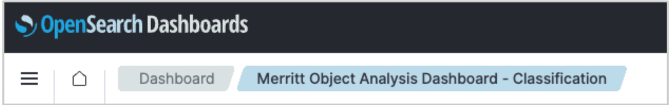
University of California
CDL
California Digital Library

UNIVERSITY
OF
CALIFORNIA

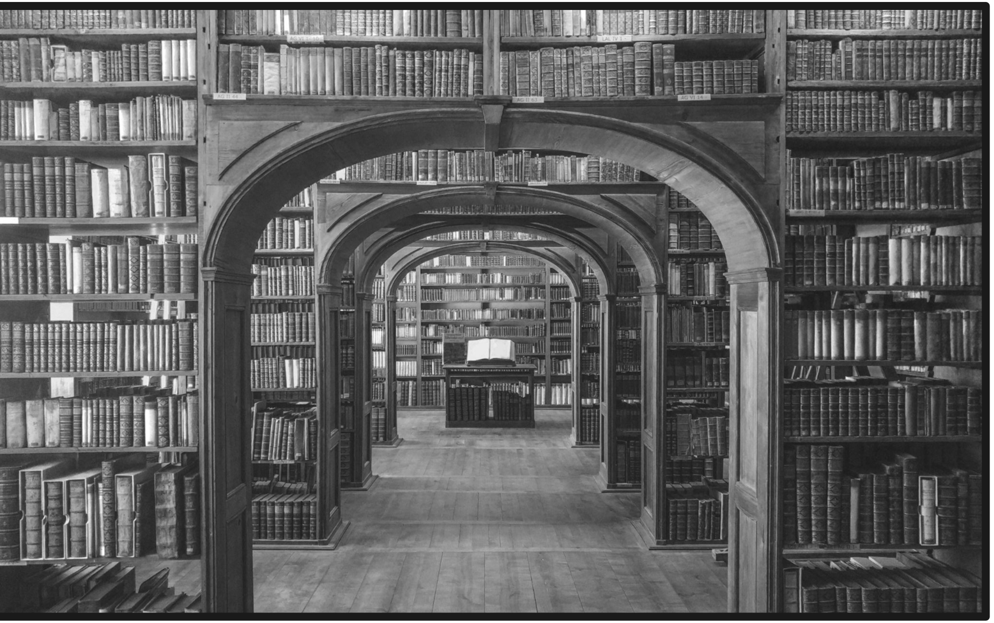
Promoting Visibility into Collections through Object Analysis

Leveraging Amazon OpenSearch



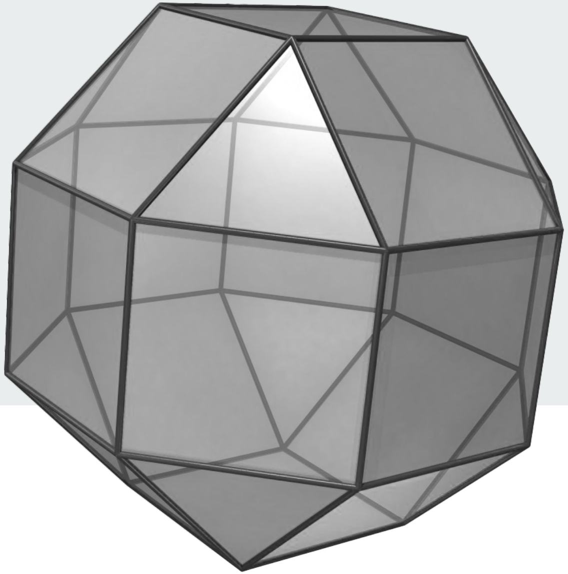
Library of Congress Designing Storage Architectures Meeting, 2024

Eric Lopatin, Digital Preservation Service Manager
Terry Brady, Sr. Developer & Technical Lead



