Image Data Storage

Josh Moore
University of Dundee

The OME Consortium
openmicroscopy.org
@openmicroscopy

University of Dundee

OME

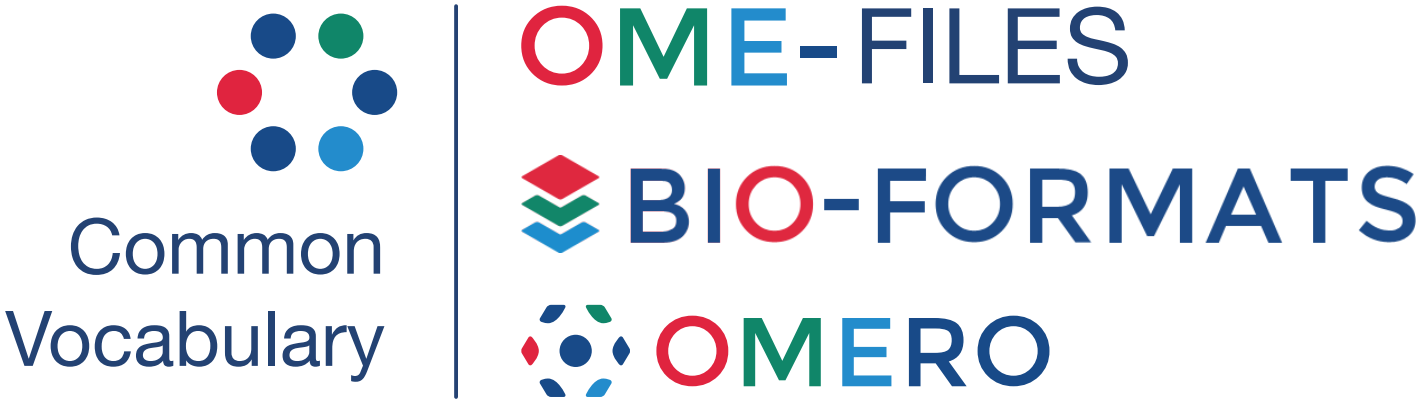
IDR

Outline

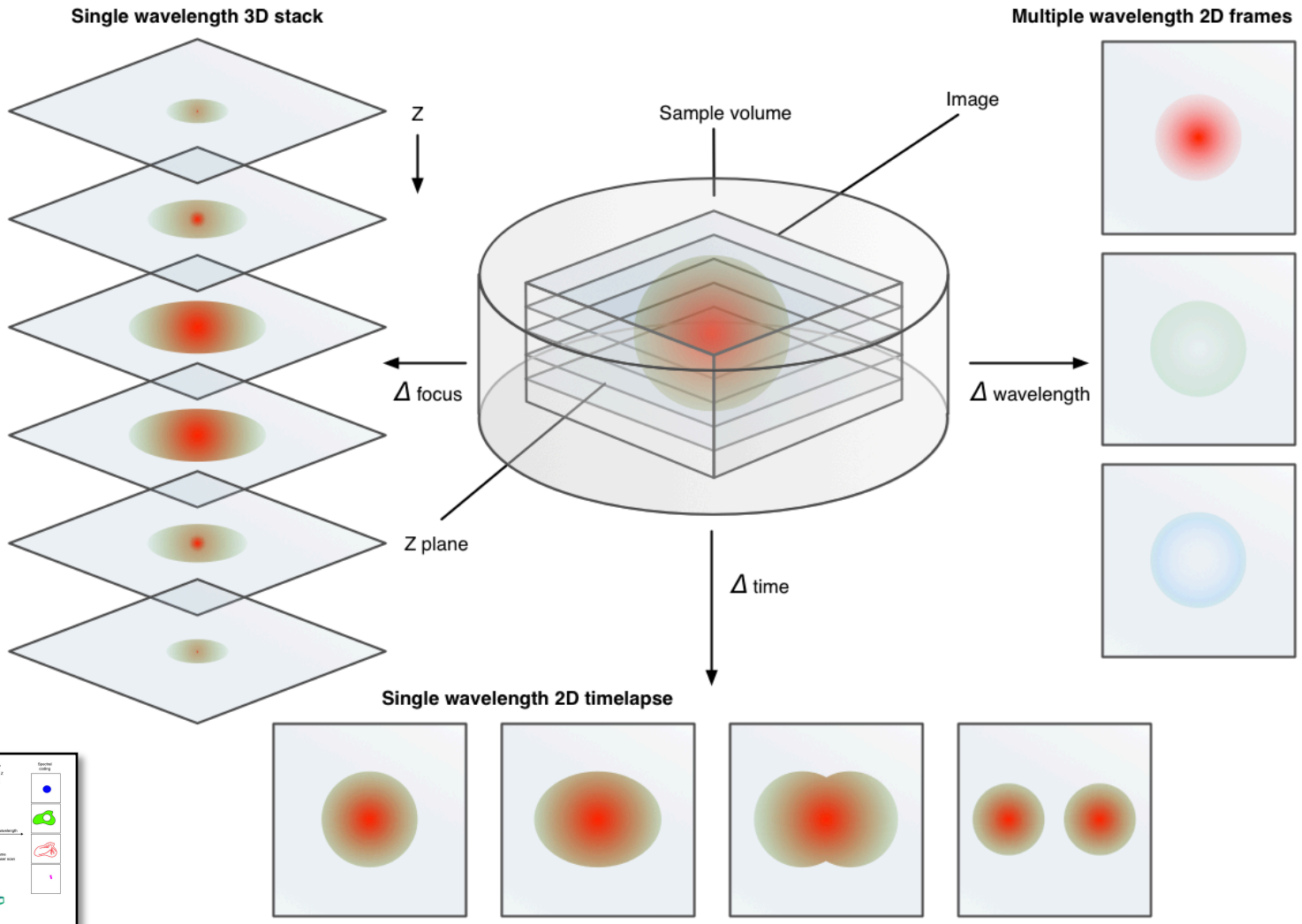
- OME definition of `Image`
- File-formats
  - Current storage
  - Challenges, etc.
- Object storage
  - Expectations
  - Cost/Value
  - Requirements

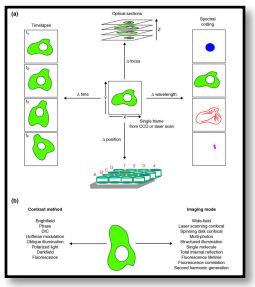
IMAGES

OME Strategy



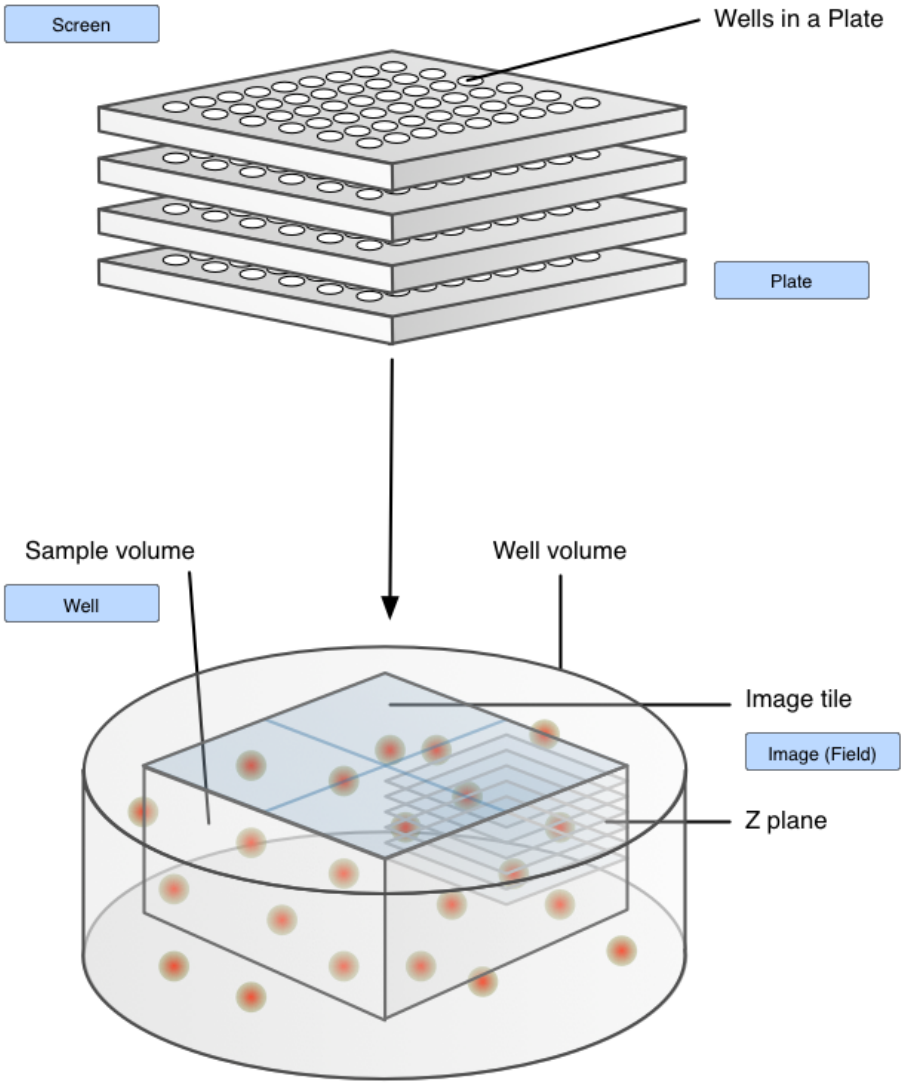
Spatio-temporal dimensions + measurements





https://genomebiology.biomedcentral.com/articles/10.1186/gb-2005-6-5-r47

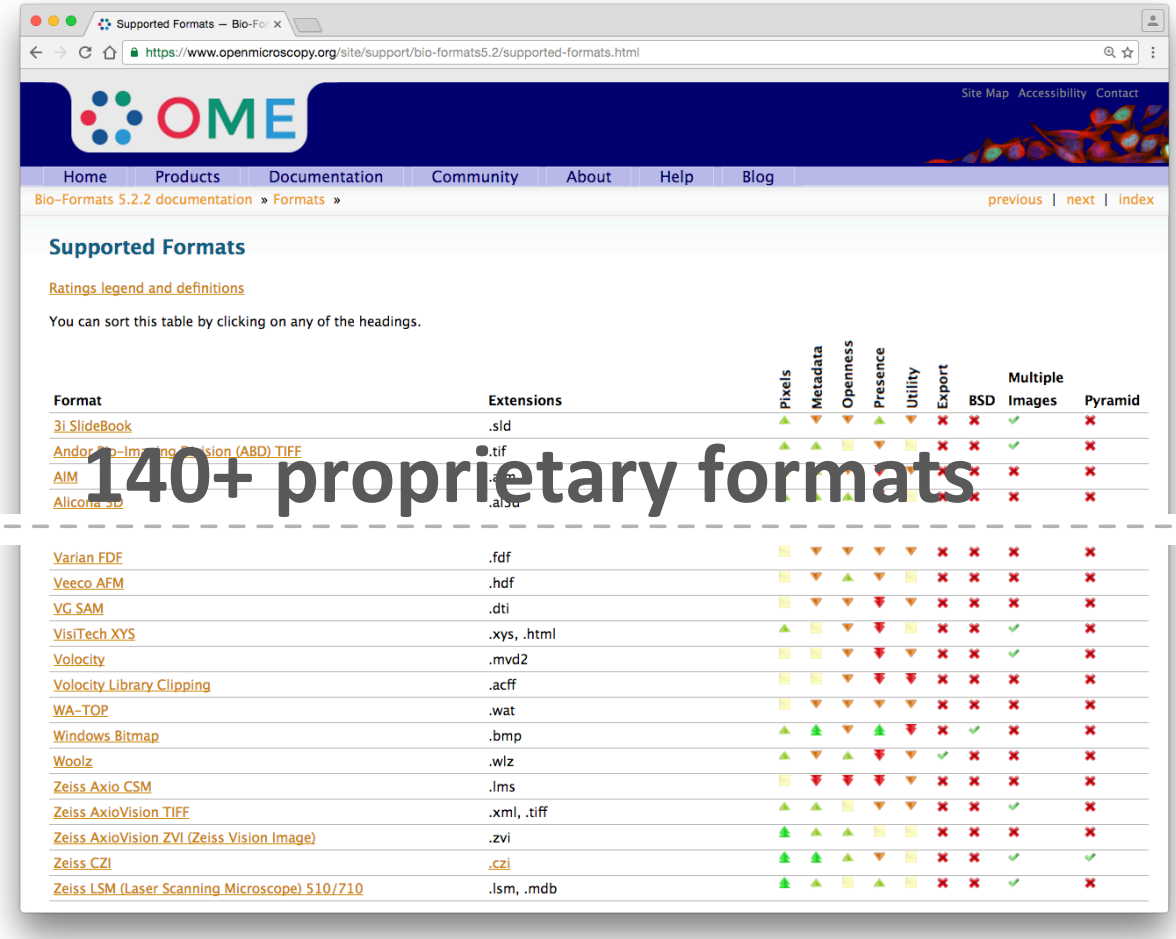
High-content screening dimensions



https://docs.openmicroscopy.org/ome-model/5.5.7/developers/screen-plate-well.html

