Enabling rich data sharing for Science Gateways via the SeedMeLab platform

Amit Chourasia
San Diego Supercomputer Center
UCSD
amit@sdsc.edu

David Nadeau
San Diego Supercomputer Center
UCSD
nadeau@sdsc.edu

Jiaping Luo
Computer Science and Engg
UCSD
luojiaping230@gmail.com

Tony Chen
San Diego Supercomputer Center
UCSD
t9chen@ucsd.edu

Mark Miller
San Diego Supercomputer Center
UCSD
mmiller@sdsc.edu

Emre Brookes
Biochemistry
UT Health San Antonio
brookes@uthscsa.edu

***Abstract*— Science Gateways provide an easily accessible and powerful computing environment for researchers. These are built around a set of software tools that are frequently and heavily used by large number of researchers in specific domains. Science Gateways have been catering to a growing need of researchers for easy to use computational tools, however their usage model is typically single user-centric. As scientific research becomes ever more team oriented, the need driven by user-demand to support integrated collaborative capabilities in Science Gateways is natural progression. Ability to share data/results with others in an integrated manner is an important and frequently requested capability. In this article we will describe and discuss our work to provide a rich environment for data organization and data sharing by integrating the SeedMeLab (formerly SeedMe2) platform with two Science Gateways: CIPRES and GenApp. With this integration we also demonstrate SeedMeLab's extensible features and how Science Gateways may incorporate and realize FAIR data principles in practice and transform into community data hubs.**

Keywords—data management, content management system, science gateways

I. Introduction

Science Gateways have been extremely successful in democratizing the use of high performance computing for any user. They have largely remained focused on providing easy and usually swift computation for a predefined set of software tools for a specific research community. Over time, the needs of Science Gateways have evolved and the users have expressed demand for a better collaborative environment. In this article we will discuss and describe how we enable an improved collaborative environment for two Science Gateways by integrating them with the SeedMeLab platform. First we provide a brief overview of each system, followed by challenges, implementation and conclusions.

A. SeedMeLab

The SeedMeLab (formerly SeedMe2) [1, 2] is a set of modular building blocks built on the Drupal content management system for creating data management and data sharing websites. It enables research teams to effectively manage, share, search, visualize, and present their data on the Web using an access-controlled, branded, and customizable website that the team owns and controls. SeedMeLab modules implement a virtual file system that supports storing, sharing and presenting data in a familiar tree hierarchy. It also supports formatted annotations, custom metadata extensions, lightweight visualizations, and threaded comments on any file/folder. It also provides REST interface and corresponding clients to automate data movement and integration with other applications (Fig. 1).

