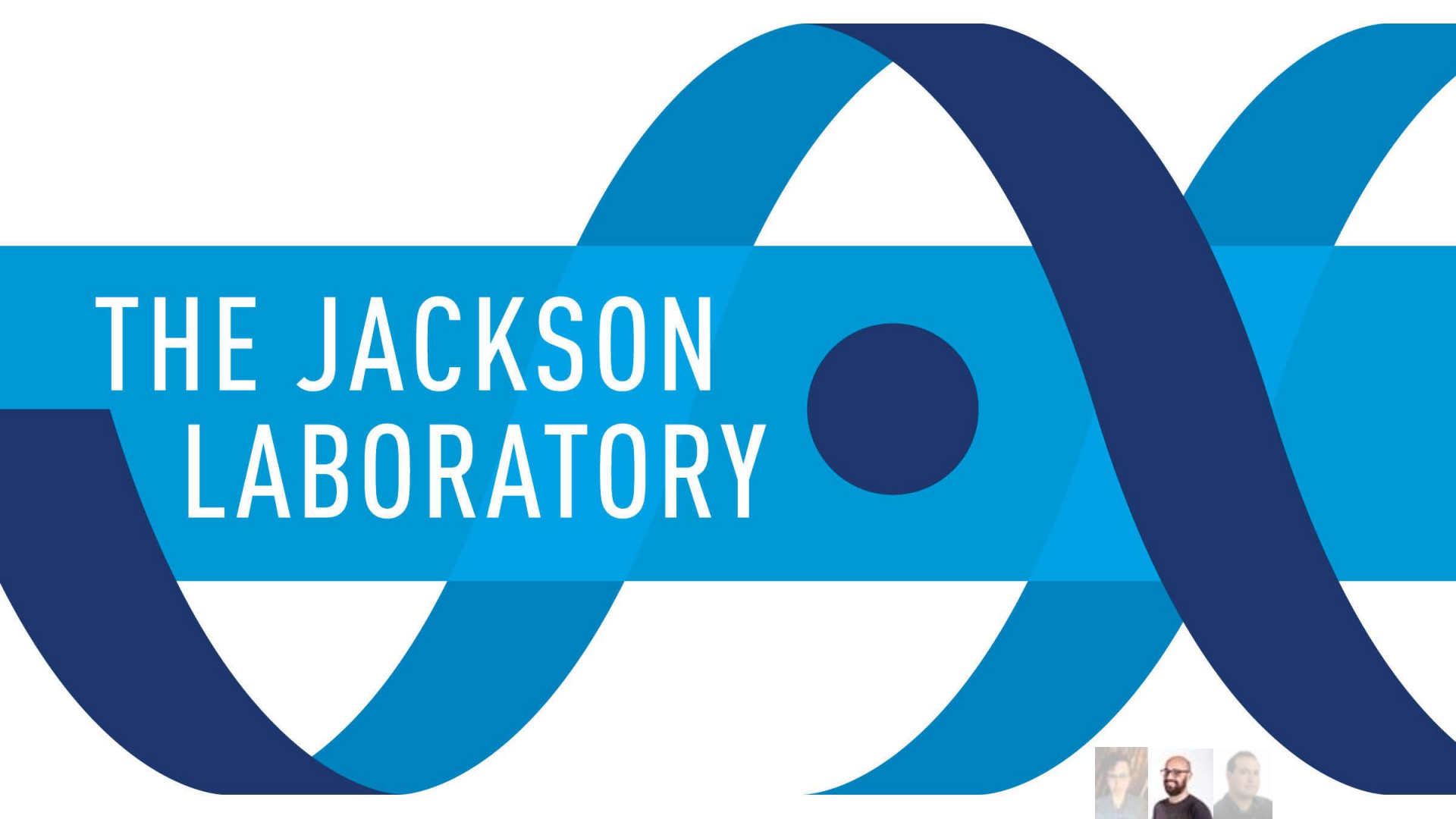


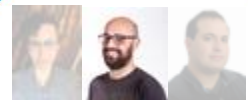
Making JAX data FAIR: the pixel access problem