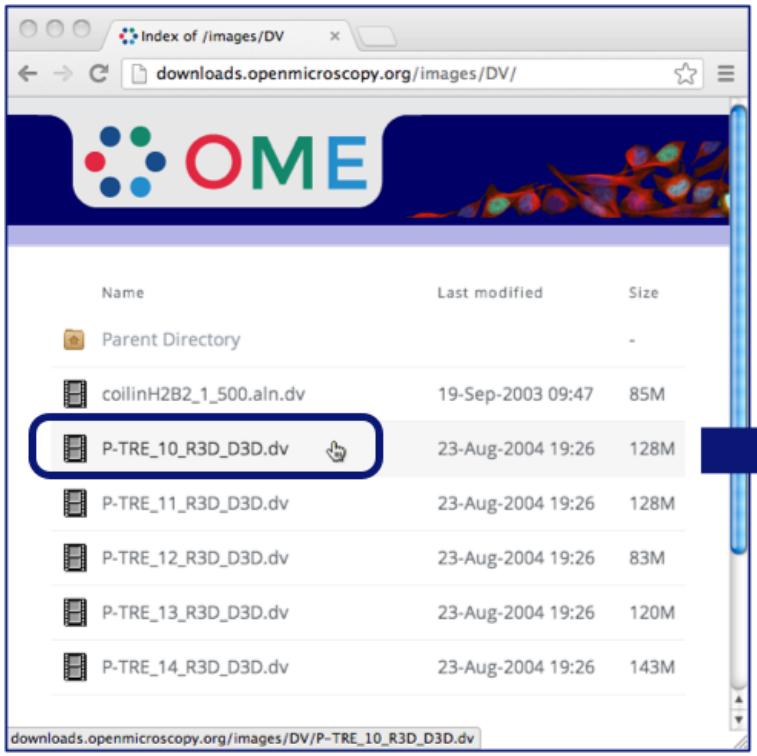
FILE-FORMATS

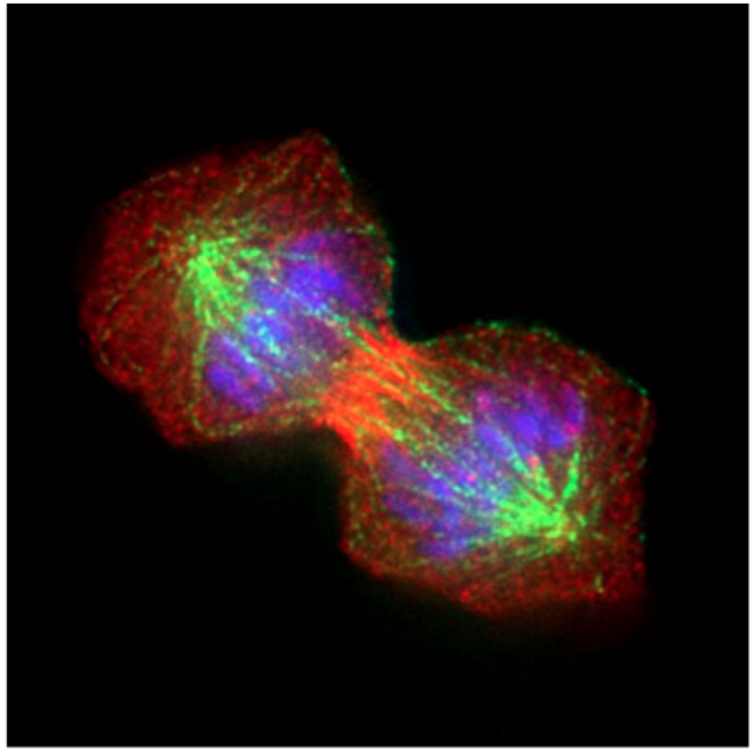
Bio-Formats



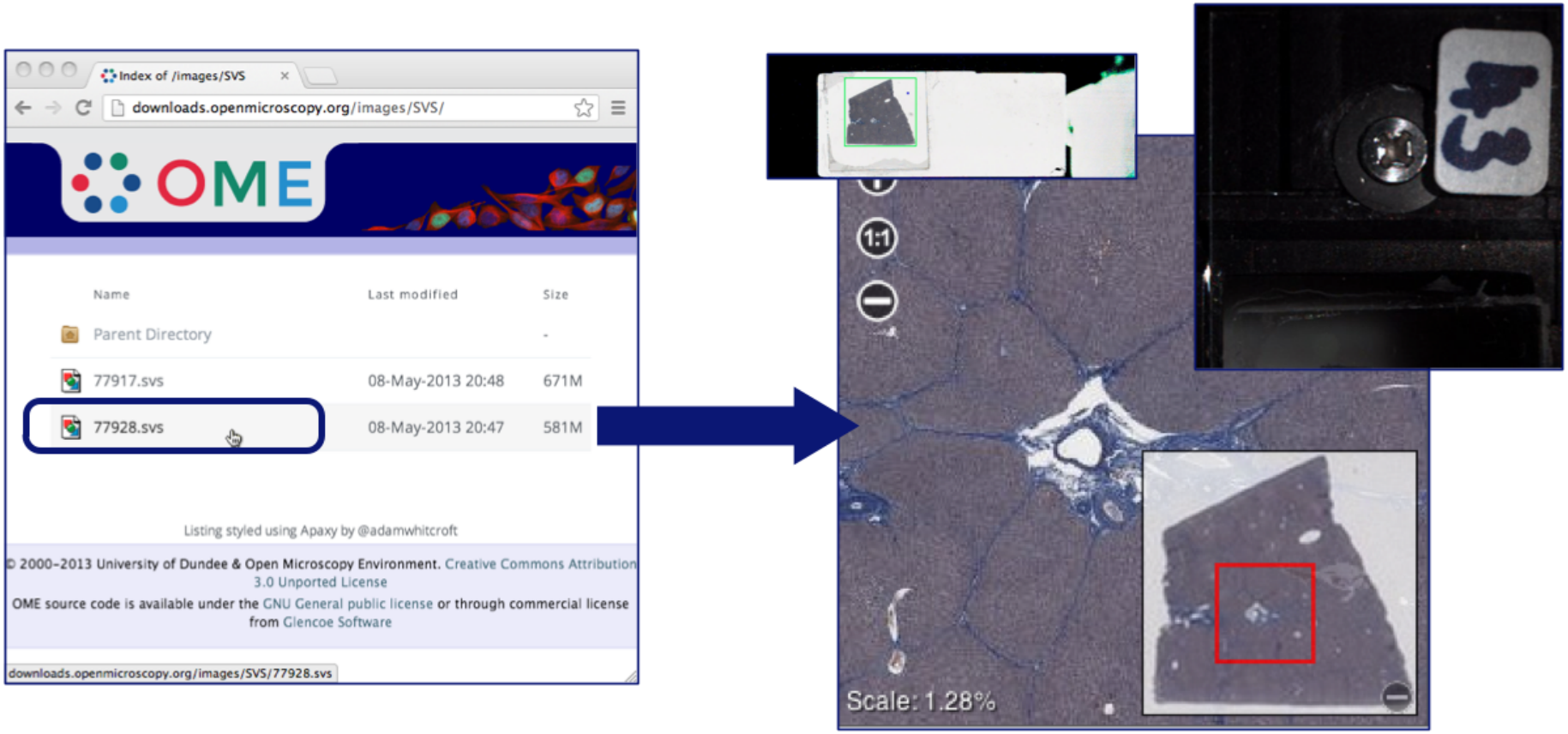
140+ proprietary formats

Filesets: 1-1


Index of /images/DV
downloads.openmicroscopy.org/images/DV/
OME
Name
Last modified
Size
Parent Directory -
coilinH2B2_1_500.aln.dv 19-Sep-2003 09:47 85M
P-TRE_10_R3D_D3D.dv 23-Aug-2004 19:26 128M
P-TRE_11_R3D_D3D.dv 23-Aug-2004 19:26 128M
P-TRE_12_R3D_D3D.dv 23-Aug-2004 19:26 83M
P-TRE_13_R3D_D3D.dv 23-Aug-2004 19:26 120M
P-TRE_14_R3D_D3D.dv 23-Aug-2004 19:26 143M
downloads.openmicroscopy.org/images/DV/P-TRE_10_R3D_D3D.dv



Filesets: 1-N


Index of /images/SVS
downloads.openmicroscopy.org/images/SVS/
OME
Name
Last modified
Size
Parent Directory
-
77917.svs
08-May-2013 20:48
671M
77928.svs
08-May-2013 20:47
581M
Listing styled using Apaxy by @adamwhitcroft
© 2000–2013 University of Dundee & Open Microscopy Environment. Creative Commons Attribution 3.0 Unported License
OME source code is available under the GNU General public license or through commercial license from Glencoe Software
downloads.openmicroscopy.org/images/SVS/77928.svs
1:1
Scale: 1.28%

Filesets: N-N


Index of /images/HCS/Op
downloads.openmicroscopy.org/images/HCS/Operetta/59548/96015bda-454a-4...
OME
Name
Last modified
Size
Parent Directory
-
r01c01-0565849973.tiff.gz 03-Apr-2013 09:17 2.0M
r01c01-0816390759.tiff.gz 03-Apr-2013 09:17 1.9M
r01c01-0991679013.tiff.gz 03-Apr-2013 09:17 1.7M
r01c02-0203004369.tiff.gz 03-Apr-2013 09:17 2.0M
r01c02-1384679954.tiff.gz 03-Apr-2013 09:17 2.0M
r01c02-1749148666.tiff.gz 03-Apr-2013 09:17 1.7M
r01c03-0847230653.tiff.gz 03-Apr-2013 09:19 2.1M
r01c03-1032313189.tiff.gz 03-Apr-2013 09:19 1.7M
r01c03-1985691398.tiff.gz 03-Apr-2013 09:19 2.2M
r01c04-0512157530.tiff.gz 03-Apr-2013 09:19 2.2M
r01c04-1271778531.tiff.gz 03-Apr-2013 09:19 1.7M
r01c04-1630181344.tiff.gz 03-Apr-2013 09:19 2.2M
r01c05-0006079694.tiff.gz 03-Apr-2013 09:21 2.2M
r01c05-0175927353.tiff.gz 03-Apr-2013 09:21 1.7M
r01c05-2128047606.tiff.gz 03-Apr-2013 09:21 2.2M

