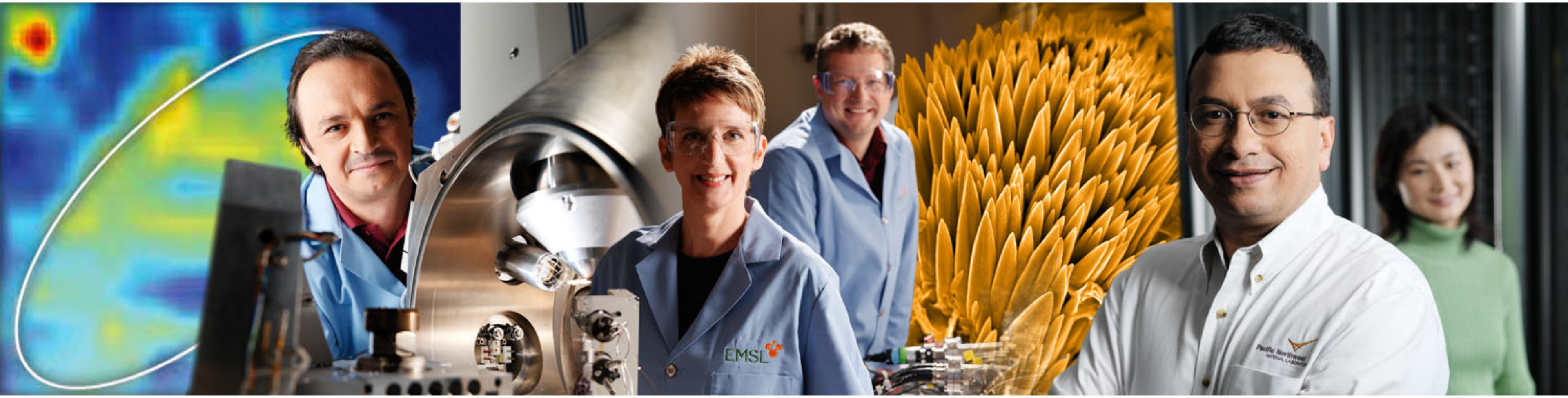
EMSL

MyEMSL & EMSLHub: Creating A flexible framework for scientific data sharing, discovery, and collaboration

PNNL-SA-96303

David Cowley



www.emsl.pnl.gov

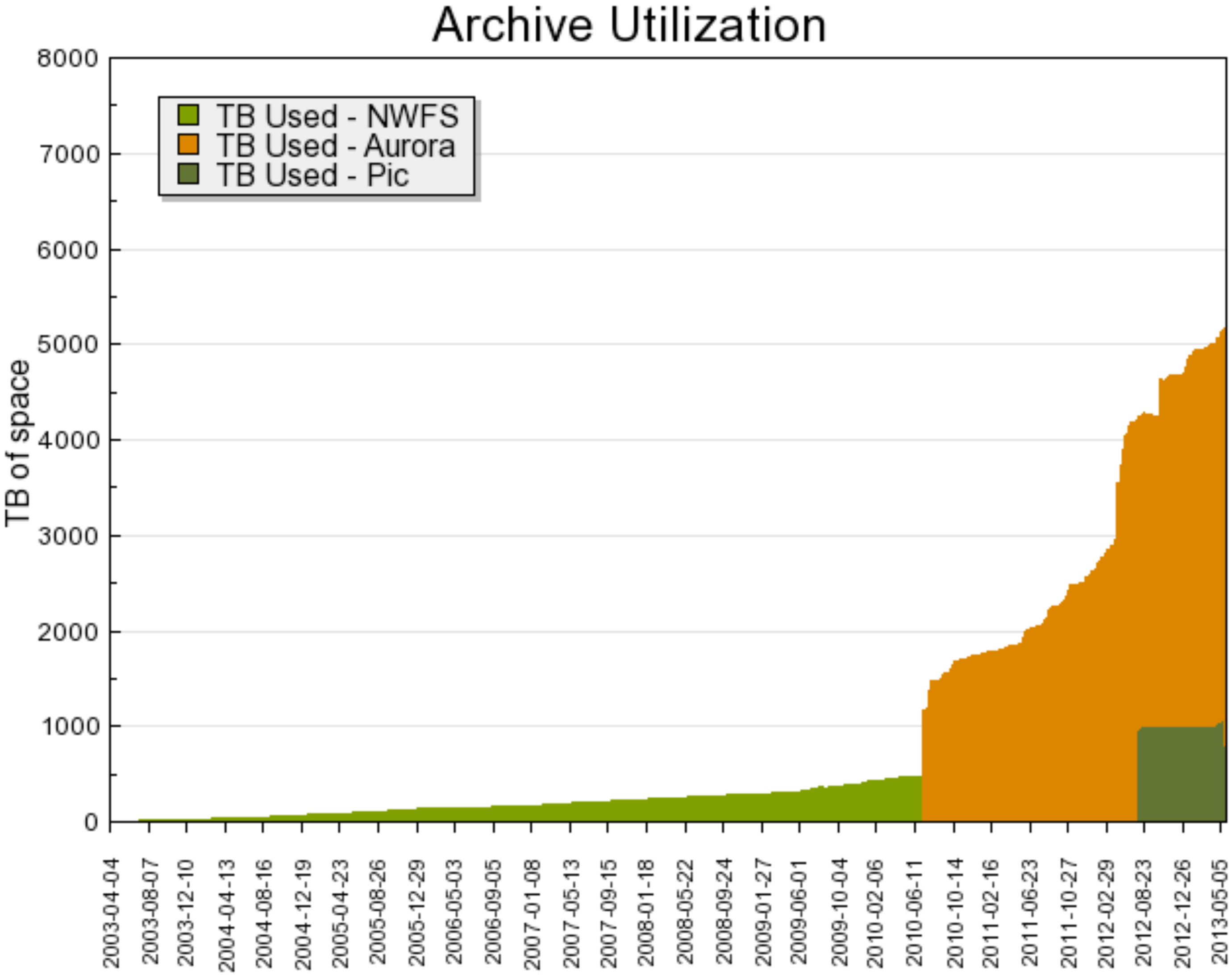
Pacific Northwest NATIONAL LABORATORY

Proudly Operated by Battelle Since 1965

U.S. DEPARTMENT OF ENERGY

Data Growth at EMSL 2002-2013

EMSL



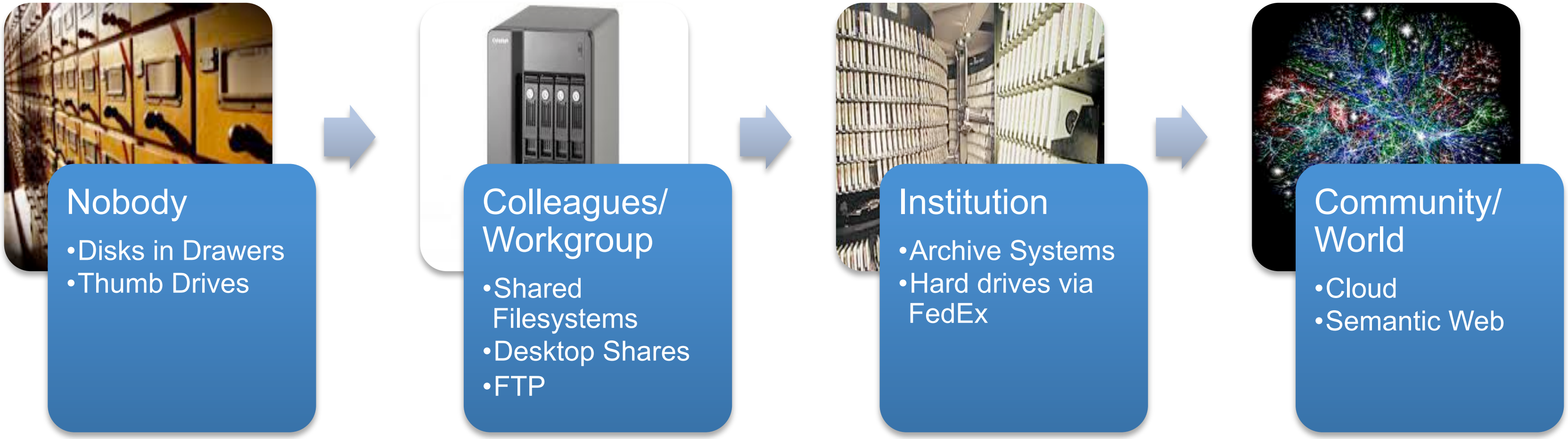
EMSL Estimated Aggregate Data Production/Transfer Rates

- At present:
  - 6 TB/day produced in EMSL
  - 5 TB/day transferred within EMSL
  - 200 GB/month transferred out of EMSL
- 2-5 years from now:
  - 20 TB/day produced
  - 40 TB/day transferred within
  - 600 TB/month transferred
  - 5 TB/month transferred into EMSL
- 5+ years from now:
  - 100 TB/day produced
  - 200 TB/day transferred within
  - 3 PB/month transferred out
  - 50 TB/month transferred in

New Directives: Expand Public Access to the Results of Federally Funded Research

- *"The Obama Administration is committed to the proposition that citizens deserve easy access to the results of scientific research their tax dollars have paid for. That's why, in a policy memorandum released today, OSTP Director John Holdren has directed Federal agencies with more than $100M in R&D expenditures to develop plans to make the published results of federally funded research freely available to the public within one year of publication and requiring researchers to better account for and manage the digital data resulting from federally funded scientific research."*
- -- Posted by Michael Stebbins on whitehouse.gov February 22, 2013 at 12:04 PM EDT

Continuum of Data Sharing



Nobody

- Disks in Drawers
- Thumb Drives

Colleagues/ Workgroup

- Shared Filesystems
- Desktop Shares
- FTP

Institution

- Archive Systems
- Hard drives via FedEx

Community/ World

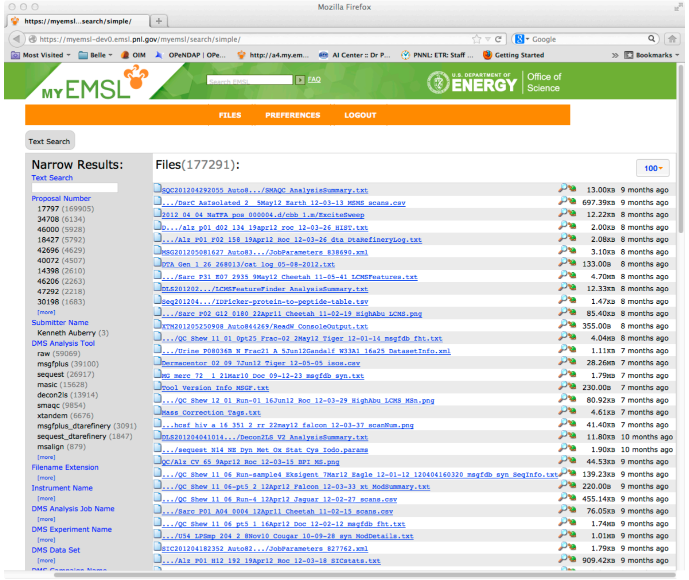
- Cloud
- Semantic Web

Limitations of our old approach

- **Aurora** is EMSL’s current archive system
- It has lots of parts:
  - Multi-petabyte robotic tape library (in CSF)
  - 1 Petabyte Disk Storage (in EMSL)
  - Many servers
  - Many services
- It’s mostly been used as a big ol’ pile of files
  - Organization (if any) was strictly DIY
  - Hard to find anything or even know it’s there!
- Access from outside EMSL is extremely limited