Erick Ratamero, Kiya Govek, Eric Perlman

OME Community Meeting
May 29, 2024



THE JACKSON LABORATORY



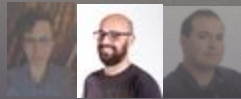


THREE LOCATIONS

BAR HARBOR, MAINE

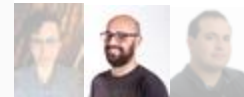
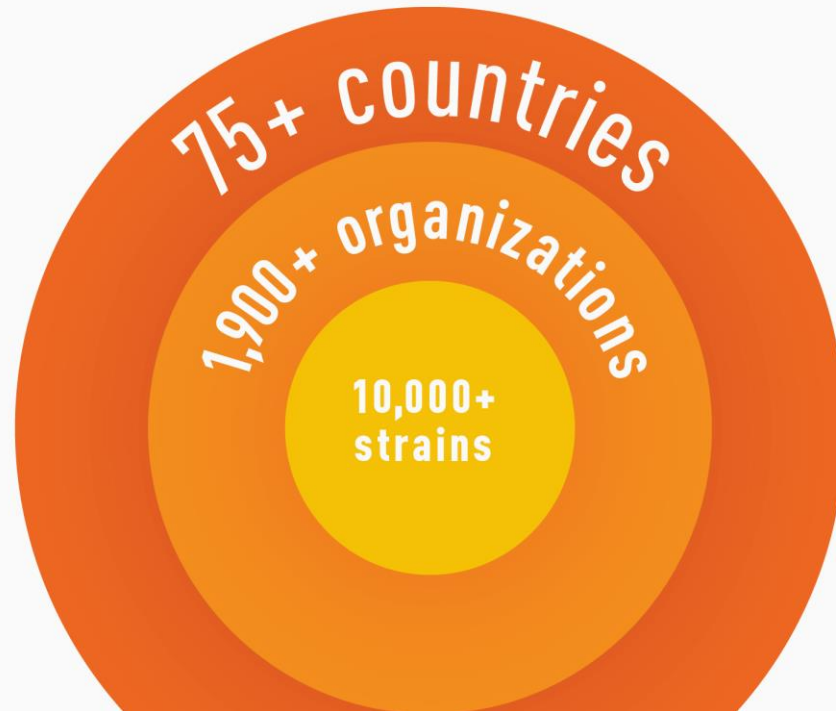
FARMINGTON, CONN.

SACRAMENTO, CALIF.



We distribute mice and provide services to organizations around the globe

JAX[®] MICE
& CLINICAL
RESEARCH
SERVICES



Motivation: mouse phenotype data

Home > Genes > A1cf > Image comparator

WT Images

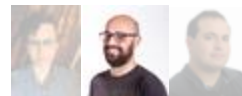
© 2007-2022 Glenoe Software Inc. All rights reserved.

Mutant Images

© 2007-2022 Glenoe Software Inc. All rights reserved.



www.mousephenotype.org



THE JACKSON LABORATORY

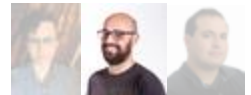
5

Viewing imaging data: embedded OMERO!

Home > Genes > A1cf > Image comparator

iframe#control_frame | 500 x 400

www.mousephenotype.org



What is images.jax.org?

