

Importing data with rich metadata @scale

Eleanor Williams, Simon Li, Josh Moore

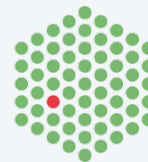
University of Dundee
The OME Consortium

@openmicroscopy #OME2016



Open Microscopy Environment
Centre for Gene Regulation & Expression
College of Life Sciences, University of Dundee
Dundee, Scotland, UK

EMBL-EBI



UNIVERSITY OF
CAMBRIDGE

Talk Outline

- Our metadata pipeline for the Image Data Repository (IDR)
- What kinds of metadata are we adding?
- How do we record the metadata?
- How we load it into OMERO?
- What are we thinking about next?

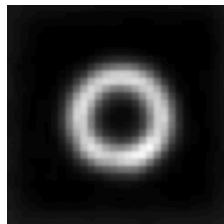
Our metadata pipeline

- Typical publication workflow
- Single processes which take more than an hour
- Groups of processes more than a dozen
- Batched imports, i.e. not real-time
- Nor automated (background) imports

Example datasets

Nuclear pore Complex

- Super resolution microscopy and particle averaging
- 7 proteins in the Nup107-160 complex, 524 image stacks
- Immunolabelling and nanobody labeling of GFP fusion proteins to determine average radial positions

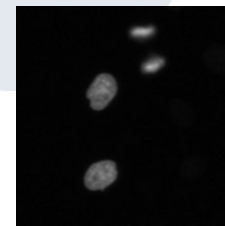


idr0023-szymborska-nucleopore

Szymborska et al. Science 2013

Mitocheck

- High Content Screen
- siRNA knockdown of ~22,000 human genes, over 500 plates, ~200,000 wells
- time-lapse imaging of fluorescently labelled chromosomes.
- Computational analysis of phenotypes

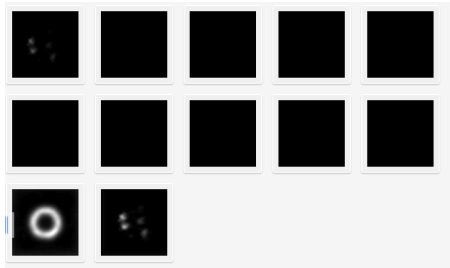


idr0013-neumann-mitocheck

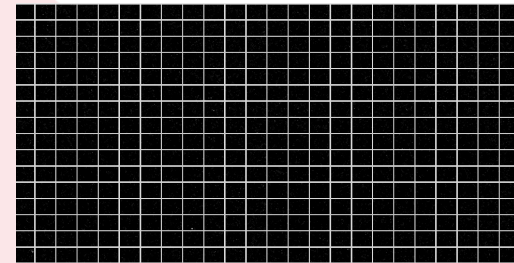
Neumann et al. Nature 2010

What kinds of metadata?

Nuclear pore Complex



Mitochcheck



Experimental Metadata

Study and experiment descriptors

Organism, cell line

Targeted protein, gene identifier, symbol

Antibody

File type

Analytic Metadata

Phenotype and Ontology annotation

NumberOfImages, NumberOfParticles

RadialDistance, SEMerror

Experimental Metadata

Study and screen descriptors

siRNA identifiers and sequences

Gene identifier and symbol

Controls, quality control at well level

What is labeled in each channel

Analytic Metadata

Manual and automatic scores

Phenotype reproducibility at siRNA level

Phenotypes and Ontology annotations

What kinds of metadata?

Adding Gene Identifiers

Image Name	Targeted Protein	Gene Identifier	Gene Symbol	Gene Symbol Synonyms
nup96_corr2_low500_6.tif	Nup96	ENSG00000110713	NUP98	NUP96, ADAR2, NUP196, ADIR2

Adding Phenotype Mappings

Phenotype 1	Phenotype 1 Term Name	Phenotype 1 Term Accession
nuclei stay close together (manual)	binuclear cell phenotype	CMPO_0000213

