SeedMe: Data sharing building blocks

Amit Chourasia*; David R. Nadeau; and Michael L. Norman
*Corresponding author: San Diego Supercomputer Center, University of California, San Diego, La Jolla, CA, 92093, USA; email: amit@sdsc.edu

Abstract: *The need for data sharing and rapid data access has become central with the rise of collaborative research in many disciplines. For the general public, several file sharing products are available that post and share files using web browsers. But for science data and research use, these products are not well suited. While consumer products get by with manual user interfaces to add and remove a few shared files, this is not practical for sharing large numbers of science data files, like those generated during and after large-scale computation. Instead, automated and scriptable mechanisms are required that can integrate into computation workflows to post files during and after computation jobs. Scientific data sharing also requires support for collaborative discussion of research results, quick rough-draft visualizations to analyze the data, and support for metadata and descriptive information that can record job and compute platform characteristics, input data, job parameters, job completion status, and other provenance information.*

Here we describe work in progress under the umbrella of the SeedMe (Stream, Encode, Explore and Disseminate My Experiments) project that is developing scientific data-sharing and data management tools that cater to the unique needs of computational scientists. These tools support automated and scriptable access to shared data, browser-based data access, secure data storage, sharing with a project workgroup, data descriptions and metadata, threaded collaborative discussion, and light-weight visualization.

1 Introduction

Collaborative research depends on data sharing and timely access to data. This is particularly challenging in computational science research where experiments are conducted by teams of distributed researchers doing different tasks at different times and using different compute resources from laptops to High Performance Computing (HPC) clusters. Keeping track of the shared pool of source and result data is challenging. The ability to share this data, metadata, job parameters, computed results, rough-draft visualizations, and discussion threads quickly and easily is essential for efficient collaboration.

While HPC infrastructures support computation well, their support for post-computation data access is generally limited. This forces projects to create their own ad hoc data sharing strategies that copy HPC data to local project-specific repositories and rely upon custom conventions and scripts for recording job parameters, saving metadata, creating rough-draft visualizations, and tracking email threads for data discussion. While this works, it can be cumbersome, labor-intensive, and hard to maintain.

2 SeedMe platform

Our interest is in developing a generic suite or robust modular building-blocks focused upon scientific data sharing, metadata tracking, collaborative discussion, and creating simple visualizations. These building blocks can be installed in *Drupal* [1] content management system (CMS) based project or lab websites to provide a lot of functionality quickly without requiring local expertise in data management, data security, or metadata handling.

The SeedMe (Stream, Encode, Explore and Disseminate My Experiments) platform is a set of modules for the open source *Drupal* CMS. *Drupal* is widely used to create academic and commercial web sites [2,3] and it provides extensive out-of-the-box functionality to customize sites, manage accounts, post articles, and present content for

Proceedings of Gateways 2017, University of Michigan, Ann Arbor, MI; October 23-25, 2017.
https://gateways2017.figshare.com/.

desktop and mobile platforms. With over 4,000 active contributors [4], *Drupal*'s extensive ecosystem includes 1,000+ extension modules [5] to add special features for target markets, including blogs, forums, news aggregators, and commerce. *Drupal* is platform-agnostic, easy to install and manage, and it is supported directly by a large number of hosting services [6] as well as internal academic and government support organizations.

For software developers, *Drupal*'s modular extension capability supports rich programmatic customization to create, manage, and present new types of structured data. The core features of *Drupal* are stable and go through an extensive review process for good security and performance practices. *Drupal* is updated and patched quickly when vulnerabilities are found, and there is a wealth of documentation available.

The worked described here is an evolution of the original SeedMe project [7, 8] that prototyped functionality that is now being built into modular data sharing building blocks for the computation community. The SeedMe building blocks provide functionality to manage, organize, search and quickly visualize data for collaborating project teams. The building blocks support data sharing, data security, access control, and web services. The goal of the project is to enable quick and convenient access to data from variety of compute (HPC/Cloud) and consumption (mobile devices) platforms.

Development of these building blocks is in progress with software going through internal and friendly-user testing now. When ready, multiple related modules will be released:

(a) A core virtual file system module stores, secures, organizes, and presents a hierarchy of files and folders available within a web environment. The module supports rich text descriptions and arbitrary metadata attached to any file or folder. Security and access control features enable users to share their data only with specific collaborators, or publish selected data for public access.

(b) A suite of data presentation modules parse, interpret, and visualize light-weight data, images, and videos.

(c) Extensions to the virtual file system support search operations on data file content along with file and folder names, descriptions, and metadata.

(d) Web service extensions support a REST API [9] to enable remote access to web-available data, including interactive clients and command-line tools for automated and scripted post and retrieve tasks that can integrate with HPC scripts and workflows.