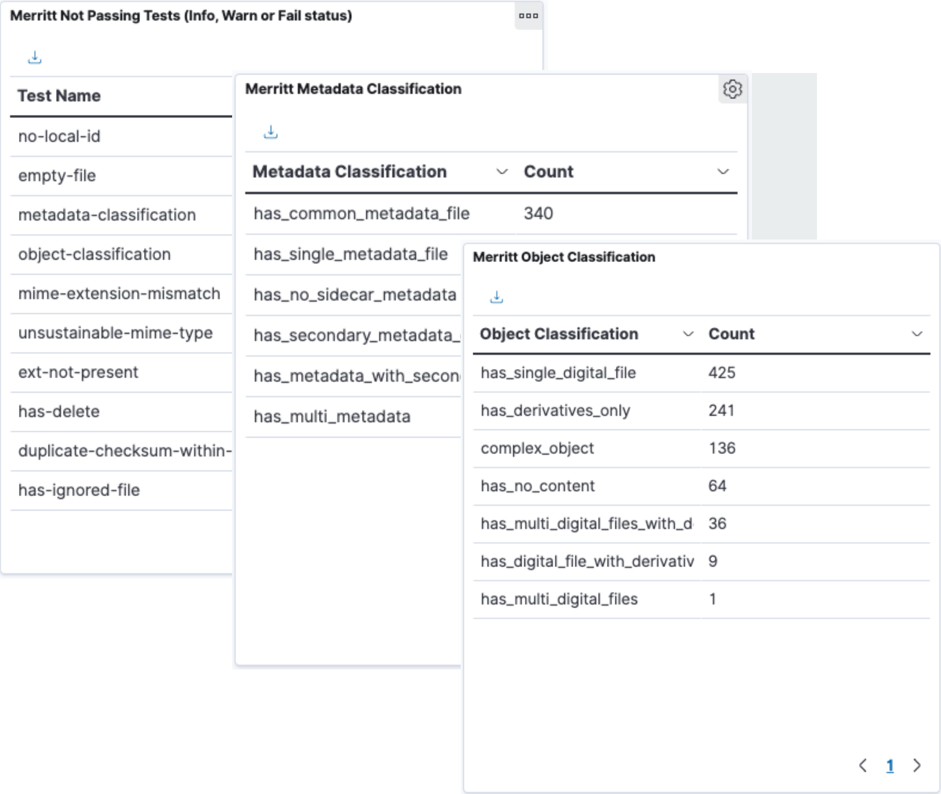
And though most were extensively familiar with the file formats and metadata of that content as it was deposited, *the overarching goal of preserving the resulting collections will, in a chronological sense, stretch* ***beyond any one group of individuals.*** Digital objects are often in an ideal state as they begin their lives in a repository, but the opinions, policies and communities sharing the work of digital preservation will inevitably effect change in their state.

One of our goals is to provide the tools that facilitate the application of content-specific requirements established by our depositors and end users. *As community practices surrounding these requirements change,* ***we considered the promotion of visibility into collections at any time to be a key operation.*** There's a need to re-examine content, so we set out to provide the means to do this, **at scale.**



We defined a limited set of **common object characteristics to analyze,** including object-level metadata, object complexity in terms of the purposes of each file, file formats and format sustainability. Criteria were defined regarding metadata.

We considered the origin and purpose of individual metadata files and their extent, as well as file naming conventions employed across collections.



Merritt Not Passing Tests (Info, Warn or Fail status)

Test Name
no-local-id
empty-file
metadata-classification
object-classification
mime-extension-mismatch
unsustainable-mime-type
ext-not-present
has-delete
duplicate-checksum-within-
has-ignored-file

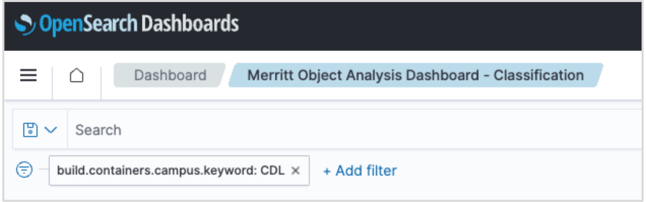
Merritt Metadata Classification

Metadata Classification	Count
has_common_metadata_file	340
has_single_metadata_file	
has_no_sidecar_metadata	
has_secondary_metadata_	
has_metadata_with_secon	
has_multi_metadata	

Merritt Object Classification

Object Classification	Count
has_single_digital_file	425
has_derivatives_only	241
complex_object	136
has_no_content	64
has_multi_digital_files_with_d	36
has_digital_file_with_derivativ	9
has_multi_digital_files	1

Given the AWS underpinnings of our repository infrastructure, we identified Amazon OpenSearch as a possible solution. OpenSearch, allowed us to create **a rich map of relationships across elements of data.** And it allowed us to filter for and **visualize** these while applying categories and bubbling up the results of object analysis.