Fileset Diversity

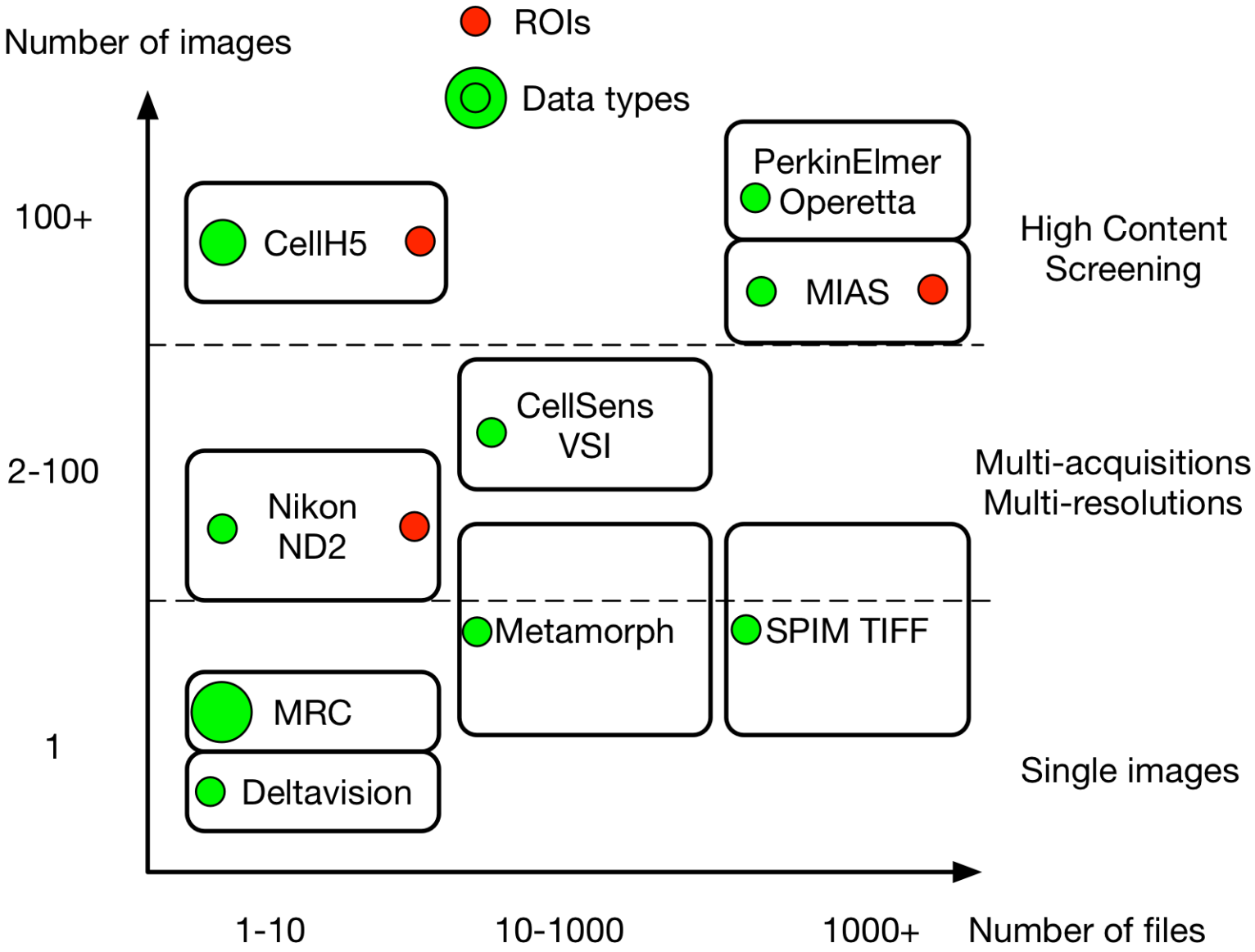
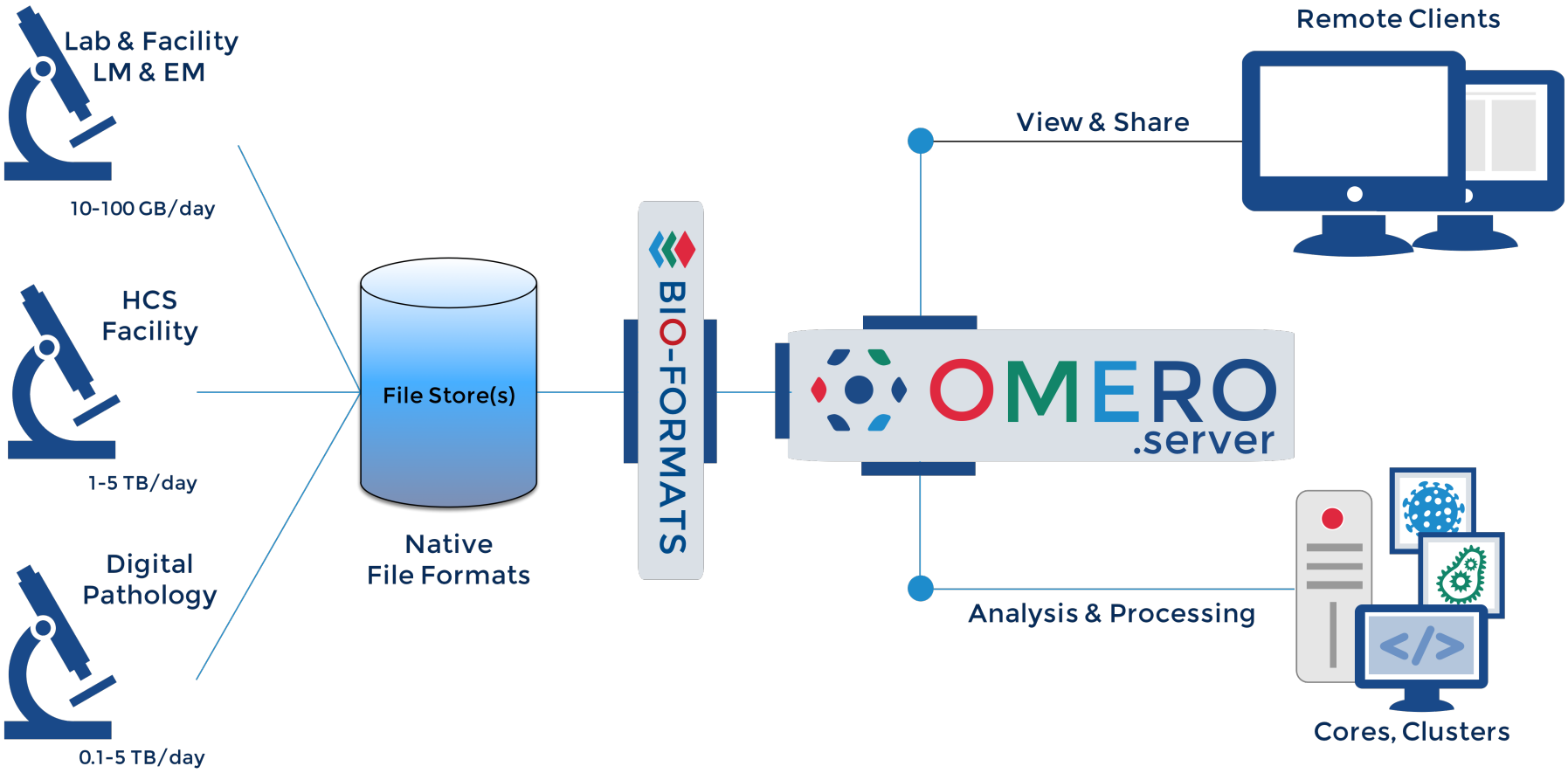
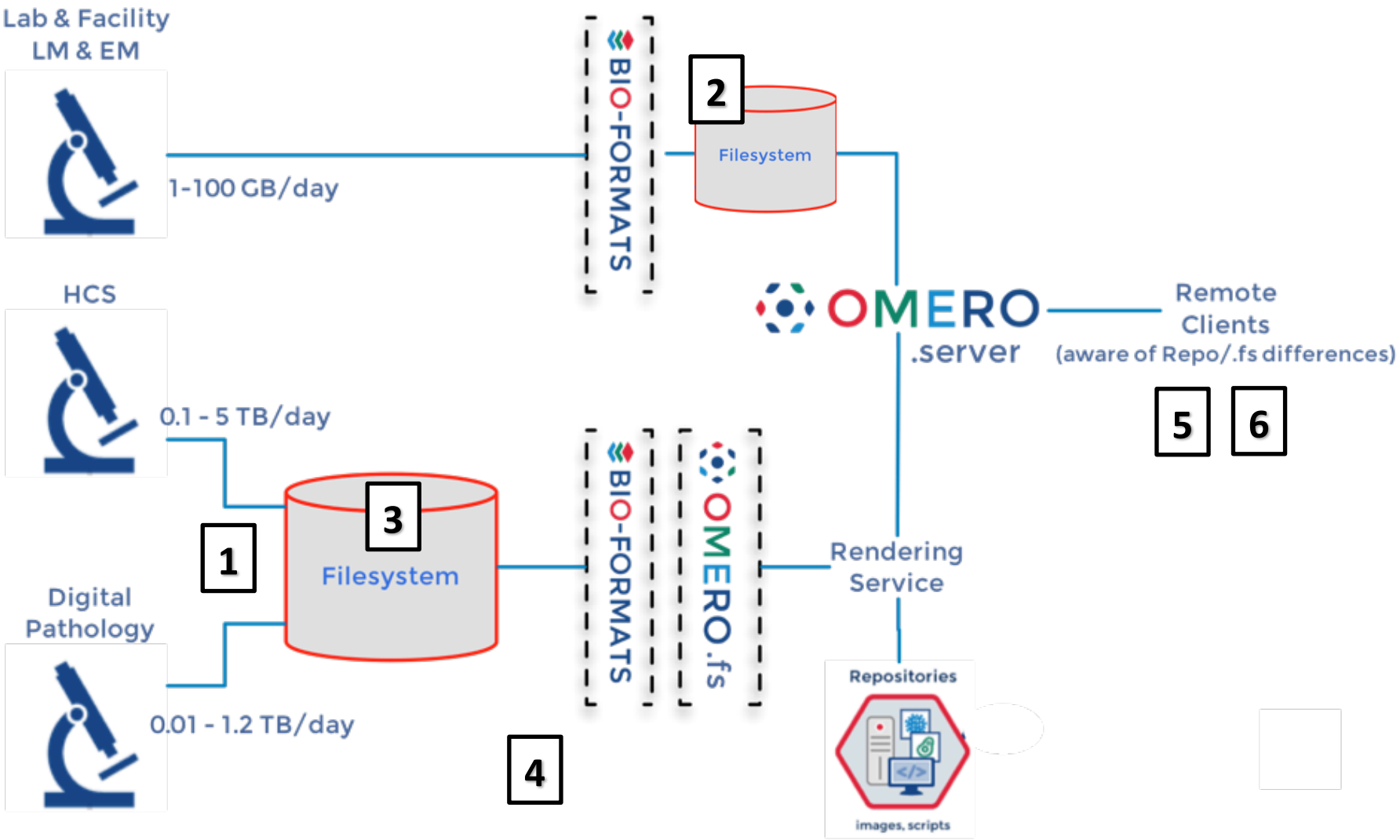


Image Data Access: Today

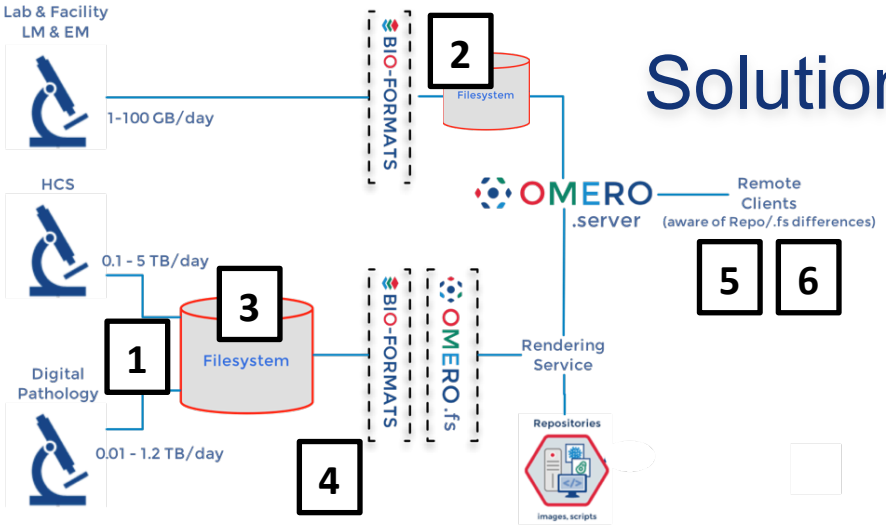


OBJECT-STORES

Image Data Access: Past & Future



Solution Trade-offs



		Scope	Latency	Storage	Value	ETA
1	Filesystem cache	Backend cache	Slow once	>1x	Fast & simple	Straight-forward once defined
2	Direct storage (“ROMIO”)	Swappable backend	Slow **always**	1	Space-efficient	Requires new import strategy
3	Reader/Writer	Supported format	Slow **always**	1	Stability	**Longest; requires spec**
4	Plane/chunk cache	Flexible layer	Slow once	>1x	Can be combined	Relatively simple
5	Client loads from S3	Full S3 integration	*Potentially fast*	1	Most “modern”	Requires significant changes
6	Client caches to S3	Frontend cache	Slow once	>1x	Trans-parent	Simple

Questions

- Will the object store be the primary storage?
  - Assuming: yes
- Will the files be deleted?
  - Assuming: yes
- Will the storage be file-based?
  - Strongly urge: **no**
- Will multi-resolutions be stored/generated?
  - Assuming: yes
- Will orthogonal views be needed? 2D needed?
  - Probably, but what's the default? e.g. XYZCT
- Will multi-fileset correlation be needed?