Recording the metadata



study file

study type
imaging method
study description
phenotype ontology
mappings



library or assay file

reagent identifiers
gene identifiers
controls
quality control



results file

measurements
scores
phenotypes

Recording the metadata



study file

study type
imaging method
study description
phenotype ontology
mappings



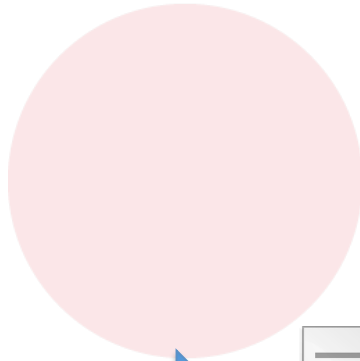
library or assay file

reagent identifiers
gene identifiers
controls
quality control



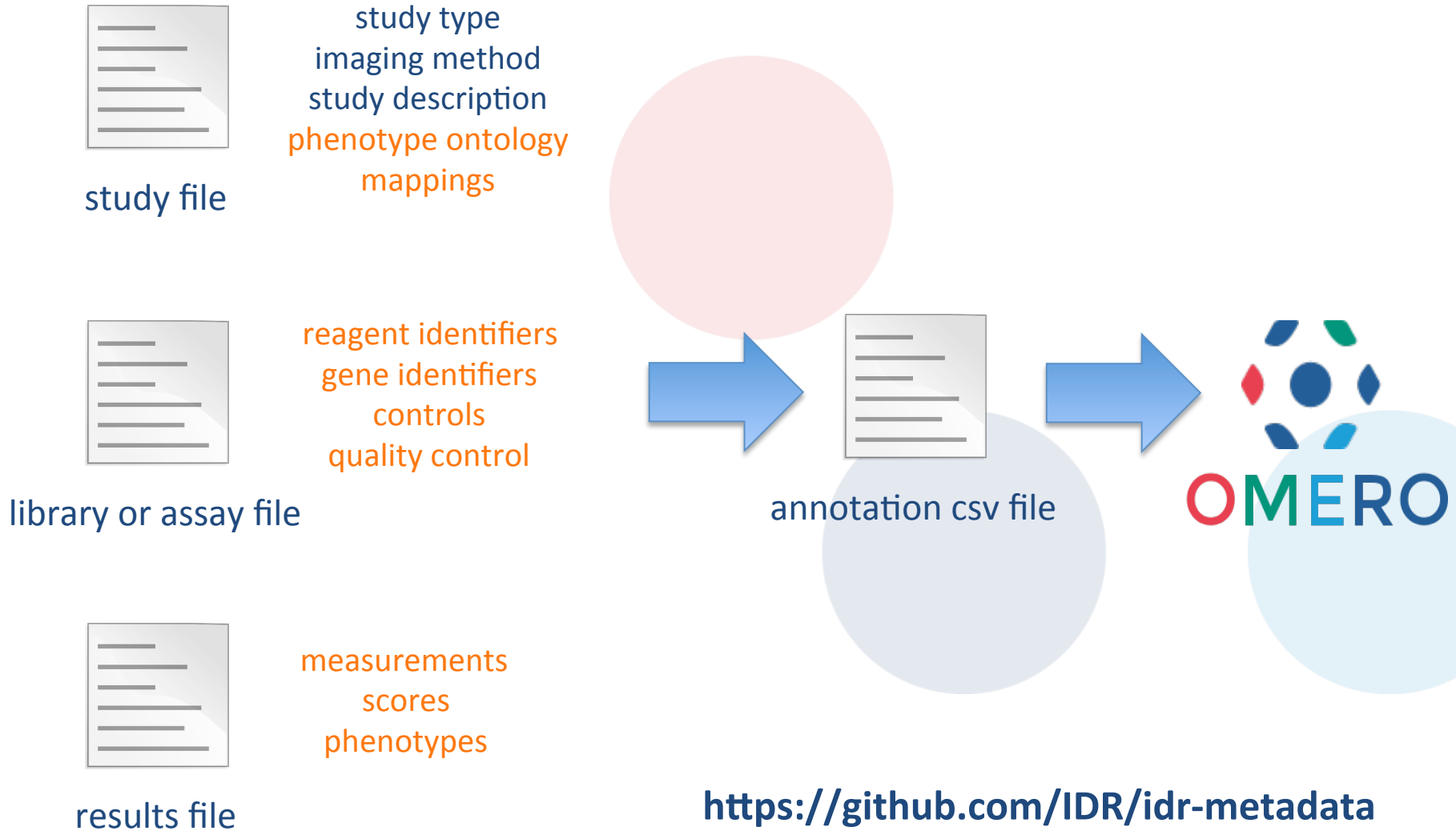
results file

measurements
scores
phenotypes



annotation csv file

Recording the metadata



Recording the metadata



annotation csv file

Plate	Well	siRNA	Gene Identifier	Gene Symbol	Control	Score	Phenotype	Phenotype Term Name	Phenotype Term Accession
HT1	A1				empty				
HT1	A2	s5673	ENSG023	ARP1		32	small cells	decreased cell size phenotype	CMPO_003
HT1	A3	s4562	ENSG023	ARP1		45	small cells	decreased cell size phenotype	CMPO_003
HT1	A4	s4567	ENSG055	KAT2		6			

