

Rich data sharing via SeedMeLab

Amit Chourasia, San Diego Supercomputer Center, UC San Diego



<https://SeedMeLab.org>

Share...

Manage...

Search...

Visualize...

...data, as never before!

- Obstacles for collaborators
 - Not all team members have/want access to compute cluster(s)
 - Command line interface adds barriers for many users
- Results spread between different compute resources
 - Local computer, clusters, other file servers
- Recurring overhead with sharing
 - First download to local computer
 - Upload to file sharing service
- Alienation of data from its context
 - email body
 - email attachments

Each group develops an adhoc data sharing solution

- Build expertise and expend time on this rather than research
- Maintain and operate software and infrastructure
- Who should do this?
- How to sustain when key personnel leaves?

Data
stumbling
blocks

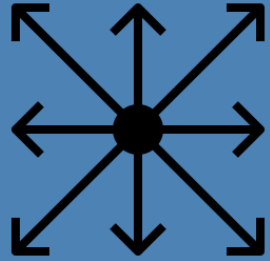
Access control

Collaboration

Storage

Transfer

Automation



Content dispersion

- Emails, meeting notes, protocols, parameters, etc.
- Lack of authoritative service to share data
- Scattered data is harder to manage and use well



Weak presentation

- Data presentation requires context to provide it depth and meaning



Dark data

- Without context, data becomes harder to find, use and reuse



Data shall be

1. Accessible on the web via Browser, API & Command line
2. Share-able with access control
3. Annotate-able with context & metadata
4. Presentable in a rich format

Over 80% of data movement (by number of files) on HPC systems is < 1MB

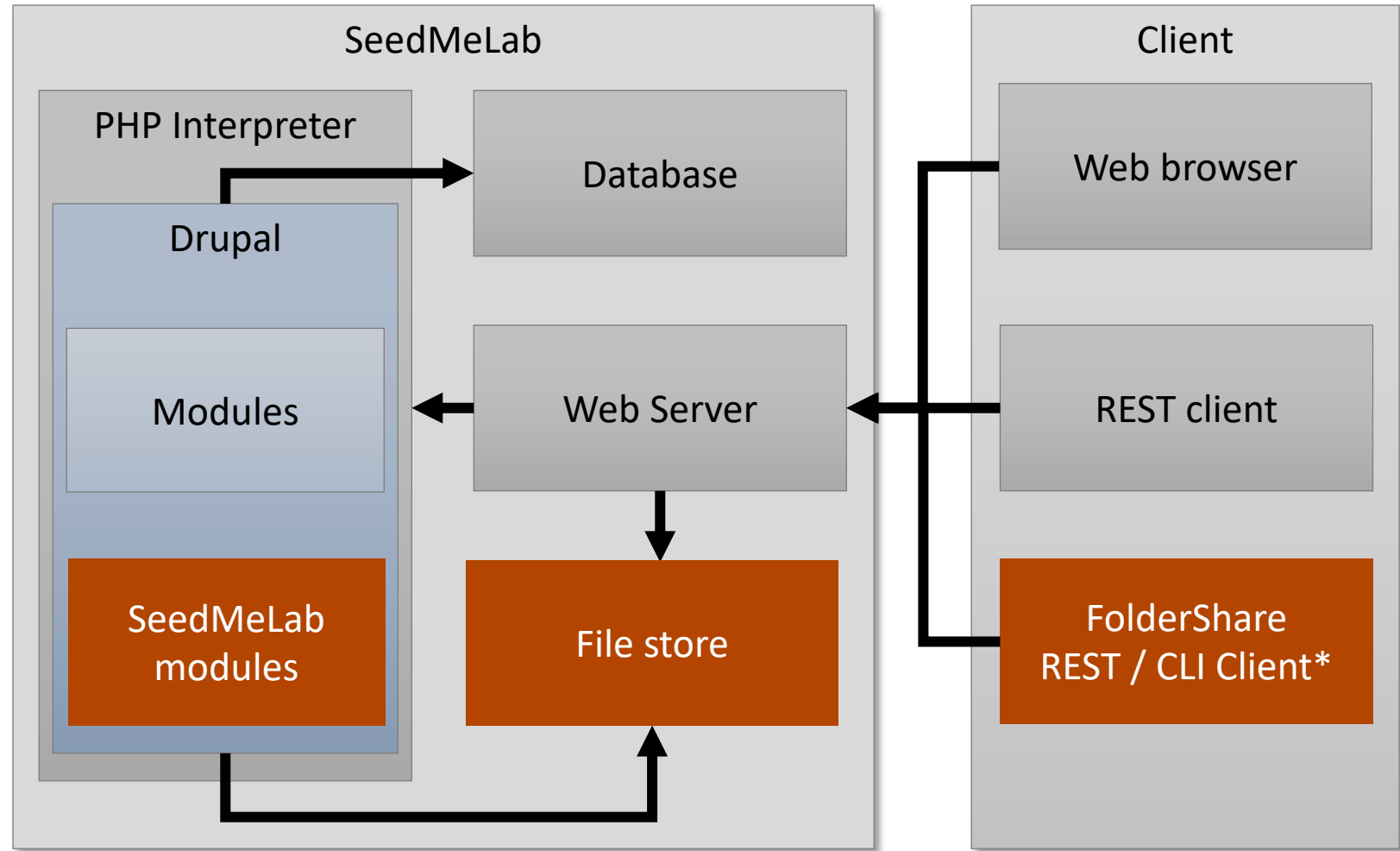
[Watch video: SeedMeLab overview](#)



