SeedMe2: Extensible data sharing websites for teams

Amit Chourasia
amit@sdsc.edu

David R. Nadeau
nadeau@sdsc.edu

Mona Wong
mona@sdsc.edu

Michael L. Norman
mlnorman@sdsc.edu

San Diego Supercomputer Center, University of California, San Diego

***Abstract*—Data is an integral part of scientific research, and data size problems have become endemic as computation and analyses are producing an increasingly large amount of data that research teams are inevitably tasked with managing these rapidly growing data collections. Existing solutions are largely focused upon providing storage space, whether local or in the cloud, and a familiar folder tree-style hierarchy. While these file system solutions work, they separate the data from essential contextual information, such as metadata, descriptive text and equations, job execution parameters, visualizations, and on-going data discussion among the researchers. Important discussions, for instance, remain in email logs or forums, while descriptive text is left in README files or embedded in those same email logs and forums. This distribution of contextual information makes it harder to keep track of it all and keep data from being orphaned or misinterpreted. A more unified approach is needed that keeps data and context together within the same storage system.**

This interactive demonstration shows key features of building blocks for data sharing and data management developed by the SeedMe2 (Stream, Encode, Explore and Disseminate My Experiments) project . It enables research teams to manage, share, search, visualize, and present their data in a web-based environment using an access-controlled, branded, and customizable website they own and control. It supports storing and viewing data in a familiar tree hierarchy, but also supports formatted annotations, lightweight visualizations, and threaded comments on any file/folder. The system can be easily extended and customized to support metadata, job parameters, and other domain and project-specific contextual items. The software is open source and available as an extension to the popular Drupal content management system.

Keywords—Data sharing, Visualization, HPC, Science Gateways

I. INTRODUCTION

Collaborative research depends timely access to data with flexible grant access controls. This is especially true in computational science research, where experiments are conducted by distributed research teams using disparate compute resources from laptops to High Performance Computing (HPC) clusters. The ability to share relevant job data, parameters, and results quickly and easily is essential for efficient collaboration.

II. MOTIVATION

Data is, of course, essential for modern research. Typically, data is generated and run through a series of well-defined and/or *ad hoc* steps to create, filter, remap, summarize, and visualize the data to help researchers understand what they have, and what they need to do next. Along the way, discussion among collaborating teams is essential as data features are found and evaluated, and potential next steps are planned. All of

TABLE 1: COMPARISON OF SEEDME2 WITH SEEDME1 AND OTHER DATA SHARING SERVICES

Features	**SeedMe2**	**SeedMe1**	**Cloud drives (Dropbox, One drive, etc.)**
Project focus	Arbitrary data & metadata	Image sequences & videos	Arbitrary data, not metadata
Upload size	Up to 2 GB per file	100 mb per file	Varies (usually <5 GB per file)
Sharing	Read and/or write	Read only	Read and/or write
Folder hierarchy	Tree	None	Tree
Extensible file system	Yes	No	No
Add description on file/folder	Yes	Yes	Varies
Threaded discussion on file/folder	Yes	Yes	For text documents only
Command line client	Yes	Yes	No (from third-party clients)
Visualization	Automatic vis for CSV and JSON files	No	No
Branding and layout	Customizable	Fixed	Fixed
Extensible as a website	Yes	Yes	No
Hosting options	On-premise or on any cloud platform	Central hosting	Central hosting
Open source	Yes	No	No

Presented at Gateways 2018, University of Texas, Austin, TX, September 25–27, 2018.
https://gateways2018.figshare.com/

these tasks, from data analysis to discussion, use a mix of tools on different platforms, from big computation clusters down to desktops and mobile devices. As workflows increase in complexity, so does the coordination required among researchers to keep track of all the data and metadata.

Current common procedures rely upon shared file systems, file permissions for access control, and ad-hoc coordination methods. These work, but they are limited. For instance, it may not be possible or practical to add project-specific metadata to an OS's file system and have it supported by folder viewers and search features. And no major OS supports attaching, managing, and searching threaded discussions on individual files and folders. General-purpose file systems and associated tools are not built to track data, context, and discussion within the same framework. Nor is this likely to change. Third-party data sharing networks have emerged that do some of this, but their general-purpose approach to gain a large target market means they do not support the customization needed for research teams with their own metadata, search, access control, and collaboration needs. In order to address these needs we have developed web-based *SeedMe2* building blocks. A brief feature comparison of *SeedMe2* with *SeedMe1* [4, 5] and other services is provided in Table 1.

III. SeedMe2 Building Blocks

SeedMe2 is a set of modular building blocks that extend the popular open source *Drupal* [1] Content Management System (CMS) to support data-centric collaborative research activities. Below are brief descriptions of our building blocks (a.k.a. modules) .

A. *FolderShare module*

FolderShare is a principal building block that implements a web-based virtual file system for Drupal. It enables users to create a folder tree similar to those in Windows, Mac, or Linux operating systems. It supports storage, management, and sharing of files and folders that each support access controls lists to limit read and write operations to specific collaborators. It also includes a rich feature set to configure how files and folders are listed, viewed, searched, and accessed on the web and via an optional command-line tool using standard web services protocols (e.g. REST). The module also supports an API and features to extend the core feature set via plugins.