Last thoughts

- Caching as alternate views
- Diversity of acquisition, usage & domains
- Cost of proprietary formats
- Single API vs. maintenance burden
- Relationship to file-based solutions (HDF5)

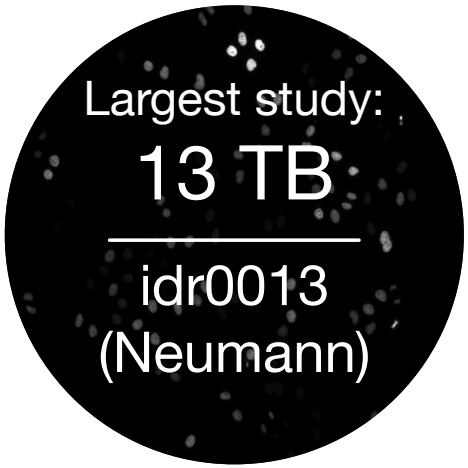
THANKS

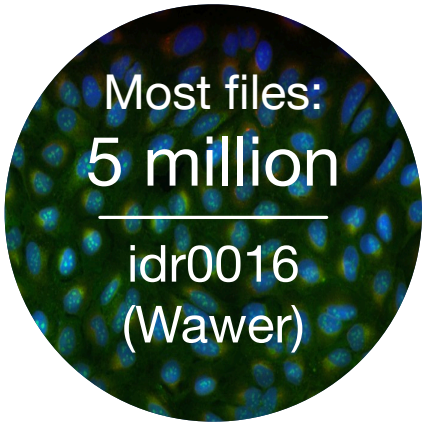
EXTRA SLIDES

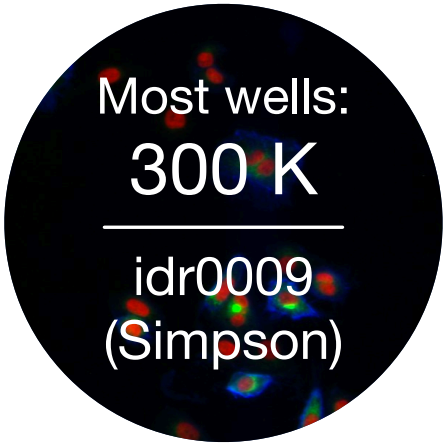
How to Source Data

IDR

Fileset challenges



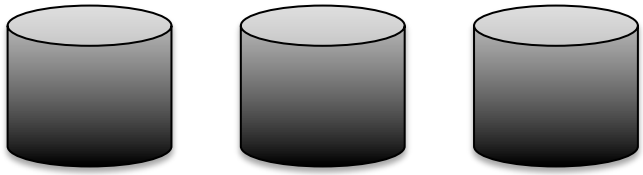




Reference images

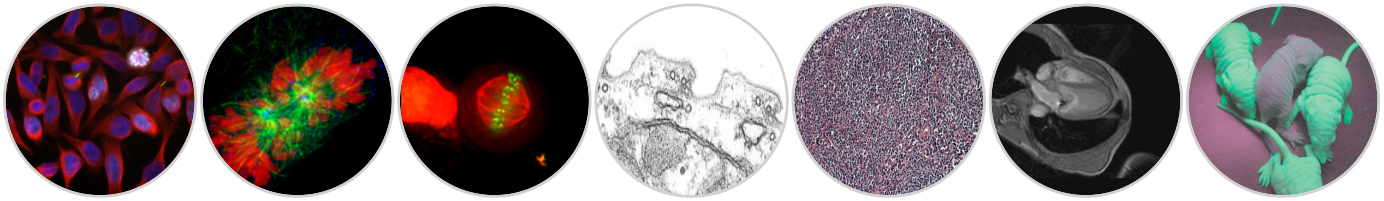
IDR

Biomolecular Resources



Genes Compounds Cell Lines

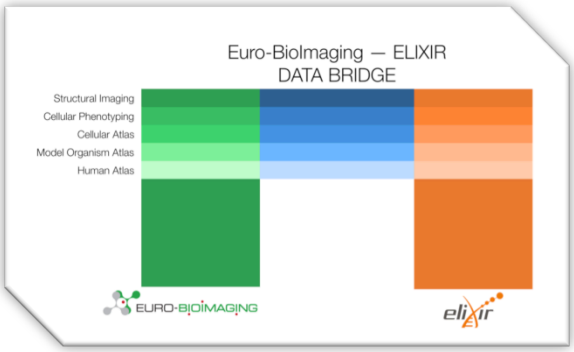
Imaging Domains



Controlled Vocabularies

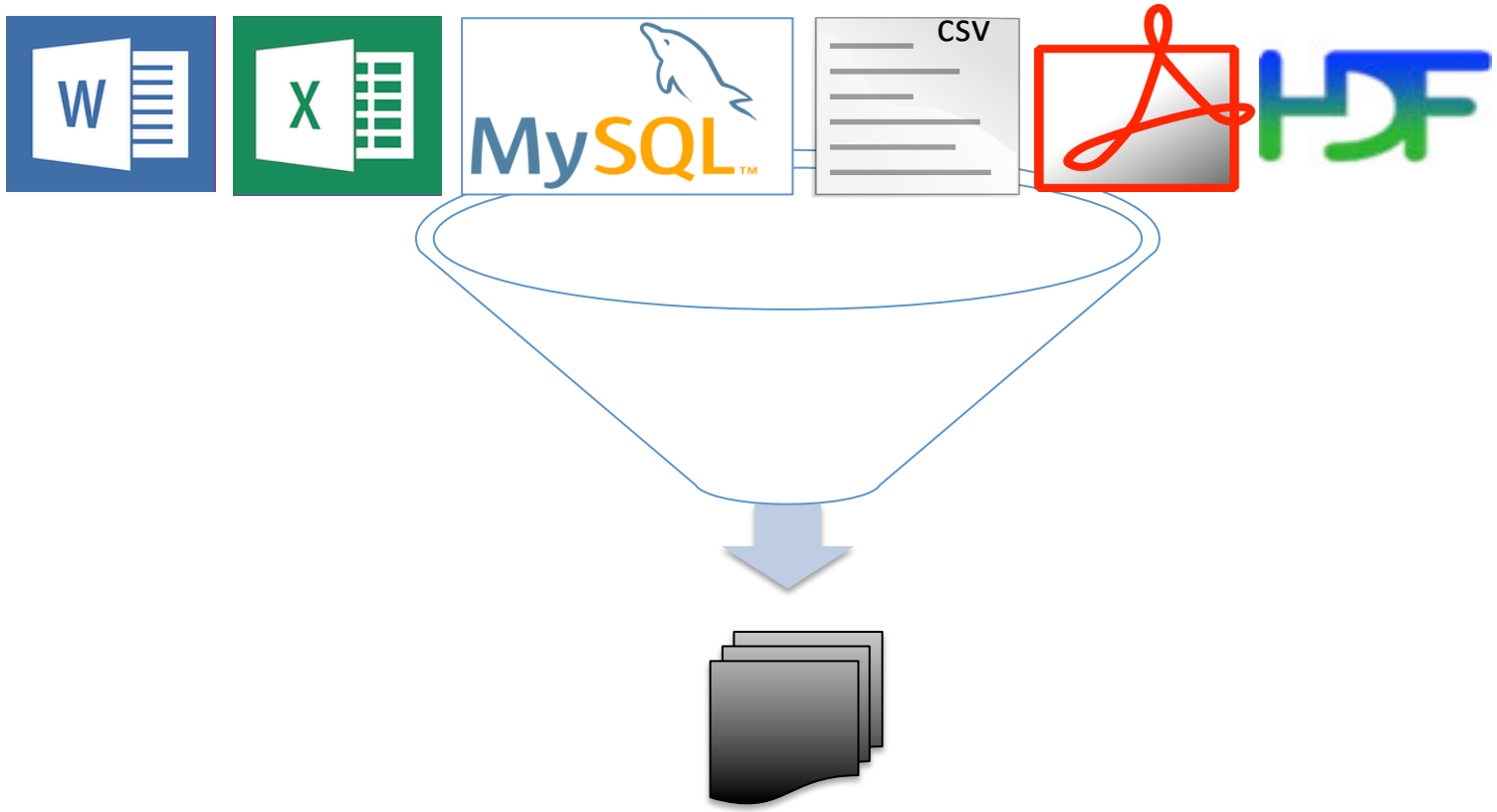
Phenotypes

Ground Truth



Google: “Euro-BioImaging/Elixir Data Strategy”

Submitted metadata



In the style of: MAGE-TAB isatab *or,* Cellular Phenotype Database

http://github.com/IDR/idr-metadata

Vital Stats: IDR @ EBI Embassy

IDR

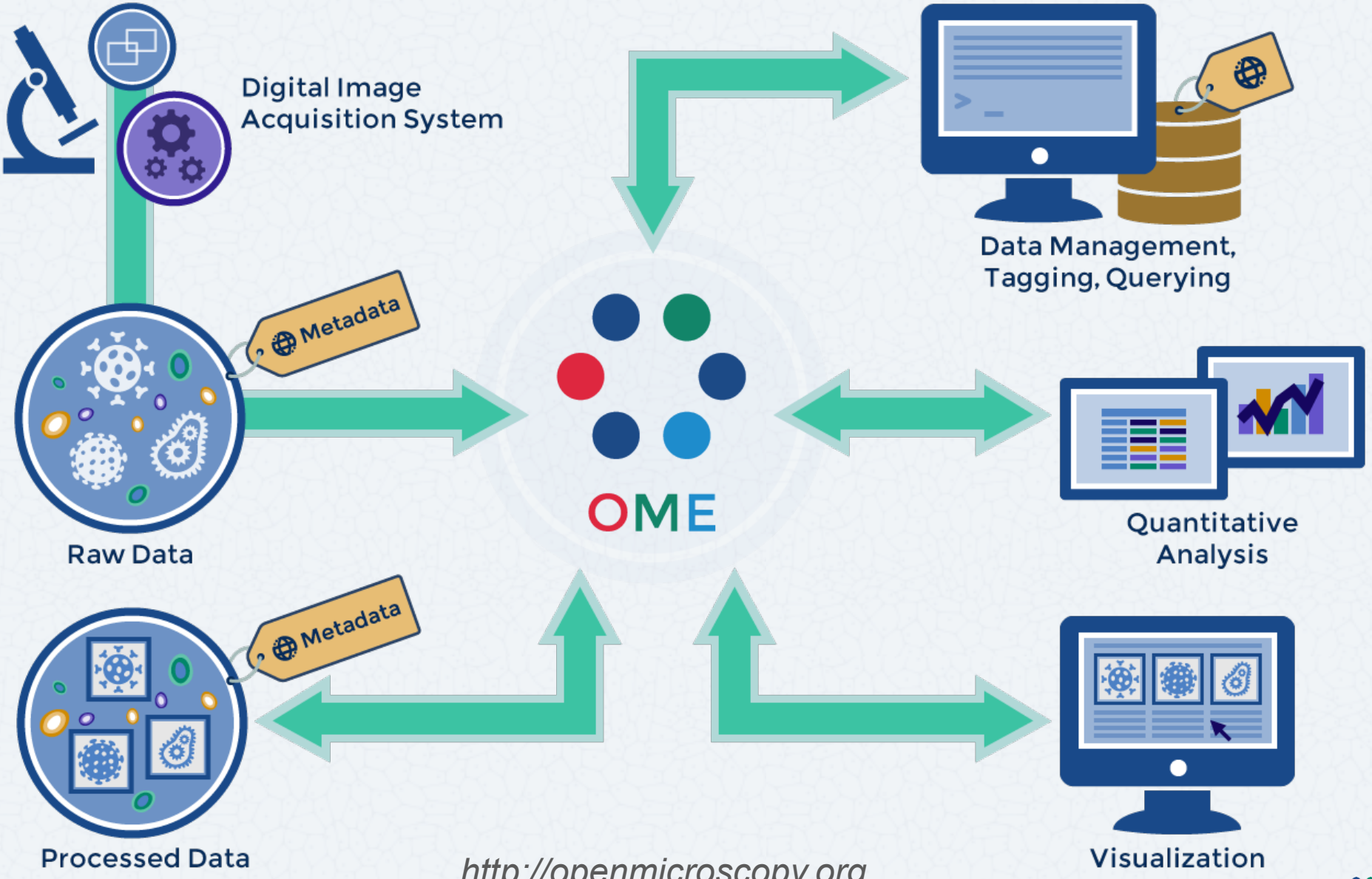
http://idr.openmicroscopy.org

Metric	Sept 2017
Image Data Size	42.8 TB
Image files	14.6 M
Datasets	3805
Total Images	2.68 M
Planes	37.3 M
Experiments	1.04M
Genes	19,605
Annotated Images	386 k
Phenotypic Classes	161
External Links	172 k