SeedMeLab

Specialized modules for Drupal

- Data management
- REST service + client
- Visualization plugins
- Single sign-on



* FolderShare client for automation/scripting

Drupal

- Mature (Version 9)
- Battle tested
- Sustained (20 years)
- Extensive built-in features
- Community contributed modules (9,700+)
- 1+ Million active websites
- Open source (GPL2 License)

Community

- 4,000+ core contributors
- 46K Active contributors
- 1.4M users on drupal.org
- 1,000+ vendors
- Drupal security bug bounty €15,000 (Funded by EU-FOSS)

Content management

- Articles, blog entries, comments, forums, lists
- Deep customization for specific content
- Search (built-in and external)
- Caching, CSS & JS aggregation
- File upload, download, WYSIWYG editors
- Content rating, favorites lists, analytics

Account management

- Log in, reset password, request account
- Delete account, block account, email confirms
- User lists & profiles, user & site contact forms
- Password strength indicators, password policies
- Repeat password block, leaked password block
- CAPTCHA, LDAP, OAuth, 2FA

Site layout

- Primary & secondary menus, hierarchical menus
- Breadcrumbs, page & form layout, formatting
- Image, audio, & video management & display
- Regional settings, time zones, UI languages
- Vocabularies, RDF, RSS, meta tags, site maps
- Themes, responsive layouts

Security

- Log in flood control, log in history, failure logging
- Automatic account block after repeated failures
- IP address block, IP blacklist block
- User input filtering, SPAM filtering
- Role-based permissions, content access control

■ Capabilities provided by contributed modules

Academia

Brown University
Cal State LA
Cal State Long Beach
Dartmouth
Georgetown U.
Georgia Tech
Oregon State U.
Oxford U.
Rutgers U.
Stanford U.
UC San Francisco
U. of Colorado
U. of Minnesota

Government

City of Austin
City of London
City of LA
Gov. of Australia
Gov. of Bermuda
London Gatwick Airport
New York State
Royal Mail
State of Colorado
State of Georgia
UN Food Program
US Commerce Dept.
US Transportation Dept.
World Economic Forum

Healthcare

Care
Doctors without Borders
Greenpeace
Field Museum
Leukemia Lymphoma Soc.
Muscular Dystrophy Ass.
Museum of Fine Arts
National Audubon Society
Oxfam
Red Cross
Rotary International

Industry

Alcatel-Lucent
Box.com
Cisco
eBay (Corp. site)
Genentech
General Electric
Grammys
Guardian Financial
Johnson & Johnson
Lush
Novartis
Pfizer
Pinterest (Corp. site)
Princess Cruises
Redhat
Tesla
Twitter (Corp. site)
Verizon

News

Al Jazeera
BBC Store
Fox
The Economist
NBC
Turner Broadcasting
Warner Music Group

Sports

Dallas Cowboys
Kentucky Derby
Major League Soccer
NBA
NCAA
Patriots
PGA

Long term benefits of building on Drupal framework

1. Easily usable, customizable and extensible
2. Deployable in turn-key fashion
3. Mature & sustainable
4. Supported by a large ecosystem
5. Open source & distributable

Data Management



Organize and manage data

Sample folder

Public content shared with everyone by you



Search folder...