OpenSearch Dashboards

Dashboard / Merritt Object Analysis Dashboard - Classification

Search

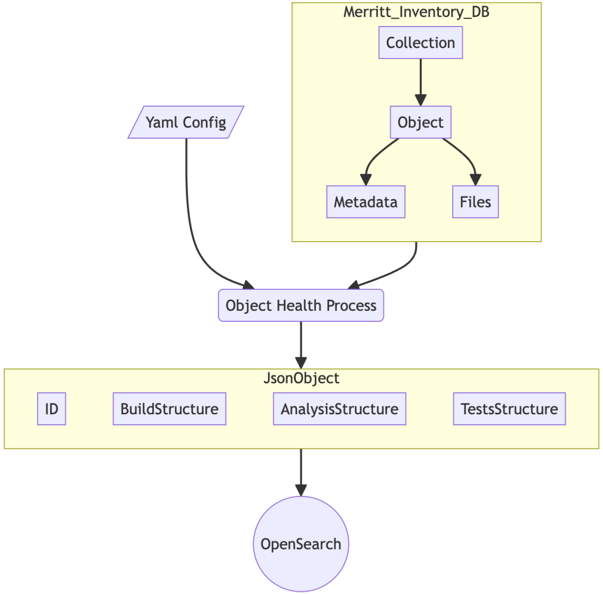
build.containers.campus.keyword: CDL ×

+ Add filter

merritt.cdlib.org/presentations

What are the key components of the system?

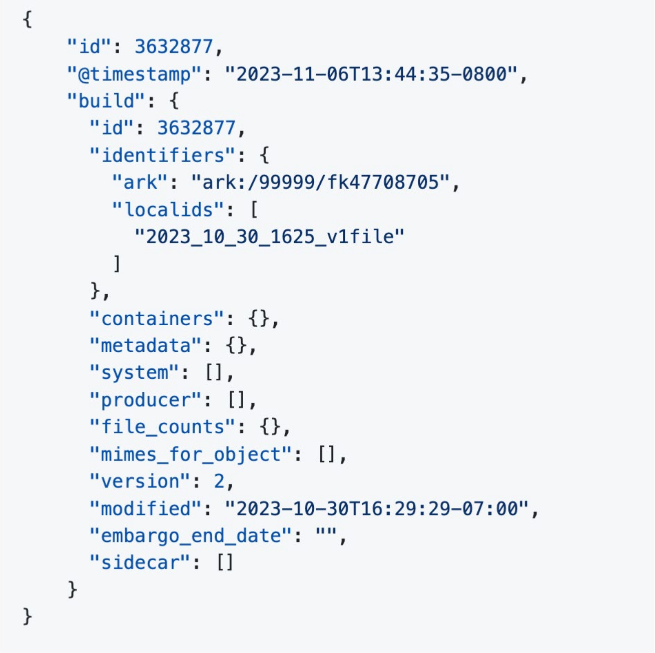
- Inventory database containing object and file data
- Analysis configuration file (yaml)
- Process: Object Build, Object Analysis, Object Test
- JSON schema designed to work with OpenSearch filters
- OpenSearch Dashboards for visualization



Build

The Build process is intended to extract and assemble known information about an object.

