Talk Outline
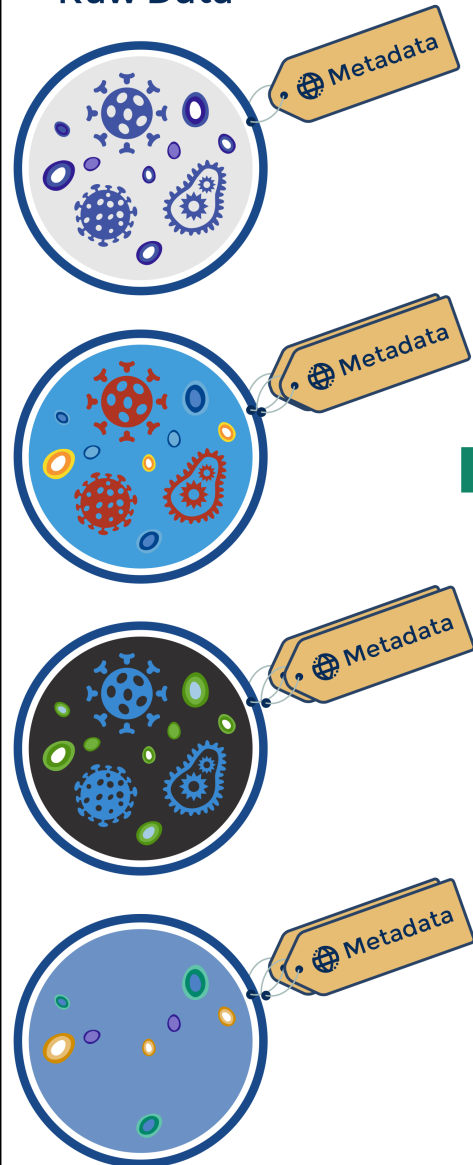
- The Problem

- Pattern files

- Screen files

- Limitations and next steps
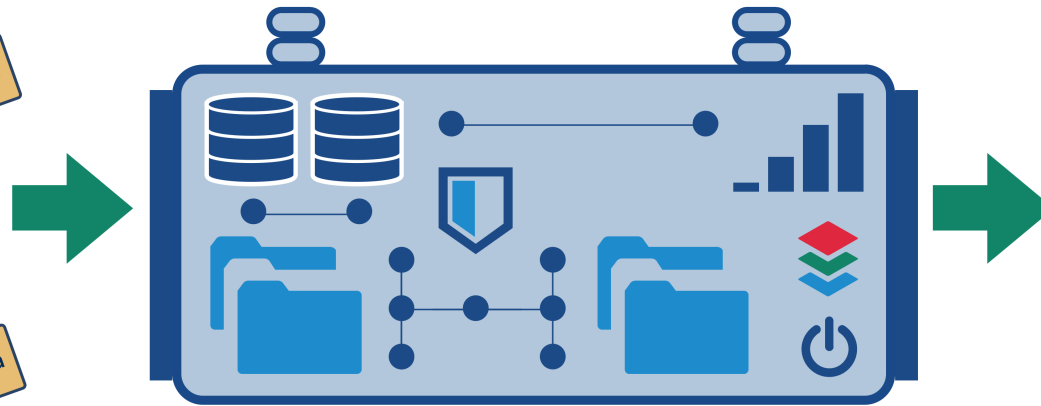
THE PROBLEM

BIO-FORMATS:
Proprietary File Conversion

Raw Data

Processed Data

BIO-FORMATS

600,000+ FILES

30,000+ DATASETS

5.7+ TERABYTES

ImageJ
Image Processing and Analysis in Java

MathWorks

Columbus™

Your Software

→ Apps using Bio-Formats were started >100,000 times per day in 2015 ←

Multi-file imaging formats

o Bio-Formats for imaging formats composed of multiple files

o Reader initialization includes the detection of all files required for the metadata and pixel data

- IFormatReader.getUsedFiles()

- IFormatReader.getSeriesUsedFiles(boolean)

o File grouping implemented per reader based on:

- the information from a file header (OME-TIFF)

- the information from a metadata file (Operetta)

- The file format specification

Patterns & screens: overview

o   Problem: acquired data with missing format metadata/specification and implicit
   file grouping conventions based on the folder hierarchy and file names

o   Requires **additional** metadata to
   1- determine the image metadata (core dimensions, HCS)
   2- map the files required for accessing the pixel data

o   Current solutions: pattern and screen files and their associated readers

o   Used in practice by **a third** of the current production IDR studies

Pattern files: introduction

- Applications: multi-series multi-dimensional (ZCT) imaging data

- Goal: group files based on their filename

- Library: Bio-Formats 5.1.0 and above

- Reader: [FilePatternReader](#)