→~$25,000/month on AWS←

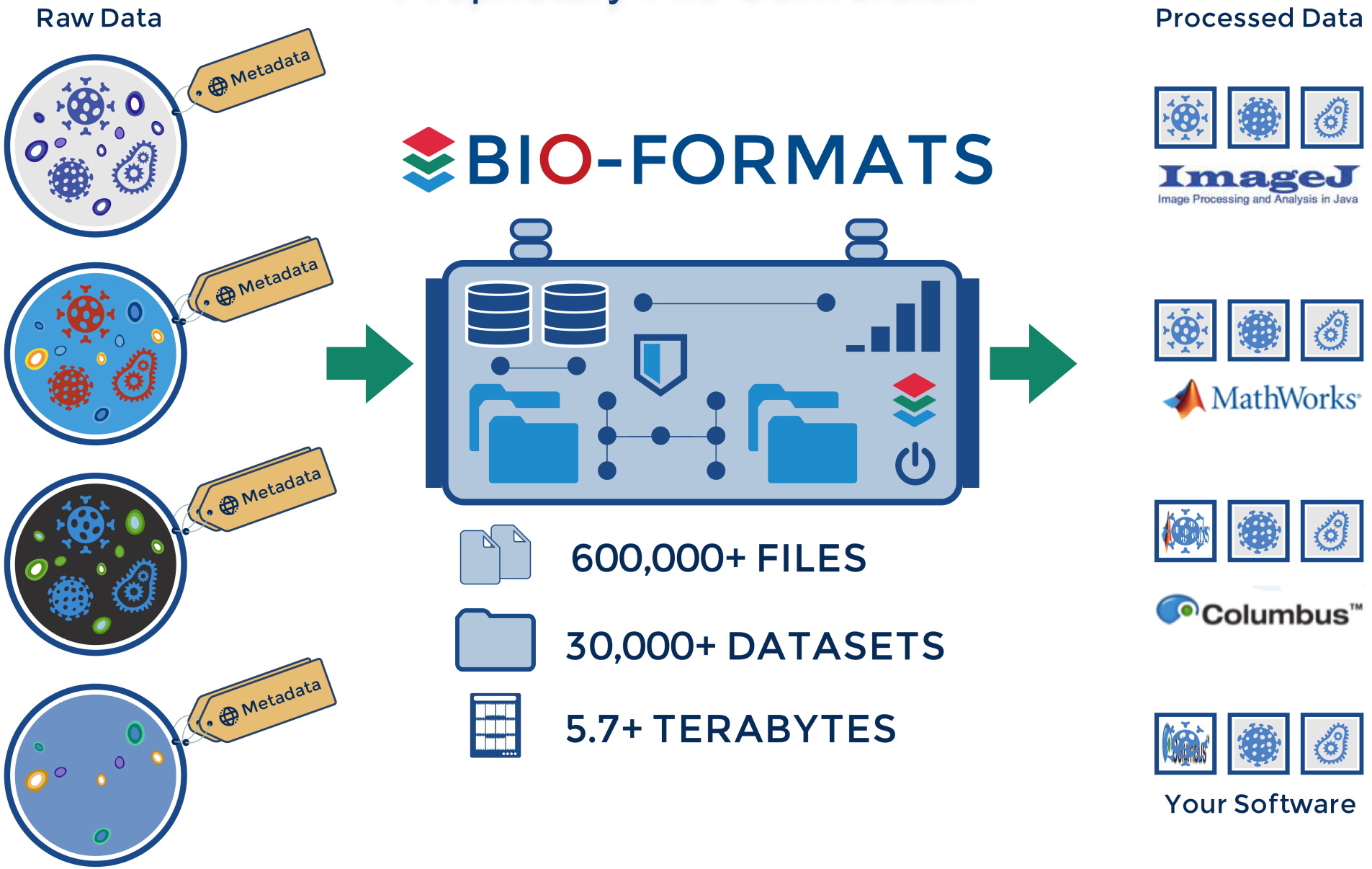
Williams et al (2017) Nature Methods

…Towards Image Informatics



http://openmicroscopy.org

BIO-FORMATS:
Proprietary File Conversion



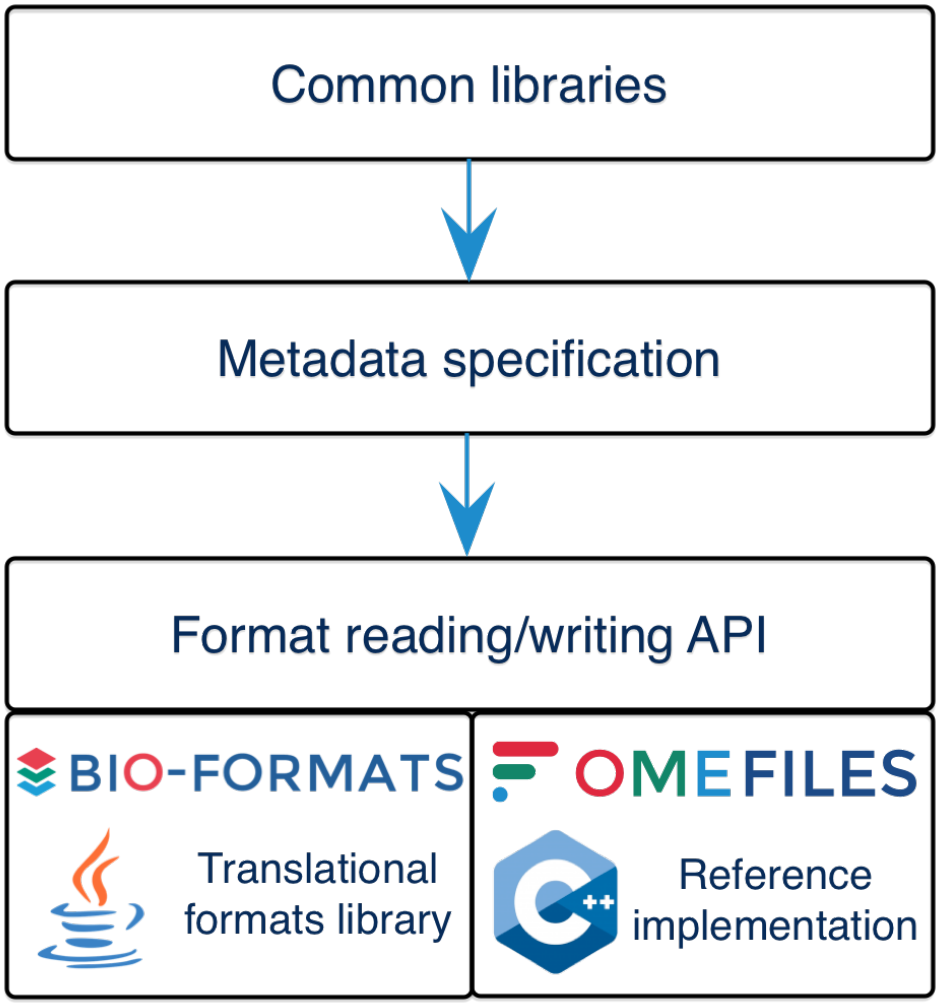
→ Apps using Bio-Formats were started >100,000 times per day in 2015 ←

BIO-FORMATS:
Opportunities for Academic/Commercial Collaborations

- 3i: Building and maintaining a SlideBook file format reader
- PerkinElmer: Commissioning Glencoe to provide open source Harmony HCS format reader
- ZEISS: Commissioning Glencoe to build open source JPEG-XR decoder

→ *See blog.openmicroscopy.org for more info*

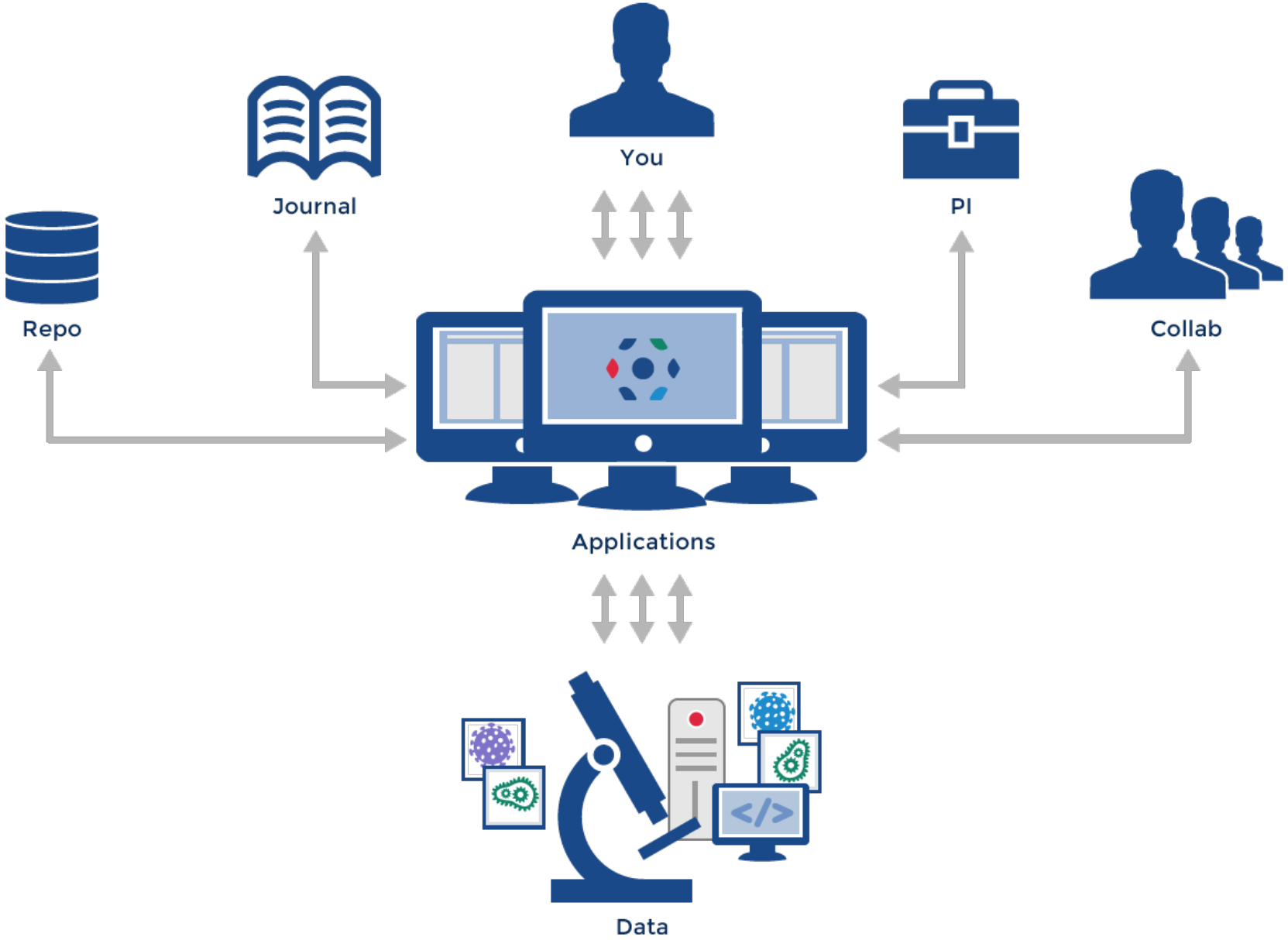
Software Stack



https://docs.openmicroscopy.org/ome-model/5.5.7/developers/screen-plate-well.html