library file

results file

study file

What it looks like in OMERO

General Acquisition Preview

Added by: Demo User
openmicroscopy.org/omero/bulk_annotations

siRNA Identifier 103860

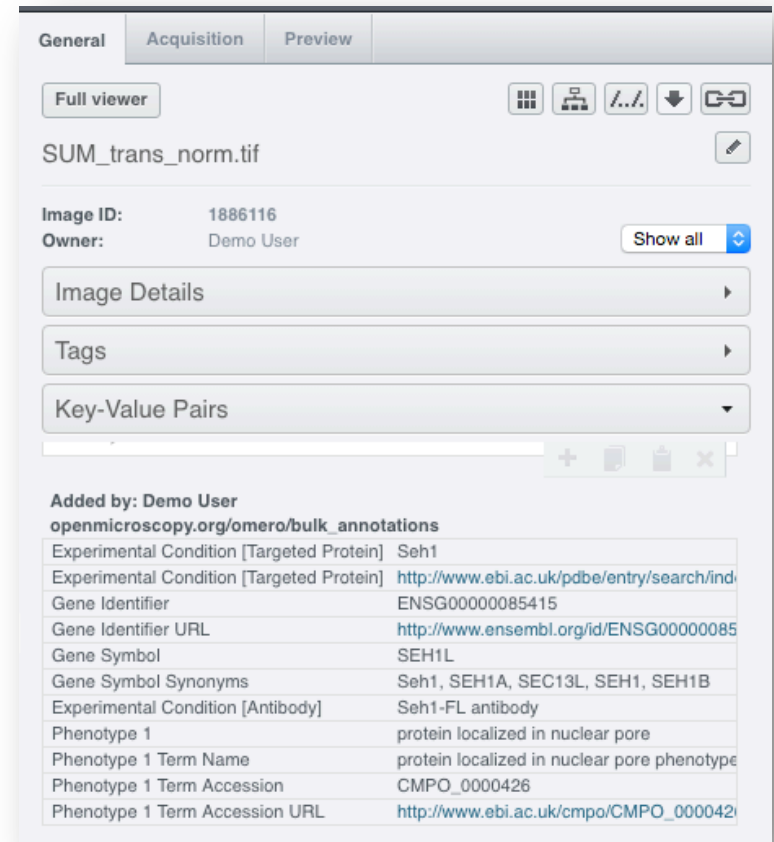
TABLES

Plate:	52
Well Number:	26
Well:	785
Plate_Well:	LT0002_02_B2
siRNA Identifier:	103860
Mitocheck siRNA Identifier:	MCO_0016403
Sense Sequence:	AGUACUGCUUACGAUACGGtt
Antisense Sequence:	CCGUAUACGUAAGCAGUACUtt
siRNA Intended Target:	negative control
Gene Identifier:	
Control Type:	negative control
Control Comments:	non-targetting siRNA
Quality Control:	TRUE
Score - cell death (automatic):	0.009548
Score - binuclear (automatic):	0.011956
Score - polylobed (automatic):	-0.001843
Score - large (automatic):	-0.002168
Score - dynamic changes (automatic):	0.008499
Score - mitotic delay/arrest (automatic):	0.000902
Score - grape (automatic):	-0.000118
Score - migration (speed) (automatic):	0.293941
Score - migration (distance) (automatic):	-1.46556
Score - increased proliferation (automatic):	-0.49314
siRNA phenotype reproducibility - nuclei stay close together (manual):	0

Mitocheck

Annotations in OMERO

- Convert chosen columns to Map Annotations (yaml)
 - Makes them searchable
 - Highlights key information (not all results details)
 - Some can be made into links



The screenshot shows the OMERO interface for an image named 'SUM_trans_norm.tif'. The interface includes tabs for 'General', 'Acquisition', and 'Preview'. Below the image name, there are fields for 'Image ID: 1886116' and 'Owner: Demo User'. A 'Show all' button is visible. The 'Image Details' section is expanded, showing a table of annotations. The annotations are organized into sections: 'Added by: Demo User' and 'openmicroscopy.org/omero/bulk_annotations'. The table lists various metadata fields such as 'Experimental Condition [Targeted Protein]', 'Gene Identifier', 'Gene Symbol', and 'Phenotype 1' with their corresponding values and links.

Added by: Demo User	
openmicroscopy.org/omero/bulk_annotations	
Experimental Condition [Targeted Protein]	Seh1
Experimental Condition [Targeted Protein]	http://www.ebi.ac.uk/pdbe/entry/search/ind
Gene Identifier	ENSG00000085415
Gene Identifier URL	http://www.ensembl.org/id/ENSG00000085415
Gene Symbol	SEH1L
Gene Symbol Synonyms	Seh1, SEH1A, SEC13L, SEH1, SEH1B
Experimental Condition [Antibody]	Seh1-FL antibody
Phenotype 1	protein localized in nuclear pore
Phenotype 1 Term Name	protein localized in nuclear pore phenotype
Phenotype 1 Term Accession	CMPO_0000426
Phenotype 1 Term Accession URL	http://www.ebi.ac.uk/cmipo/CMPO_0000426

idr0023-szymborska-nucleopore

Annotations in OMERO

General Acquisition Preview

Full viewer

SUM_trans_norm.tif

Image ID: 1886116
Owner: Demo User

Show all

Image Details

Tags

Key-Value Pairs

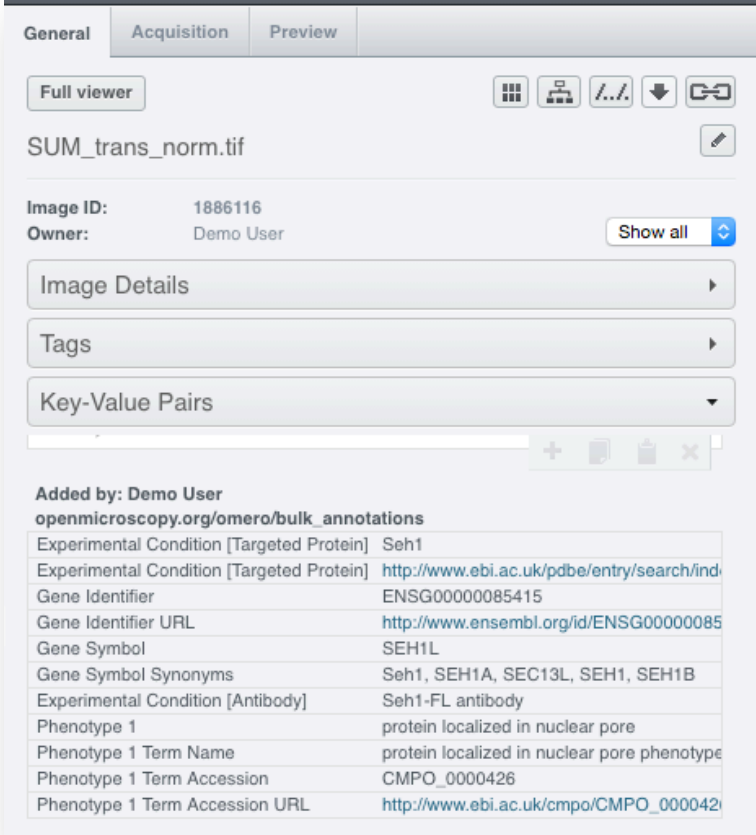
Added by: Demo User
openmicroscopy.org/omero/bulk_annotations