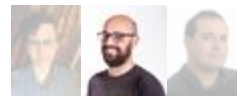
The screenshot displays the 'Public' section of the images.jax.org website. The left sidebar shows a file tree under 'All Members' with a folder 'Anapc7 JR34803 150' containing a list of files. The main area shows a grid of image thumbnails. The right sidebar contains metadata for the selected image 'Anapc7_A_5542_OV_01.ndpi [0]'. The metadata includes:

- Image ID: 187859
- Owner: Data Importer
- Image Details: 000231
- Acquisition Date: 2022-12-13 03:21:55
- Import Date: 2023-04-05 17:33:50
- Dimensions (XY): 199680 x 90112
- Pixels Type: uint8
- Pixels Size (XYZ) (µm): 0.22 x 0.22 x -
- Z-sections/Timepoints: 1 x 1
- Channels: 0, 1, 2
- ROI Count: 30

Additional sections include 'Tags' (0), 'Key-Value Pairs' (3), and a table of key-value pairs:

jax.org/omeroutils/user_submitted/v0	
Added by: Data Importer	
species	Mouse
Strain	C57BL/6NJ-Anapc7- <i>em1</i> (MPC) Jw/Mmjax
genotype	Homozygous
antibodies/stains	PAS
Tissue	Ovaries
Tissue 2	Oviducts
DOB	2022-08-28 00:00:00
Harvest date	2022-11-02 00:00:00
Age (days)	67
JR	34083
Mouse ID	A-5542

Other sections include 'Tables' (0), 'Attachments' (1), and 'Comments' (0). The attachment 'Figure_2023-1-23_9-4-51.tiff (33.17 MB)' is visible.



Purpose of two OMERO instances



- Publishing at images.jax.org
- Data hosting
- NGFF

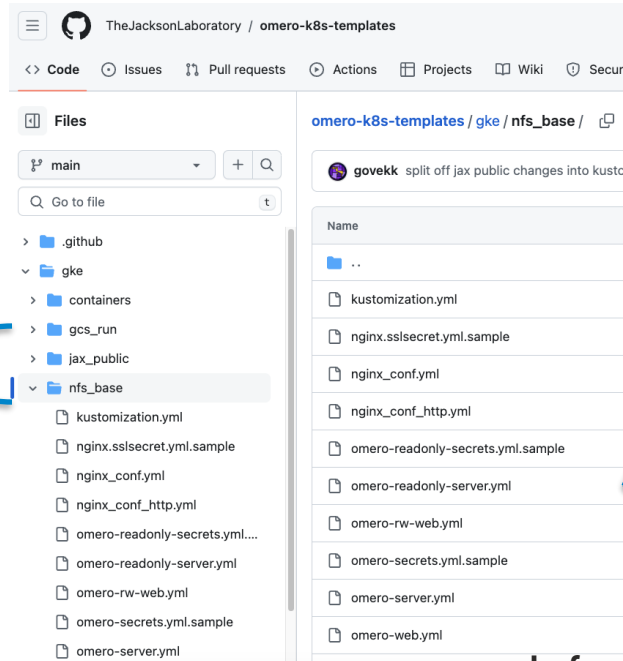


- Data ingestion
- Collaboration within JAX
- Annotating images

Kubernetes: Infrastructure as Code



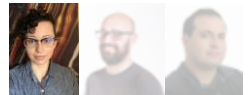
Kustomize
templates
for different
OMERO
instances



```
36 apiVersion: apps/v1
37 kind: Deployment
38 metadata:
39   name: omero-readonly-server
40   labels:
41     app: omero-readonly-server
42 spec:
43   replicas: 1
44   selector:
45     matchLabels:
46       app: omero-readonly-server
47   template:
48     metadata:
49       labels:
50         app: omero-readonly-server
51   spec:
52     automountServiceAccountToken: false
53     nodeSelector:
54       cloud.google.com/gke-nodepool: <POOL_NAME>
55     containers:
56     - name: omero-readonly-server
57       image: <CONTAINER_IMAGE>
58       imagePullPolicy: Always
59     env:
60     - name: CONFIG_omero_db_name
61       value: <DB NAME>
62     - name: CONFIG_omero_db_host
63       value: <CLOUD SQL IP>
64     - name: CONFIG_omero_data_dir
65       value: /OMERO # this is the default
```