(e) Additional modules enable federated authentication using CILogon [10], notifications, collaborator messaging, and server-side batch processing for data filtering and transcoding for images and videos.

2.1 Architecture

The SeedMe architecture (Fig. 1) is based on *Drupal* and its underlying webserver (typically *Apache* [11]) and database (typically *MySQL* [12]). SeedMe building blocks are implemented as modules plugged into Drupal and customized using web forms for a site administrator.

3 Discussion

In this section we discuss and describe use cases and our software distribution plans.

3.1 Use cases

This work targets a variety of use cases, including:

a) **On-demand monitoring**
A research group needs to monitor and share progress for HPC simulations. Shared data includes job status, statistics, preliminary results and analysis, rough-draft visualizations, and discussion of simulation results. All collaborators may not have direct access to the HPC resource. Using the SeedMe modules, the project team sets up a web site and shared data storage. HPC job scripts periodically post status updates, data, and metadata to the site, triggering update notifications to the team's collaborators. Threaded discussion associated with specific job data supports the team's review of job results and preparations for the

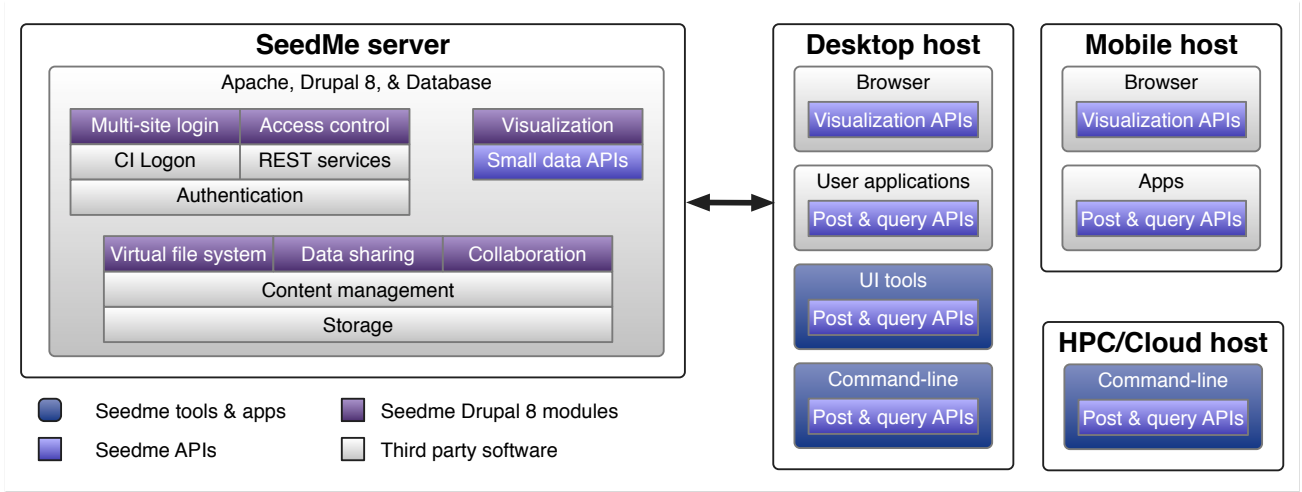

Fig. 1. SeedMe architecture: Modules (purple) and APIs (blue), interaction from mobile, desktop and HPC/Cloud hosts.

next simulation run.

b) **Interface with data analysis, visualization, and ad hoc tasks**
A research group needs to go beyond SeedMe's light-weight visualizations to use sophisticated visualization tools for large-scale 3D surface or volume data and create publication-ready images and animations. Team members using the SeedMe modules installed at their project web site use the module's web services to download relevant data for use in visualization workflows. Newly computed single images and animation sequences are posted back to the site using command-line post tools. Annotated with appropriate metadata, the visualizations are available for review by the research team before iterating on the next round of visualization refinement. When the results are suitable, access control settings are updated to publish the images for wider review or public access.

c) **Application integration**
Scientific applications and workflow systems can easily support the easy flow of data into and out of their systems using generic plugins for SeedMe's web services to a project's data sharing web site. The posted data can automatically trigger notifications to project team members, or transcoding tasks that convert file formats or build videos from animation sequences. The workflows themselves may be posted to the site as part of the data's provenance.

d) **Share reusable content**
Along with workflows, project teams create scripts and configuration files for a wide variety of software tools for computation, visualization, and data formatting. With SeedMe's data sharing modules, all of these may be posted to a project's web site or to a central repository of reusable content. When shared publically, these provide reusable templates and starting points for further analysis and visualization..

e) **Transient data locker**
Science gateways need to share computed results with users who do not have access to the underlying HPC resources supported by the gateway. Instead of leaving these data in a gateway's limited temporary storage, these data may be automatically forwarded to a project's data sharing site using the SeedMe modules and web services. Project web sites may elect to save all of this data indefinitely or set up automatic expiration dates. Notifications may alert data owners as expiration dates approach, giving them time to decide if data is important enough for longer-term storage at the same site or another.

f) **Data repositories**
When data is important, it needs to be saved along with the metadata and threaded discussions that give it context. The same

SeedMe modules can manage longer-term storage. Such data remains on-line, secure, searchable, and browsable. Data access controls can lock the data to prevent further modifications of archived data.

g) **Project, group, and personal data clouds** While SeedMe's modules target collaboration during the creation, analysis, and visualization of new data, the same modules may be used for a general-purpose data cloud for projects, groups, and individuals. Data may be posted, secured, and optionally published for public access.