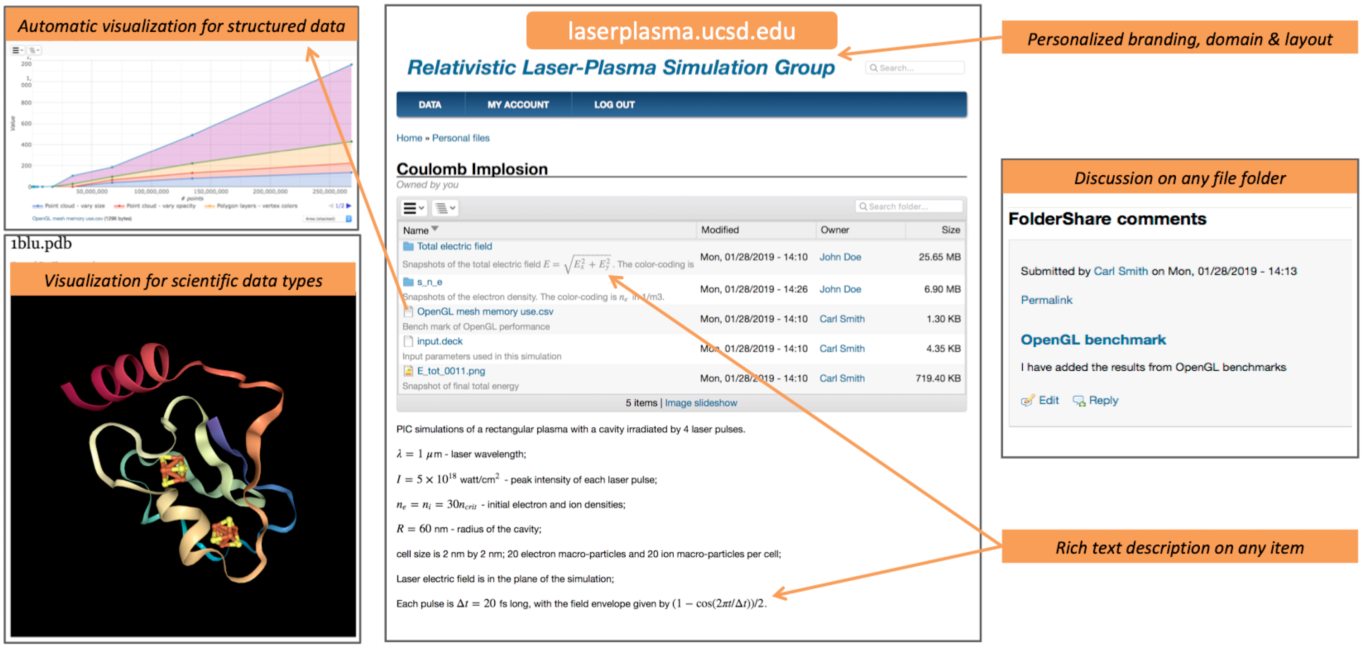
SeedMeLab's capability allows users to add descriptions in rich text format (HTML) to any file or folder, these text annotations are subsequently indexed for search based discovery. The administrator may also configure if text file content should be indexed for search. The search indexing is performed on a periodic basis and updated as content changes. Visualizations and file display can be further customized via extension modules. Existing JavaScript libraries can be easily integrated to generate visualizations of supported file types (Fig. 1 left top and bottom insets).

B. CIPRES Gateway

The CIPRES Science Gateway [3, 4] is a web portal designed to provide researchers with transparent access to the fastest available community codes for inference of phylogenetic relationships, and implementation of these codes on scalable computational resources. Meeting the needs of the community has included developing infrastructure to provide access, working with the community to improve existing community codes, developing infrastructure to insure the portal is scalable to the entire systematics community, and adopting strategies that make the project sustainable by the community. CIPRES has allowed more than 33,000 unique users to run jobs that required 120 million Service Units since its release in December 2009 on XSEDE [5]. CIPRES was created more than a decade ago, following the single-user paradigm that was considered de rigeur at the time of its creation.

Following grant awards supported this work
Chourasia: NSF OAC-1443083
Miller: NSF DBI-1759844, NSF OCI-932251
& NIH 5R01GM126463-01
Brookes: NSF CHE-1265817, NSF OAC-1740097, NSF OAC-1912444
& NIH GM120600

Presented at Gateways 2019, San Diego, CA, USA, September 23–25, 2019. https://osf.io/meetings/gateways2019/


Automatic visualization for structured data
1, 200
1, 000
800
600
400
200
0
Value
50,000,000
100,000,000
150,000,000
200,000,000
250,000,000
# points
Point cloud - vary size
Point cloud - vary opacity
Polygon layers - vertex colors
1/2
OpenGL mesh memory use.csv (1296 bytes)
Area (stacked)
1blu.pdb
Visualization for scientific data types
laserplasma.ucsd.edu
Personalized branding, domain & layout
Discussion on any file folder
Rich text description on any item
Search...
DATA
MY ACCOUNT
LOG OUT
Search folder...
Edit
Reply

Relativistic Laser-Plasma Simulation Group

Home » Personal files

Coulomb Implosion

Owned by you

Name	Modified	Owner	Size
Total electric field Snapshots of the total electric field $E = \sqrt{E_x^2 + E_y^2}$. The color-coding is	Mon, 01/28/2019 - 14:10	John Doe	25.65 MB
s_n_e Snapshots of the electron density. The color-coding is n_e in 1/m3.	Mon, 01/28/2019 - 14:26	John Doe	6.90 MB
OpenGL mesh memory use.csv Bench mark of OpenGL performance	Mon, 01/28/2019 - 14:10	Carl Smith	1.30 KB
input.deck Input parameters used in this simulation	Mon, 01/28/2019 - 14:10	Carl Smith	4.35 KB
E_tot_0011.png Snapshot of final total energy	Mon, 01/28/2019 - 14:10	Carl Smith	719.40 KB

5 items | Image slideshow

PIC simulations of a rectangular plasma with a cavity irradiated by 4 laser pulses.