Kubernetes
automates
deployment
and recovery

Infrastructure configuration
is recorded in yaml files



Configuration differences

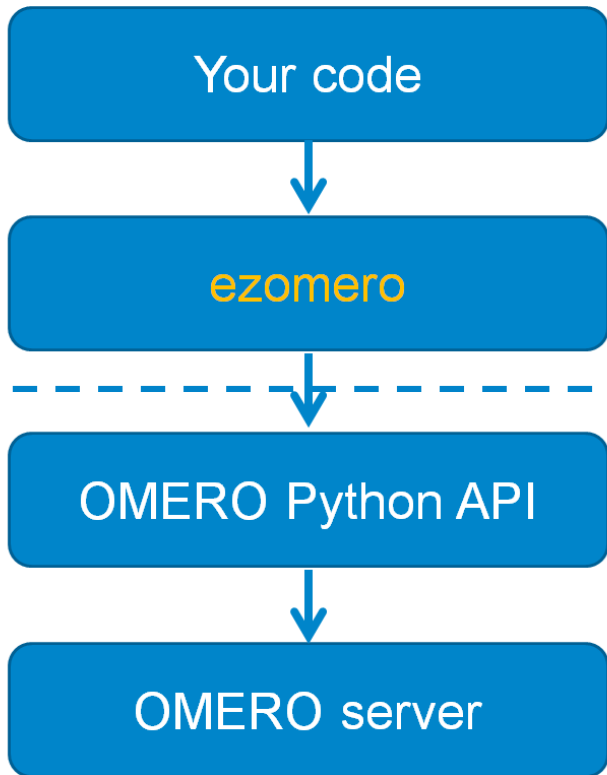


- Image data files on **object storage**
- **Read-only** access to data and OMERO system files



- Allows connections to server for **Python API**, QuPath, Fiji, OMERO CLI, etc

ezomero: making OMERO easier

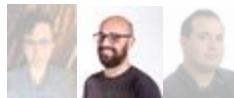


```
1 import ezomero
2 conn = ezomero.connect(USERNAME, PWD, host=HOSTNAME, port=PORT, secure=True)
3 ds_id = ezomero.post_dataset(conn, "New Dataset", project_id=projectId)
```

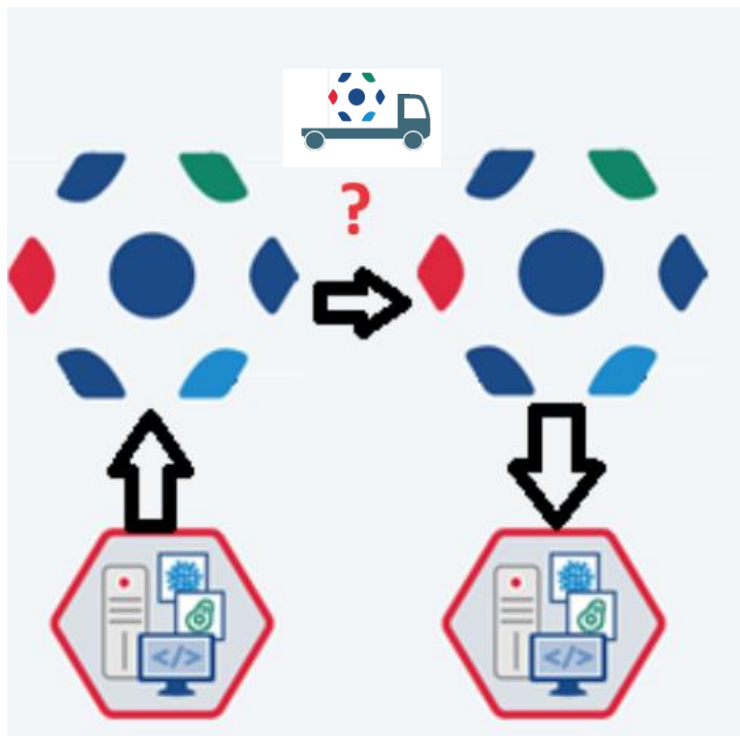
You focus on
this part...

