illinois.edu>
- Bing Zhang <bing@illinois.edu>

External Services

- geoserver

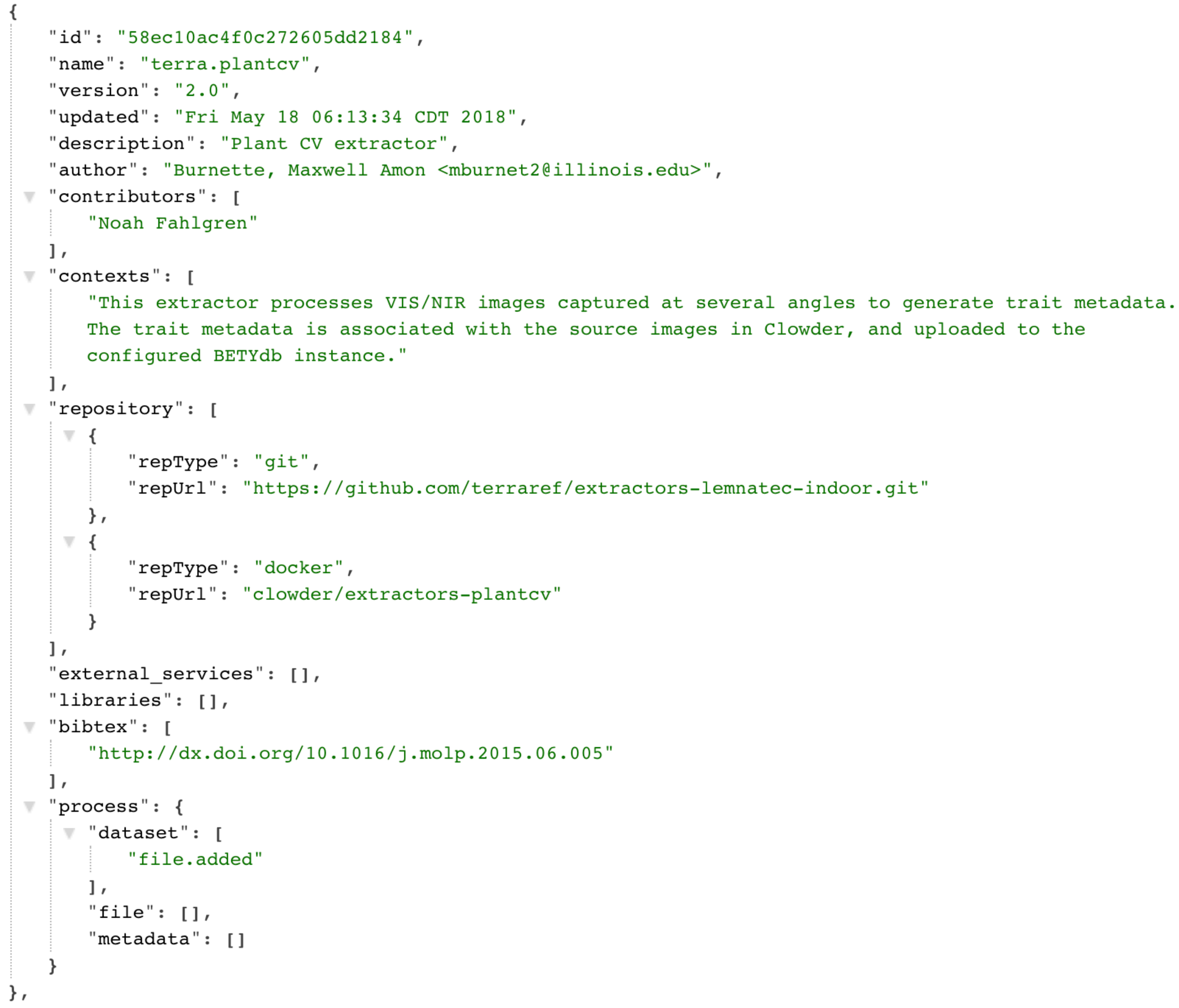
Dataset Triggers

- file.removed

File Triggers

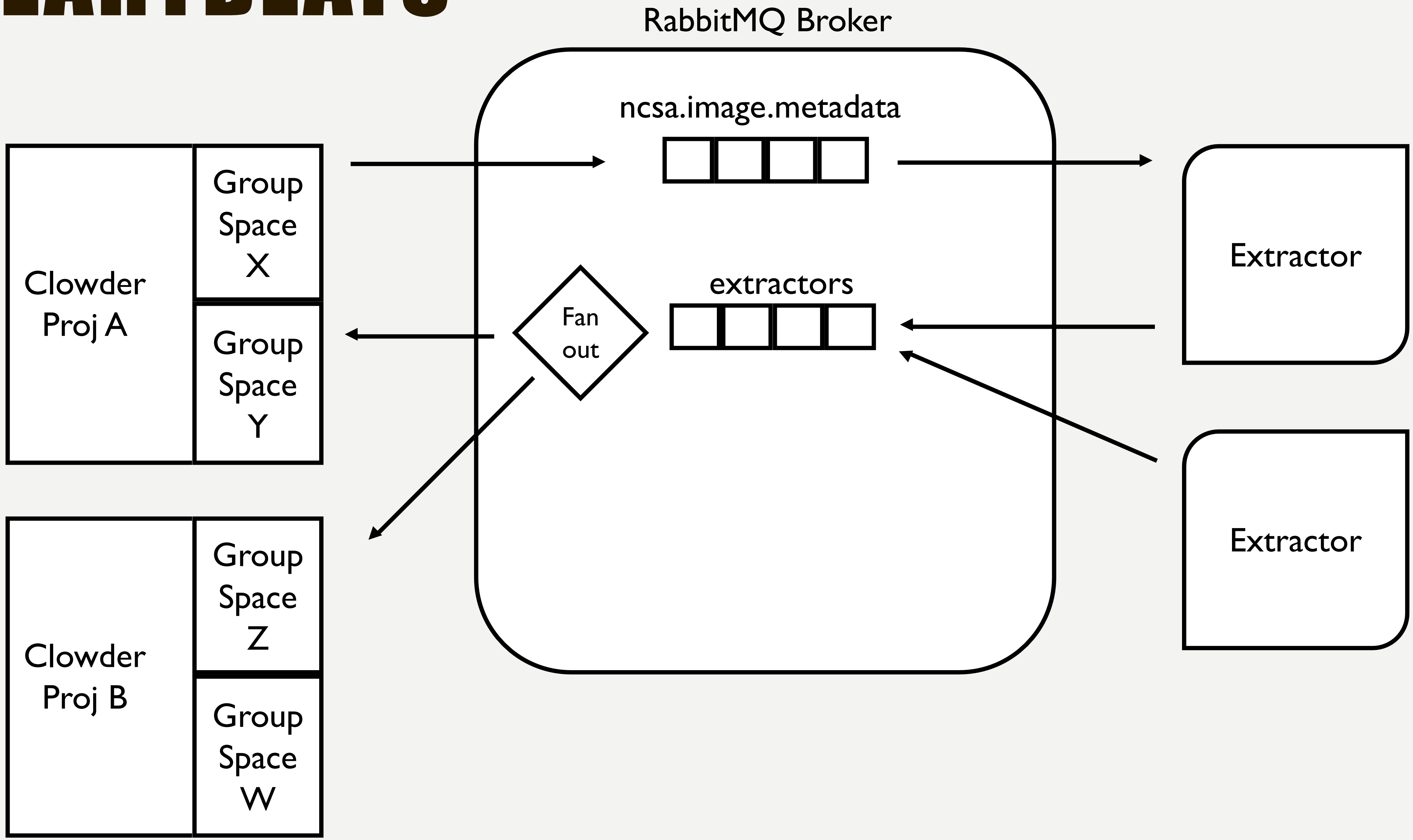
- image/tiff
- image/tif

EXTRACTOR MANIFEST NOW REQUIRED

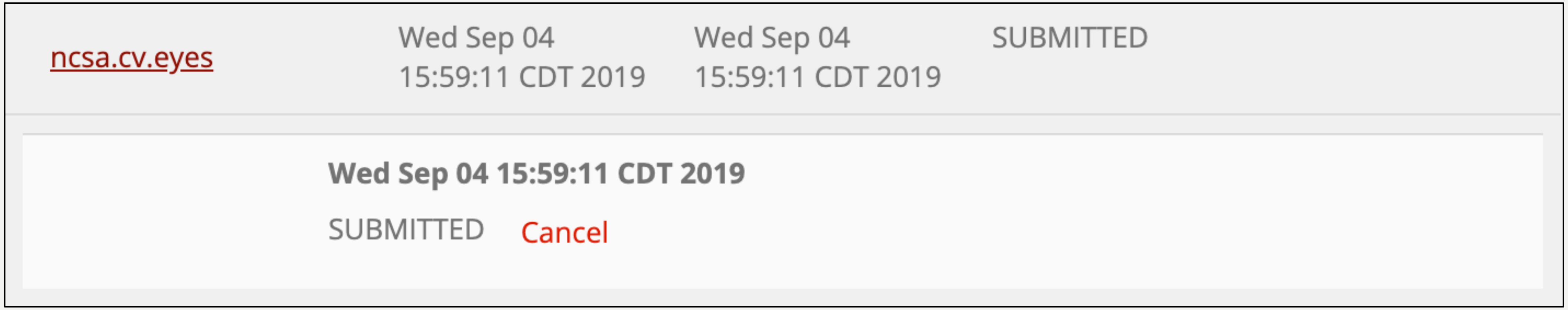


```
{
   "id": "58ec10ac4f0c272605dd2184",
   "name": "terra.plantcv",
   "version": "2.0",
   "updated": "Fri May 18 06:13:34 CDT 2018",
   "description": "Plant CV extractor",
   "author": "Burnette, Maxwell Amon <mburnet2@illinois.edu>",
 ▼ "contributors": [
      "Noah Fahlgren"
   ],
 ▼ "contexts": [
      "This extractor processes VIS/NIR images captured at several angles to generate trait metadata.
      The trait metadata is associated with the source images in Clowder, and uploaded to the
      configured BETYdb instance."
   ],
 ▼ "repository": [
    ▼ {
         "repType": "git",
         "repUrl": "https://github.com/terraref/extractors-lemnatec-indoor.git"
      },
    ▼ {
         "repType": "docker",
         "repUrl": "clowder/extractors-plantcv"
      }
   ],
   "external_services": [],
   "libraries": [],
 ▼ "bibtex": [
      "http://dx.doi.org/10.1016/j.molp.2015.06.005"
   ],
 ▼ "process": {
    ▼ "dataset": [
         "file.added"
      ],
      "file": [],
      "metadata": []
   }
},
```