Favorites

 Site Admin

 Shared

 Public

 All

Name	Modified	Size	Owner
 Images	11/26/20 1:07 AM	1.14 MB	Site Admin
 Vis Showcase	11/26/20 1:07 AM	2.87 MB	Site Admin
 Visible spectrum.png	12/13/20 10:52 PM	607 bytes	Site Admin
 mandelbrot.mp4	11/26/20 1:07 AM	744.01 KB	Site Admin
 seedmelab-overview.pdf	11/26/20 1:08 AM	1.87 MB	Site Admin

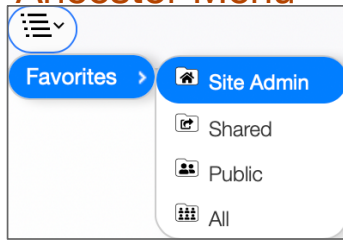
5 items

SeedMeLab File Manager

Builds upon status quo

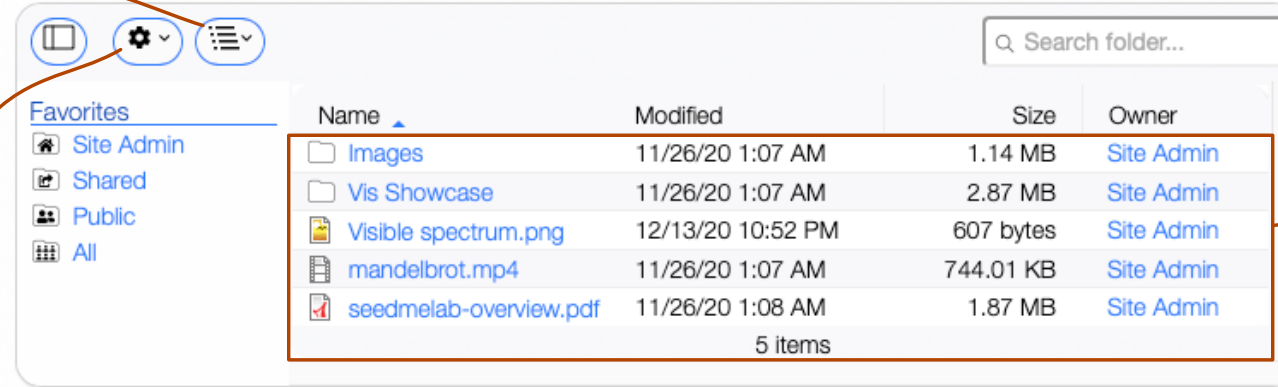
Standard tree hierarchy for files and folders

Ancestor Menu



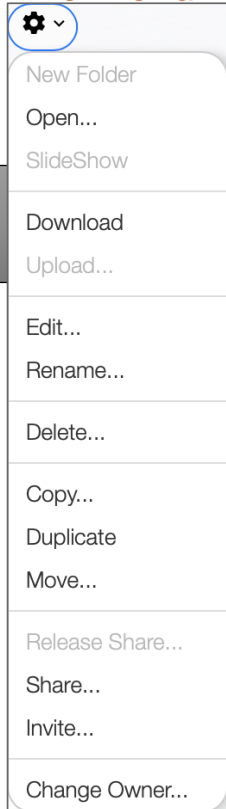
Sample folder

Public content shared with everyone by you



Upload files via
drag & drop

File Menu



Standard file operations

Data Expressiveness



Describe & Discuss Data

Sample folder

Public content shared with everyone by you

Search folder...

Favorites

Site Admin

Shared

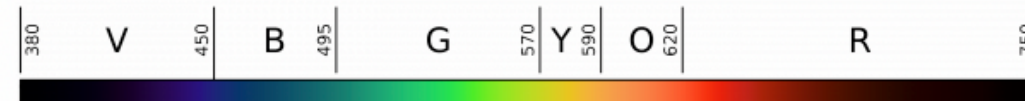
Public

All

Name	Modified	Size	Owner
Images	11/26/20 1:07 AM	1.14 MB	Site Admin
Vis Showcase	11/26/20 1:07 AM	2.87 MB	Site Admin
Visible spectrum.png	12/13/20 10:52 PM	607 bytes	Site Admin
mandelbrot.mp4	11/26/20 1:07 AM	744.01 KB	Site Admin
seedmelab-overview.pdf	11/26/20 1:08 AM	1.87 MB	Site Admin

5 items

The [visible spectrum](#) is the portion of the electromagnetic spectrum that is visible to the human eye.