Experimental Condition [Targeted Protein]	Seh1
Experimental Condition [Targeted Protein]	http://www.ebi.ac.uk/pdbe/entry/search/ind
Gene Identifier	ENSG00000085415
Gene Identifier URL	http://www.ensembl.org/id/ENSG00000085415
Gene Symbol	SEH1L
Gene Symbol Synonyms	Seh1, SEH1A, SEC13L, SEH1, SEH1B
Experimental Condition [Antibody]	Seh1-FL antibody
Phenotype 1	protein localized in nuclear pore
Phenotype 1 Term Name	protein localized in nuclear pore phenotype
Phenotype 1 Term Accession	CMPO_0000426
Phenotype 1 Term Accession URL	http://www.ebi.ac.uk/cmipo/CMPO_0000426

- name: Gene Identifier
- clientname: Gene Identifier URL
- clientvalue: <http://.../result/{{value}}>
- include: yes

idr0023-szymborska-nucleopore

Annotations in OMERO



General Acquisition Preview

Full viewer

SUM_trans_norm.tif

Image ID: 1886116
Owner: Demo User

Show all

Image Details

Tags

Key-Value Pairs

Added by: Demo User

openmicroscopy.org/omero/bulk_annotations

Experimental Condition [Targeted Protein]	Seh1
Experimental Condition [Targeted Protein]	http://www.ebi.ac.uk/pdbe/entry/search/ind
Gene Identifier	ENSG00000085415
Gene Identifier URL	http://www.ensembl.org/id/ENSG00000085415
Gene Symbol	SEH1L
Gene Symbol Synonyms	Seh1, SEH1A, SEC13L, SEH1, SEH1B
Experimental Condition [Antibody]	Seh1-FL antibody
Phenotype 1	protein localized in nuclear pore
Phenotype 1 Term Name	protein localized in nuclear pore phenotype
Phenotype 1 Term Accession	CMPO_0000426
Phenotype 1 Term Accession URL	http://www.ebi.ac.uk/cmipo/CMPO_0000426

○ bin/omero metadata populate

- Screens, Plates
- Projects, Datasets

idr0023-szymborska-nucleopore

Other types of metadata

○ Bulk import (yml + tsv)

- Config:

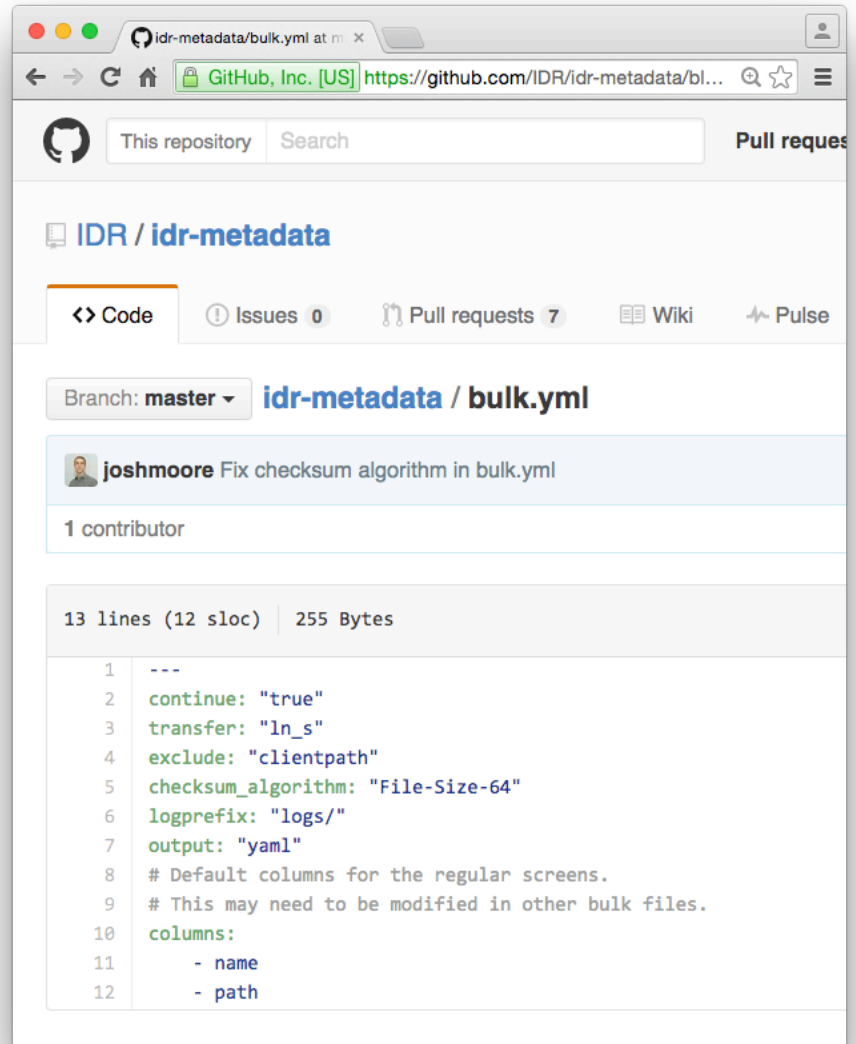
```
exclude: "clientpath"
```

```
transfer: "ln_s"
```