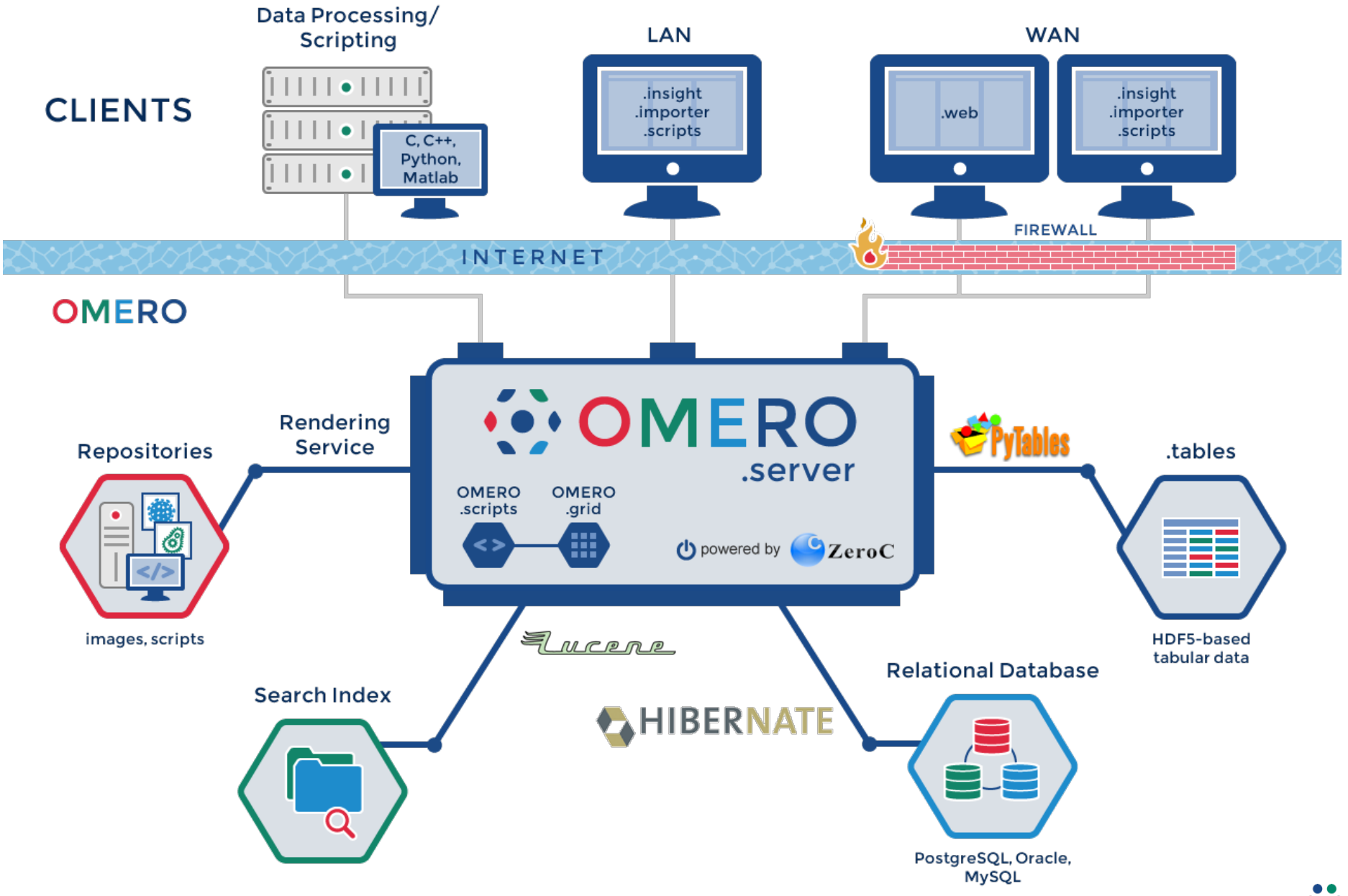
The Standard Paradigm

The “Scientific Data” Paradigm



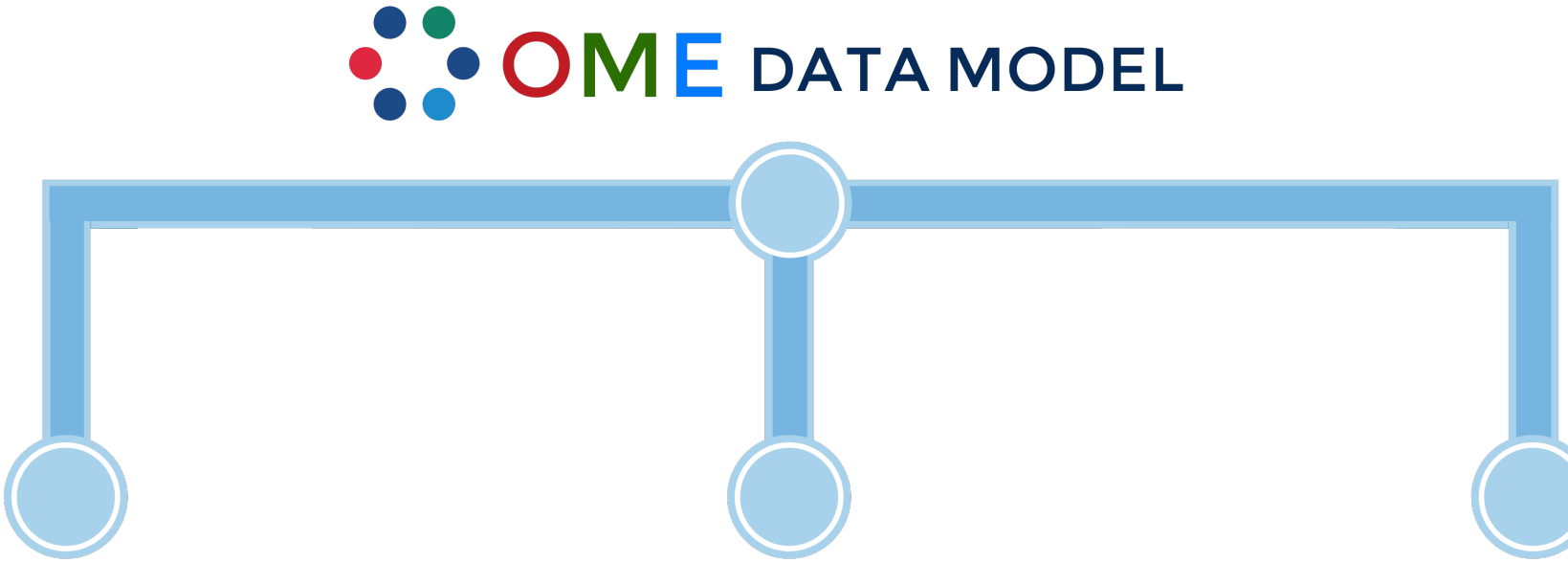
Gray et al, 2005, Scientific Data Management in the Coming Decade, Microsoft Research

The OMERO Platform

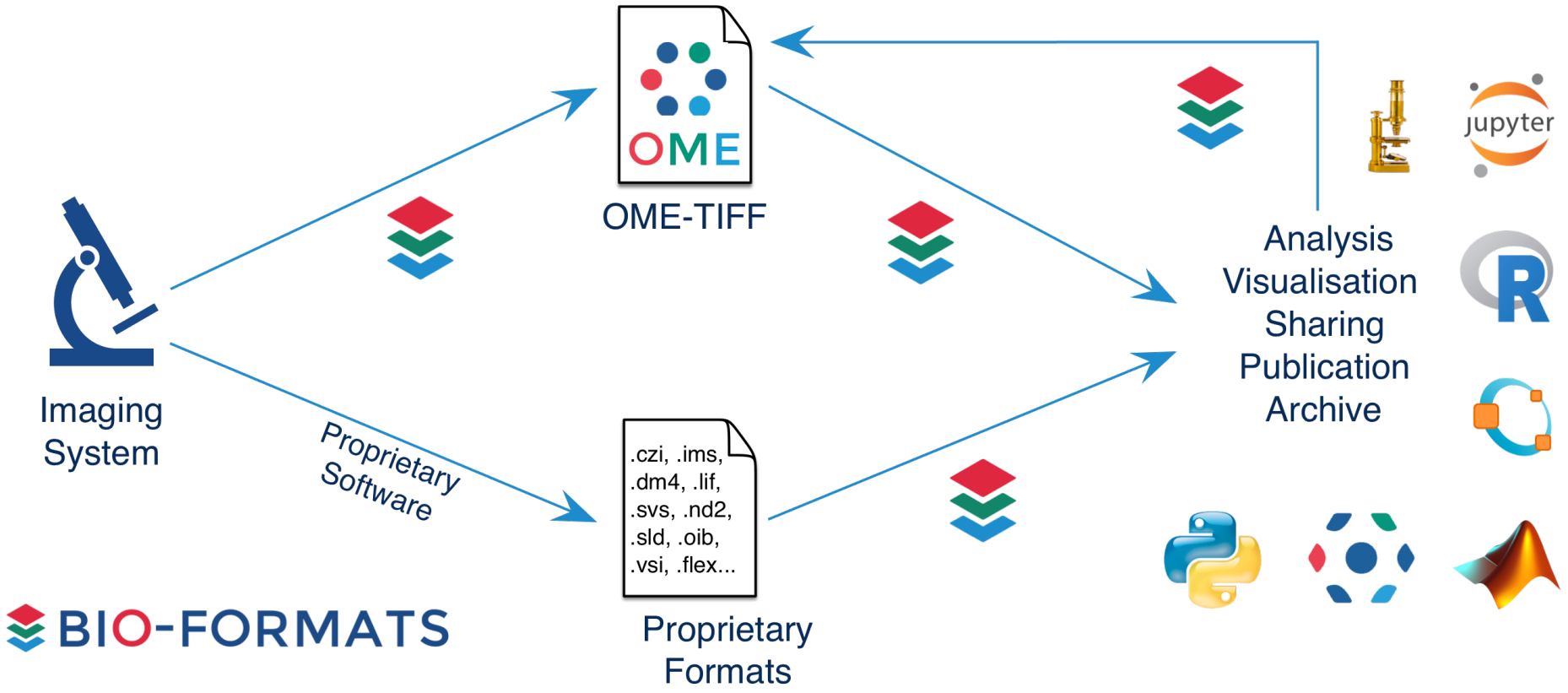


Allan et al, 2012, Nature Methods

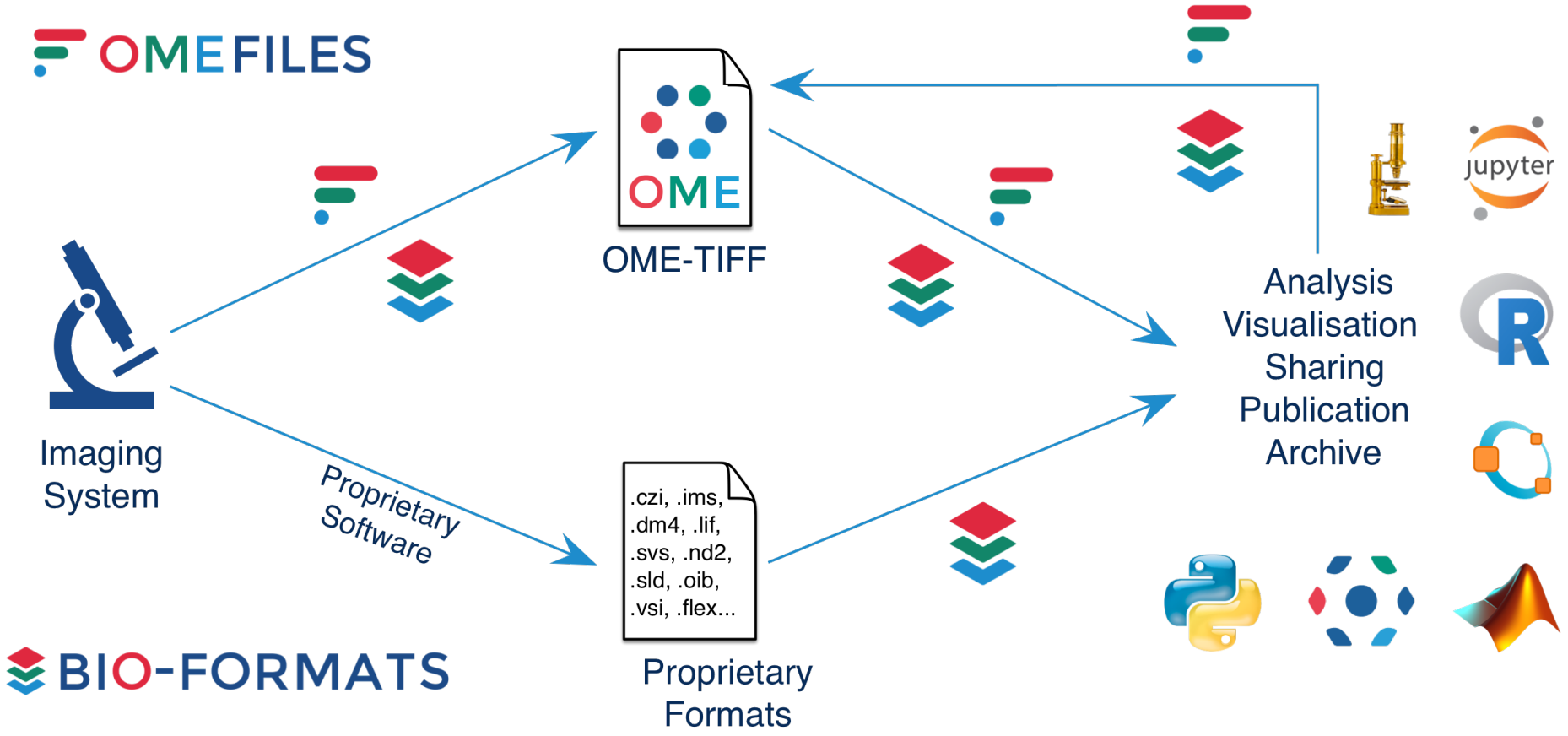
What We Do


OME DATA MODEL
OME-XML
OME-TIFF
OME FILES
Open exchangeable file formats & metadata
BIO-FORMATS
Open Proprietary File Format Translation
OMERO
Open Image Data Management