HEARTBEATS


RabbitMQ Broker
ncsa.image.metadata
extractors
Fan out
Clowder Proj A
Group Space X
Group Space Y
Clowder Proj B
Group Space Z
Group Space W
Extractor
Extractor

CANCEL EXTRACTION



ncsa.cv.eyes	Wed Sep 04 15:59:11 CDT 2019	Wed Sep 04 15:59:11 CDT 2019	SUBMITTED

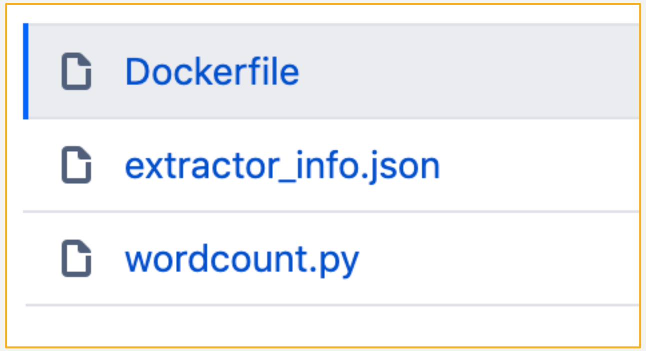
Wed Sep 04 15:59:11 CDT 2019

SUBMITTED Cancel

DEVELOPING EXTRACTORS

- PyClowder
  - https://opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/pyclowder2
  - A Python library to simplify the process
  - Modules to call API
  - HPC Extractor
  - Simple Extractor
- JClowder
  - A Java library to simplify the process
  - Experimental, few functions
- From Scratch
  - RabbitMQ client library
  - HTTP/JSON client libraries