- Identifiers
- Metadata
- Digital files
- Ownership/collection taxonomies



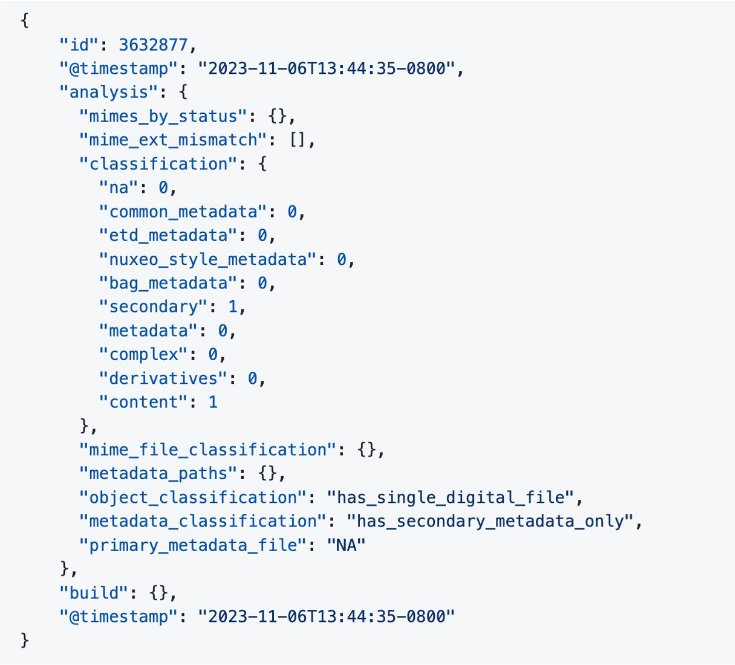
```
{
    "id": 3632877,
    "@timestamp": "2023-11-06T13:44:35-0800",
    "build": {
      "id": 3632877,
      "identifiers": {
        "ark": "ark:/99999/fk47708705",
        "localids": [
          "2023_10_30_1625_v1file"
        ]
      },
      "containers": {},
      "metadata": {},
      "system": [],
      "producer": [],
      "file_counts": {},
      "mimes_for_object": [],
      "version": 2,
      "modified": "2023-10-30T16:29:29-07:00",
      "embargo_end_date": "",
      "sidecar": []
    }
}
```

Analysis

The Analysis process is driven by a set of Tasks defined in the project's yaml config file.

An Analysis Task analyzes Build information and creates new JSON structures that may be queried by one or more Tests.

- *Categorization*
- *Relationships*

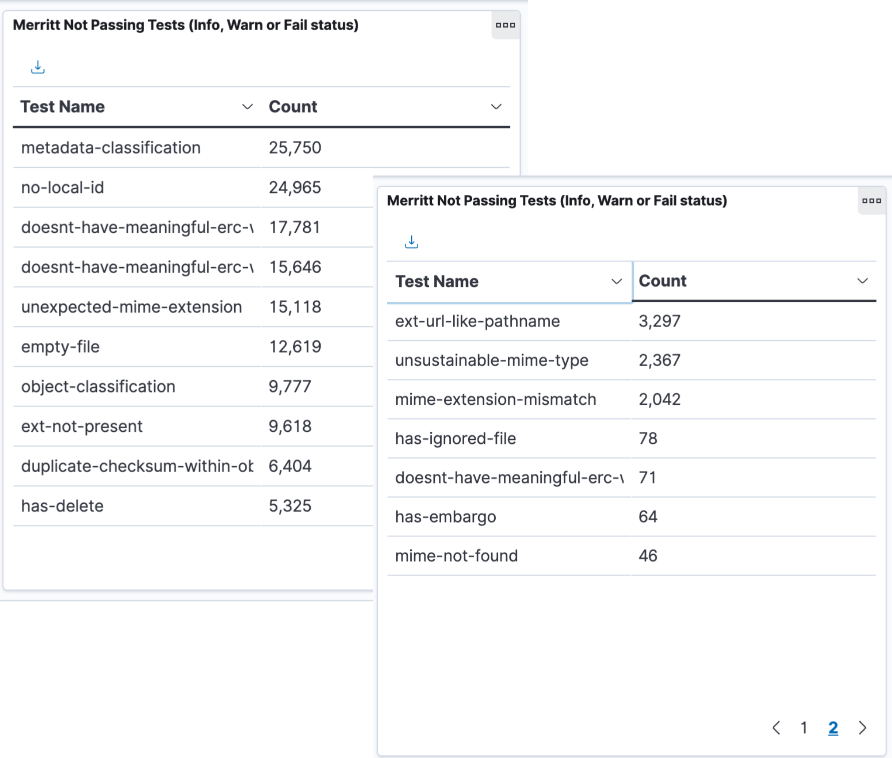


```
{
    "id": 3632877,
    "@timestamp": "2023-11-06T13:44:35-0800",
    "analysis": {
      "mimes_by_status": {},
      "mime_ext_mismatch": [],
      "classification": {
        "na": 0,
        "common_metadata": 0,
        "etd_metadata": 0,
        "nuxeo_style_metadata": 0,
        "bag_metadata": 0,
        "secondary": 1,
        "metadata": 0,
        "complex": 0,
        "derivatives": 0,
        "content": 1
      },
      "mime_file_classification": {},
      "metadata_paths": {},
      "object_classification": "has_single_digital_file",
      "metadata_classification": "has_secondary_metadata_only",
      "primary_metadata_file": "NA"
    },
    "build": {},
    "@timestamp": "2023-11-06T13:44:35-0800"
}
```