- Format specification: see [documentation page](#)

- Generalized to any format supported by Bio-Formats, most typically TIFF or PNG files
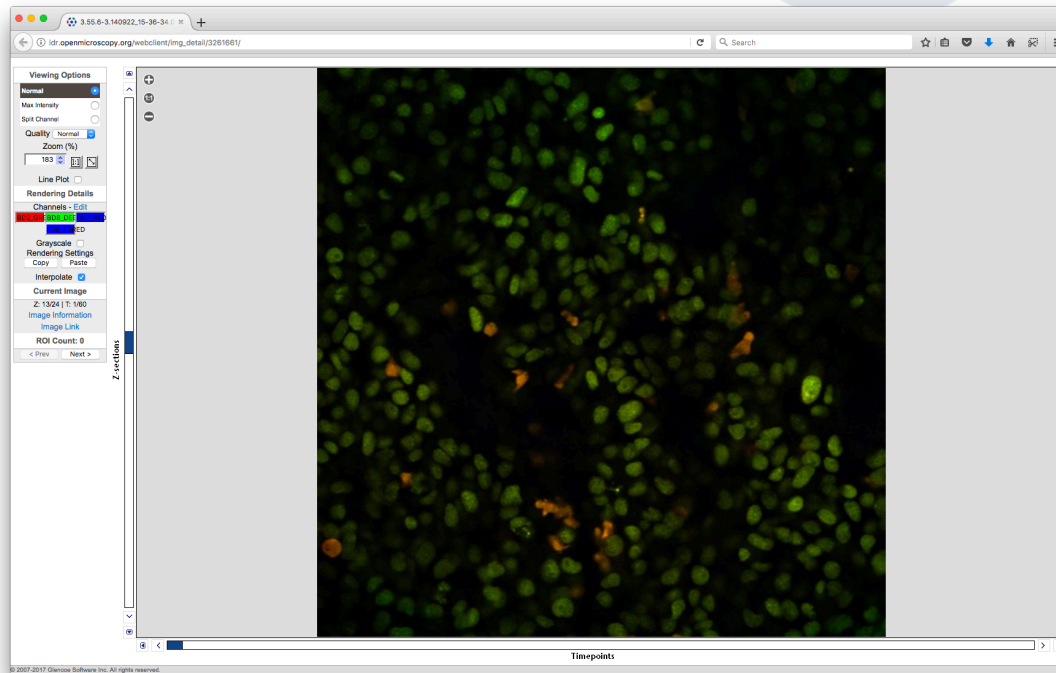
Pattern files: examples

○ See

https://github.com/sbesson/idr0000-lastname-example/tree/screens_training/patterns

○ Requires Bio-Formats 5.x (command-line tools, ImageJ/Fiji...)

○ Time-series experiment

○ Multi-channel experiment

○ Multi-dimensional experiment

Pattern files: format specification

o See https://docs.openmicroscopy.org/bio-formats/5.7.1/formats/pattern-file.html

o Single file ending with .pattern containing a single line of text

o Patterns must be grouped between <> symbols

  • Range of integers incl. step increments e.g. <0-10>, <0-10:2>

  • List of values e.g. <red,green,blue>

o Assumes dimensions (XYZCT) can be concatenated

o Pattern prefix forces the dimension - see AxisGuesser

  • Channel: c, ch, w, wavelength

  • Timepoint: t, tl, tp, time

  • Z sections: fp, sec, z, zs, focal, focalplane

  • Series: s, series, sp

Pattern files in production (IDR 0026)

https://idr.openmicroscopy.org/webclient/?show=image-3261661

- 4 channels, 60 timepoints, 24 Z-sections distributed across 240 TIFF files

- Pattern file

Screen files: introduction

○ Application: High-Content Screening datasets e.g.

https://data.broadinstitute.org/bbbc/BBBC021/

○ Goal: specify the HCS layout in machine-readable format

○ Library: IDR Bio-Formats (metadata branches)

○ Reader: ScreenReader

○ Format specification: see README.md for a minimal description of the format and some minimal examples

Screen file: examples

○ See

https://github.com/sbesson/idr0000-lastname-example/tree/screens_training/screens

○ Use IDR Bio-Formats (0.4.1) JAR with the existing clients (command-line tools, ImageJ/Fiji)