The module's feature set has benefits for users, administrators, and developers:

i. For users

- Manage hierarchical structure of files and folders with a familiar user interface (Fig. 1)
- Share any folder tree with access controls (from top level folder including all content inside it) with

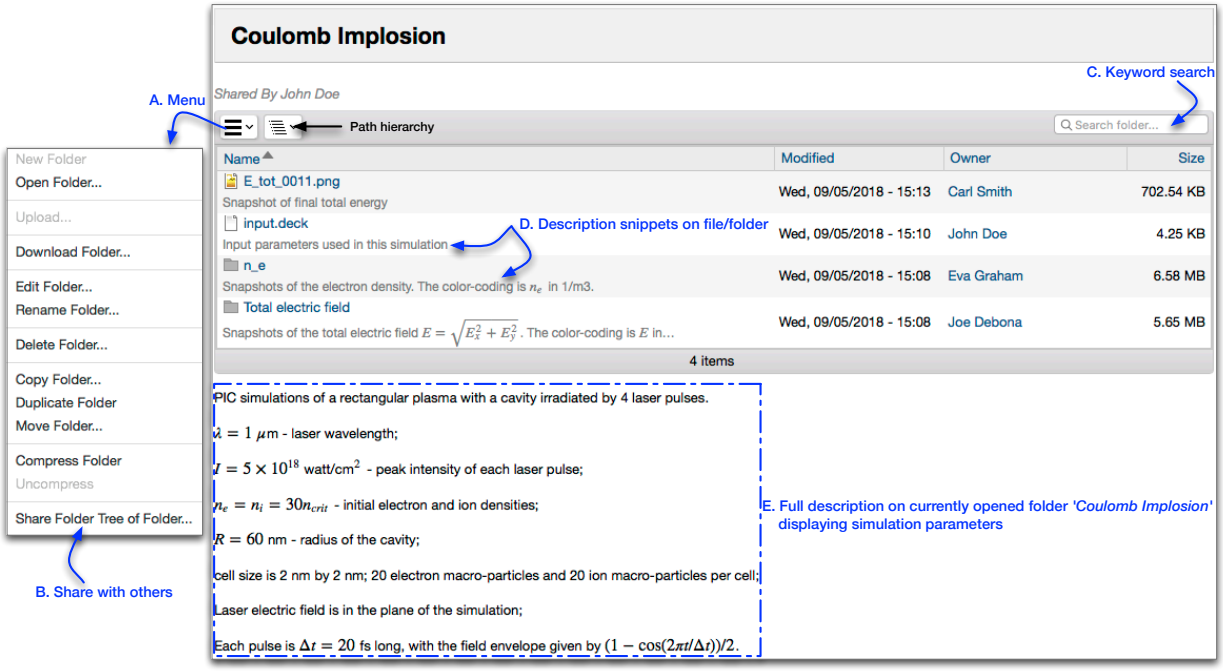

Fig. 1. Sample screenshot showing a folder listing from usage by a physics group (people's names are changed). (A) A menu showing various operations. (B) A menu item for sharing, (C) A search field to search for items by their names, descriptions, and file content. (D) Trimmed description summaries associated with files/folders. (E) A full description of the currently open folder, including text and equations.

any website users. Each top level folder may grant following privileges to site users
    - Public
    - View (read only)
    - Author (write)
- Perform common file management operations such as copy, compress, upload, and download
- Add description to any file and folder
- Add threaded comments to any file or folder
- Use command line utility from remote resources, in workflows or for automation
- View and interact with automatically generated light visualizations of CSV/JSON data

ii. For administrators

- Configure site-wide data storage, share permissions, set restrictions on permitted operations, enable web services, etc.
- Extend *FolderShare's* virtual file system with additional fields such as key value pairs, persistent identifiers, taxonomy, license, etc.
- Configure presentation and data entry tailored to research needs
- Configure search and indexing to enhance discovery of content
- View usage report for site users

iii. For developers

- Extend *FolderShare's* core capabilities through a Plugin API for e.g. add visualization formatter for a data type
- Integrate exiting workflow or scientific tools via REST or Command line clients

B. Chart suite module

Chart suite provides a suite of file field formatters to create interactive charts from file data for tables, trees, and graphs. Charts are shown as line plots, area plots, pie charts, bar charts, tree diagrams, and more. A variety of well-known file formats are supported, including comma-separated values (CSV), tab-separated values (TSV), HTML tables, and JSON tables, trees, and graphs in common schemas. This module can be configured with the *FolderShare* module to automatically visualize file content.

C. Formatter suite module

Formatter suite provides a suite of field formatters to help present numbers, dates, times, text, images, and more. Collectively these support formatting needed for scientific data, including scientific notation; multi-value lists; binary, octal, hex, and arbitrary bases; value range indicators; progress bars; and configurable dates, times, and timestamps.

D. SeedMe2 Architecture

SeedMe2 is based on the *Drupal* CMS, which requires a compatible webserver (typically *Apache* [2]) and a database (typically *MySQL* [3]). Our building blocks are implemented as modules that can be plugged into the Drupal and customized as desired (Fig. 2). *SeedMe2* provides a rich set of collaboration capabilities. A comparative overview of key features is provided in Table 1.