- Run:

```
bin/omero import
```

```
--bulk config.yml
```

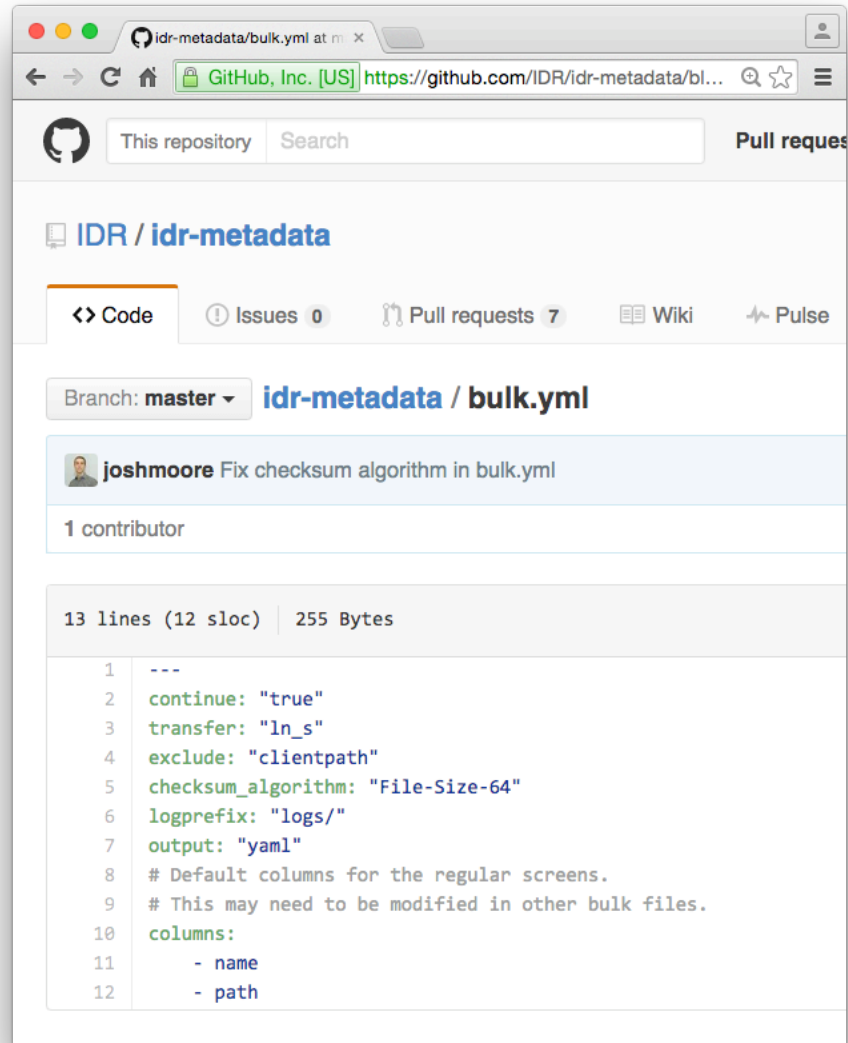


The screenshot shows a GitHub repository page for IDR/idr-metadata. The file bulk.yml is selected, showing its content. The file has 13 lines (12 sloc) and 255 Bytes. The content is as follows:

```
1 ---
2 continue: "true"
3 transfer: "ln_s"
4 exclude: "clientpath"
5 checksum_algorithm: "File-Size-64"
6 logprefix: "logs/"
7 output: "yaml"
8 # Default columns for the regular screens.
9 # This may need to be modified in other bulk files.
10 columns:
11   - name
12   - path
```

Other types of metadata

- Bulk import (yml + tsv)
- Rendering settings (yml)
 - channels:
 - 1:
 - label: "DAPI"
 - min: 0
 - max: 750
 - color: "0000FF"