...we take care
of this one!

```
1 import omero
2 conn = omero.gateway.BlitzGateway(USERNAME, PWD, host=HOSTNAME, port=PORT, secure=True)
3 conn.connect()
4 dataset_obj = omero.model.DatasetI()
5 dataset_obj.setName(rstring("New Dataset"))
6 dataset_obj = conn.getUpdateService().saveAndReturnObject(dataset_obj, conn.SERVICE_OPTS)
7 dataset_id = dataset_obj.getId().getValue()
8 link = omero.model.ProjectDatasetLinkI()
9 link.setChild(omero.model.DatasetI(dataset_obj.id.val, False))
10 link.setParent(omero.model.ProjectI(projectId, False))
11 conn.getUpdateService().saveObject(link)
```



omero-cli-transfer

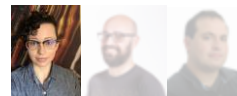


```
omero transfer pack Image:123 transfer_pack.zip  
omero transfer pack Dataset:1111 /home/user/new_folder/new_pack.zip  
omero transfer pack 999 zipfile.zip # equivalent to Project:999
```

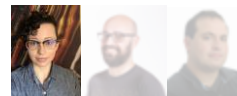
```
omero transfer unpack transfer_pack.zip  
omero transfer unpack --output /home/user/optional_folder --ln_s
```



Data lifecycle: from microscope to public

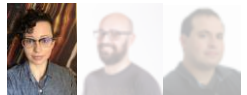


Data lifecycle: from microscope to public



Tackling access to pixels

- Current options through OMERO aren't great
 - OMERO provides download links which often fail
 - OMERO APIs are indirect ways of accessing pixel data and are slow
- We need a way to find and directly access pixels from cloud object storage



What are we doing *today*?

- We are embracing OME-NGFF as an alternative method for accessing image data
- All image data on images.jax.org is now available in Zarr (v2) with NGFF v0.4 metadata
- We will update all data with each major revision to the NGFF spec



How to find http NGFF URLs?

- Navigate to an image and expand "Key-Value Pairs"
- Details under the `jax.org/public/zarr` namespace
- Functional but rough

`jax.org/public/zarr`

Added by: Data Importer

`zarr_path`

`https://storage.googleapis.com/jax-public-ngff/data/0.4/public_data/3814/kjp_513/2023-04/03/12-50-52.057/Defb20_A_5622_Con_Test_02.zarr/0/`

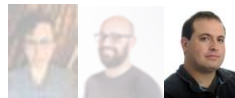
`ngff_version`

0.4

`zarr_conversion_date`

2023-12-10T05:53:50.737001

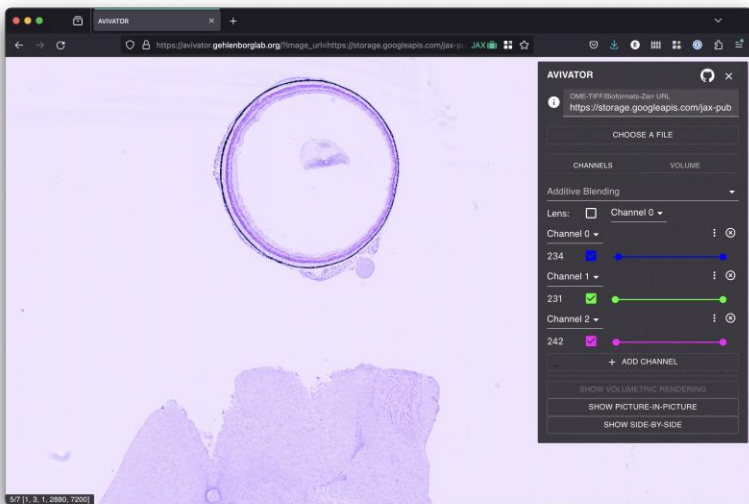
The screenshot shows the JAX Browser interface. The 'Public' namespace is selected, and a list of files is displayed. The file 'Defb20_A_5622_Con_Test_02.ndpi' is selected. The 'Key-Value Pairs' section is expanded, showing the URL and other metadata. A blue arrow points from the URL in the 'Key-Value Pairs' section to the URL in the 'Details' section.