$\lambda = 1\ \mu$m - laser wavelength;

$I = 5 \times 10^{18}$ watt/cm^2 - peak intensity of each laser pulse;

$n_e = n_i = 30 n_{crit}$ - initial electron and ion densities;

$R = 60$ nm - radius of the cavity;

cell size is 2 nm by 2 nm; 20 electron macro-particles and 20 ion macro-particles per cell;

Laser electric field is in the plane of the simulation;

Each pulse is $\Delta t = 20$ fs long, with the field envelope given by $(1 - \cos(2\pi t/\Delta t))/2$.

FolderShare comments

Submitted by Carl Smith on Mon, 01/28/2019 - 14:13

Permalink

OpenGL benchmark

I have added the results from OpenGL benchmarks

Fig. 1. Sample screenshot of SeedMeLab. The center portion displays file/folder listing with custom branding and descriptive text annotations on any f ile or folder. The right inset shows comment on a data item. The top left inset shows when a CSV file is opened it is parsed by our custom library and uses Google Charts JavaScript library to generate an automatic visualization. The bottom left inset shows when a Protein Data Bank file is opened it uses the NGL viewer JavaScript Library to generate the 3D visualization of protein structure.

The single user model had no concept of collaborative interactions, and CIPRES currently uses a single user-specific file system that does not support data sharing. In annual user surveys, 30-40% of responding CIPRES users expressed a desire for the ability to share data selectively with others they were working with, and the ability to make their work publicly available to others. Because of the costs involved in creating these capabilities, CIPRES management chose to adopt the SeedMeLab platform.

C. GenApp Framework

GenApp [6, 7] provides a framework for rapidly building and deploying GUI and web based graphical front end wrapping command line applications and executing them on variety of compute resources in an extensible manner. For example, the GenApp generated SASSIE-web gateway supports researchers in the small angle scattering community [8]. SASSIE-web has over 600 registered users that ran over 20k jobs in year 2018. Other GenApp generated gateways and learning materials are available at its project website [7]. GenApp provides users with a rudimentary "cloud" file system used primarily for inputs to and outputs from jobs. Users can download and delete these files, but currently there is no support in GenApp generated applications for direct upload other than via attachment at the job submission time. GenApp users have requested advancements to the cloud file system.

II. CHALLENGES

Data sharing between two independent systems (Gateway and SeedMeLab) required overcoming a set of challenges as outlined below.

A. Account synchronization and authentication

Both the Gateway and SeedMeLab support and include a their own user management and authentication system. The two systems need to be synchronized in way that provides a seamless user experience.

B. Data synchronization, organization & sharing

Ideally, data should reside at a single location and be exposed to both systems. While such a setup is possible by sharing a common file system between the two system, this approach is fraught with issues, such as security and authority on which system is permitted to change the data for a given operation and instance. For example, one system may allow the user to delete files/folders that are being used by a process in another system causing a string of failures. On the other hand, using an independent file systems minimizes the issues due to usage contention, but requires that data must be duplicated and kept synchronized between the two systems in a well-defined manner.

In addition to data transfer the Gateway needs to identify and implement when the data will be transferred and how it will be kept synchronized. Should its organization be same or different between both systems?

C. Theme and Layout

The Gateway(s) and SeedMeLab are different systems have differing and distinct theme and layout. However the users are familiar with these as offered by the Gateway, therefore SeedMeLab needs to incorporate theme and layout elements where possible such that the user has a sense of continuity between the integrated yet two distinct underlying systems.

III. IMPLEMENTATION

In this section we describe our implementation, and how we overcame the challenges set forth earlier to support the gateway. Each gateway runs a private, gateway-specific instance of the SeedMeLab platform that is further customized to their needs.