Properties of few colors in visible spectrum

Color	Wavelength	Frequency	Photon energy
Violet	380–450 nm	670–790 THz	2.75–3.26 eV
Blue	450–485 nm	620–670 THz	2.56–2.75 eV

Roots of quadratic equations

When $a \neq 0$, there are two solutions to $ax^2 + bx + c = 0$ and they are

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

Add rich description
to any file/folder with

- * Formatted text
- * Link, list & table
- * Images & videos
- * Equations

Add custom metadata fields filesystem wide
e.g. Identifiers, dates, keywords, taxonomy, etc.

Sample folder

Public content shared with everyone by you

Search folder...

Favorites

Site Admin

Shared

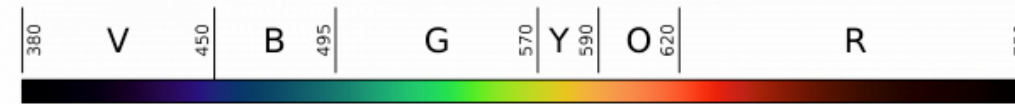
Public

All

Name	Modified	Size	Owner
Images	11/26/20 1:07 AM	1.14 MB	Site Admin
Vis Showcase	11/26/20 1:07 AM	2.87 MB	Site Admin
Visible spectrum.png	12/13/20 10:52 PM	607 bytes	Site Admin
mandelbrot.mp4	11/26/20 1:07 AM	744.01 KB	Site Admin
seedmelab-overview.pdf	11/26/20 1:08 AM	1.87 MB	Site Admin

5 items

The [visible spectrum](#) is the portion of the electromagnetic spectrum that is visible to the human eye.



Properties of few colors in visible spectrum

Color	Wavelength	Frequency	Photon energy
Violet	380–450 nm	670–790 THz	2.75–3.26 eV
Blue	450–485 nm	620–670 THz	2.56–2.75 eV

Roots of quadratic equations

When $a \neq 0$, there are two solutions to $ax^2 + bx + c = 0$ and they are

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

Comments

Submitted by [Site Admin](#) on Fri, 12/27/2019 - 14:58 [Permalink](#)

Comments may be posted to discuss any file or folder.

Discuss any
file/folder with
comments

Data Presentation & Branding



Present, brand and visualize data

Sample folder

Public content shared with everyone by you



Q Search folder...

[Site Admin](#)

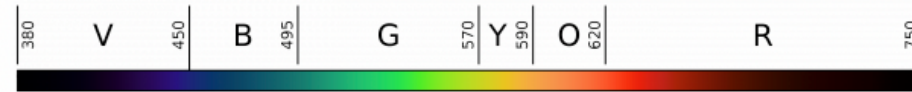
[Shared](#)

[Public](#)

[All](#)

Name	Modified	Size	Owner
 Images	11/26/20 1:07 AM	1.14 MB	Site Admin
 Vis Showcase	11/26/20 1:07 AM	2.87 MB	Site Admin
 Visible spectrum.png	12/13/20 10:52 PM	607 bytes	Site Admin
 mandelbrot.mp4	11/26/20 1:07 AM	744.01 KB	Site Admin
 seedmelab-overview.pdf	11/26/20 1:08 AM	1.87 MB	Site Admin
5 items			

The [visible spectrum](#) is the portion of the electromagnetic spectrum that is visible to the human eye.



Properties of few colors in visible spectrum

Color	Wavelength	Frequency	Photon energy
Violet	380–450 nm	670–790 THz	2.75–3.26 eV
Blue	450–485 nm	620–670 THz	2.56–2.75 eV

Roots of quadratic equations

When $a \neq 0$, there are two solutions to $ax^2 + bx + c = 0$ and they are

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

Comments

Submitted by [Site Admin](#) on Fri, 12/27/2019 - 14:58 [Permalink](#)

Comments may be posted to discuss any file or folder.

Add custom metadata fields filesystem wide
e.g. Identifiers, dates, keywords, taxonomy, etc.