MyEMSL: Data Management for a diverse user facility

- Provide a system to capture, store and share data & metadata from EMSL's instruments and computers
- Create a flexible data framework that supports multiple scientific fields & approaches
- Allow scientific collaboration around data via EMSLHub (aka HubZero):
  - Workflow-based computation
  - Data integration and analysis
- Leverage high quality open source software
  - Form open source development groups
  - Easily shared with community
  - Highly efficient support and maintenance model
- Long term, create a framework to federate data sources across institutions



Narrow Results:

Text Search

Proposal Number
17797 (169905)
34708 (6134)
46000 (5928)
18427 (5792)
42696 (4629)
40072 (4507)
14398 (2610)
46206 (2263)
47292 (2218)
30198 (1683)
[more]

Submitter Name
Kenneth Auberry (3)

DMS Analysis Tool
raw (59069)
msgfplus (39100)
sequest (26917)
masic (15628)
decon2ls (13914)
smaqc (9854)
xtandem (6676)
msgfplus_dtarefinery (3091)
sequest_dtarefinery (1847)
msalign (879)
[more]

Filename Extension
[more]

Instrument Name
[more]

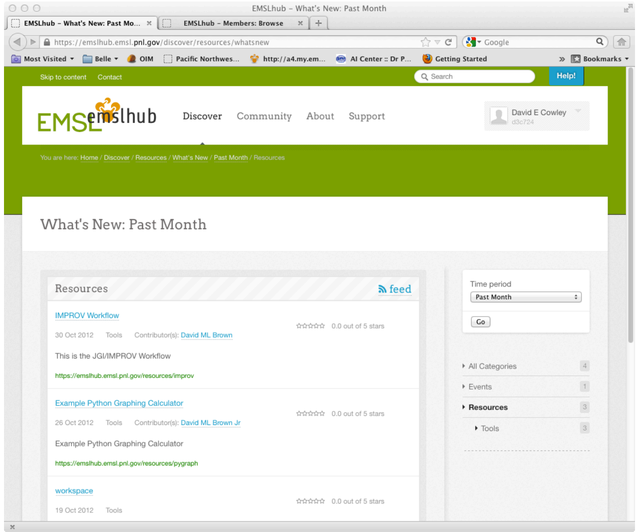
DMS Analysis Job Name
[more]

DMS Experiment Name
[more]

DMS Data Set
[more]

Files(177291):

SQC201204292055 Auto8.../SMAQC AnalysisSummary.txt	13.00KB	9 months ago
.../DsrC AsIsolated 2 5May12 Earth 12-03-13 MSMS scans.csv	697.39KB	9 months ago
2012 04 04 NaTFA pos 000004.d/cbb 1.m/ExciteSweep	12.22KB	8 months ago
D.../alz p01 d02 134 19apr12 roc 12-03-26 HIST.txt	2.00KB	8 months ago
.../Alz P01 F02 158 19Apr12 Roc 12-03-26 dta DtaRefineryLog.txt	2.08KB	8 months ago
MSG201205081627 Auto83.../JobParameters 838690.xml	3.10KB	8 months ago
DTA Gen 1 26 268013/cat log 05-08-2012.txt	133.00B	8 months ago
.../Sarc P31 E07 2935 9May12 Cheetah 11-05-41 LCMSFeatures.txt	4.70MB	8 months ago
DLS201202.../LCMSFeatureFinder AnalysisSummary.txt	12.33KB	8 months ago
Seq201204.../IDPicker-protein-to-peptide-table.tsv	1.47KB	8 months ago
.../Sarc P02 G12 0180 22Apr11 Cheetah 11-02-19 HighAbu LCMS.png	85.40KB	8 months ago
XTM201205250908 Auto844269/ReadW ConsoleOutput.txt	355.00B	8 months ago
.../QC Shew 11 01 0pt25 Frac-02 2May12 Tiger 12-01-14 msgfdb fht.txt	4.04MB	8 months ago
.../Urine P08036B N Frac21 A 5Jun12Gandalf W33A1 16a25 DatasetInfo.xml	1.11KB	7 months ago
Dermacentor 02 09 7Jun12 Tiger 12-05-05 isos.csv	28.26MB	7 months ago
MG merc 72 1 21Mar10 Doc 09-12-23 msgfdb syn.txt	1.79MB	7 months ago
Tool Version Info MSGF.txt	230.00B	7 months ago
.../QC Shew 12 01 Run-01 16Jun12 Roc 12-03-29 HighAbu LCMS MSn.png	80.92KB	7 months ago
Mass Correction Tags.txt	4.61KB	7 months ago
...hcsf hiv a 16 351 2 rr 22may12 falcon 12-03-37 scanNum.png	41.40KB	7 months ago
DLS201204041014.../Decon2LS V2 AnalysisSummary.txt	11.80KB	10 months ago
.../sequest N14 NE Dyn Met Ox Stat Cys Iodo.params	1.90KB	10 months ago
QC/Alz CV 65 9Apr12 Roc 12-03-15 BPI MS.png	44.53KB	9 months ago
.../QC Shew 11 06 Run-sample4 Eksigent 7Mar12 Eagle 12-01-12 120404160320 msgfdb syn SeqInfo.txt	139.23KB	9 months ago
.../QC Shew 11 06-pt5 2 12Apr12 Falcon 12-03-33 xt ModSummary.txt	220.00B	9 months ago
.../QC Shew 11 06 Run-4 12Apr12 Jaguar 12-02-27 scans.csv	455.14KB	9 months ago
.../Sarc P01 A04 0004 12Apr11 Cheetah 11-02-15 scans.csv	76.05KB	9 months ago
.../QC Shew 11 06 pt5 1 16Apr12 Doc 12-02-12 msgfdb fht.txt	1.74MB	9 months ago
.../U54 LPSmp 204 2 8Nov10 Cougar 10-09-28 syn ModDetails.txt	1.01MB	9 months ago
SIC201204182352 Auto82.../JobParameters 827762.xml	1.79KB	9 months ago
.../Alz P01 H12 192 19Apr12 Roc 12-03-18 SICstats.txt	909.42KB	9 months ago



You are here: Home / Discover / Resources / What's New / Past Month / Resources

What's New: Past Month

Resources — feed

IMPROV Workflow — ☆☆☆☆☆ 0.0 out of 5 stars
30 Oct 2012 Tools Contributor(s): David ML Brown
This is the JGI/IMPROV Workflow
https://emslhub.emsl.pnl.gov/resources/improv

Example Python Graphing Calculator — ☆☆☆☆☆ 0.0 out of 5 stars
26 Oct 2012 Tools Contributor(s): David ML Brown Jr
Example Python Graphing Calculator
https://emslhub.emsl.pnl.gov/resources/pygraph

workspace — ☆☆☆☆☆ 0.0 out of 5 stars
19 Oct 2012 Tools

Time period: Past Month
Go

All Categories	4
Events	1
Resources	3
Tools	3

Supporting Scientific Collaboration with MyEMSL

- Provide search and easy access to data & metadata for authorized users
- Facilitate and enforce data release policies
- Find enthusiastic scientists who know they have data needs and partner with them
- Provide data interaction framework to facilitate “Bring Your Own”:
    - Data catalogs & definitions
    - Workflow based computation
    - Data integration and analysis
    - Web and mobile apps

Why MyEMSL?

- **MyEMSL** aims to solve a number of problems with the old archive approach:
  - Make data easy to find by good use of metadata
  - Make files easy to get via the web
  - Set up good access control
  - Make it easy to share data with collaborators
- Leverage existing HPSS archive for bulk data storage
- Augment legacy archive with other parts:
  - Metadata Database
  - Automated uploader for instruments
  - Web portal @ https://my.emsl.pnl.gov
  - Links to EMSL's User/Resource allocation systems & EMSLHub

The importance of Metadata

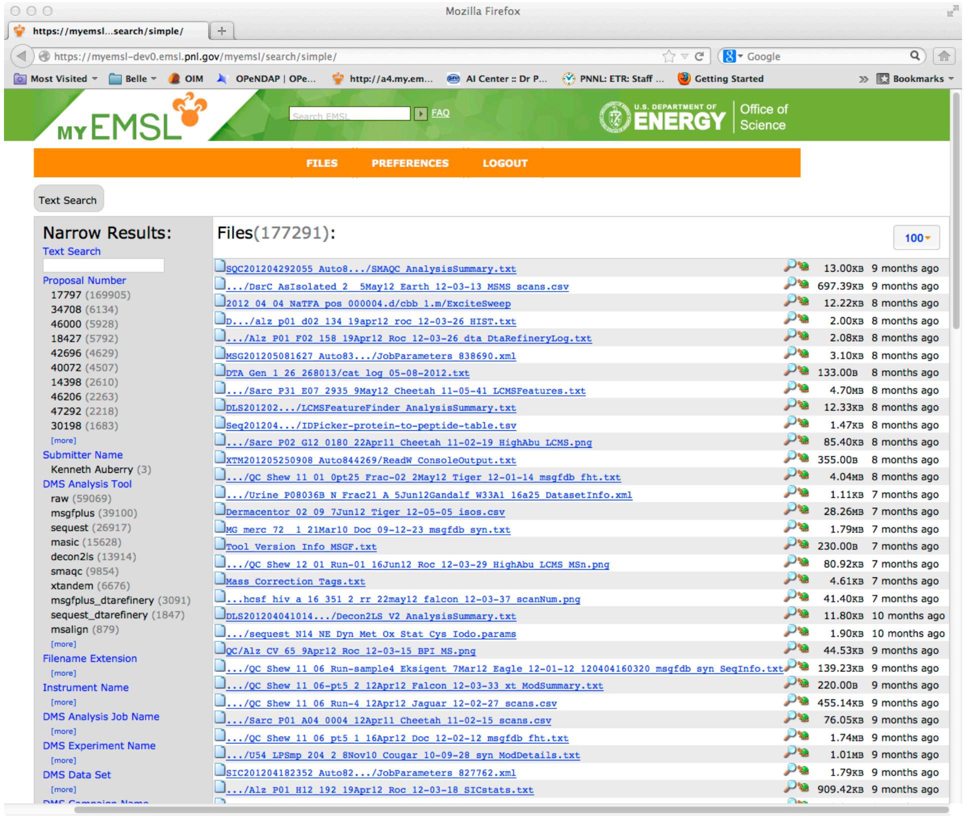
- We have a big pile of files going back 15 years, but nobody knows what all is in there!
- The key to discovering and understanding data is to have good **metadata!**
- Metadata tells you what is known about the data
- Can be trivial:
  - File name
  - File owner
  - Time and date stamps
- Can be much more interesting:
  - Instrument settings & environmental parameters
  - Sample information
  - Registration data for multimodal analyses

What we need to do about metadata

- The more we can get of it, the more powerful MyEMSL becomes
- We have defined a little essential metadata that all data must have
- MyEMSL must be **flexible and non-restrictive** about metadata and **enable metadata to evolve** over time
- We don't know whatever other metadata people may need – maybe they don't know yet
- Science partners will need to tell us what metadata will be useful to them and MyEMSL will need to support it
- How can we get it?
  - Our processes (i.e. autouploader) can capture or generate a little of it
  - Humans may need to enter it
  - Code can be written to extract/generate it, but this will require funding and people time!