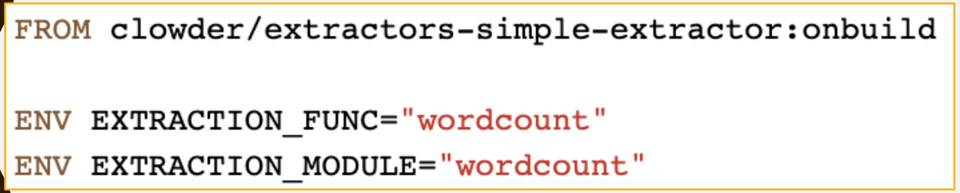
PYTHON SIMPLE EXTRACTOR



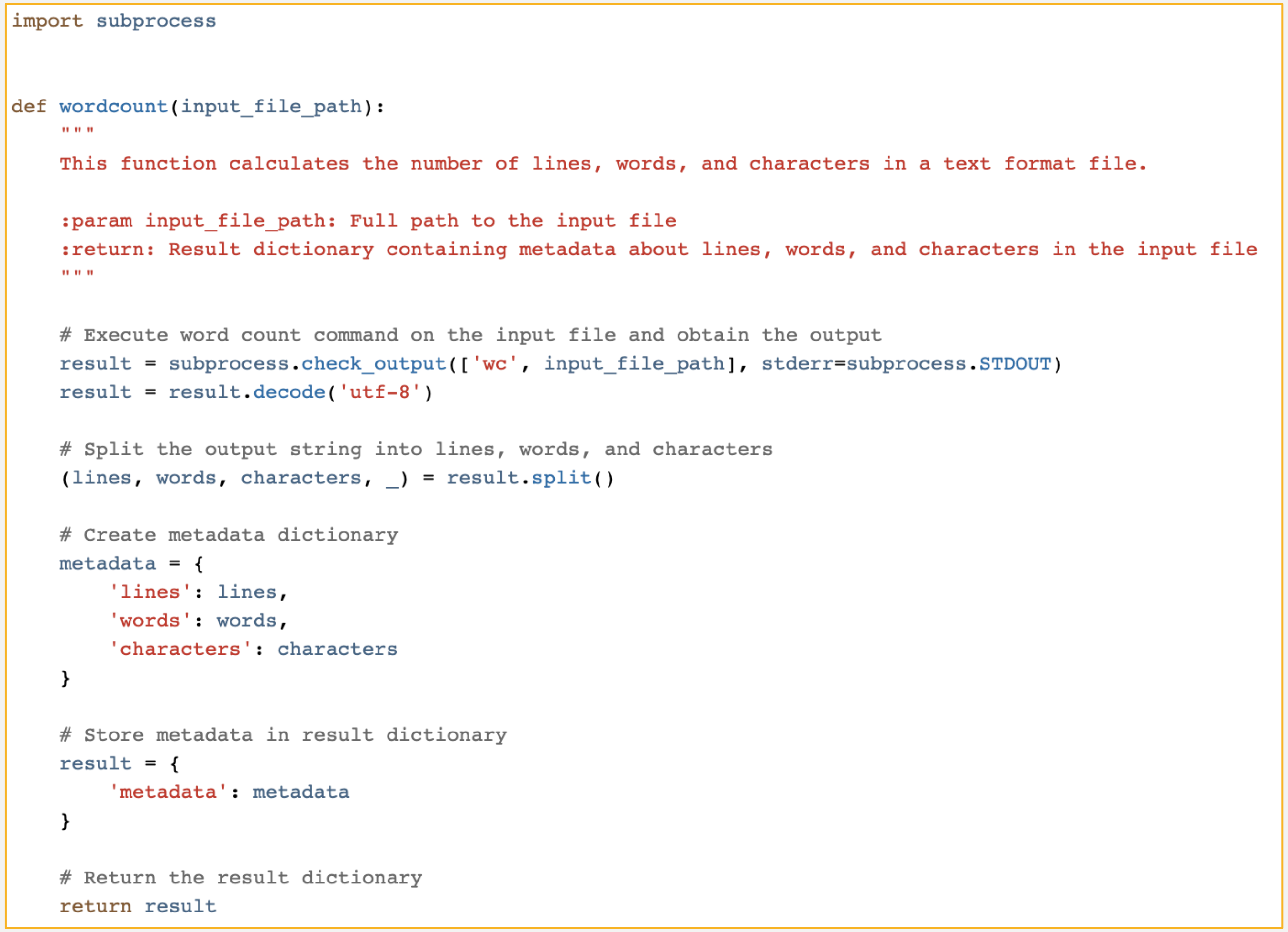
Dockerfile

extractor_info.json

wordcount.py

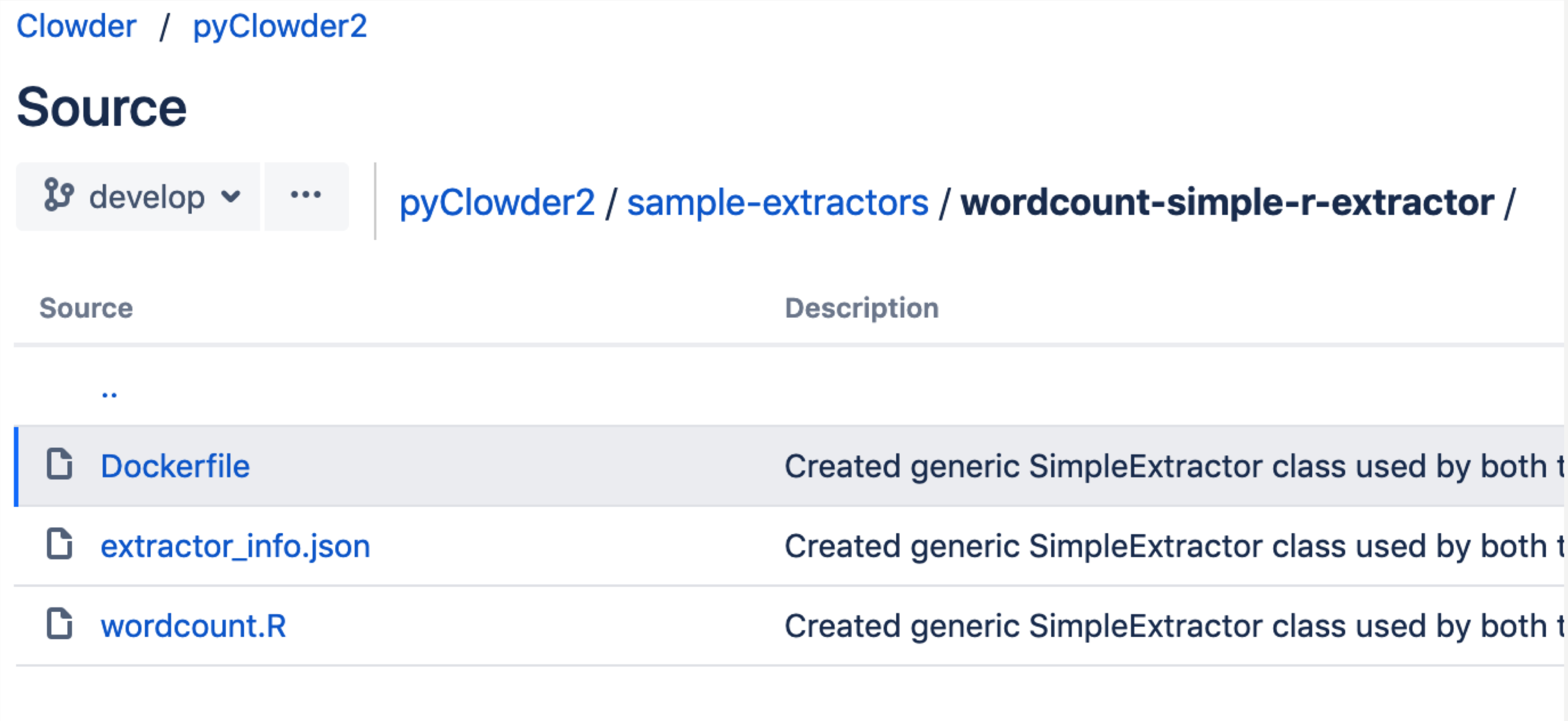


```
FROM clowder/extractors-simple-extractor:onbuild

ENV EXTRACTION_FUNC="wordcount"
ENV EXTRACTION_MODULE="wordcount"
```



```python
import subprocess


def wordcount(input_file_path):
    """
    This function calculates the number of lines, words, and characters in a text format file.

    :param input_file_path: Full path to the input file
    :return: Result dictionary containing metadata about lines, words, and characters in the input file
    """

    # Execute word count command on the input file and obtain the output
    result = subprocess.check_output(['wc', input_file_path], stderr=subprocess.STDOUT)
    result = result.decode('utf-8')

    # Split the output string into lines, words, and characters
    (lines, words, characters, _) = result.split()

    # Create metadata dictionary
    metadata = {
        'lines': lines,
        'words': words,
        'characters': characters
    }

    # Store metadata in result dictionary
    result = {
        'metadata': metadata
    }

    # Return the result dictionary
    return result
```