○ 1 well, 1 field

○ 4 wells, 1 field

○ 4 wells, 2 fields

○ 48 wells, 2 fields, sparse

○ BBBC021

Screen files in production (IDR 0033)

https://idr.openmicroscopy.org/webclient/?show=plate-5958
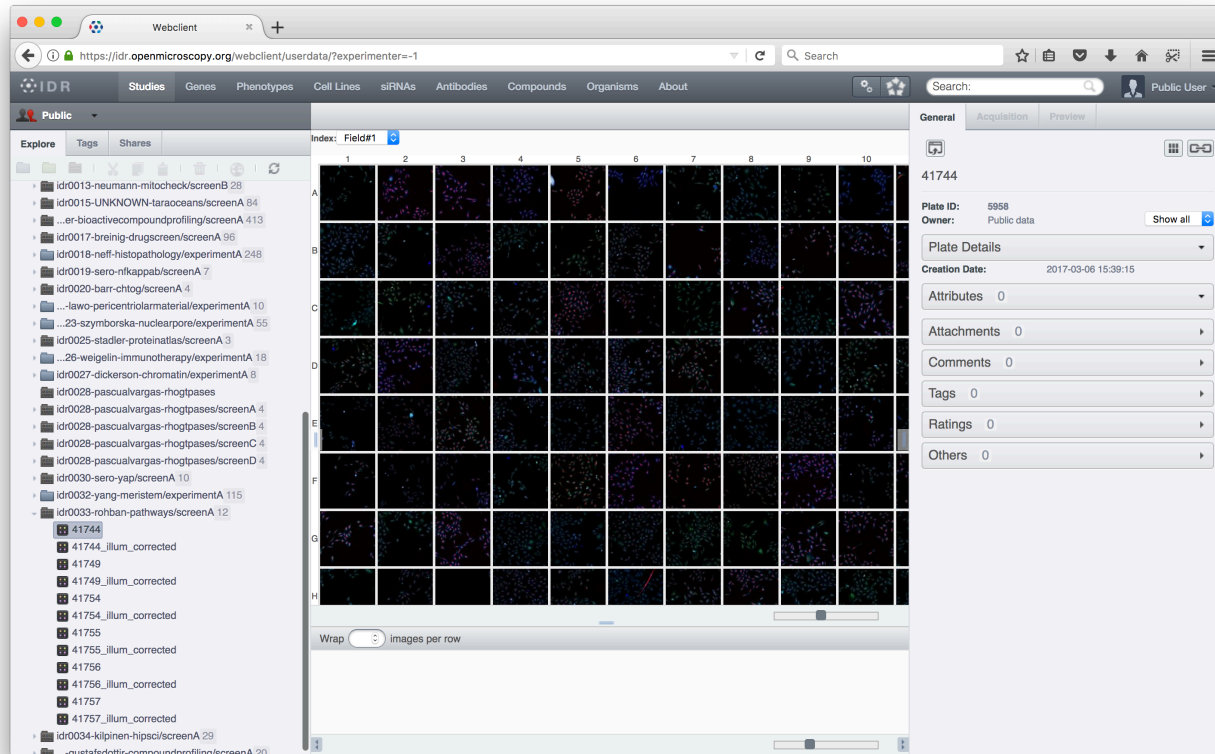
- 16 rows, 24 columns, 9 fields (each field image is a pattern file)

- Screen file

Screen & pattern files: generation at scale

○ [idr-metadata](#) repository contains both the tools/scripts and the metadata files

○ 1.3 K screen files, 150 pattern files (0.4.3 tag)

○ Tools

  • Top-level [Python library](#): pyidr.file_pattern, pyidr.screenio

  • Per-study file generation scripts: make_patterns.py, generate_screens/

  • Screen format validation script: [check_screen.py](#)

○ screenY/idrXXX-screenY-plates.tsv: additional level for grouping plate specified by screen files in a screem for each study

○ IDR import step

Next steps

- Screen reader/format currently IDR-only

  - Solution for real-life multi-dimensional HCS datasets

  - Existing reader + utility library/scripts

  - Cost of maintained a Bio-Formats fork/branch

- Backporting ScreenReader to the mainline

  - Another metadata file format to support

  - Based on poorly specified INI format

  - Versioning

  - Licensing

  - Performance issues

Next steps

○ ScreenReader translates .screen metadata into OME-XML

○ IDR: most of the pixel data is expressed in TIFF or HDF5 format

○ Possibility: convert .screen files into companion OME-XML files

○ Current limitations of the OME companion format

- only works with strict OME-TIFF file

- metadata file must be suffixed as .companion.ome

- binary files must link back to the companion file

○ Design requirements

- Relaxed version of the [companion OME file specification](#) – initial investigation by Simone

- Support for HDF data at the model level