What exists today

- Web and service interfaces for moving data and metadata into and out of MyEMSL
- Authentication mechanisms
- Search and "shopping cart" interfaces
- EMSLHub collaboration site with pilot workflows
- Proteomics pipeline adapted to MyEMSL interfaces
- Increasing numbers of scientific instruments uploading data



Narrow Results:

Text Search

Proposal Number
17797 (169905)
34708 (6134)
46000 (5928)
18427 (5792)
42696 (4629)
40072 (4507)
14398 (2610)
46206 (2263)
47292 (2218)
30198 (1683)
[more]

Submitter Name
Kenneth Auberry (3)

DMS Analysis Tool
raw (59069)
msgfplus (39100)
sequest (26917)
masic (15628)
decon2ls (13914)
smaqc (9854)
xtandem (6676)
msgfplus_dtarefinery (3091)
sequest_dtarefinery (1847)
msalign (879)
[more]

Filename Extension
[more]

Instrument Name
[more]

DMS Analysis Job Name
[more]

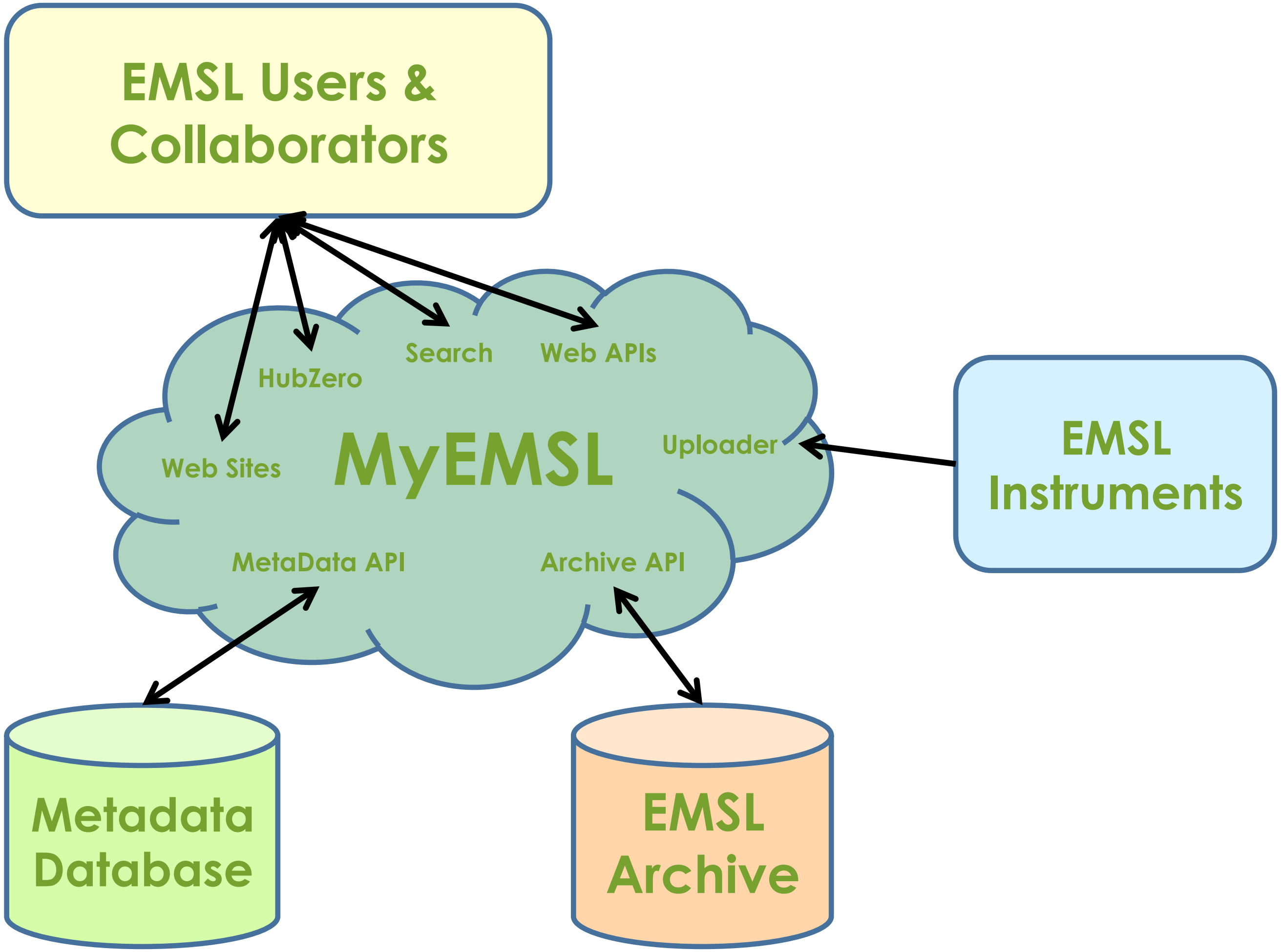
DMS Experiment Name
[more]

DMS Data Set
[more]

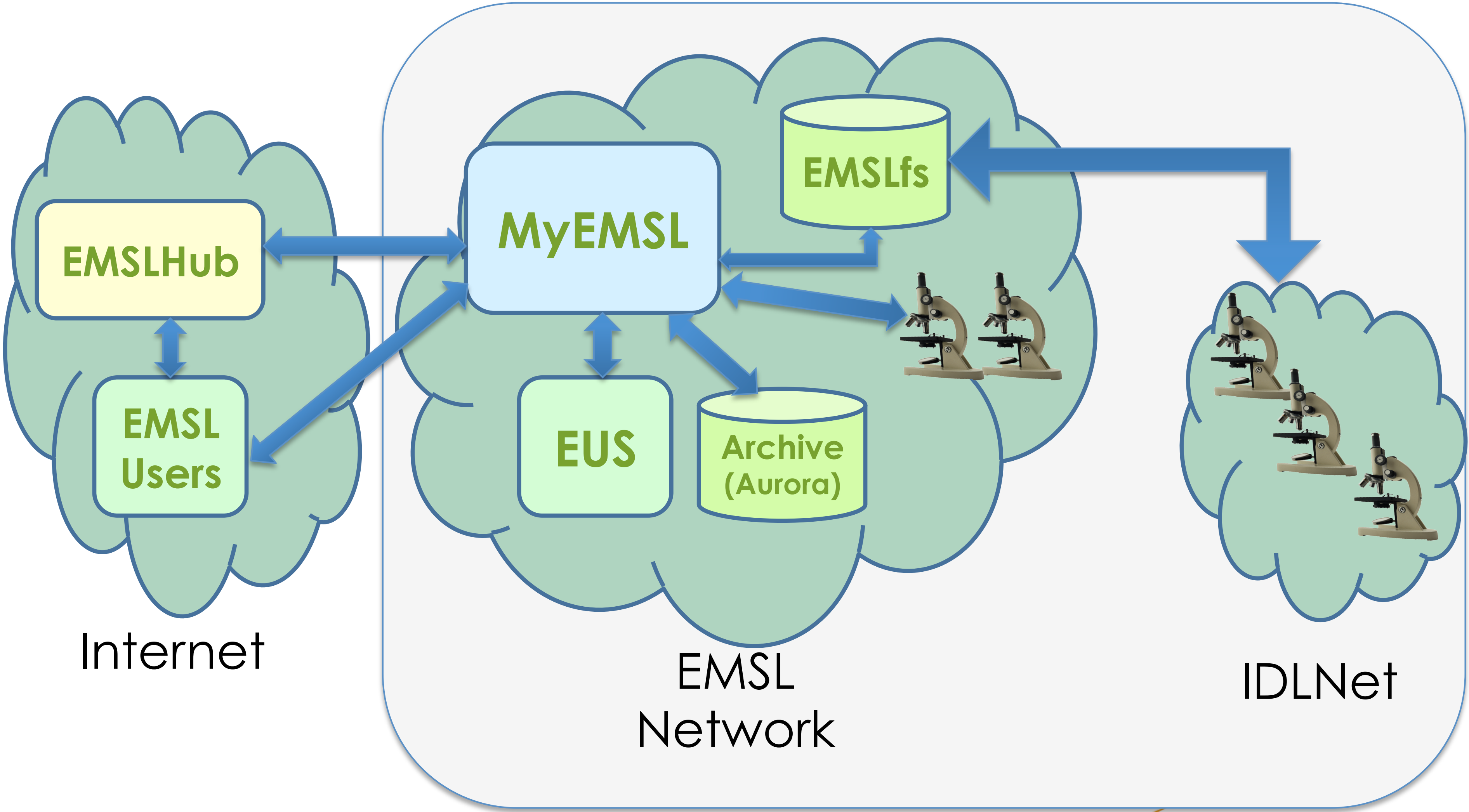
Files(177291):

SQC201204292055 Auto8.../SMAQC AnalysisSummary.txt	13.00KB	9 months ago
.../DsrC AsIsolated 2 5May12 Earth 12-03-13 MSMS scans.csv	697.39KB	9 months ago
2012 04 04 NaTFA pos 000004.d/cbb 1.m/ExciteSweep	12.22KB	8 months ago
D.../alz p01 d02 134 19apr12 roc 12-03-26 HIST.txt	2.00KB	8 months ago
.../Alz P01 F02 158 19Apr12 Roc 12-03-26 dta DtaRefineryLog.txt	2.08KB	8 months ago
MSG201205081627 Auto83.../JobParameters 838690.xml	3.10KB	8 months ago
DTA Gen 1 26 268013/cat log 05-08-2012.txt	133.00B	8 months ago
.../Sarc P31 E07 2935 9May12 Cheetah 11-05-41 LCMSFeatures.txt	4.70MB	8 months ago
DLS201202.../LCMSFeatureFinder AnalysisSummary.txt	12.33KB	8 months ago
Seq201204.../IDPicker-protein-to-peptide-table.tsv	1.47KB	8 months ago
.../Sarc P02 G12 0180 22Apr11 Cheetah 11-02-19 HighAbu LCMS.png	85.40KB	8 months ago
XTM201205250908 Auto844269/ReadW ConsoleOutput.txt	355.00B	8 months ago
.../QC Shew 11 01 0pt25 Frac-02 2May12 Tiger 12-01-14 msgfdb fht.txt	4.04MB	8 months ago
.../Urine P08036B N Frac21 A 5Jun12Gandalf W33A1 16a25 DatasetInfo.xml	1.11KB	7 months ago
Dermacentor 02 09 7Jun12 Tiger 12-05-05 isos.csv	28.26MB	7 months ago
MG merc 72 1 21Mar10 Doc 09-12-23 msgfdb syn.txt	1.79MB	7 months ago
Tool Version Info MSGF.txt	230.00B	7 months ago
.../QC Shew 12 01 Run-01 16Jun12 Roc 12-03-29 HighAbu LCMS MSn.png	80.92KB	7 months ago
Mass Correction Tags.txt	4.61KB	7 months ago
...hcsf hiv a 16 351 2 rr 22may12 falcon 12-03-37 scanNum.png	41.40KB	7 months ago
DLS201204041014.../Decon2LS V2 AnalysisSummary.txt	11.80KB	10 months ago
.../sequest N14 NE Dyn Met Ox Stat Cys Iodo.params	1.90KB	10 months ago
QC/Alz CV 65 9Apr12 Roc 12-03-15 BPI MS.png	44.53KB	9 months ago
.../QC Shew 11 06 Run-sample4 Eksigent 7Mar12 Eagle 12-01-12 120404160320 msgfdb syn SeqInfo.txt	139.23KB	9 months ago
.../QC Shew 11 06-pt5 2 12Apr12 Falcon 12-03-33 xt ModSummary.txt	220.00B	9 months ago
.../QC Shew 11 06 Run-4 12Apr12 Jaguar 12-02-27 scans.csv	455.14KB	9 months ago
.../Sarc P01 A04 0004 12Apr11 Cheetah 11-02-15 scans.csv	76.05KB	9 months ago
.../QC Shew 11 06 pt5 1 16Apr12 Doc 12-02-12 msgfdb fht.txt	1.74MB	9 months ago
.../U54 LPSmp 204 2 8Nov10 Cougar 10-09-28 syn ModDetails.txt	1.01MB	9 months ago
SIC201204182352 Auto82.../JobParameters 827762.xml	1.79KB	9 months ago
.../Alz P01 H12 192 19Apr12 Roc 12-03-18 SICstats.txt	909.42KB	9 months ago