R SIMPLE EXTRACTOR



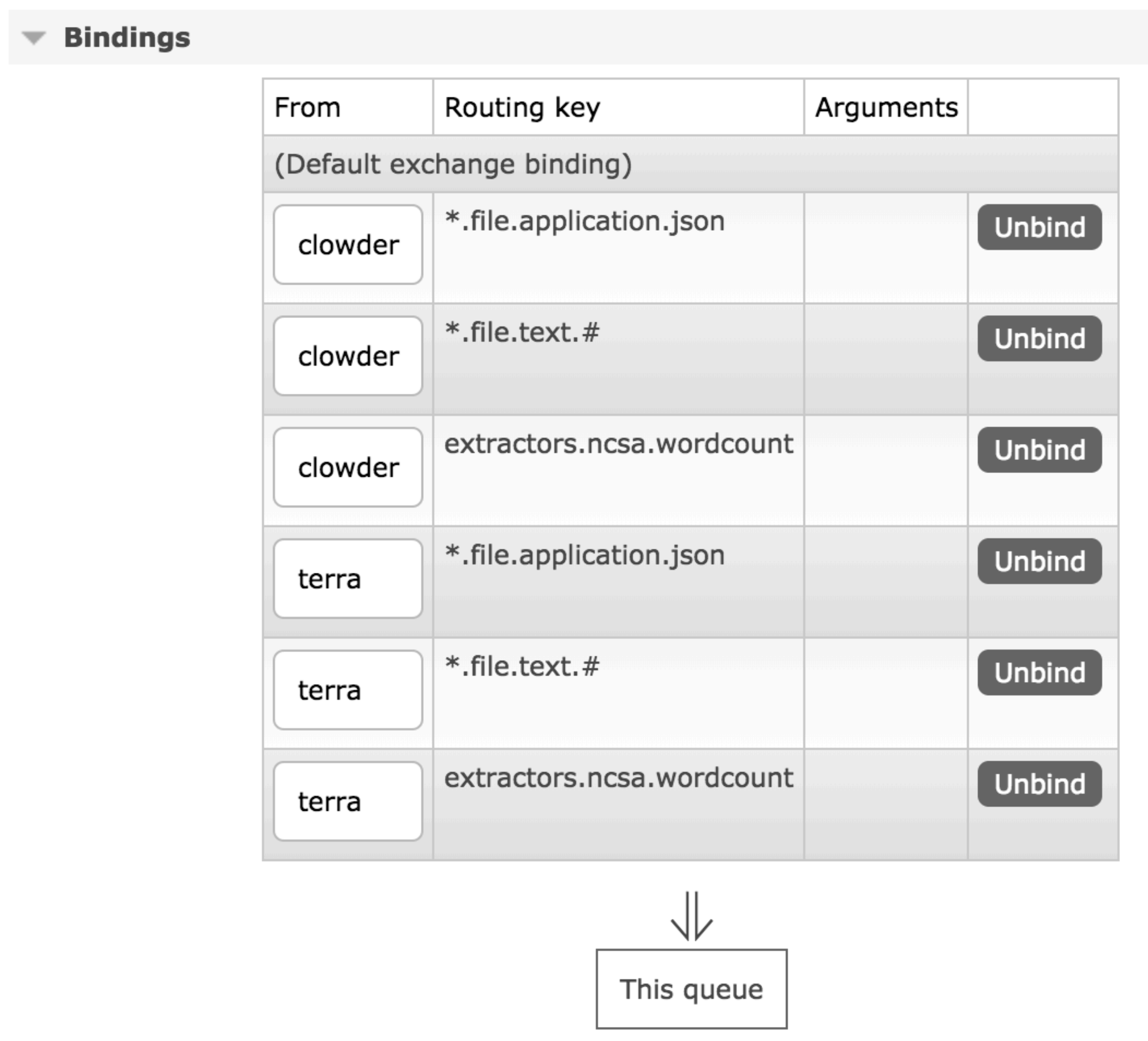
Clowder / pyClowder2

Source

pyClowder2 / sample-extractors / **wordcount-simple-r-extractor** /

Source	Description
..	
Dockerfile	Created generic SimpleExtractor class used by both t
extractor_info.json	Created generic SimpleExtractor class used by both t
wordcount.R	Created generic SimpleExtractor class used by both t

[OLD] RABBITMQ QUEUES



Bindings

From	Routing key	Arguments	
(Default exchange binding)			
clowder	*.file.application.json		Unbind
clowder	*.file.text.#		Unbind
clowder	extractors.ncsa.wordcount		Unbind
terra	*.file.application.json		Unbind
terra	*.file.text.#		Unbind
terra	extractors.ncsa.wordcount		Unbind