How to access NGFF data?

Visualization

Multiple web-based and desktop clients



Programmatically

Python, C/C++, Java, Julia, Javascript

```
[1]: from ome_zarr.io import parse_url
from ome_zarr.reader import Reader
import matplotlib.pyplot as plt
import numpy as np

url = "https://storage.googleapis.com/jax-public-ngff/KOMP/adult_lacZ/omero/control/B6N_FBrainEye.zarr/0/"

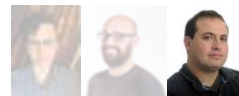
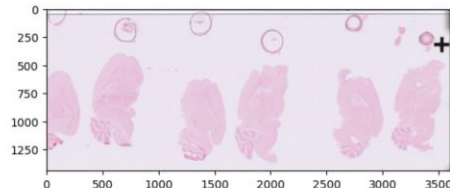
[2]: # Open the NGFF URL
store = parse_url(url, mode="r").store
zarr_reader = Reader(parse_url(url)).zarr

[3]: # Get a dask_array for the data at resolution 5 (1/2^5)
res5 = zarr_reader.load("5")

[4]: # Grab the data & reshape to c,y,x to x,y,c
image = np.transpose(res5[0, :, 0, :, :], (1,2,0))
image.shape

[4]: (1440, 3600, 3)

[5]: plt.imshow(image)
plt.show()
```



Cloud-based NGFF Conversion

- Bulk conversion using Google Batch and `bioformats2raw`
- ~17,000 images converted so far (~100%)
- Wall-clock conversion time of ~2 days
 - Total CPU time of 860 days for all JAX public data



Current challenges

- Data size explosion
 - 20TB NDPI → 200TB Zarr (with default BLOSC parameters)
- File count explosion
 - ~20K files → ~90M objects
 - Zarr v3 sharding will help!
- Excess metadata operations with `gcsfuse`

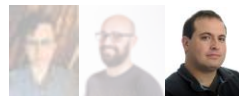
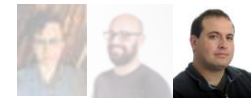