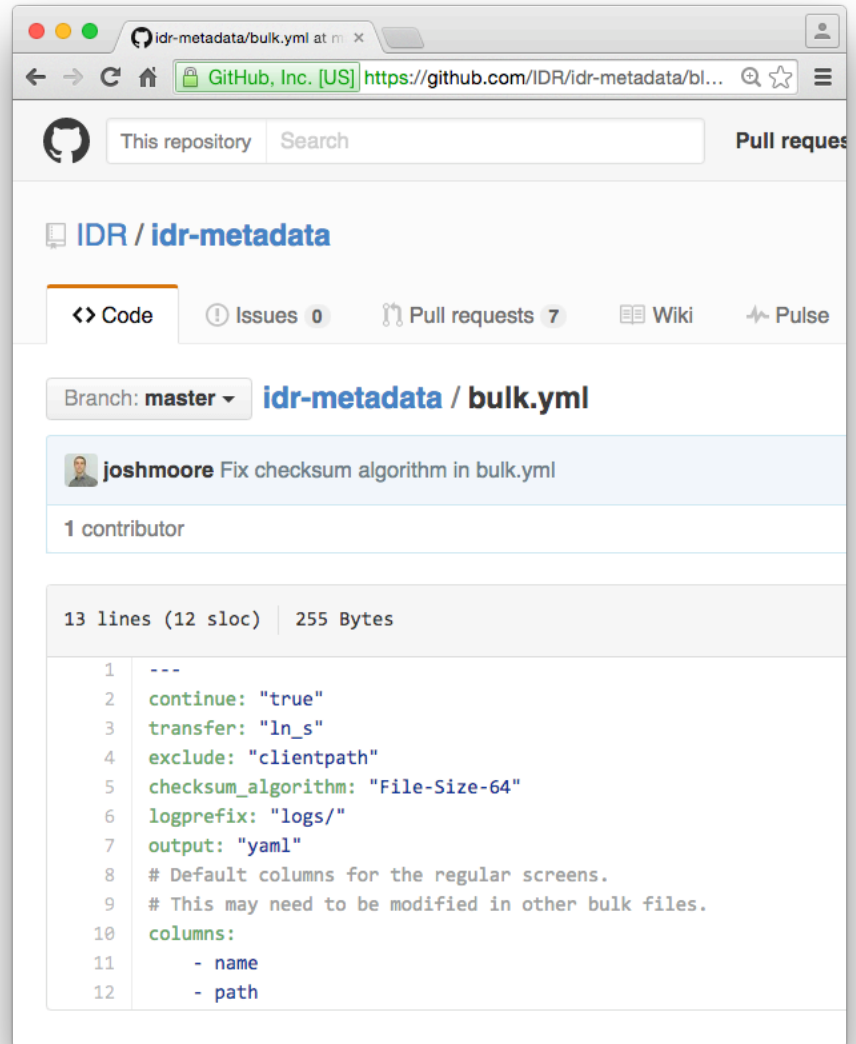


The screenshot shows a GitHub repository page for IDR/idr-metadata. The file being viewed is bulk.yml, which contains the following content:

```
1 ---
2 continue: "true"
3 transfer: "ln_s"
4 exclude: "clientpath"
5 checksum_algorithm: "File-Size-64"
6 logprefix: "logs/"
7 output: "yaml"
8 # Default columns for the regular screens.
9 # This may need to be modified in other bulk files.
10 columns:
11   - name
12   - path
```


Other types of metadata

- Bulk import (yml + tsv)
- Rendering settings (yml)
- TBD: Regions-of-interest



The screenshot shows a GitHub repository page for 'IDR / idr-metadata'. The file 'bulk.yml' is selected, showing a commit by 'joshmoore' with the message 'Fix checksum algorithm in bulk.yml'. The file content is as follows:

```
1 ---
2 continue: "true"
3 transfer: "ln_s"
4 exclude: "clientpath"
5 checksum_algorithm: "File-Size-64"
6 logprefix: "logs/"
7 output: "yaml"
8 # Default columns for the regular screens.
9 # This may need to be modified in other bulk files.
10 columns:
11   - name
12   - path
```

What next?