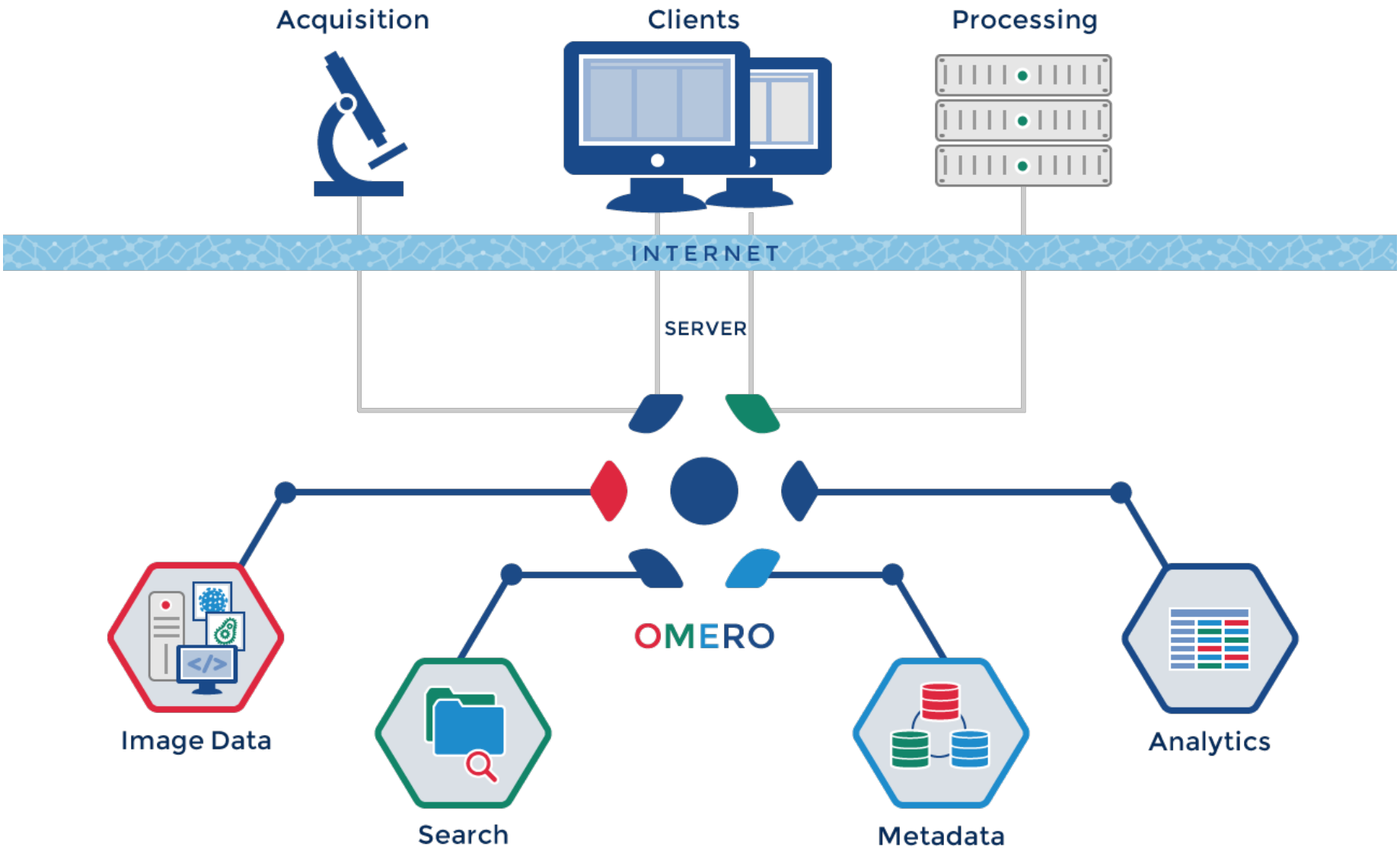
Bio-Formats


OME
OME-TIFF
Imaging System
Proprietary Software
.czi, .ims, .dm4, .lif, .svs, .nd2, .sld, .oib, .vsi, .flex...
Proprietary Formats
Analysis
Visualisation
Sharing
Publication
Archive
jupyter
BIO-FORMATS

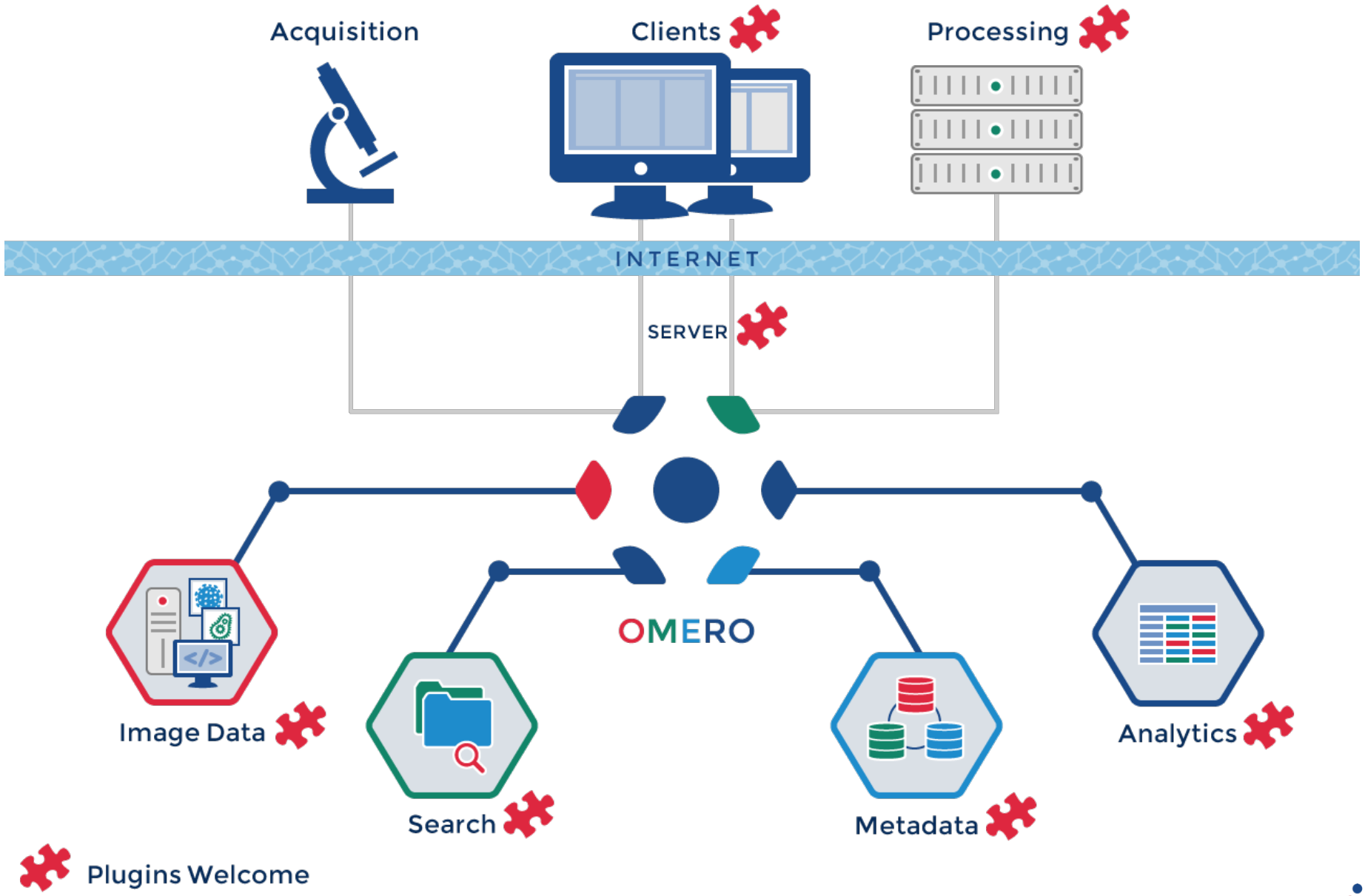
Bio-Formats


OMEFILES
OME
OME-TIFF
Imaging System
Proprietary Software
.czi, .ims, .dm4, .lif, .svs, .nd2, .sld, .oib, .vsi, .flex...
Proprietary Formats
Analysis
Visualisation
Sharing
Publication
Archive
jupyter
BIO-FORMATS