IV. USE CASES

Use cases for *SeedMe2* building blocks include:

1. Group websites – A research group could set up a website for project news, publications, and documentation as well as shared data using these building blocks. This enables the group to discuss, document, and disseminate their data from a single website instead of using separate sites and tools for each task.
2. Application content gallery – Scientific applications may integrate these building blocks to share

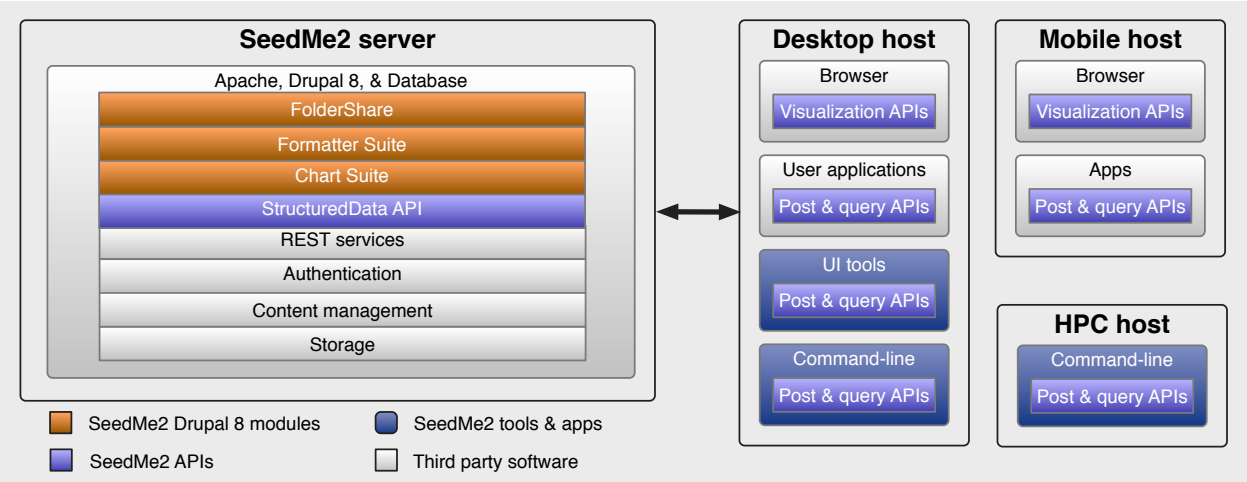

Fig. 2. SeedMe2 architecture: Showing various components and interaction from mobile, desktop and HPC/Cloud hosts.

demonstration data, test data sets, and highlighted content. They also may provide a central data service for their users to contribute content that may be searched/discovered and re-used by others, such as configuration files, parameter sets, color maps, scripts, and so forth.

3. Gateway data locker – Science Gateways provide central computation resources, but increasingly they also need to provide shared data repositories related to that computation. The building blocks we provide support data storage, annotation, and collaboration features as well as the necessary access controls to insure that only the appropriate users gain access to stored data.
4. Data repositories – Project, group, department, and domain data repositories may use these building blocks to store, document, discuss, and search contributed content.

V. Conclusion

We have discussed a set of general-purpose customizable web-based data sharing and data management building blocks for research teams. These address a common need for scientific data-centric data storage, documentation, access controls, and collaboration by the research community. The building blocks are in early user phase and are being actively used by friendly users to facilitate research and collaboration among distributed researchers (Fig 1).

A. Deployment

We anticipate the following deployment scenarios for the *SeedMe2*

1. Platform-as-a-service: A central supported and maintained service, which is open and available to researchers.
2. Vendor hosted: Academic computing centers or commercial vendors could provide an instance for their clients.
3. Self-hosted: Researchers self host an instance.

B. Distribution

The source code is open source and available as iterative releases at the project website dibbs.seedme.org [6].

C. Trial & Demonstration

A self-serve trial demonstration of the latest software is available at the project's sandbox website sandbox.seedme.org [7]. Trial web sites provide a sample configuration and sample data to illustrate the software's features.

Acknowledgments

We would like to thank all users of *SeedMe1*, early adopters of *SeedMe2, Dmitry Mishin* for creating the sandbox website, and collaborators that have provided valuable feedback on features and bugs. 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."

References

1. Drupal. 2018. *Drupal – Open Source CMS*. Retrieved Sep 2, 2018 from http://drupal.org/
2. Apache. 2018. *The Apache HTTP Server Project*. Retrieved Sep, 2018 from http://httpd.apache.org/
3. MySQL. 2018. *MySQL*. Retrieved Sep 2, 2016 from http://www.mysql.com/
4. SeedMe. 2018. SeedMe (Stream Encode, Explore and Disseminate My Experiments) Retrieved Sep 2, 2018 from https://www.seedme.org
5. 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
6. SeedMe2 building blocks. Retrieved Sep 2, 2018 from http://dibbs.seedme.org
7. SeedMe's data sharing sandbox builder. Retrieved Sep 2, 2018 from http://sandbox.seedme.org