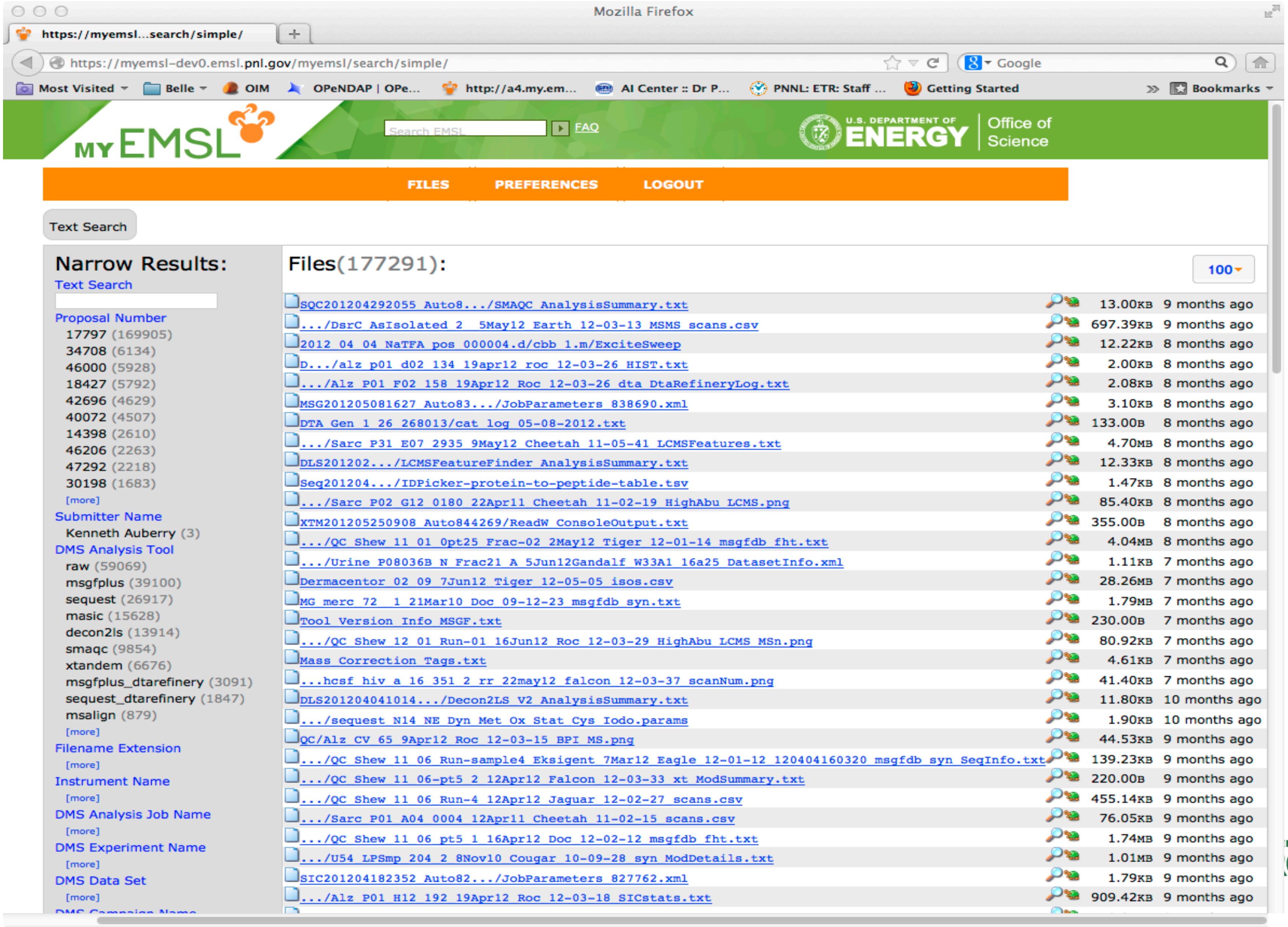
MyEMSL Conceptual Diagram



Data, Instruments in EMSL context



http://my.emsl.pnl.gov



Narrow Results:

Text Search

Proposal Number

17797	(169905)
34708	(6134)
46000	(5928)
18427	(5792)
42696	(4629)
40072	(4507)
14398	(2610)
46206	(2263)
47292	(2218)
30198	(1683)

[more]

Submitter Name

Kenneth Auberry	(3)

DMS Analysis Tool

raw	(59069)
msgfplus	(39100)
sequest	(26917)
masic	(15628)
decon2ls	(13914)
smaqc	(9854)
xtandem	(6676)
msgfplus_dtarefinery	(3091)
sequest_dtarefinery	(1847)
msalign	(879)

[more]

Filename Extension

[more]

Instrument Name

[more]

DMS Analysis Job Name

[more]

DMS Experiment Name

[more]

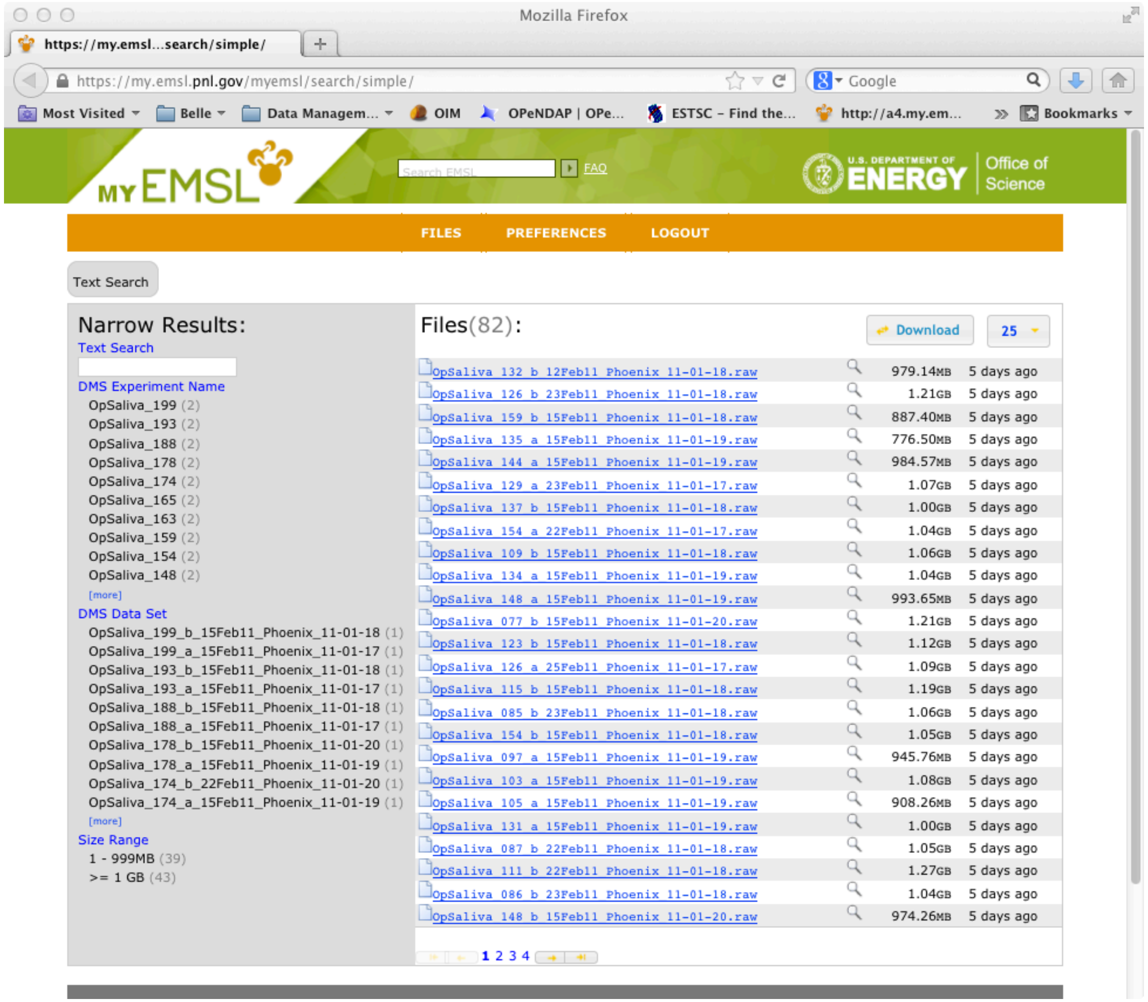
DMS Data Set

[more]

Files(177291):

SQC201204292055 Auto8.../SMAQC AnalysisSummary.txt	13.00KB	9 months ago
.../DsrC AsIsolated 2 5May12 Earth 12-03-13 MSMS scans.csv	697.39KB	9 months ago
2012 04 04 NaTFA pos 000004.d/cbb 1.m/ExciteSweep	12.22KB	8 months ago
D.../alz p01 d02 134 19apr12 roc 12-03-26 HIST.txt	2.00KB	8 months ago
.../Alz P01 F02 158 19Apr12 Roc 12-03-26 dta DtaRefineryLog.txt	2.08KB	8 months ago
MSG201205081627 Auto83.../JobParameters 838690.xml	3.10KB	8 months ago
DTA Gen 1 26 268013/cat log 05-08-2012.txt	133.00B	8 months ago
.../Sarc P31 E07 2935 9May12 Cheetah 11-05-41 LCMSFeatures.txt	4.70MB	8 months ago
DLS201202.../LCMSFeatureFinder AnalysisSummary.txt	12.33KB	8 months ago
Seq201204.../IDPicker-protein-to-peptide-table.tsv	1.47KB	8 months ago
.../Sarc P02 G12 0180 22Apr11 Cheetah 11-02-19 HighAbu LCMS.png	85.40KB	8 months ago
XTM201205250908 Auto844269/ReadW ConsoleOutput.txt	355.00B	8 months ago
.../QC Shew 11 01 0pt25 Frac-02 2May12 Tiger 12-01-14 msgfdb fht.txt	4.04MB	8 months ago
.../Urine P08036B N Frac21 A 5Jun12Gandalf W33A1 16a25 DatasetInfo.xml	1.11KB	7 months ago
Dermacentor 02 09 7Jun12 Tiger 12-05-05 isos.csv	28.26MB	7 months ago
MG merc 72 1 21Mar10 Doc 09-12-23 msgfdb syn.txt	1.79MB	7 months ago
Tool Version Info MSGF.txt	230.00B	7 months ago
.../QC Shew 12 01 Run-01 16Jun12 Roc 12-03-29 HighAbu LCMS MSn.png	80.92KB	7 months ago
Mass Correction Tags.txt	4.61KB	7 months ago
...hcsf hiv a 16 351 2 rr 22may12 falcon 12-03-37 scanNum.png	41.40KB	7 months ago
DLS201204041014.../Decon2LS V2 AnalysisSummary.txt	11.80KB	10 months ago
.../sequest N14 NE Dyn Met Ox Stat Cys Iodo.params	1.90KB	10 months ago
QC/Alz CV 65 9Apr12 Roc 12-03-15 BPI MS.png	44.53KB	9 months ago
.../QC Shew 11 06 Run-sample4 Eksigent 7Mar12 Eagle 12-01-12 120404160320 msgfdb syn SeqInfo.txt	139.23KB	9 months ago
.../QC Shew 11 06-pt5 2 12Apr12 Falcon 12-03-33 xt ModSummary.txt	220.00B	9 months ago
.../QC Shew 11 06 Run-4 12Apr12 Jaguar 12-02-27 scans.csv	455.14KB	9 months ago
.../Sarc P01 A04 0004 12Apr11 Cheetah 11-02-15 scans.csv	76.05KB	9 months ago
.../QC Shew 11 06 pt5 1 16Apr12 Doc 12-02-12 msgfdb fht.txt	1.74MB	9 months ago
.../U54 LPSmp 204 2 8Nov10 Cougar 10-09-28 syn ModDetails.txt	1.01MB	9 months ago
SIC201204182352 Auto82.../JobParameters 827762.xml	1.79KB	9 months ago
.../Alz P01 H12 192 19Apr12 Roc 12-03-18 SICstats.txt	909.42KB	9 months ago