A. *Account synchronization and authentication*

We have implemented an extensible module such that the independent user management systems between Gateway and SeedMeLab work in unison. In this module SeedMeLab is provided read-only access to the Gateway database tables that contain user and user session information. At setup stage all Gateway users are imported into SeedMeLab. Subsequently a periodic process checks for new users on Gateway and adds them to SeedMeLab. If a new Gateway user happens to log on to SeedMeLab before their accounts were imported by periodic process, their accounts are provisioned during the login process (just in time) from the Gateway. A user may log on to SeedMeLab in two ways:

1) *Direct login on SeedMeLab website:* User enters the same authentication credentials they use on the Science Gateway on the SeedMeLab login page, a password hash is constructed by SeedMeLab using a matching algorithm used by the Science Gateway. Subsequently the username and hash are matched in the Gateway's database to determine authenticity. On success, the user is logged on to SeedMeLab.

2) *Visiting SeedMeLab website from the Science Gateway website:* In this scenario web browser cookies are used to determine whether a user is already logged in to the gateway. This is achieved by setting up SeedMeLab instance on a subdomain used by the Science Gateway and permitting SeedMeLab instance to read the cookies and extract the session id from them, this session id, when available, is subsequently looked up on the Gateway database and if the session is active, the user corresponding to the session id is matched and subsequently logged on to SeedMeLab. If no session is found, the user would use a direct login as noted above.

B. *Data synchronization, organization & sharing*

SeedMeLab provides a virtual file system where the Science Gateway can transfer and manage data as deemed necessary. The data movement between the two systems is accomplished by using a REST client that provides rich capabilities for data transfer and its management. The two gateways implement data synchronization with a different design

Data synchronization requires an ability for the Gateway to exchange data with SeedMeLab in userspace such that any data transferred to/from SeedMeLab is assigned ownership for the user and not by Gateway administrator. This is accomplished by implementing another REST based authorization scheme that allows any Gateway administrative accounts to manage data in userspace. This authentication scheme is strictly limited to roles that have Gateway administrator privileges and not available to other users.

GenApp assigns a distinct root directory for each user which contains a set of project subdirectories created by the user. When a user submits a job, it is run in the selected project subdirectory. This subdirectory may contain files uploaded during job submission or the results of a prior job runs. To integrate with SeedMeLab, we add two steps to the job processing. First, before executing a job, the project subdirectory is synchronized between Gateway and SeedMeLab to account for any data changes to either system. Secondly, after job completion, the project subdirectory is synchronized again, so the compted results are transferred to SeedMeLab for immediate availability. GenApp's separation of user project subdirectories allows relatively efficient synchronization, as the user's entire file tree need not be synchronized. GenApp and SeedMeLab are both written in PHP; an integration script was developed that handles the data synchronization via the SeedMeLab's REST client.

CIPRES Gateway integration follows a different data movement pattern. CIPRES user interface adds a new capability for the user to trigger data transfer only the selected files/ from the job outputs to SeedMeLab on-demand. The data organization is kept intact between the two systems so the user can easily find the data. CIPRES is written in JAVA and accomplishes this by implementing an orchestration to use SeedMeLab's command line client that is written in PHP using system calls.

On SeedMeLab, a user can share any existing data with other users either publicly or privately. The user may upload additional data, delete or reorganize existing data, they may also add description to any file or folder to provide richer context for collaboration.

C. *Theme and Layout*

SeedMeLab is built on Drupal content management system, that offers a very rich and extensible theme and layout system. For both gateways we customize the theme and layout to closely match the principal elements of menu interface design used by Gateways and their branding. Both gateways add a top level menu for users to switch back and forth between two systems.

IV. CONCLUSION

SeedMeLab integration with Science Gateways offers new and rich data-centric collaboration capabilities previously not available in an integrated fashion to their users, filling an important feature and efficiency gap that was frequently requested. With this integration, the Science Gateway users can not only share their data, but also annotate them with context and discuss any file/folder using a commenting system. By leveraging SeedMeLab, Gateways saved a considerable amount of effort and associated costs by not developing these capabilities in-house.