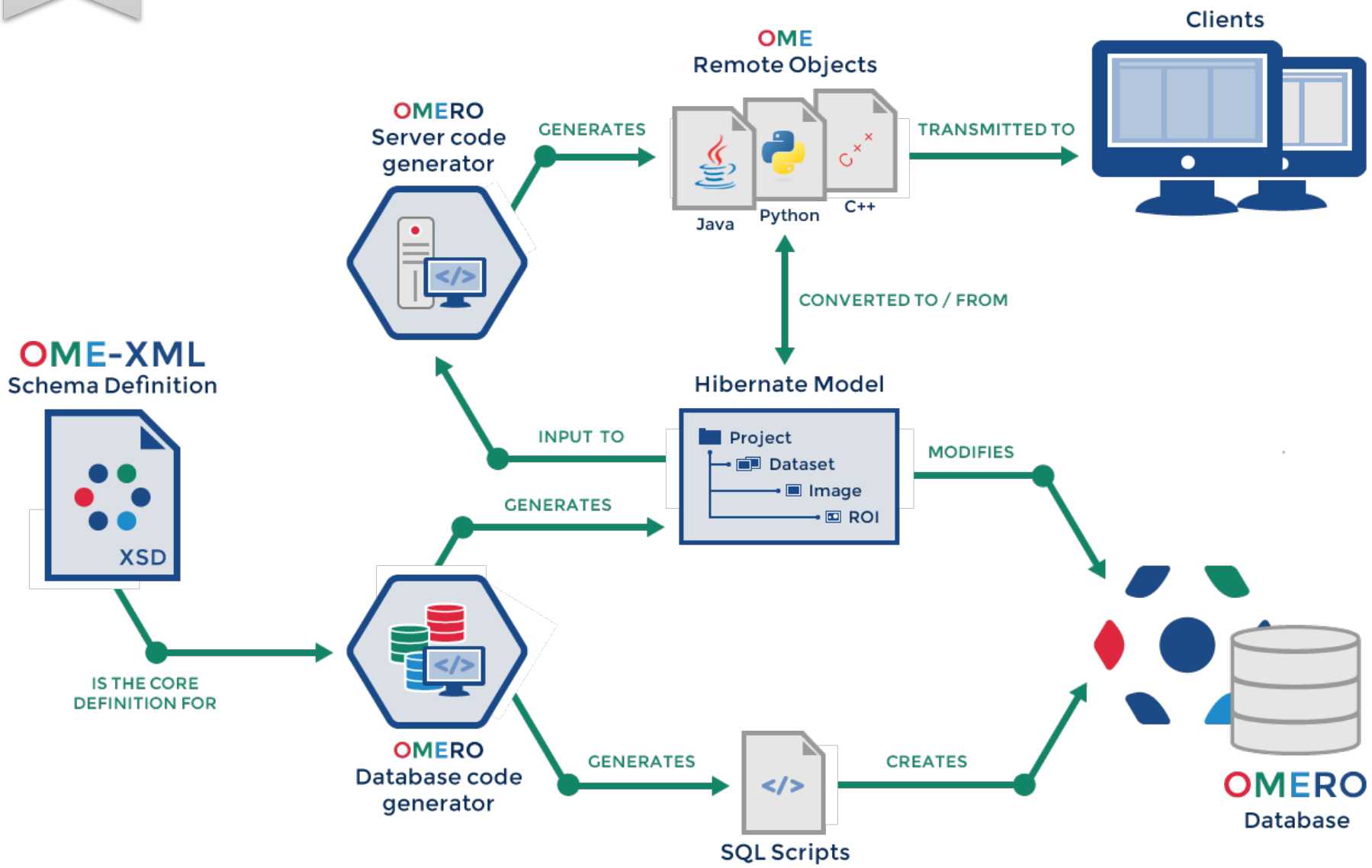
The OMERO Platform



The *Extensible* OMERO Platform



Leveraging the OME Data Model

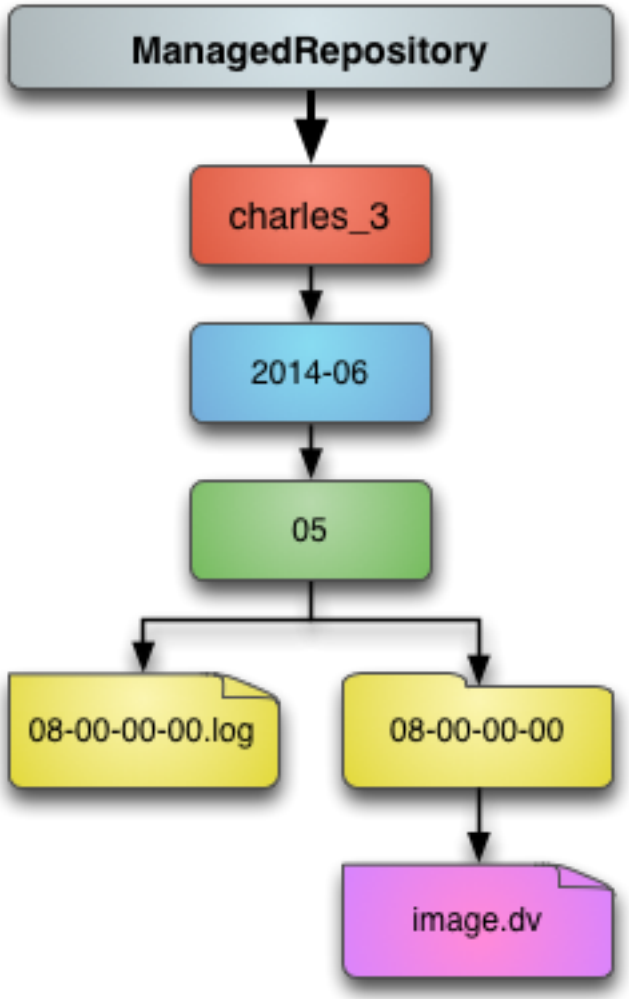


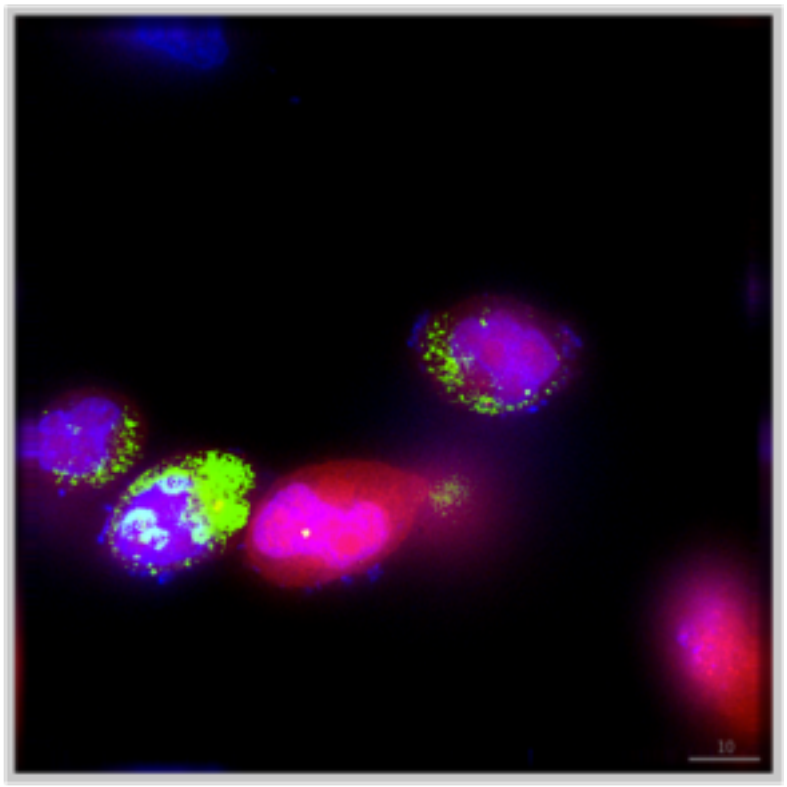
Google: ‘Extending OMERO’

Security Model

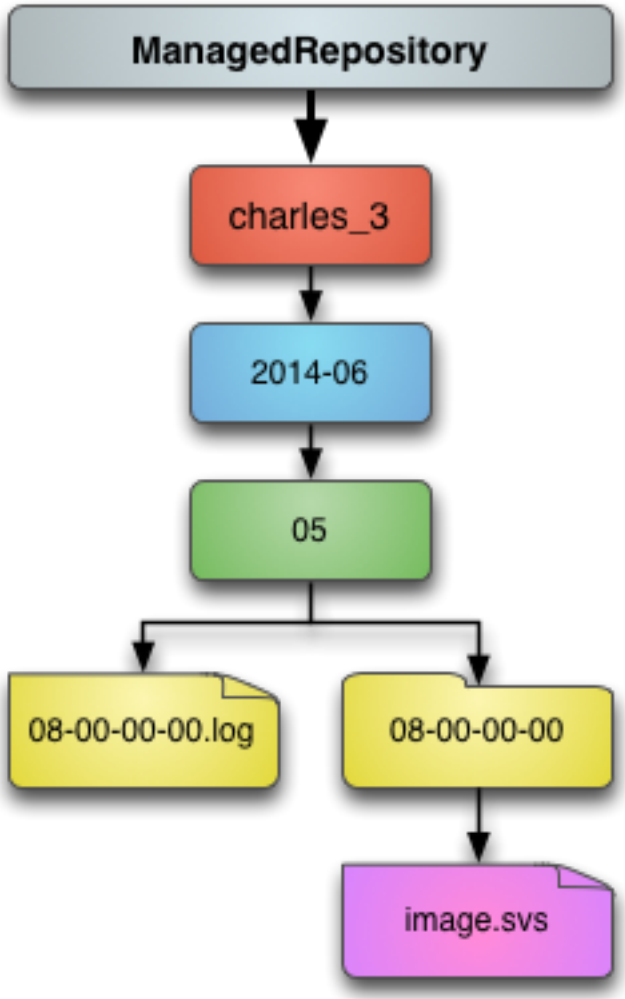
PERMISSIONS	Read	Annotate	Write	Privacy
Private	✓	✓	✓	You
Group-Read	✓	✗	✗	
Group-Annotate	✓	✓	✗	Your Group
Group-Write	✓	✓	✓	
Public-Read	✓	✗	✗	
Public-Annotate	✓	✓	✗	Anyone
Public-Write	✓	✓	✓	

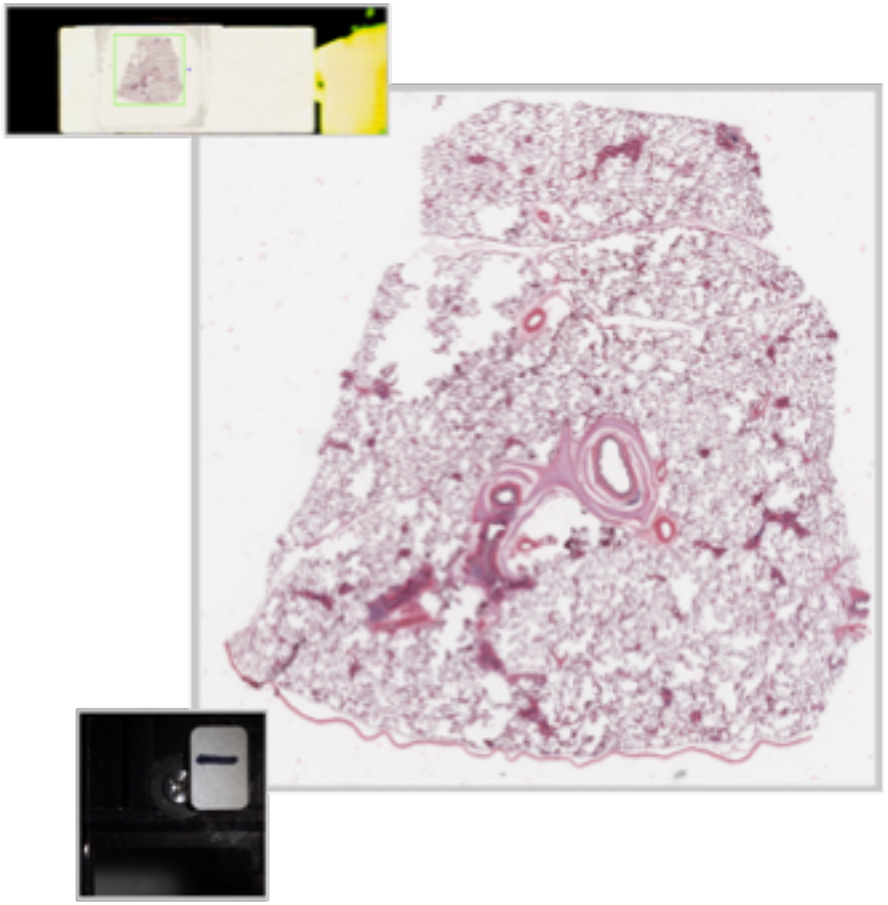
Filesets: 1-1


ManagedRepository
charles_3
2014-06
05
08-00-00-00.log
08-00-00-00
image.dv

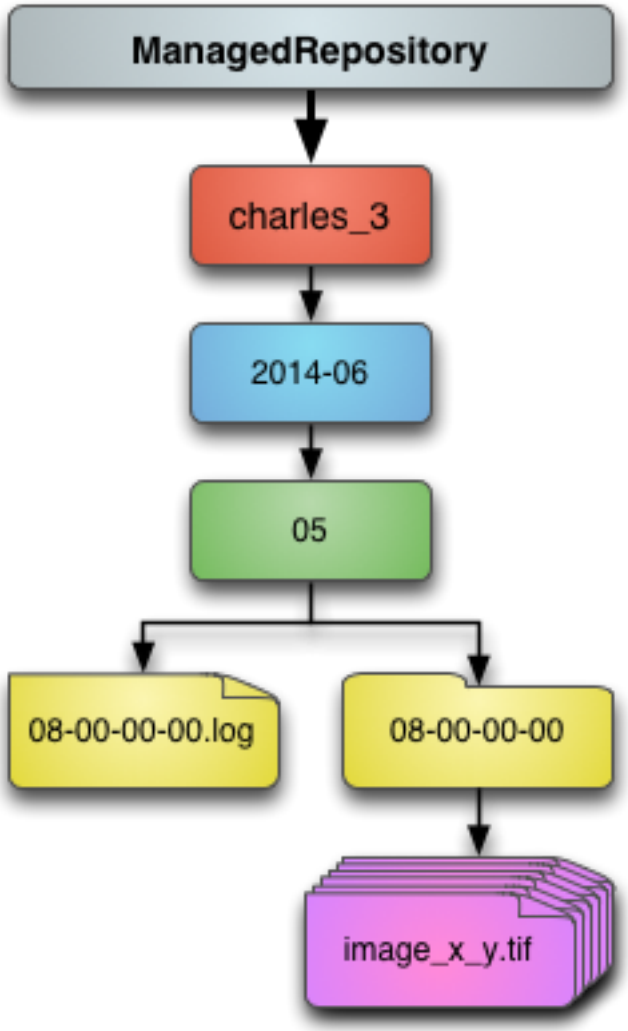


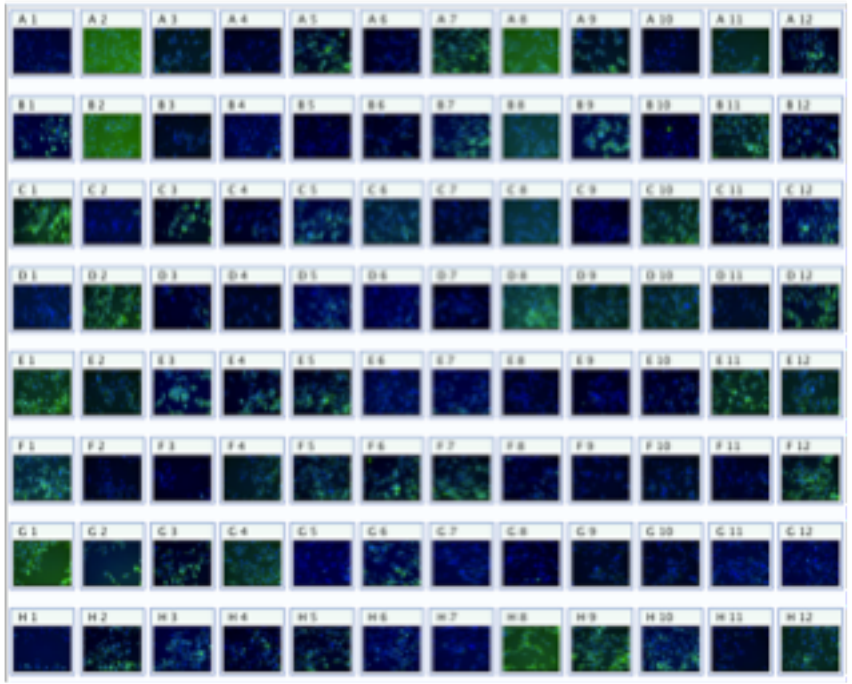
Filesets: 1-N


ManagedRepository
charles_3
2014-06
05
08-00-00-00.log
08-00-00-00
image.svs



Filesets: N-N


ManagedRepository
charles_3
2014-06
05
08-00-00-00.log
08-00-00-00
image_x_y.tif



Modulo Dimensions

```
<SA:StructuredAnnotations>
  <SA:XMLAnnotation ID="Annotation:3" Namespace="openmicroscopy.org/omero/dimension/modulo">
    <SA:Value>
      <Modulo namespace="http://www.openmicroscopy.org/Schemas/Additions/2011-09">
        <ModuloAlongZ Type="angle" Unit="degree">
          <Label>45</Label>
          <Label>90</Label>
        </ModuloAlongZ>
        <ModuloAlongT Type="lifetime" TypeDescription="TCSPC" Start="0" Step="2" End="128"/>
        <ModuloAlongC Type="phase" Start="0" Step="1" End="255"/>
      </Modulo>
    </SA:Value>
  </SA:XMLAnnotation>
</SA:StructuredAnnotations>
```

https://docs.openmicroscopy.org/ome-model/5.5.7/developers/6d-7d-and-8d-storage.html