Example: Public data in MyEMSL web pages



Narrow Results:

Text Search

DMS Experiment Name

- OpSaliva_199 (2)
- OpSaliva_193 (2)
- OpSaliva_188 (2)
- OpSaliva_178 (2)
- OpSaliva_174 (2)
- OpSaliva_165 (2)
- OpSaliva_163 (2)
- OpSaliva_159 (2)
- OpSaliva_154 (2)
- OpSaliva_148 (2)
- [more]

DMS Data Set

- OpSaliva_199_b_15Feb11_Phoenix_11-01-18 (1)
- OpSaliva_199_a_15Feb11_Phoenix_11-01-17 (1)
- OpSaliva_193_b_15Feb11_Phoenix_11-01-18 (1)
- OpSaliva_193_a_15Feb11_Phoenix_11-01-17 (1)
- OpSaliva_188_b_15Feb11_Phoenix_11-01-18 (1)
- OpSaliva_188_a_15Feb11_Phoenix_11-01-17 (1)
- OpSaliva_178_b_15Feb11_Phoenix_11-01-20 (1)
- OpSaliva_178_a_15Feb11_Phoenix_11-01-19 (1)
- OpSaliva_174_b_22Feb11_Phoenix_11-01-20 (1)
- OpSaliva_174_a_15Feb11_Phoenix_11-01-19 (1)
- [more]

Size Range

- 1 - 999MB (39)
- >= 1 GB (43)

Files(82):

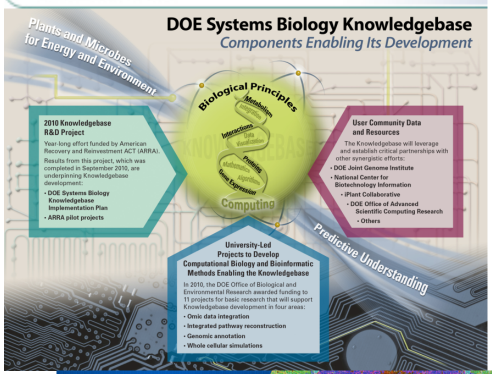
OpSaliva_132_b_12Feb11_Phoenix_11-01-18.raw	979.14MB	5 days ago
OpSaliva_126_b_23Feb11_Phoenix_11-01-18.raw	1.21GB	5 days ago
OpSaliva_159_b_15Feb11_Phoenix_11-01-18.raw	887.40MB	5 days ago
OpSaliva_135_a_15Feb11_Phoenix_11-01-19.raw	776.50MB	5 days ago
OpSaliva_144_a_15Feb11_Phoenix_11-01-19.raw	984.57MB	5 days ago
OpSaliva_129_a_23Feb11_Phoenix_11-01-17.raw	1.07GB	5 days ago
OpSaliva_137_b_15Feb11_Phoenix_11-01-18.raw	1.00GB	5 days ago
OpSaliva_154_a_22Feb11_Phoenix_11-01-17.raw	1.04GB	5 days ago
OpSaliva_109_b_15Feb11_Phoenix_11-01-18.raw	1.06GB	5 days ago
OpSaliva_134_a_15Feb11_Phoenix_11-01-19.raw	1.04GB	5 days ago
OpSaliva_148_a_15Feb11_Phoenix_11-01-19.raw	993.65MB	5 days ago
OpSaliva_077_b_15Feb11_Phoenix_11-01-20.raw	1.21GB	5 days ago
OpSaliva_123_b_15Feb11_Phoenix_11-01-18.raw	1.12GB	5 days ago
OpSaliva_126_a_25Feb11_Phoenix_11-01-17.raw	1.09GB	5 days ago
OpSaliva_115_b_15Feb11_Phoenix_11-01-18.raw	1.19GB	5 days ago
OpSaliva_085_b_23Feb11_Phoenix_11-01-18.raw	1.06GB	5 days ago
OpSaliva_154_b_15Feb11_Phoenix_11-01-18.raw	1.05GB	5 days ago
OpSaliva_097_a_15Feb11_Phoenix_11-01-19.raw	945.76MB	5 days ago
OpSaliva_103_a_15Feb11_Phoenix_11-01-19.raw	1.08GB	5 days ago
OpSaliva_105_a_15Feb11_Phoenix_11-01-19.raw	908.26MB	5 days ago
OpSaliva_131_a_15Feb11_Phoenix_11-01-19.raw	1.00GB	5 days ago
OpSaliva_087_b_22Feb11_Phoenix_11-01-18.raw	1.05GB	5 days ago
OpSaliva_111_b_22Feb11_Phoenix_11-01-18.raw	1.27GB	5 days ago
OpSaliva_086_b_23Feb11_Phoenix_11-01-18.raw	1.04GB	5 days ago
OpSaliva_148_b_15Feb11_Phoenix_11-01-20.raw	974.26MB	5 days ago

1 2 3 4

Establishing MyEMSL Collaborations

- We are establishing key collaborations with other institutions to:
  - Define tools and interfaces for data exchange
  - Leverage existing tools and frameworks
  - Establish open source consortia
- Partner organizations and activities:
  - DOE Joint Genome Institute (JGI)
  - DOE Systems Biology Knowledgebase (KBASE)
  - National Center for Microscopy and Imaging Research (NCMIR):
  - Semantic Physical Sciences (SPS) workshops

JGI
DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE



DOE Systems Biology Knowledgebase
Components Enabling Its Development

2010 Knowledgebase R&D Project

Year-long effort funded by American Recovery and Reinvestment ACT (ARRA).

Results from this project, which was completed in September 2010, are underpinning Knowledgebase development:

- DOE Systems Biology Knowledgebase Implementation Plan
- ARRA pilot projects

User Community Data and Resources

The Knowledgebase will leverage and establish critical partnerships with other synergistic efforts:

- DOE Joint Genome Institute
- National Center for Biotechnology Information
- iPlant Collaborative
- DOE Office of Advanced Scientific Computing Research
- Others

University-Led Projects to Develop Computational Biology and Bioinformatic Methods Enabling the Knowledgebase

In 2010, the DOE Office of Biological and Environmental Research awarded funding to 11 projects for basic research that will support Knowledgebase development in four areas:

- Omic data integration
- Integrated pathway reconstruction
- Genomic annotation
- Whole cellular simulations

NCMIR

University of BRISTOL

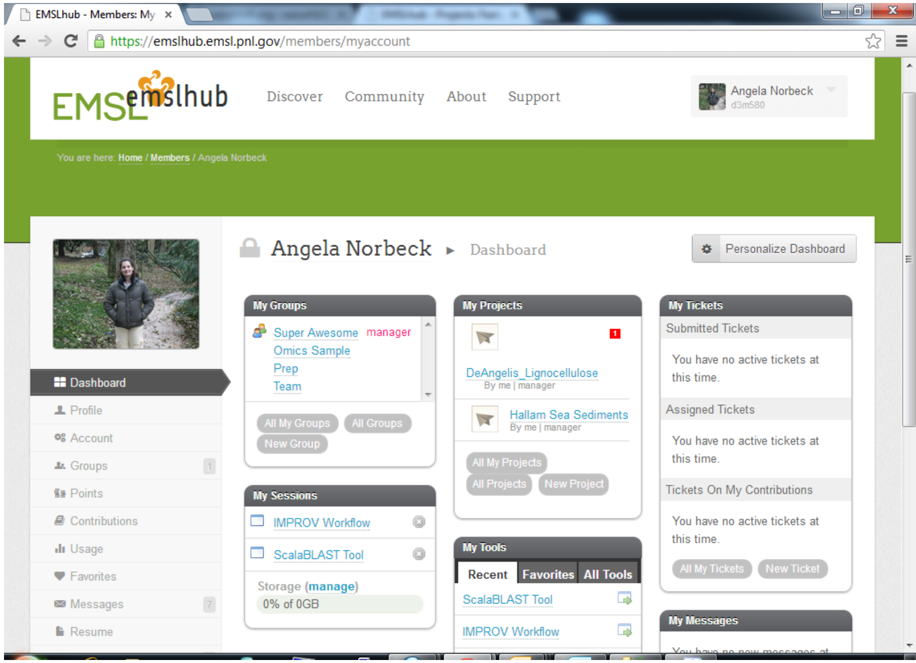
UNIVERSITY OF CAMBRIDGE

CSIRO

PENNSTATE
1855

EMSLHub

- **EMSLHub** is a collaborative web site at https://emslhub.emsl.pnl.gov
- Purpose: provide a system to allow scientific collaboration around *data* with a rich toolset
- Early collaborative efforts:
  - JGI – EMSL Data Integration
  - Proteomics project coordination & data dissemination
  - Integration of NWChem & NMR
  - Scientific Workflows



EMSLhub - Members: My

https://emslhub.emsl.pnl.gov/members/myaccount

You are here: Home / Members / Angela Norbeck

Angela Norbeck ▸ Dashboard

My Groups: Super Awesome manager, Omics Sample, Prep, Team

My Projects: DeAngelis_Lignocellulose (By me | manager); Hallam Sea Sediments (By me | manager)

My Tickets

Submitted Tickets: You have no active tickets at this time.

Assigned Tickets: You have no active tickets at this time.

Tickets On My Contributions: You have no active tickets at this time.