3.2 Software distribution

SeedMe's modules for *Drupal* will be shared as open source and distributed via the project's own web site [13] (which already uses our modules). The modules will also be available as contributed modules distributed via Drupal.org. All software will follow Drupal's best practices for security, code structure, and documentation.

While the modules may be installed individually into any existing *Drupal* web site, SeedMe will also distribute packages that bundle the modules into a *Drupal* profile that includes everything needed to quickly start up a new web site. Profiles will be provided for a variety of use cases.

3.3 Deployment

Drupal is an easily-installed and very flexible content management system for building custom project, group, department, or company web sites. It is certainly possible for project teams to install *Drupal* themselves and add our modules to support collaboration and data sharing. No programming is required and all configuration is done through self-explanatory web forms.

We also envision *platform-as-a-service* scenarios where an academic, government, or commercial IT service provider or hosting service offers a standardized, supported, and maintained *Drupal* service that integrates the SeedMe modules with other common *Drupal* modules. Project groups would contract with the service to build and maintain custom web sites supporting data sharing. This is already a common service model for *Drupal* sites.

3.4 Demonstration & trail

A demonstration of this project's modules is available at demo.seedme.org [14]. Short-term trial web sites with administrator access may be built automatically at try.seedme.org [15].

4 Conclusion

We have presented a suite of *Drupal* modules that support scientific data sharing and collaboration. They provide a customizable suite or building blocks that provide data management, data security, access control, metadata handling, threaded collaborator discussion, and light-weight visualization tools. These modules address critical community needs for better data sharing and data-focused collaboration.

5 Acknowledgments

This work is supported by the National Science Foundation under Grant No. 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 NSF."

6 References

[1] Drupal. 2016. *Drupal – Open Source CMS*. Retrieved Sep 2, 2016 from http://drupal.org/

[2] Market share trends for content management systems. Retrieved Jun 11, 2017 from https://w3techs.com/technologies/history_overview/content_management

[3] Who uses Drupal? | Drupal.com. http://www.drupal.com/showcases

[4] A very basic table of all contributors to Drupal 8 Core. Retrieved Jun 11, 2017 from http://drupalcores.com

[5] Modules project | Drupal.org. Retrieved Jun 11, 2017 from https://www.drupal.org/project/project_module

[6] Drupal Hosting | Drupal.org. Retrieved Jun 11, 2017 from https://www.drupal.org/hosting

[7] SeedMe. 2016. SeedMe (Stream Encode, Explore and Disseminate My Experiments) Retrieved Sep 2, 2016 from https://www.seedme.org

[8] Amit Chourasia, Mona Wong-Barnum, Dmitry Mishin, David R. Nadeau and Michael L. Norman. SeedMe: A scientific data sharing and collaboration platform. In Proceedings of the XSEDE16 Conference on Diversity, Big Data, and Science at Scale (XSEDE16). ACM, New York, NY, USA, Article 48 , 6 pages. DOI=10.1145/2949550.2949590

[9] Basney, J. Fleury, T. and Gaynor, J. 2014 "CILogon: A Federated X.509 Certification Authority for CyberInfrastructure Logon," Concurrency and Computation: Practice and Experience, Volume 26, Issue 13, pages 2225- 2239, September 2014. http://dx.doi.org/10.1002/cpe.3265

[10] Fielding, R. T. and Taylor, R. N. 2002. "Principled Design of the Modern Web Architecture", *ACM Transactions on Internet Technology (TOIT)* (New York: Association for Computing Machinery) **2** (2): 115–150, May 2002

[11] Apache. 2016. *The Apache HTTP Server Project*. Retrieved Jun 11, 2017 from http://httpd.apache.org/

[12] MySQL. 2016. *MySQL*. Retrieved Sep 2, 2016 from http://www.mysql.com/

[13] SeedMe Science. Retrieved Jun 11, 2017 from http://dibbs.seedme.org

[14] Demo for SeedMe. Retrieved Jun 11, 2017 from http://demo.seedme.org

[15] Trail for SeedMe. Retrieved Jun 11, 2017 from http://try.seedme.org