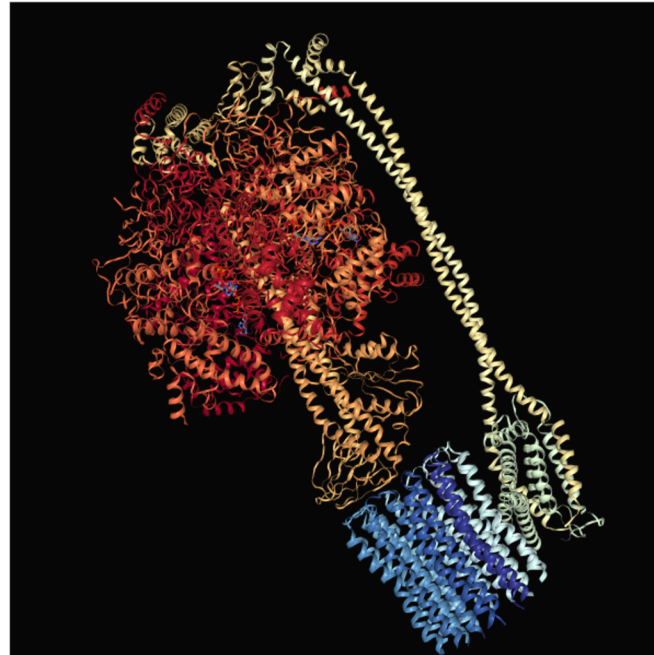
Customizable
layout, theme
& branding

structure-of-protein-6n2y.pdb

Public content shared with everyone by you



structure-of-protein-6n2y.pdb (2841399 bytes)



6N2Y

[Bacillus PS3 ATP synthase class 1](#)

- **DOI:** [10.2210/pdb6N2Y/pdb](https://doi.org/10.2210/pdb6N2Y/pdb)
- **EMDataResource:** [EMD-9333](#)
- **Classification:** [HYDROLASE](#)
- **Organism(s):** [Bacillus sp. PS3](#)
- **Expression System:** [Escherichia coli](#)
- **Mutation(s):** No
- **Deposited:** 2018-11-14 **Released:** 2019-02-20
- **Deposition Author(s):** [Guo, H.](#), [Rubinstein, J.L.](#)
- **Funding Organization(s):** Canadian Institutes of Health Research (CIHR), Japan Society for the Promotion of Science (JSPS)

Experimental Data Snapshot

- **Method:** ELECTRON MICROSCOPY
- **Resolution:** 3.00 Å
- **Aggregation State:** PARTICLE
- **Reconstruction Method:** SINGLE PARTICLE

Auto generate and present
visualization for any filetype
(via plugins)

Data Discovery



Search data and context

Sample folder

Public content shared with everyone by you

Q Search folder...

Favorites

Site Admin

Shared

Public

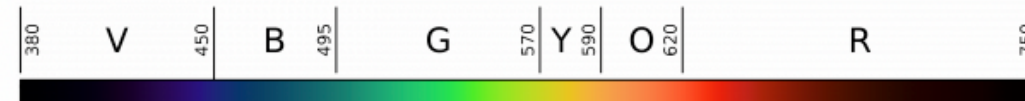
All

Name	Modified	Size	Owner
Images	11/26/20 1:07 AM	1.14 MB	Site Admin
Vis Showcase	11/26/20 1:07 AM	2.87 MB	Site Admin
Visible spectrum.png	12/13/20 10:52 PM	607 bytes	Site Admin
mandelbrot.mp4	11/26/20 1:07 AM	744.01 KB	Site Admin
seedmelab-overview.pdf	11/26/20 1:08 AM	1.87 MB	Site Admin

5 items

Search filenames
& description

The [visible spectrum](#) is the portion of the electromagnetic spectrum that is visible to the human eye.



Properties of few colors in visible spectrum

Color	Wavelength	Frequency	Photon energy
Violet	380–450 nm	670–790 THz	2.75–3.26 eV
Blue	450–485 nm	620–670 THz	2.56–2.75 eV

Roots of quadratic equations

When $a \neq 0$, there are two solutions to $ax^2 + bx + c = 0$ and they are

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

Comments

Submitted by [Site Admin](#) on Fri, 12/27/2019 - 14:58 [Permalink](#)

Comments may be posted to discuss any file or folder.

SeedMeLab



User Management



Built in

- User management
- User roles

Single Sign-On with addon modules

- Oauth2
- LDAP

Access Controls



Access grants on file/folder (set by user)

Adjust shared access to "Sample folder" ✕

View access enables other users to view, copy, and download. *Author* access enables other users to edit, delete, move, and upload too.