My Sessions: IMPROV Workflow; ScalaBLAST Tool

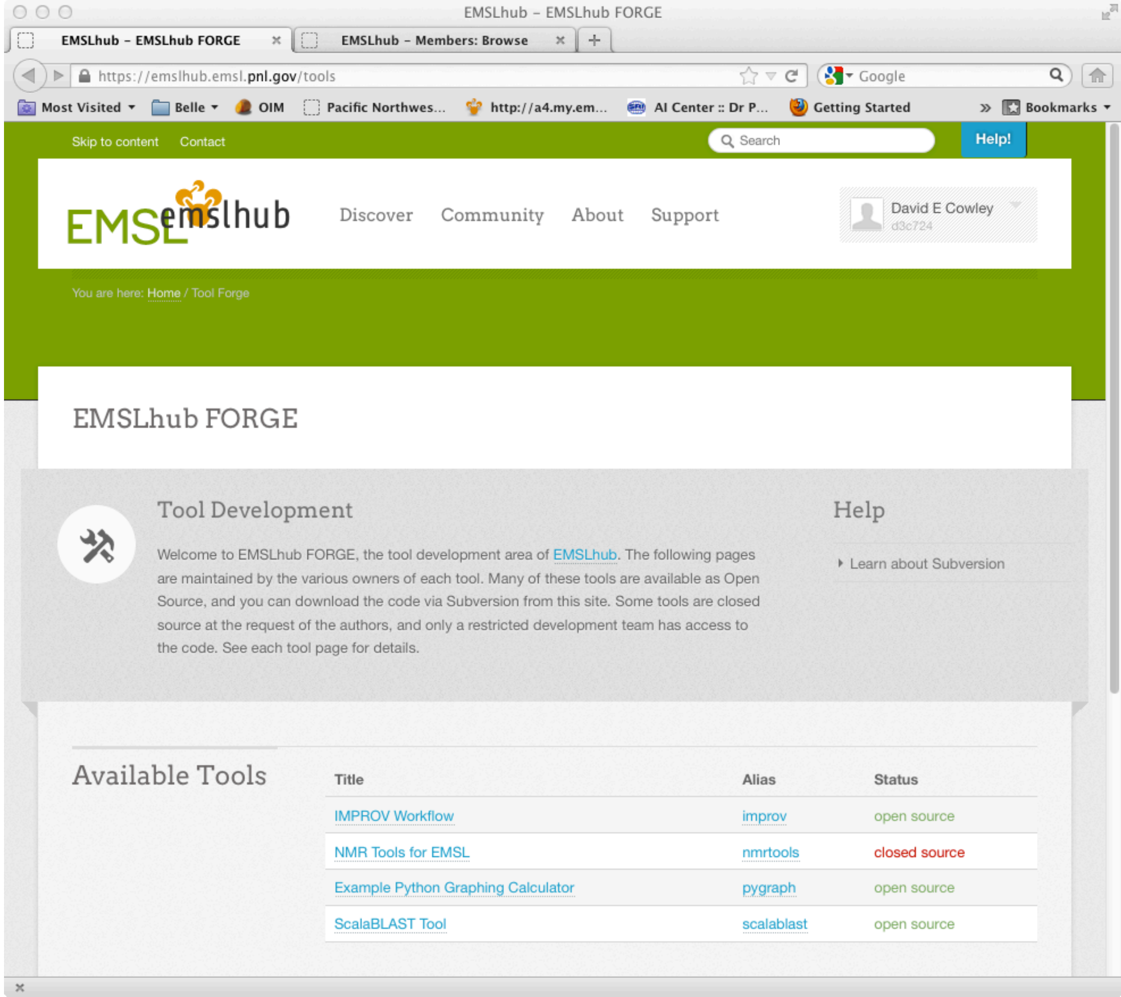
Storage (manage) 0% of 0GB

My Tools (Recent): ScalaBLAST Tool; IMPROV Workflow

My Messages

Tools Development in EMSLHub

- HubZero framework allows programs to be developed, shared, used within as “tools” in EMSLHub
- Anyone with access can submit tools
- Approved tools can be shared with groups
- First examples are:
  - NMR analysis pipeline
  - ScalaBLAST job submission
  - IMPROV in EMSLHub



EMSLhub FORGE

Tool Development

Welcome to EMSLhub FORGE, the tool development area of EMSLhub. The following pages are maintained by the various owners of each tool. Many of these tools are available as Open Source, and you can download the code via Subversion from this site. Some tools are closed source at the request of the authors, and only a restricted development team has access to the code. See each tool page for details.

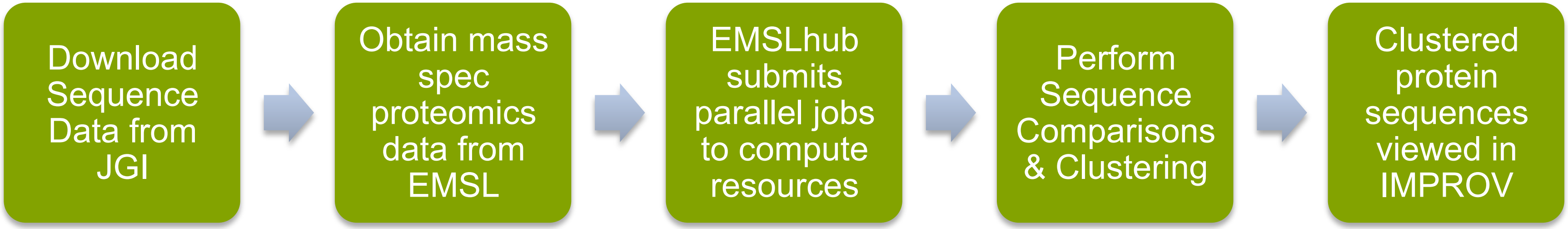
Help

- Learn about Subversion

Available Tools

Title	Alias	Status
IMPROV Workflow	improv	open source
NMR Tools for EMSL	nmrtools	closed source
Example Python Graphing Calculator	pygraph	open source
ScalaBLAST Tool	scalablast	open source

Example: Metaproteomics pipeline implementation in EMSLHub



Download Sequence Data from JGI

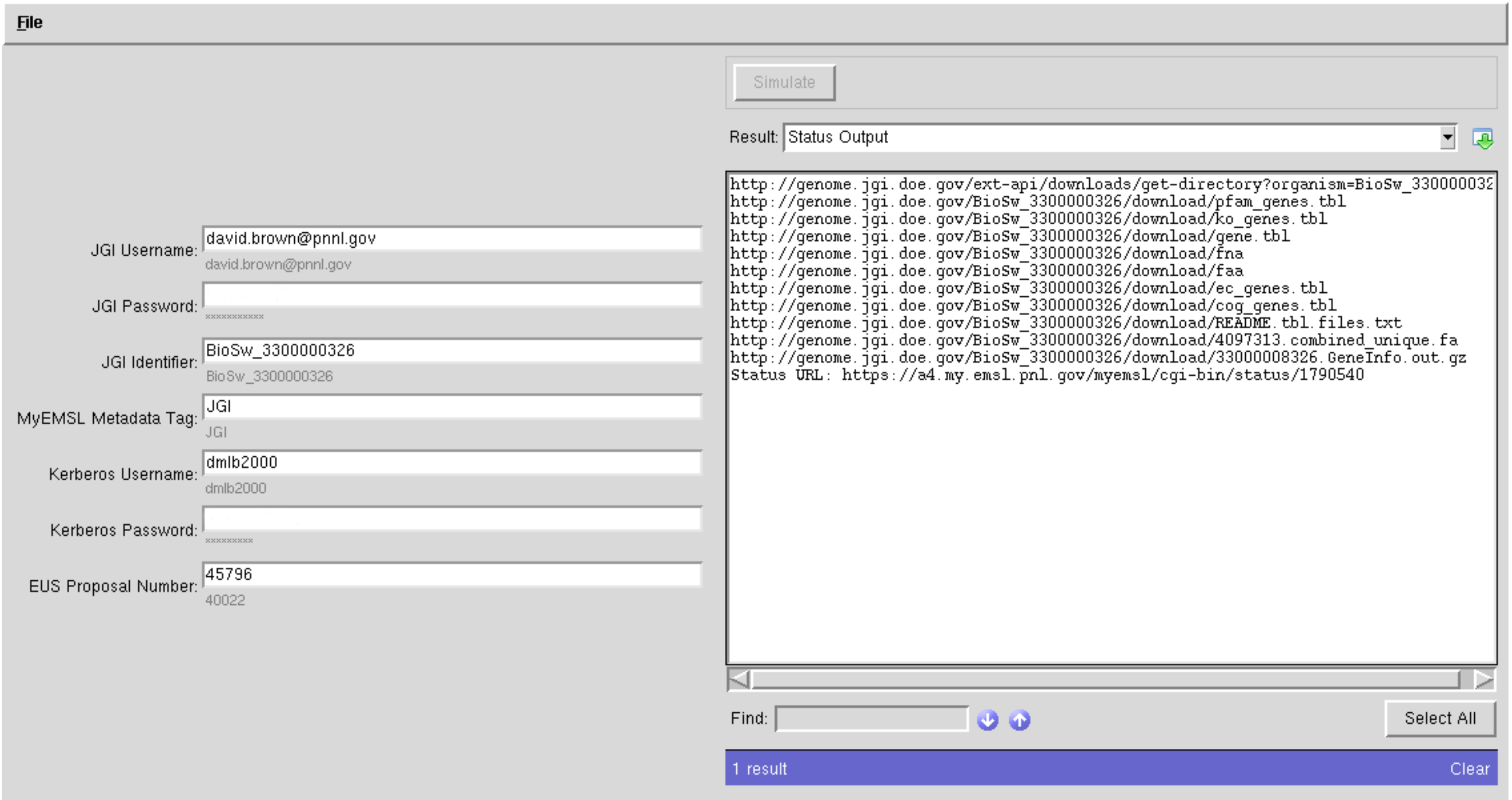
Obtain mass spec proteomics data from EMSL