Tests

Candidate Tests are enumerated in an easily editable yaml-based schema which defines conditions for test results.

Each test can be enabled or disabled for specific collections or taxonomy nodes.



Merritt Not Passing Tests (Info, Warn or Fail status)

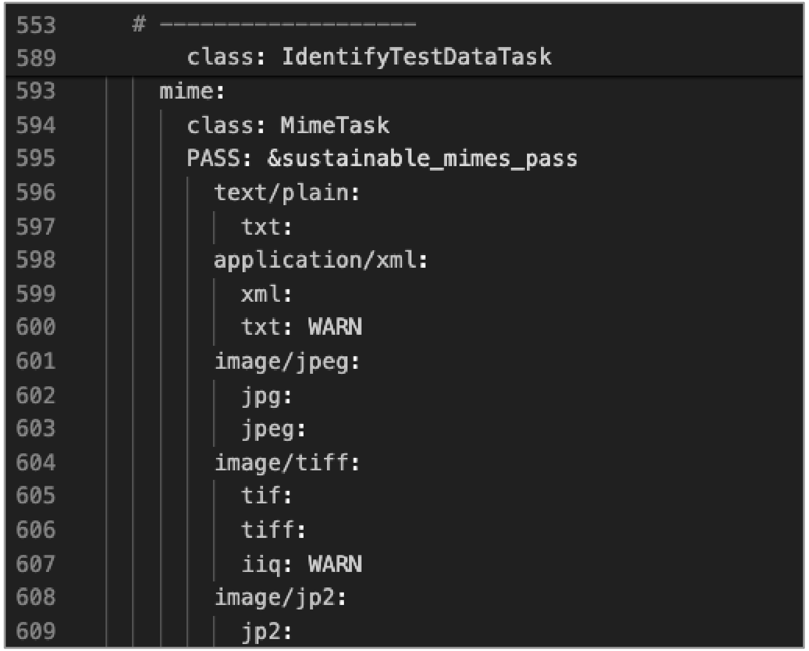
Test Name	Count
metadata-classification	25,750
no-local-id	24,965
doesnt-have-meaningful-erc-\	17,781
doesnt-have-meaningful-erc-\	15,646
unexpected-mime-extension	15,118
empty-file	12,619
object-classification	9,777
ext-not-present	9,618
duplicate-checksum-within-ob	6,404
has-delete	5,325

Merritt Not Passing Tests (Info, Warn or Fail status)

Test Name	Count
ext-url-like-pathname	3,297
unsustainable-mime-type	2,367
mime-extension-mismatch	2,042
has-ignored-file	78
doesnt-have-meaningful-erc-\	71
has-embargo	64
mime-not-found	46