Such integration is not limited to two Gateways presented here. This process can be easily adopted by other Gateways with minor customization in SeedMeLab to support account synchronization, branding and layout; and minor extension by the Gateway to incorporate data synchronization using SeedMeLab's REST client.

In conclusion, SeedMeLab provides rich data capabilities that are complimentary to Science Gateways and their integration enables users to easily collaborate, reuse and publish their experiments, data and results. This integration presents new opportunities for Science Gateways to serve as data publishers and provide access to a very large corpus of reusable data products to their scientific communities thus transforming into community data hubs which promotes and realizes values of Findable, Accessible, Interoperable and Reusable (FAIR) data principles in practice [9].

V. FUTURE WORK

Production deployment for this integration work in underway. Based on end user feedback and demand, additional capabilities of SeedMeLab such as adding comments on any data item, custom metadata fields for data items, sharing data by inviting non gateway users and visualizations will be rolled out to users. We will also investigate support for external file systems that are used by the Gateway to avoid data duplication. Lastly, we will explore capabilities to assign globally unique persistent identifiers to enable data publishing.

ACKNOWLEDGMENTS

SeedMeLab (formerly SeedMe2) is supported by the National Science Foundation under Grant No. OAC-1443083 to Chourasia. CIPRES Science Gateway was supported in part by National Science Foundation awards NSF DBI-1759844 and OCI-932251 as well as National Institutes of Health award 5R01GM126463-01 to Miller. GenApp is supported by NSF grants CHE-1265817, OAC-1740097 and OAC-1912444 and NIH grant GM120600 to Brookes. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562. 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.

REFERENCES

[1] A. Chourasia, D. Nadeau, and M. Norman. 2017. SeedMe2: Data Sharing Building Blocks. In Proceedings of the Practice and Experience in Advanced Research Computing 2017 on Sustainability, Success and Impact (PEARC17). ACM, New York, NY, USA, Article 69, 1 pages. DOI: https://doi.org/10.1145/3093338.3104153

[2] Create extensible data sharing websites powered by SeedMe2 building blocks. Retrieved Feb 19, 2019 from https://dibbs.SeedMe2.org

[3] Miller, M.A., Pfeiffer, W., and Schwartz, T. (2010) "Creating the CIPRES Science Gateway for inference of large phylogenetic trees" in Proceedings of the Gateway Computing Environments Workshop (GCE), 14 Nov. 2010, New Orleans, LA pp 1 - 8.

[4] Portal | CIPRES. Retrieved Feb 19, 2019 from https://www.phylo.org

[5] John Towns, Timothy Cockerill, Maytal Dahan, Ian Foster, Kelly Gaither, Andrew Grimshaw, Victor Hazlewood, Scott Lathrop, Dave Lifka, Gregory D. Peterson, Ralph Roskies, J. Ray Scott, and Nancy Wilkins-Diehr. 2014. XSEDE: Accelerating Scientific Discovery. Computing in Science & Engineering 16, 5 (2014), 62–74. https://doi.org/10. 1109/MCSE.2014.80

[6] Emre H. Brookes. 2014. An Open Extensible Multi-Target Application Generation Tool for Simple Rapid Deployment of Multi-Scale Scientific Codes. In Proceedings of the 2014 Annual Conference on Extreme Science and Engineering Discovery Environment (XSEDE '14). ACM, New York, NY, USA, Article 53, 6 pages. DOI: https://doi.org/10.1145/2616498.2616560

[7] GenApp. https://genapp.rocks. Retrieved Feb 19, 2019 from https://genapp.rocks

[8] Stephen Perkins, David Wright, Hailiang Zhang, Emre Brookes, Jianhan Chen, Thomas Irving, Susan Krueger, David Barlow, Karen Edler, David Scott, and N. Terrill. 2016. Atomistic modelling of scattering data in the Collaborative Computational Project for Small Angle Scattering (CCP-SAS). Journal of Applied Crystallography 49, 6 (2016) https://doi.org/10.1107/S160057671601517X

[9] Wilkinson, M. D., M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, et al. 2016. “The FAIR Guiding Principles for scientific data management and stewardship.” Scientific Data 3 (1): 160018. doi:10.1038/sdata.2016.18. http://dx.doi.org/10.1038/sdata.2016.18.