Image Data Resource

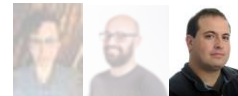
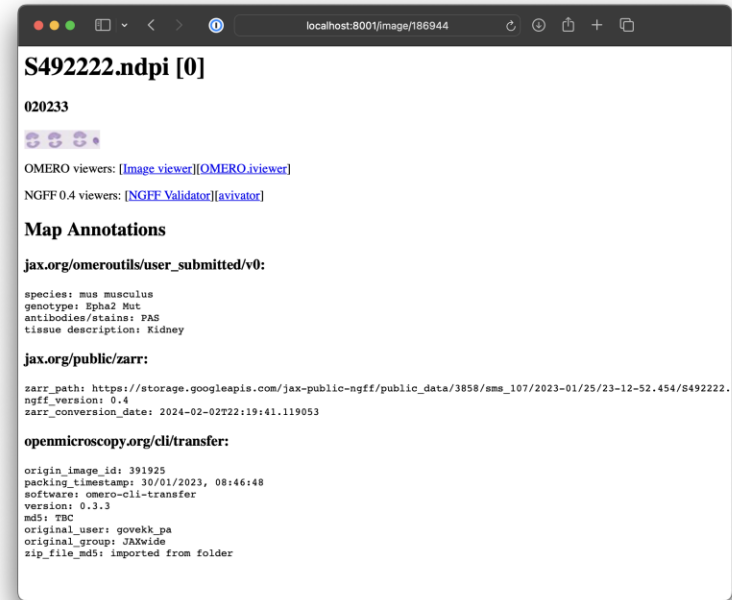
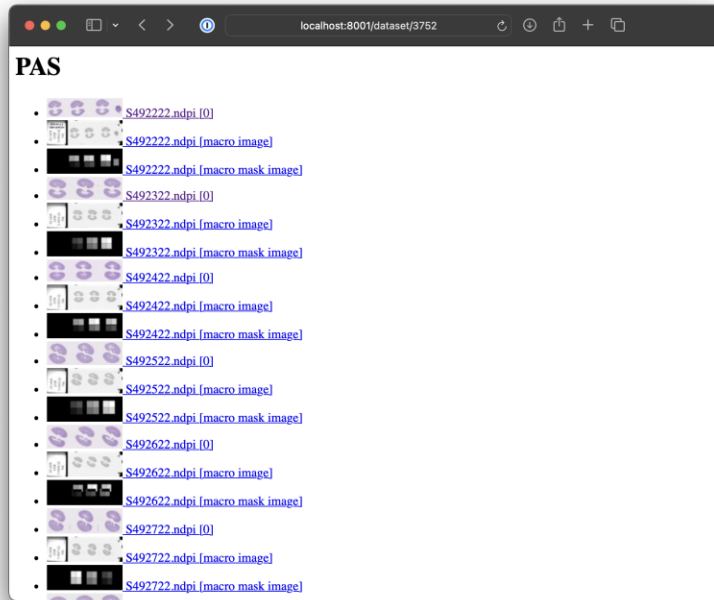


A screenshot of the IDR website interface. The browser address bar shows "idr.openmicroscopy.org/cell". The page features a navigation bar with "IDR", "CELL - IDR", and "TISSUE - IDR" tabs, and a menu with "ABOUT", "RESOURCES", and "SUBMISSIONS". The main content area has a large background image of a cell with colorful filaments. The IDR logo is centered, with a description: "The Image Data Resource (IDR) is a public repository of image datasets from published scientific studies, where the community can submit, search and access high-quality bio-image data." Below this are two orange buttons: "Cell - IDR" and "Tissue - IDR". A search bar contains the text "Search for anything...". A statistics bar shows "82 Studies", "8,050,408 Images", and "162 TB". A study card for "Feldman D et. al" is highlighted, showing "6 Experiments" and "218,673 Images" with a description: "Optical pooled screens in human cells". The card includes a grid of small image thumbnails.



THE JACKSON LABORATORY

In progress: querying OMERO



Towards a FAIRer future...

- OMERO provides a way to make imaging data FAIR(ish) and we are improving accessibility and interoperability
- We see OMERO as the source of truth for data, but want pixel access and visualization to be available outside of it
- OME-NGFF presents a path to combine the OME data model and efficient, cloud-friendly pixel access



Acknowledgments

JAX Imaging Applications

- Fernando Cervantes
- Peter Sobolewski

JAX Research IT

- Dave McKenzie
- Jason Macklin
- Vishal Thummala

OME & Friends

- Josh Moore
- Will Moore
- Seb Besson

Zarr, NGFF & Others

- Juan Nunez-Iglesias
- Davis Bennett
- Jeremy Maitin-Shepard (neuroglancer)
- Trevor Manz (vizarr)

The Open Source community

- OpenCV
- Scikit-image
- ImageJ/Fiji
- OMERO & BioFormats
- Zarr-python
- N5-lib



Presentations available @

**[https://downloads.openmicroscopy.org/
presentations/2024/Dundee](https://downloads.openmicroscopy.org/presentations/2024/Dundee)**