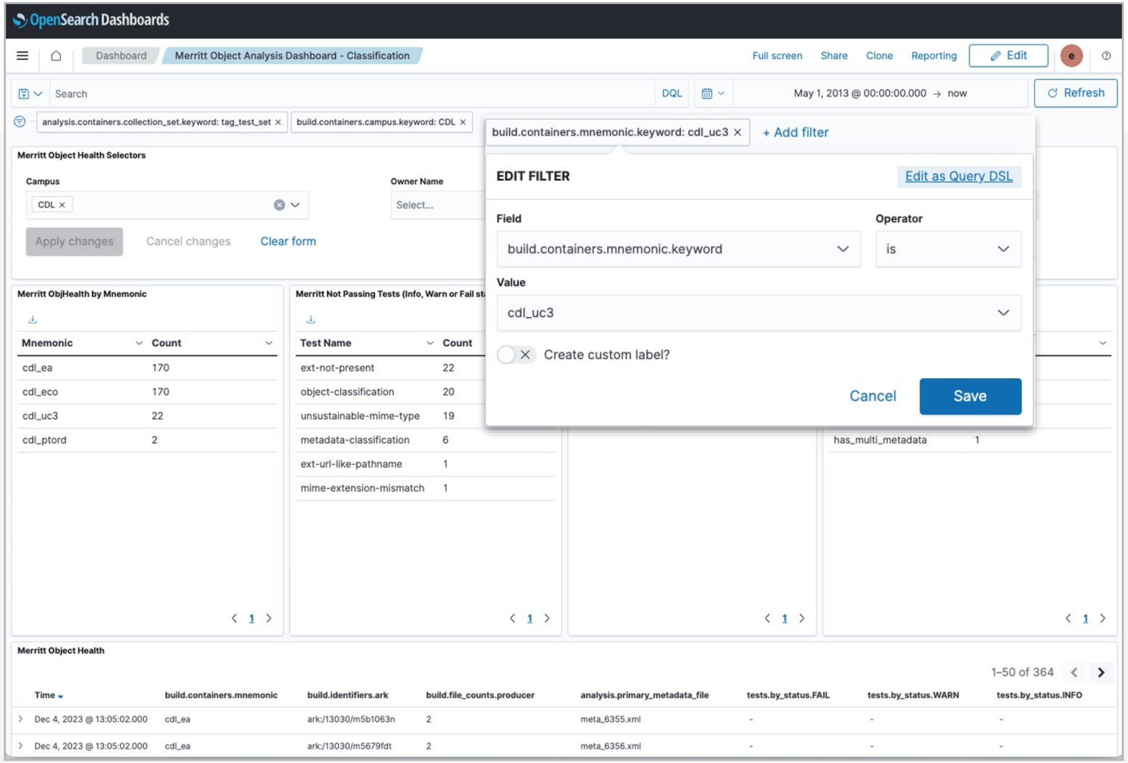
Customization

Customization is provided through the organization of file types in the yaml schema, such that all types may be assigned the desired test outcome status.



```
  # -------------------
      class: IdentifyTestDataTask
    mime:
      class: MimeTask
      PASS: &sustainable_mimes_pass
        text/plain:
          txt:
        application/xml:
          xml:
          txt: WARN
        image/jpeg:
          jpg:
          jpeg:
        image/tiff:
          tif:
          tiff:
          iiq: WARN
        image/jp2:
          jp2:
```

Conclusion



Merritt Object Analysis Dashboard - Classification

May 1, 2013 @ 00:00:00.000 → now

analysis.containers.collection_set.keyword: tag_test_set ×
build.containers.campus.keyword: CDL ×
build.containers.mnemonic.keyword: cdl_uc3 ×

EDIT FILTER

Field: build.containers.mnemonic.keyword

Operator: is

Value: cdl_uc3

○ Create custom label?

Merritt Object Health Selectors

Campus: CDL ×

Owner Name: Select...

Merritt ObjHealth by Mnemonic

Mnemonic	Count
cdl_ea	170
cdl_eco	170
cdl_uc3	22
cdl_ptord	2

Merritt Not Passing Tests (Info, Warn or Fail st

Test Name	Count
ext-not-present	22
object-classification	20
unsustainable-mime-type	19
metadata-classification	6
ext-url-like-pathname	1
mime-extension-mismatch	1

has_multi_metadata 1

Merritt Object Health

1–50 of 364

Time	build.containers.mnemonic	build.identifiers.ark	build.file_counts.producer	analysis.primary_metadata_file	tests.by_status.FAIL	tests.by_status.WARN	tests.by_status.INFO
Dec 4, 2023 @ 13:05:02.000	cdl_ea	ark:/13030/m5b1063n	2	meta_6355.xml	-	-	-
Dec 4, 2023 @ 13:05:02.000	cdl_ea	ark:/13030/m5679fdt	2	meta_6356.xml	-	-	-

merritt.cdlib.org/presentations

Are you interested? Have you tried something similar?

merritt.cdlib.org/presentations

Eric Lopatin
eric dot lopatin at ucop dot edu

Terrence Brady
terrence dot brady at ucop dot edu

University of California Curation Center – UC3

University of California
CDL
California Digital Library