EMSLhub submits parallel jobs to compute resources

Perform Sequence Comparisons & Clustering

Clustered protein sequences viewed in IMPROV

Example: Metaproteomics pipeline implementation in EMSLHub



JGI Username: david.brown@pnnl.gov
david.brown@pnnl.gov

JGI Password:
xxxxxxxxxxxx

JGI Identifier: BioSw_3300000326
BioSw_3300000326

MyEMSL Metadata Tag: JGI
JGI

Kerberos Username: dmlb2000
dmlb2000

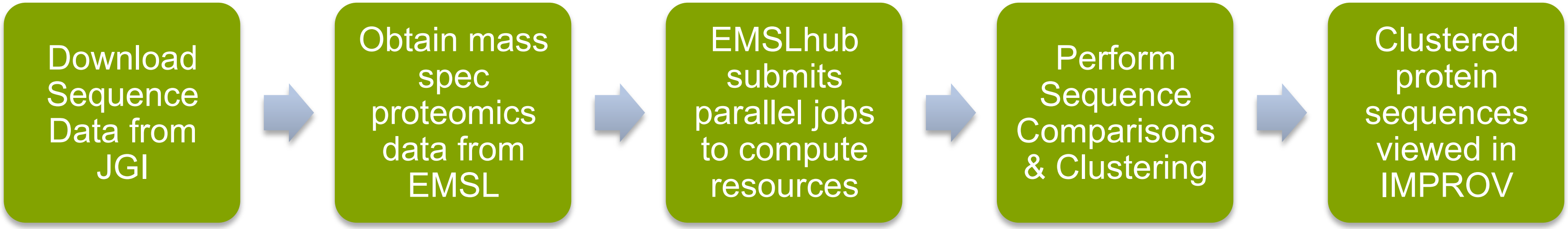
Kerberos Password:
xxxxxxxxxx

EUS Proposal Number: 45796
40022

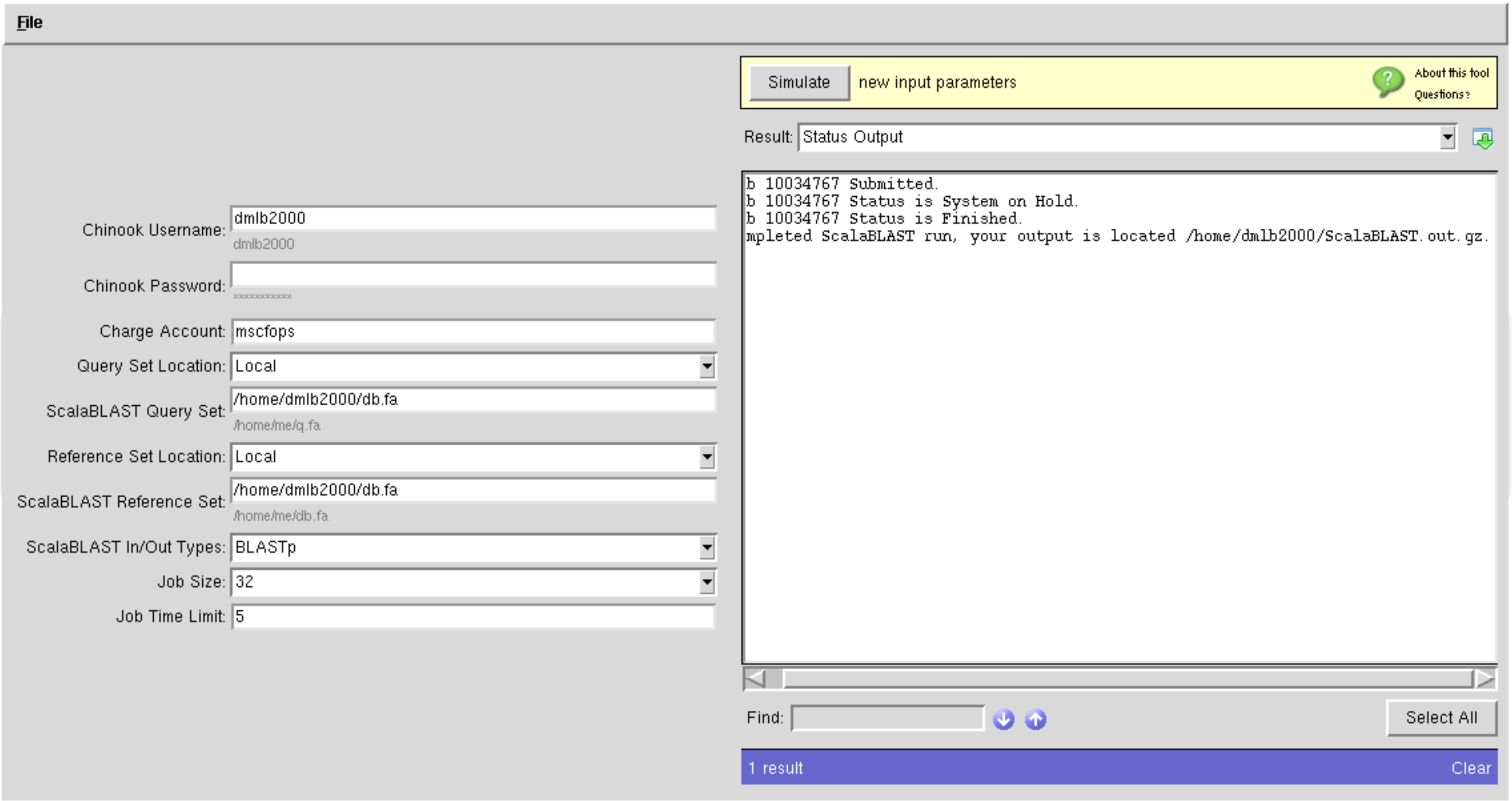
Result: Status Output

```
http://genome.jgi.doe.gov/ext-api/downloads/get-directory?organism=BioSw_330000032
http://genome.jgi.doe.gov/BioSw_3300000326/download/pfam_genes.tbl
http://genome.jgi.doe.gov/BioSw_3300000326/download/ko_genes.tbl
http://genome.jgi.doe.gov/BioSw_3300000326/download/gene.tbl
http://genome.jgi.doe.gov/BioSw_3300000326/download/fna
http://genome.jgi.doe.gov/BioSw_3300000326/download/faa
http://genome.jgi.doe.gov/BioSw_3300000326/download/ec_genes.tbl
http://genome.jgi.doe.gov/BioSw_3300000326/download/cog_genes.tbl
http://genome.jgi.doe.gov/BioSw_3300000326/download/README.tbl.files.txt
http://genome.jgi.doe.gov/BioSw_3300000326/download/4097313.combined_unique.fa
http://genome.jgi.doe.gov/BioSw_3300000326/download/3300008326.GeneInfo.out.gz
Status URL: https://a4.my.emsl.pnl.gov/myemsl/cgi-bin/status/1790540
```

Example: Metaproteomics pipeline implementation in EMSLHub


Download Sequence Data from JGI
Obtain mass spec proteomics data from EMSL
EMSLhub submits parallel jobs to compute resources
Perform Sequence Comparisons & Clustering
Clustered protein sequences viewed in IMPROV

Example: Metaproteomics pipeline implementation in EMSLHub

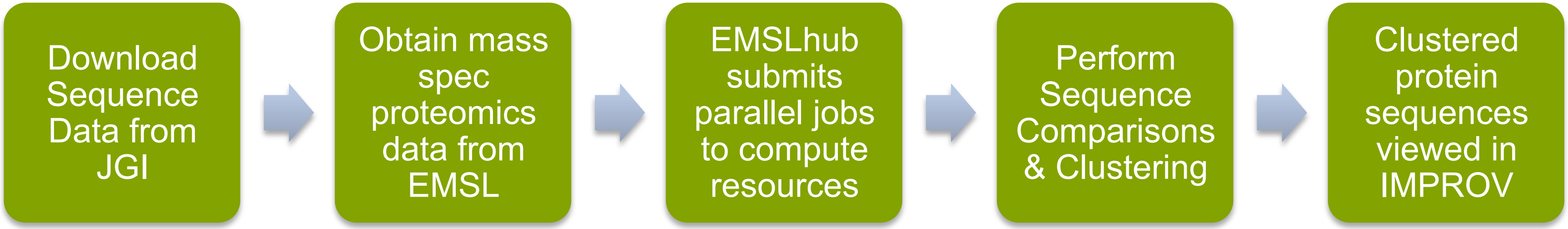


Result: Status Output

```
b 10034767 Submitted.
b 10034767 Status is System on Hold.
b 10034767 Status is Finished.
mpleted ScalaBLAST run, your output is located /home/dmlb2000/ScalaBLAST.out.gz.
```

Chinook Username:	dmlb2000 dmlb2000
Chinook Password:	××××××××××
Charge Account:	mscfops
Query Set Location:	Local
ScalaBLAST Query Set:	/home/dmlb2000/db.fa /home/me/q.fa
Reference Set Location:	Local
ScalaBLAST Reference Set:	/home/dmlb2000/db.fa /home/me/db.fa
ScalaBLAST In/Out Types:	BLASTp
Job Size:	32
Job Time Limit:	5

Example: Metaproteomics pipeline implementation in EMSLHub



Download Sequence Data from JGI

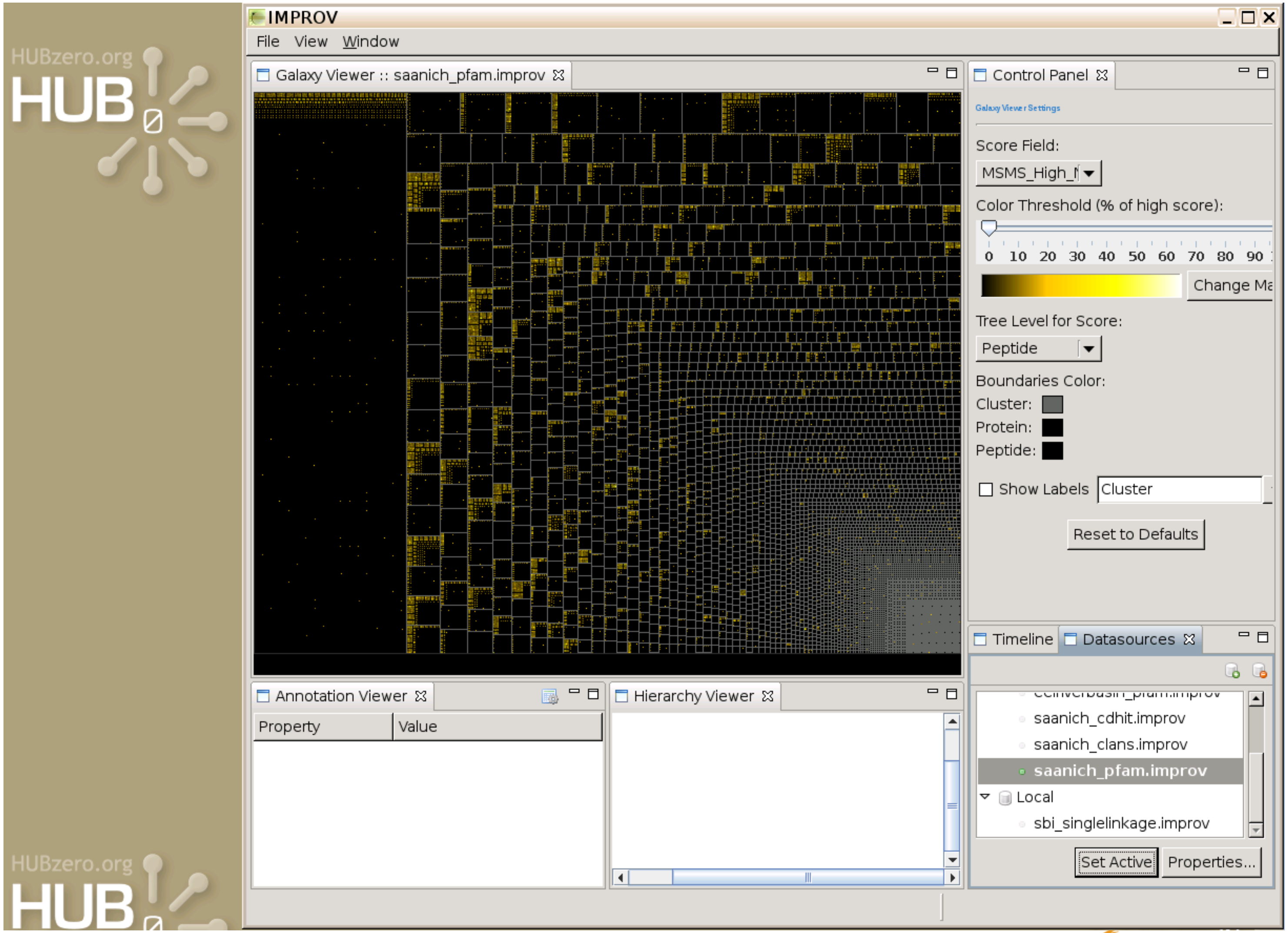
Obtain mass spec proteomics data from EMSL

EMSLhub submits parallel jobs to compute resources

Perform Sequence Comparisons & Clustering

Clustered protein sequences viewed in IMPROV

Example: Metaproteomics pipeline implementation in EMSLHub



Score Field: MSMS_High_I

Color Threshold (% of high score): 0 10 20 30 40 50 60 70 80 90

Tree Level for Score: Peptide

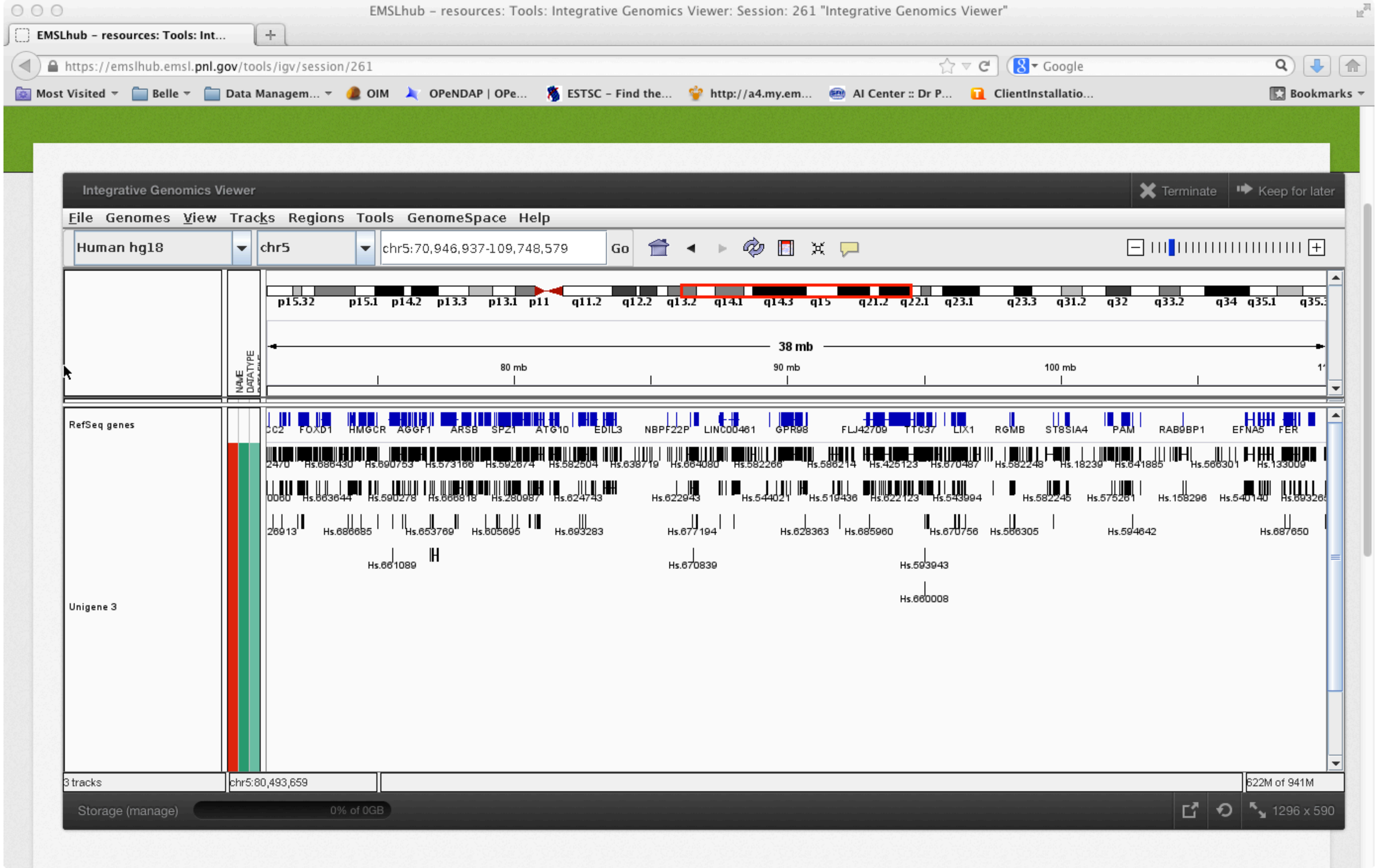
Boundaries Color:
Cluster:
Protein:
Peptide:

□ Show Labels Cluster

Property	Value

- cellverbasin_pfam.improv
- saanich_cdhit.improv
- saanich_clans.improv
- saanich_pfam.improv
- Local
  - sbi_singlelinkage.improv

Example: Genomics in EMSLHub



EMSLhub – resources: Tools: Integrative Genomics Viewer: Session: 261 "Integrative Genomics Viewer"

https://emslhub.emsl.pnl.gov/tools/igv/session/261

Human hg18

chr5

chr5:70,946,937-109,748,579

3 tracks

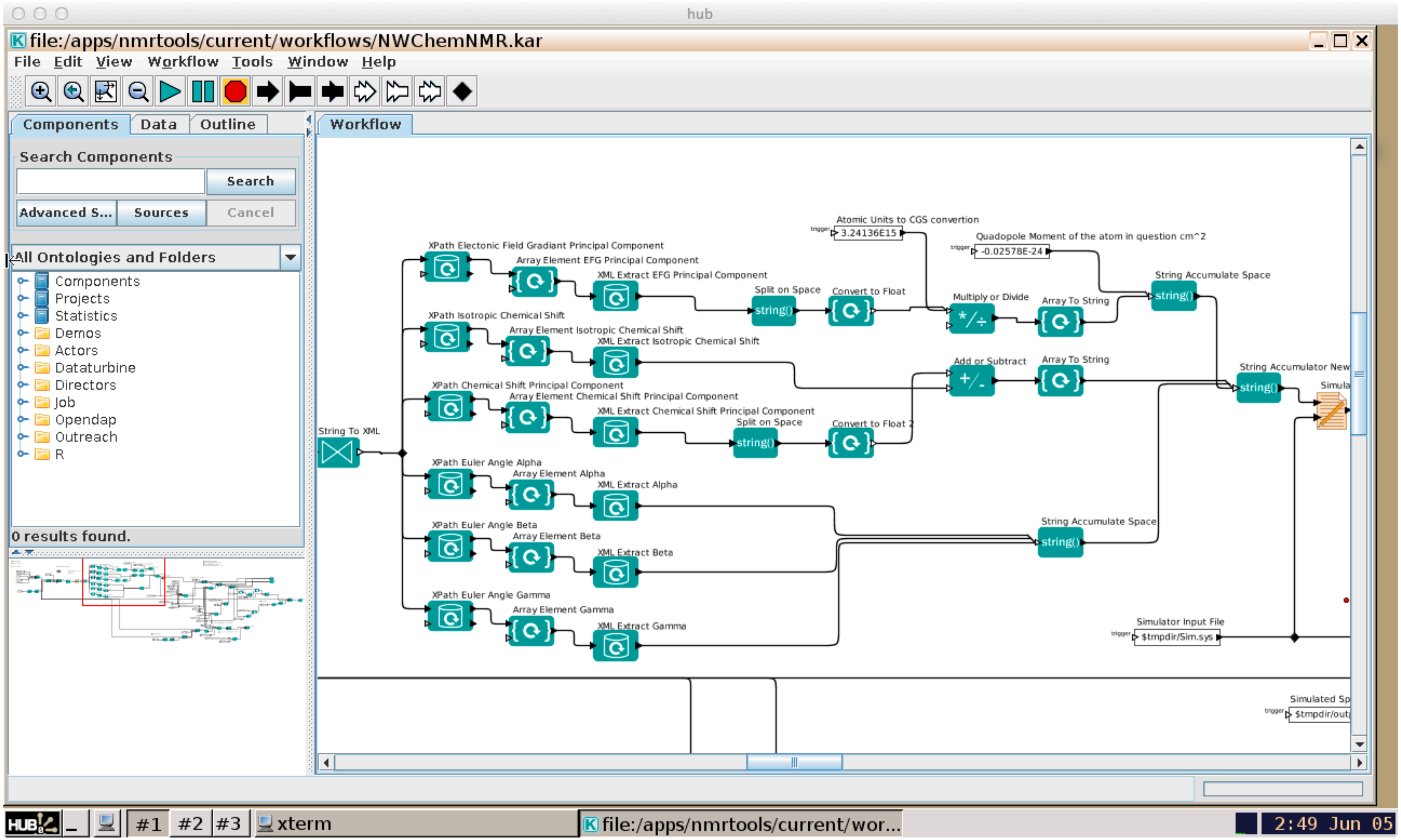
chr5:80,493,659

622M of 941M

Storage (manage) 0% of 0GB

1296 x 590

Example: Kepler workflow-based computing in EMSLHub



Pacific Northwest NATIONAL LABORATORY

Proudly Operated by **Battelle** *Since 1965*

U.S. DEPARTMENT OF ENERGY

Supporting the need for new approaches to data & collaboration

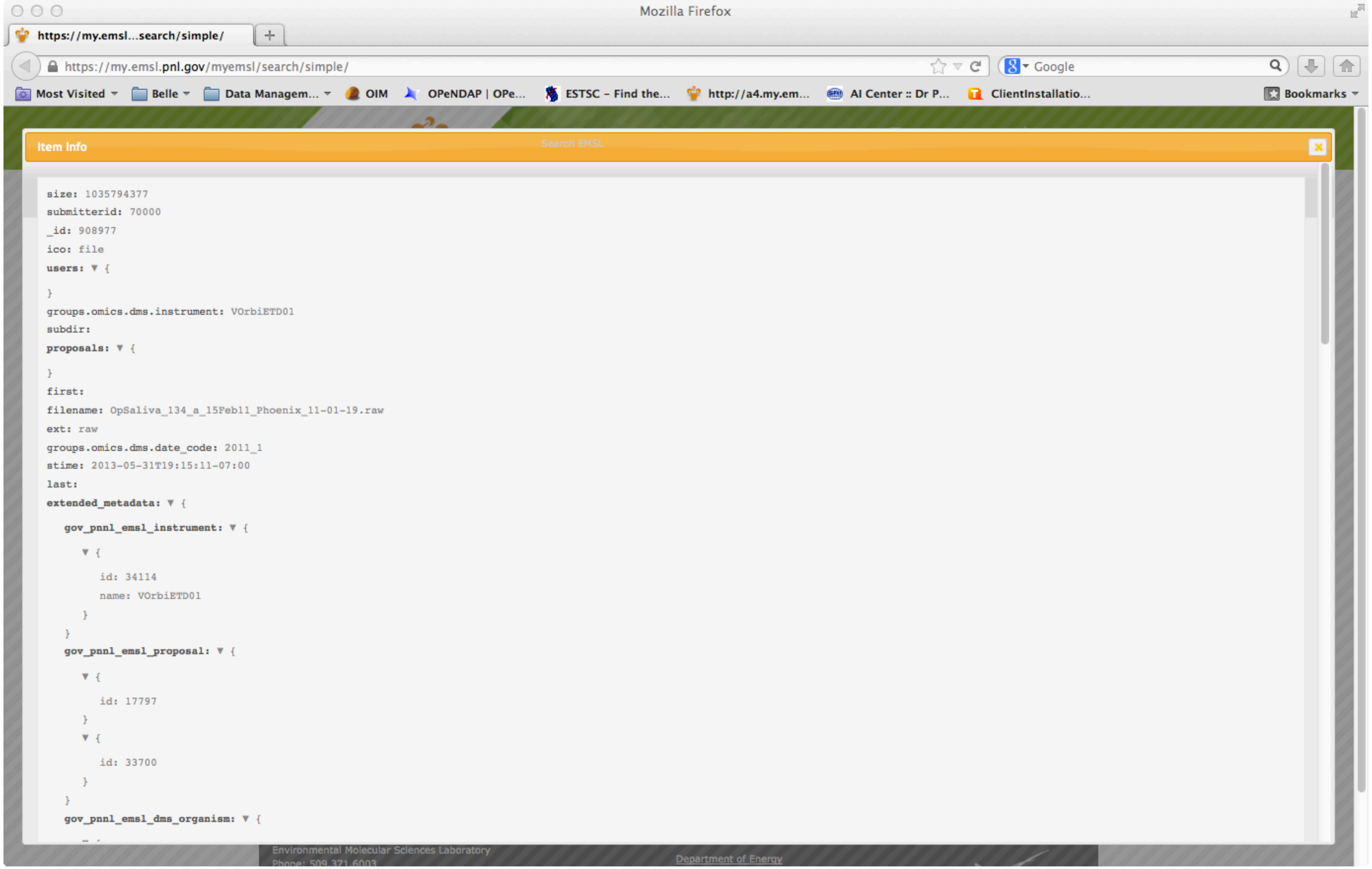
- Data is being generated both experimentally and computationally at an very rapid pace
- Islands of data are growing at institutions all over the world
- The leading edge of scientific discovery in the 21st century will be attained by collaboration and combining data in new ways
- MyEMSL & EMSLHub will help us
  - Cultivate the exchange of ideas
  - Leverage shared, searchable data
  - Foster collaborations, sharing & discovery

Team Members

- Project Management
  - William Shelton
  - David Cowley
- Core Project Team
  - Kevin Glass
  - David Brown
  - Brock Erwin
  - Kevin Fox
  - Nate Trimble
- EMSLHub
  - David Brown
  - Brock Erwin
  - Silvia Hoisie
- Metagenomics partners
  - Angela Norbeck
  - Ken Auberry
  - David Brown
- NMR partners
  - Karl Mueller
  - David Brown
  - Herman Cho
  - Bert DeJong
  - Brock Erwin
  - Nancy Washton
- Microscopy Partners
  - Nigel Browning
  - James Evans
  - James Bouwer

Thank You!

Metadata Sample



```
size: 1035794377
submitterid: 70000
_id: 908977
ico: file
users: ▼ {

}
groups.omics.dms.instrument: VOrbiETD01
subdir:
proposals: ▼ {

}
first:
filename: OpSaliva_134_a_15Feb11_Phoenix_11-01-19.raw
ext: raw
groups.omics.dms.date_code: 2011_1
stime: 2013-05-31T19:15:11-07:00
last:
extended_metadata: ▼ {

   gov_pnnl_emsl_instrument: ▼ {

      ▼ {

         id: 34114
         name: VOrbiETD01
      }
   }
   gov_pnnl_emsl_proposal: ▼ {

      ▼ {

         id: 17797
      }
      ▼ {

         id: 33700
      }
   }
   gov_pnnl_emsl_dms_organism: ▼ {
```