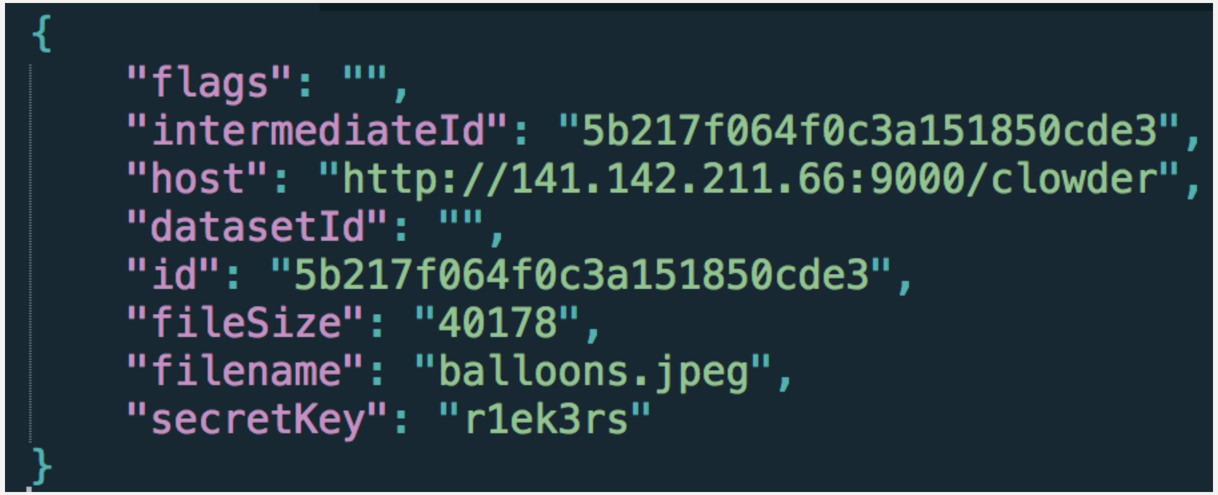
⇓

This queue

*.dataset.file.removed

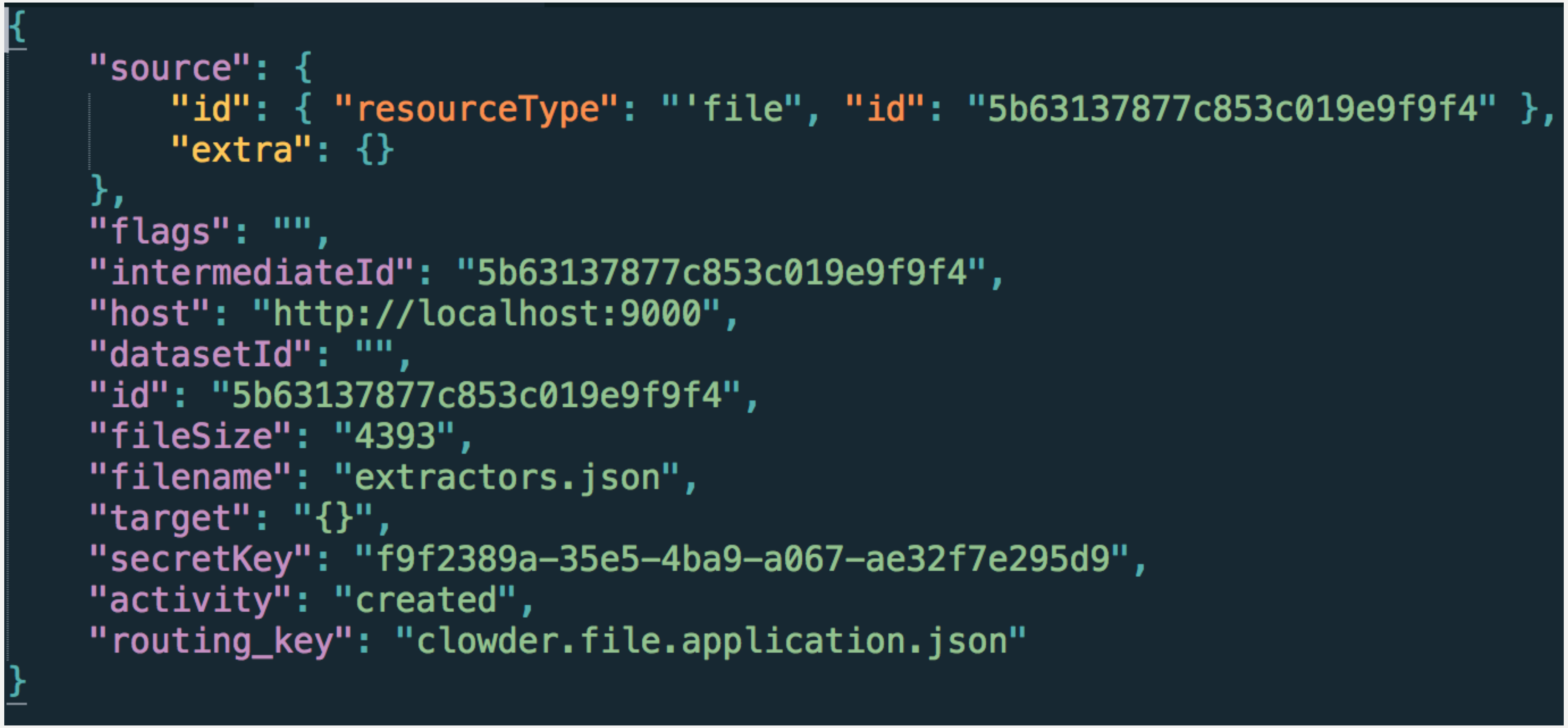
*.metadata.added

[OLD] RABBITMQ MESSAGE PAYLOAD



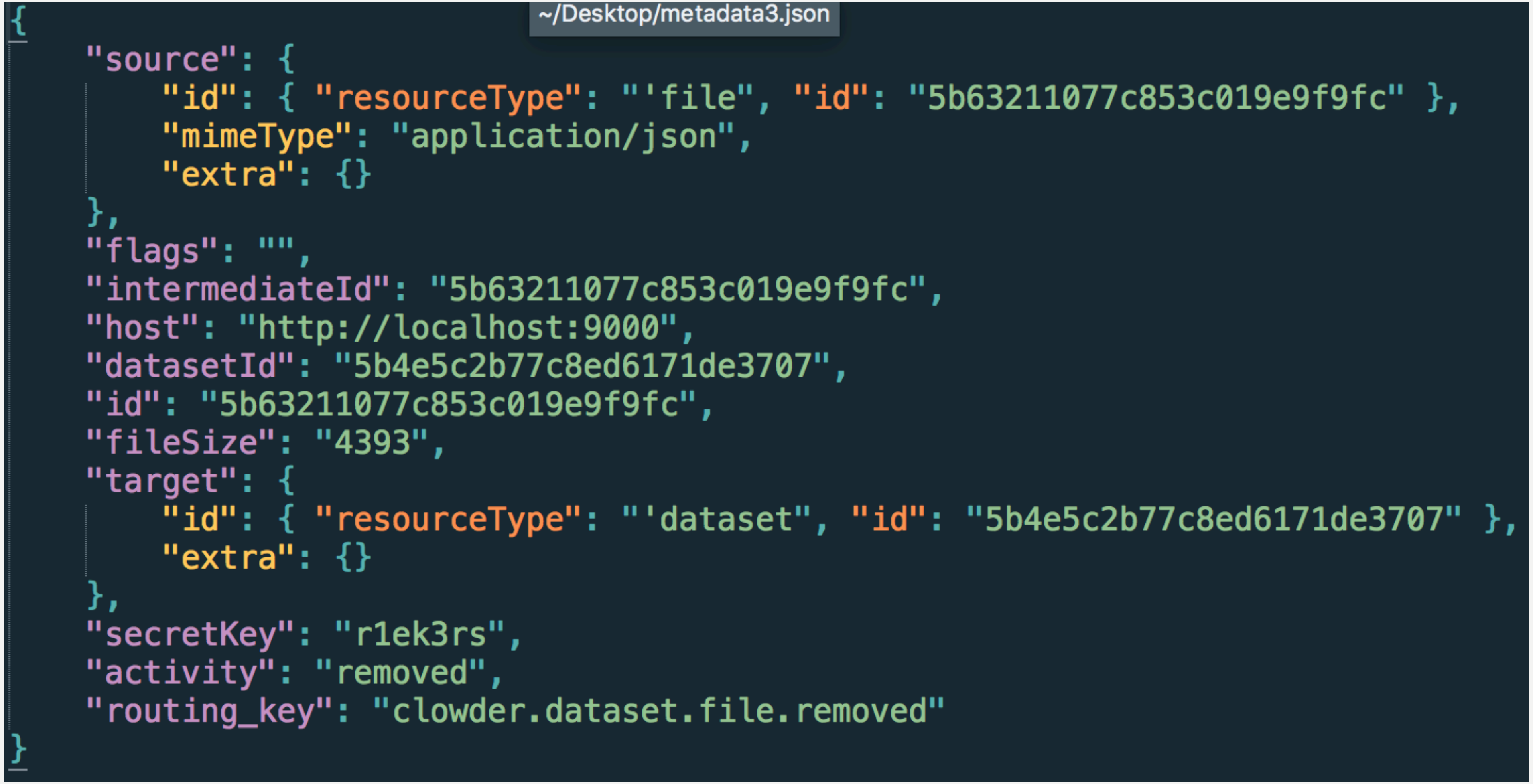
```
{
    "flags": "",
    "intermediateId": "5b217f064f0c3a151850cde3",
    "host": "http://141.142.211.66:9000/clowder",
    "datasetId": "",
    "id": "5b217f064f0c3a151850cde3",
    "fileSize": "40178",
    "filename": "balloons.jpeg",
    "secretKey": "r1ek3rs"
}
```

[NEW] RABBITMQ MESSAGE PAYLOAD



```
{
    "source": {
        "id": { "resourceType": "'file", "id": "5b63137877c853c019e9f9f4" },
        "extra": {}
    },
    "flags": "",
    "intermediateId": "5b63137877c853c019e9f9f4",
    "host": "http://localhost:9000",
    "datasetId": "",
    "id": "5b63137877c853c019e9f9f4",
    "fileSize": "4393",
    "filename": "extractors.json",
    "target": "{}",
    "secretKey": "f9f2389a-35e5-4ba9-a067-ae32f7e295d9",
    "activity": "created",
    "routing_key": "clowder.file.application.json"
}
```

[NEW] RABBITMQ MESSAGE PAYLOAD



```
{
    "source": {
        "id": { "resourceType": "'file", "id": "5b63211077c853c019e9f9fc" },
        "mimeType": "application/json",
        "extra": {}
    },
    "flags": "",
    "intermediateId": "5b63211077c853c019e9f9fc",
    "host": "http://localhost:9000",
    "datasetId": "5b4e5c2b77c8ed6171de3707",