User	None	View	Author
Everyone that can access this website	☐	☒	☐
Demo Guy	☐	☒	☐
John Doe	☐	☐	☒

Users without View or Author access are automatically removed from this list when the form is saved.

User account name or email address: ○ Add

Save Cancel

Simplified sharing: Supported from top level only

User Role Based Permissions



Global role-based permissions (set by admin)

Permission	Anonymous user	Authenticated user	Site Manager	Administrator
FolderShare				
Administer files and folders	☐	☐	☒	☒
Author files and folders	☐	☒		
Share files and folders with other users	☐	☒		
Share files and folders with the public	☐	☒		
View files and folders	☒	☒		
FolderShare Invite				
Administer invitations	☐	☐	☐	☒
Use FolderShare invite operations	☐	☒		

Versatile Customization Capabilities



Customization options

- Add filesystem fields
- Add data list views
- Visualization plugins
- Processing plugins (validation, reports, ...)
- Layout
- Theme
- Branding

Automation



REST Client

- Client includes SDK (skip API as its complex)
- Command line utility

Demonstration
videos



[Watch video: Managing data on SeedMeLab](#)



[Watch video: Customizing SeedMeLab](#)



Watch video: Managing data via REST client on SeedMeLab

```
$ foldershare --help
```

```
Usage is: foldershare [HOST_OPTIONS] [COMMAND] [OPTIONS]
```

Create, delete, change, and manage files and folders on a remote web site that uses the FolderShare module.

Commands

chown	Change ownership on files and folders.
config	Show the host administrator configuration.
cp	Copy files and folders.
download	Download files and folders.
hostname	Show the current host name.
ls	List files and folders.
mkdir	Make new folders.
mv	Move files and folders.
rm	Remove files or folders.
rmdir	Remove empty folders.
stat	Show the status of files and folders.
update	Update fields on files and folders.
upload	Upload files and folders.
usage	Show the user's storage use.
version	Show version numbers.
wait	Wait until remote files and folders are enabled.
whoami	Show the current user name.
xattr	Get extended attributes of files and folders.

Research groups

- Use a branded data management system
 - Private collaboration
 - Public dissemination

Examples: Relativistic Laser Plasma Group (UCSD), Structural Engineering Group (UCSD), Material Science Group (U Buffalo)

Gateways and scientific applications

- Integrate with existing systems
- Develop new applications

Examples: FlowGate portal (JCVI) , CIPRES Gateway (UCSD), GenApp Gateway (U Montana)

Service providers

- Customize and offer SeedMeLab to their users

Examples: HPCShare @ SDSC

SeedMeLab core modules are available at Drupal.org

<https://seedmelab.org/community>

Title	Active usage ¹
FolderShare	170
FolderShare REST	17
Formatter Suite	270
Chart Suite	78
CILogon Auth	12
Cache Utility	3
Grand total	550

¹Active usage is underreported, as many sites disable telemetry. Updated May 29, 2021

SeedMeLab Docker'ized recipe for technical adopters/evaluators

<https://github.com/sdsc/seedmelab>

Quotes

“My inbox is decongested... swift feedback to team ... easier to write research papers.”

- Alexey Arefiev, UC San Diego

“Fulfills user demand for data sharing ... easy integration with little development effort that leaves us time left over to focus on other opportunities.”

- Mark Miller, CIPRES project, UC San Diego

“Provides advanced file system to GenApp users and applications ... saving our team’s time to focus on other development priorities.”

- Emre Brookes, GenApp Framework, UT Health Science Center San Antonio

Accolade



Enabling researchers

- Proposals
 - Guidance/help on data management plan document
 - Letter of collaboration/commitment
- SeedMeLab instance for research groups
 - Host Rich data + Website (Optional)
 - Provide training
 - Customizations
- Development
 - Visualization plugins and extensions
 - Application integration

Users and integrators for their support & feedback.
Masters, undergraduate & high school interns

Support

Science Gateways Community Institute
Focus Week + Additional engagements including Usability, Marketing & Value proposition

Sponsors

National Science Foundation

This material is based upon work supported by the National Science Foundation under Grant No. OAC-1443083

"Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation."

San Diego Supercomputer Center, UC San Diego

Follow up

Contact: Amit Chourasia

Email: amit@sdsc.edu

Website: SeedMeLab.org

Try – Share – Adopt

SeedMeLab Docker'ized recipe for technical adopters/evaluators

<https://github.com/sdsc/seedmelab>