Publishing Integrated Data at Scale: Update on the IDR Prototype

Josh Moore and Eleanor Williams
University of Dundee
The OME Consortium
IDR: Outline

- The vision
- Collection & integration
- Published & re-usable?
- The future
THE VISION
The Image Problem…
is Ubiquitous

A pretty picture?  A measurement?  A resource?
<table>
<thead>
<tr>
<th>name</th>
<th>value</th>
</tr>
</thead>
<tbody>
<tr>
<td>X_11022.S1 [Well A-1; Field #1]</td>
<td>Strain Name</td>
</tr>
<tr>
<td>X_11022.S1 [Well C-11; Field #1]</td>
<td>Gene Identifier</td>
</tr>
<tr>
<td>X_11022.S1 [Well C-11; Field #1]</td>
<td>Strain Name</td>
</tr>
<tr>
<td>X_11022.S1 [Well C-11; Field #1]</td>
<td>Gene Symbol</td>
</tr>
<tr>
<td>X_11022.S1 [Well C-11; Field #1]</td>
<td>Gene Identifier</td>
</tr>
<tr>
<td>X_11022.S1 [Well C-11; Field #1]</td>
<td>Gene Symbol</td>
</tr>
<tr>
<td>X_11022.S1 [Well E-4; Field #1]</td>
<td>Gene Identifier</td>
</tr>
<tr>
<td>X_11022.S1 [Well E-4; Field #1]</td>
<td>Gene Identifier URL</td>
</tr>
<tr>
<td>X_11022.S1 [Well E-4; Field #1]</td>
<td>Strain Name</td>
</tr>
<tr>
<td>X_11022.S1 [Well E-4; Field #1]</td>
<td>Gene Identifier</td>
</tr>
<tr>
<td>X_11022.S1 [Well E-4; Field #1]</td>
<td>Gene Identifier URL</td>
</tr>
<tr>
<td>X_11022.S1 [Well E-4; Field #1]</td>
<td>Strain Name</td>
</tr>
<tr>
<td>X_11022.S1 [Well E-4; Field #1]</td>
<td>Gene Symbol</td>
</tr>
<tr>
<td>X_11022.S1 [Well C-9; Field #1]</td>
<td>Gene Identifier</td>
</tr>
<tr>
<td>X_11022.S1 [Well C-9; Field #1]</td>
<td>Gene Identifier URL</td>
</tr>
<tr>
<td>X_11022.S1 [Well C-9; Field #1]</td>
<td>Strain Name</td>
</tr>
<tr>
<td>X_11022.S1 [Well C-9; Field #1]</td>
<td>Gene Symbol</td>
</tr>
<tr>
<td>X_11022.S1 [Well C-9; Field #1]</td>
<td>Phenotype 1</td>
</tr>
<tr>
<td>X_11022.S1 [Well C-9; Field #1]</td>
<td>Phenotype 1 Term Name a</td>
</tr>
<tr>
<td>X_11022.S1 [Well C-9; Field #1]</td>
<td>Phenotype 1 Term Accession a</td>
</tr>
</tbody>
</table>
Bringing Rad52 foci into focus

Peter H. Thorpe, David Alvaro, Michael Lisby, Rodney Rothstein

JCB vol. 194 no. 5 665-667 Article DOI: 10.1083/jcb.201106695 Data/Viewer DOI: 10.1083/jcb.201106695.dv

Screen: Rad52 :: 46 plates

Colored boxes indicate hits for the specified phenotype. Use the mouse scroll to zoom. Click on a colored box to select or deselect.

<table>
<thead>
<tr>
<th>Gene ID</th>
<th># Foci</th>
<th>Plate</th>
<th>Well</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The goal

Acquisition

Experimental

Analytic
Combining existing data
In particular we will build links with established molecular and structural resources and work towards a seamless integration of these data, so that any scientist can easily browse, query and compute on genomic, structural and phenotypic data across several scales.

Based on hardware infrastructure located at EMBL-EBI and integrated with its datasets to the world’s scientific community. These resources will serve as the basis and OMERO will be used to read, manage, serve, and link the data to enable computational re-analysis of the data, we will extend OMERO’s distributed compute capacity and make use of EMBL-EBI’s Embassy system to allow virtual access to IDR data. This virtual resource will provide a ‘sandbox’ for performing processing and reanalysis of data deposited in the IDR and provide a working example of a next generation data repository that stores and manages data, but also provides community services for scientific data.
IDR Vision

Integrated studies

Experimental metadata

Ontological annotations

Feature vectors

Download local analysis

Cloud analysis

Cross-data browsing

Gene Product Targeting HCS

Genetic HCS

Geographic HCS

Chemical HCS

Histopathology

3D-SIM

Super-resolution
Another 20TB of #BigData for the @openmicroscopy @BBSRC @emblebi image repository

http://j.mp/idr-disks
COLLECTION & INTEGRATION
Data Integration

study 1

study 2

study 3

study 4
Data Integration

study 1

study 2

study 3

study 4
Data Integration

study 1

study 2

study 3

study 4
Study Metadata and Ontology Annotation

Your records and analyses

output

- study file
- library or assay file
- results file

experimental metadata
analytic metadata

---

study	ile
library	
or	
assay	
file
results	
file
Your	
records
and	
analyses
output
study file
library or assay file
results file
experimental metadata
analytic metadata
Study Metadata and Ontology Annotation