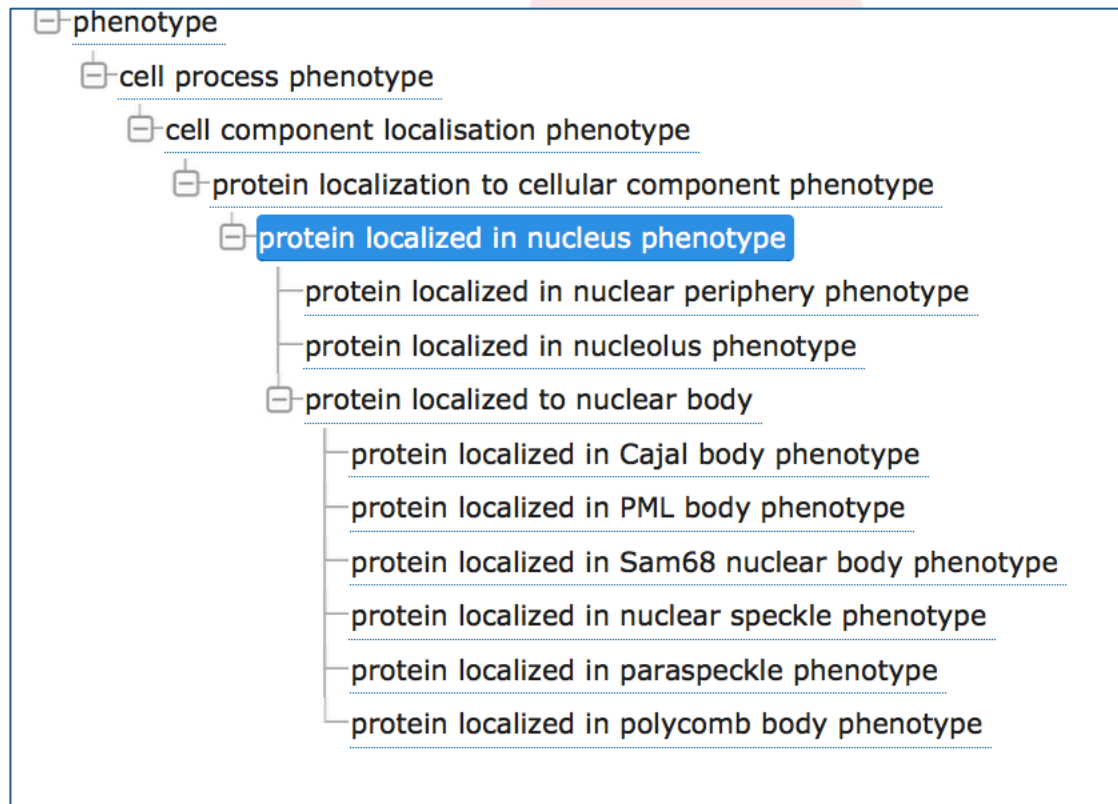
- Better display of search results

	0108-05--2006-03-10 [Well 85, Field 1 (Spot 85)]	2015-09-25 06:21:59	2015-09-25 06:21:59	demo	Browse
	0108-05--2006-03-10 [Well 55, Field 1 (Spot 55)]	2015-09-25 06:21:59	2015-09-25 06:21:59	demo	Browse
	0108-32--2006-03-06 [Well 85, Field 1 (Spot 85)]	2015-09-25 06:36:51	2015-09-25 06:36:51	demo	Browse
	0108-32--2006-03-06 [Well 55, Field 1 (Spot 55)]	2015-09-25 06:36:51	2015-09-25 06:36:51	demo	Browse
	11059 [Well H06 Field #1]	2011-12-13 17:18:26	2015-10-01 00:31:08	demo	Browse
	11059 [Well H06 Field #2]	2011-12-13 17:18:26	2015-10-01 00:31:08	demo	Browse
	11059 [Well H06 Field #3]	2011-12-13 17:18:26	2015-10-01 00:31:08	demo	Browse
	11059 [Well H06 Field #4]	2011-12-13 17:18:26	2015-10-01 00:31:08	demo	Browse
	11059 [Well H06 Field #5]	2011-12-13 17:18:26	2015-10-01 00:31:08	demo	Browse
	11059 [Well H06 Field #6]	2011-12-13 17:18:26	2015-10-01 00:31:08	demo	Browse
	11059 [Well G11 Field #1]	2011-12-13 17:18:26	2015-10-01 00:31:08	demo	Browse

Search
for
“Pipox”

What next?

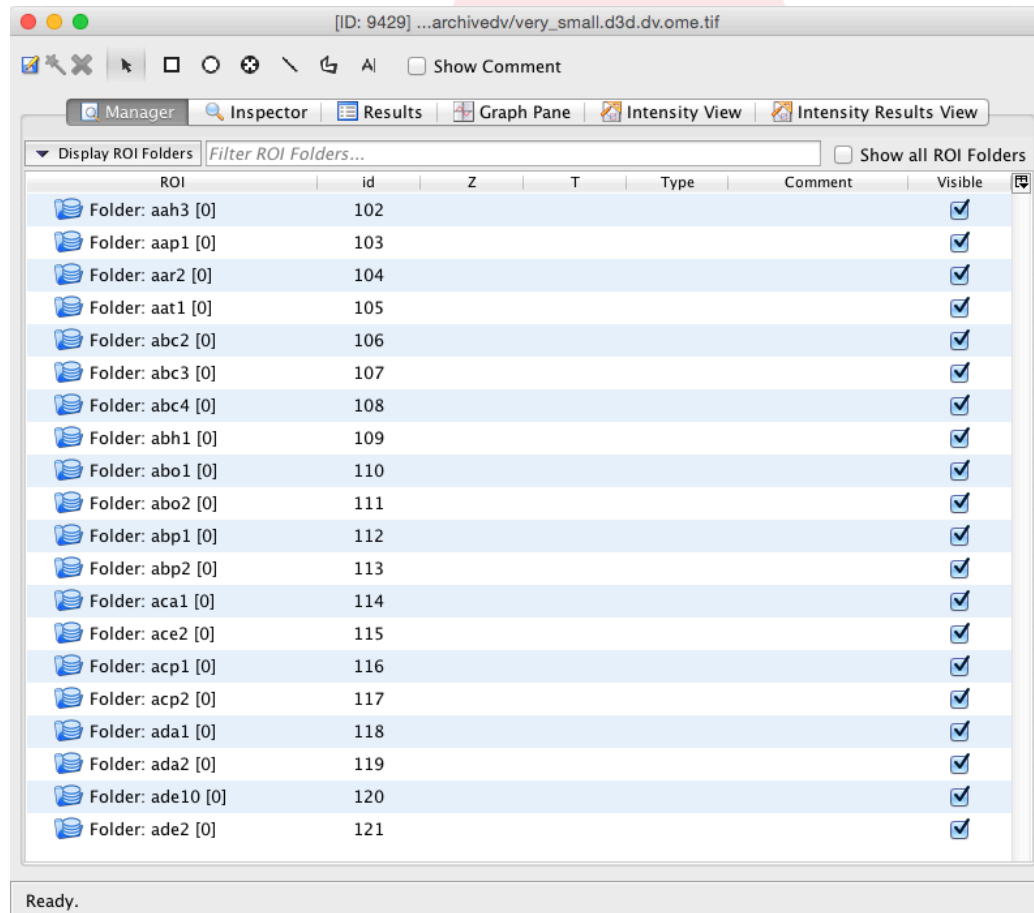
- Make more use of ontologies – integrate OLS



Ontology Look Up Service, www.ebi.ac.uk/ols

What next?