    "id": "5b63211077c853c019e9f9fc",
    "fileSize": "4393",
    "target": {
        "id": { "resourceType": "'dataset", "id": "5b4e5c2b77c8ed6171de3707" },
        "extra": {}
    },
    "secretKey": "r1ek3rs",
    "activity": "removed",
    "routing_key": "clowder.dataset.file.removed"
}
```

PROCESSING

(AKA EXTRACTORS)

- File uploaded
- File added to dataset
- File remove from dataset
- Metadata added to file
- Metadata remove from file
- Metadata added to dataset
- Metadata removed from dataset
- File/Dataset manual submission to extractor
- File Batch uploaded

“NOT EXTRACTORS”

- EXTRACT - traditional extractor, typically adds metadata or derived outputs to the triggering file/dataset; default
- *PREVIEW* - create preview for to show on Web UI
- CONVERT - primary function is to convert file(s) from source format to another format, combine files, etc.
- ARCHIVE - eligible for triggering using Archive/Unarchive buttons, should expect one of those two parameters
- PUBLISH - intended to publish files or datasets to external repositories
- WORKFLOW - primarily manages workflows, submits external jobs, triggers other extractors, e.g. extractors-rulechecker
- SILENT - if in this category, extractor will not send common status messages (e.g. STARTED)

QUESTIONS?

NEXT WEBINAR (OCTOBER)

- Phone App (Flutter) (Todd Nicholson)
- ???

HTTP://CLOWDERFRAMEWORK.ORG/