Your records and analyses

output

study file

library or assay file

results file

study type
imaging method
study description
contacts

reagent identifiers
gene identifiers
controls
quality control

measurements
reproducibility
phenotypes

analytic metadata

experimental metadata
Study Metadata and Ontology Annotation

Your records and analyses

Output

study file

library or assay file

results file

MAGE-TAB

Cellular Phenotype Database

study type
imaging method
study description
contacts

reagent identifiers
gene identifiers
controls
quality control

measurements
reproducibility
phenotypes

analytic metadata

experimental metadata

isatab
### Experimental Metadata

#### Screen Type
- primary screen
- secondary screen
- validation screen

#### Screen Technology Type
- RNAi screen
- gene deletion screen
- protein screen
- compound screen

#### Library Type
- siRNA library
- diploid homozygous deletion library
- haploid deletion library
- tag protein fusion library
  - GFP protein fusion library
  - YFP protein fusion library
  - HA-Flag protein fusion library
- compound library

---

[Image of a high content screen with grid of wells]
Heterogeneity of Experimental Metadata

Controls

**Negative control** – no effect expected  
e.g scramble, wild type cell line, DMSO, mock

**Positive control** – effect expected  
- often not mentioned or listed in publication only

**No reagent** – no effect expected  
- no treatment, often no annotation

**Empty well**  
- not listed, listed but no annotation, images/no images
Analytic Metadata - Phenotypes

• Over 200 phenotypes over all the studies
• Average of 12 phenotypes per study
• Range between 1 and 48 phenotypes per study
Heterogeneity of Analytic Metadata

- cell shape
  - round or non-adherent
  - circular
- orb
- low eccentricity
- spheroid

Rohn et al, 2011
siRNA knockdown of Act5C

round cell phenotype
CMPO_0000118

89% of phenotypes mapped to ontology term

20 ontology terms appear in more than one study

http://www.ebi.ac.uk/cmpo/
## Heterogeneity of Analytic Metadata

### Level of annotation of results and phenotypes

<table>
<thead>
<tr>
<th>Gene</th>
<th>Phenotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>INTS1</td>
<td>shorter prophase, longer prophase</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gene</th>
<th>siRNA</th>
<th>Phenotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>INTS1</td>
<td>s25212</td>
<td>shorter prophase</td>
</tr>
<tr>
<td>INTS1</td>
<td>s25213</td>
<td>longer prophase</td>
</tr>
</tbody>
</table>

[https://github.com/IDR/idr-metadata](https://github.com/IDR/idr-metadata)
Data collected so far..

20 studies
24 screens or experiments

- HCS - genetic
- HCS - gene product targeting
- HCS - chemical
- Non HCS

- Human
- Fly
- Budding Fission yeast
- Yeast
- Mouse
- Multispecies
Each human gene or its ortholog
• in an average of 4 screens
• linked to an average of 28 different samples (excluding controls)
PUBLISHED & RE-USABLE?
IDR Reality

Integrated studies

Experimental metadata

Ontological annotations

Feature vectors
IDR Reality

Integrated studies

Experimental metadata

Ontological annotations

Feature vectors

http://j.mp/idr0001
IDR Reality

Integrated studies

Experimental metadata

Ontological annotations

Feature vectors

http://j.mp/idr0015
Panel: IDR Reality

- Integrated studies
- Thumbnails
- Experimental metadata
- Ontological annotations
- Feature vectors
13 studies across 4 OMERO 5.1 servers → 20 studies in 1 OMERO 5.2 server
<table>
<thead>
<tr>
<th></th>
<th>Nov ‘15</th>
<th>May ‘16 (Today)</th>
<th>Jun ‘16</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw image data</td>
<td>37 TB</td>
<td>35 TB</td>
<td>42 TB</td>
</tr>
<tr>
<td>(4 servers)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Files</td>
<td>8.2 M</td>
<td>5.9 M</td>
<td>16.5 M</td>
</tr>
<tr>
<td>Studies</td>
<td>13</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>Plates &amp; datasets</td>
<td>2500</td>
<td>2800</td>
<td>3400</td>
</tr>
<tr>
<td>Images</td>
<td>1.3 M</td>
<td>1.4 M</td>
<td>2.8 M</td>
</tr>
<tr>
<td>Planes</td>
<td>28.9 M</td>
<td>26.0 M</td>
<td>36.1 M</td>
</tr>
</tbody>
</table>
IDR Reality

- Thumbnails
- Experimental metadata
- Ontological annotations
- Feature vectors
- Integrated studies
IDR Reality

- Experimental metadata
- Ontological annotations
- Feature vectors
- Integrated studies
- Thumbnails
- Download local analysis
- Cloud analysis
- Cross-data browsing
THE FUTURE
This week: how we did it?

IDR Future

Integrated studies

Experimental metadata

Ontological annotations

Feature vectors

Download

local analysis

Cloud analysis

Cross-data browsing
IDR Future

- Experimental metadata
- Ontological annotations
- Feature vectors

Integrated studies

Thumbnail

Download
local analysis

Cloud analysis

Cross-data browsing
Thanks to the team

Jason Swedlow  Josh Moore  Simon Li  Eleanor Williams  Richard Ferguson  Simone Leo  Ola Tarkowska

Alvis Brazma  Ugis Sarkans  Simon Jupp  Tony Burdett

Rafael Carazo-Salas  Balint Antal  Anatole Chessel
Thanks to funders

- Wellcome Trust
- Horizon 2020
- BBSRC
- BioMedBridges
- CORBEL
- Global BioImaging
- MultiMOT
- Euro-BioImaging
Thanks for the data

http://j.mp/idr-map