- Look at bulk annotation of ROIs – folders



What next?

- Providing a “standardized” template



study file



library or assay file



results file

- Incorporate other types of imaging study e.g. light sheet microscopy, plant studies, medical imaging
- Alignment with other efforts at standardization e.g. Multimot

What next?

- Handling annotation server-side
- Error-detection and feedback on background steps

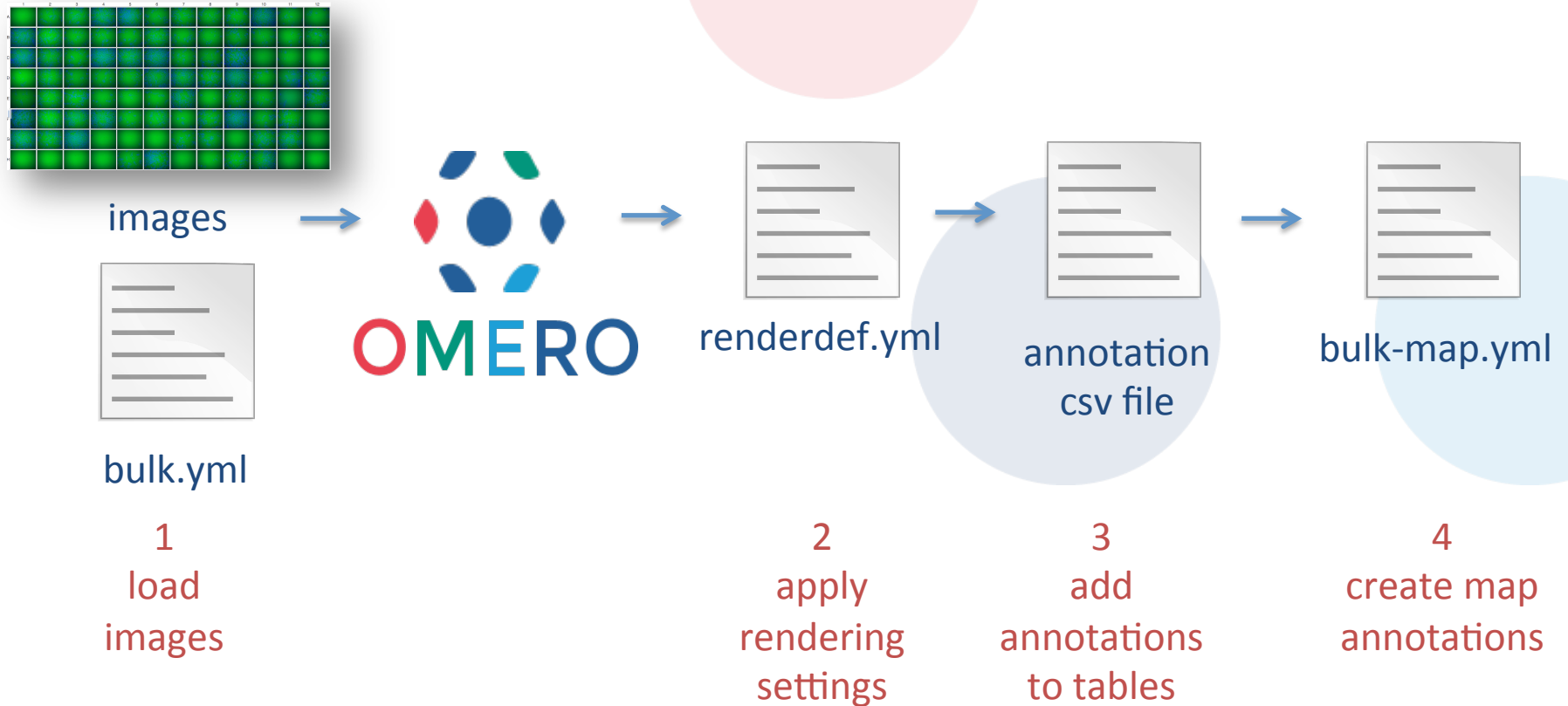


Image Data Repository URL

<http://idr-demo.openmicroscopy.org/>

Thank you

- OME group - Dundee – Colin Blackburn, Jean-Marie Burel, Mark Carroll, Gus Ferguson, Helen Flynn, David Gault, Kenny Gillen, Roger Leigh, Simone Leo, Simon Li, Dominik Lindner, June Matthew, Josh Moore, Will Moore, Balaji Ramalingam, Gabriella Rustici, Aleksandra Tarkowska, Petr Walczysko, Harald Waxenegger, Eleanor Williams
- Cambridge – Rafael Carazo-Salas, Bálint Antal, Anatole Chessel
- EMBL-EBI –Alvis Brazma, Ugis Sarkans
- Glencoe Software – Chris Allan, Joshua Ballanco, Andreas Knab, Melissa Linkert, Chris MacLeod, Josh Moore, Emil Rozbicki, Liza Unson, Rebecca Walker, Wilma Woudenberg

