SeedMe : Stream Encode Explore and Disseminate My Experiments

Amit Chourasia
University of California, San Diego
amit@sdsc.edu

David R. Nadeau
University of California, San Diego
nadeau@sdsc.edu

Mona Wong
University of California, San Diego
mona@sdsc.edu

Dmitry Mishin
University of California, San Diego
dmishin@sdsc.edu

Michael Norman
University of California, San Diego
mlnorman@sdsc.edu

ABSTRACT
Visualization plays an important role in exploring large data sets, whether they are created via computational simulations or a result of data gathered from observational sensors. Analysis and visualization of this data is often conducted on High Performance Computing (HPC) resources. While interactive visualization is desirable to explore this data, it is often impractical due to the data's size or because of the labor cost to set up and visualize the data manually. Furthermore, HPC resources typically shared with many other users and run jobs in a batch manner, this is not conducive for conducting interactive visualization on demand. Batch visualization is therefore a preferred method as it can be pipelined along with other simulation and processing steps.

Batch visualization results may include images, animations, plots, surfaces, and so forth. Animation image sequences need to be encoded to video for easy sharing and review by a science team, and the other result data types need to be accessible and viewable as well. However, data sharing from within an HPC infrastructure is minimal due to security and technical constraints. And video generation tools are not widely available within HPC environments. The lack of good data sharing mechanisms forces science teams to use manual effort and ad hoc batch scripts. A better approach is needed to help smooth out these tasks for large data and HPC environments.

In this work we describe and demonstrate the Stream Encode Explore and Disseminate My Experiments (*SeedMe*) platform that helps to solve the "last mile" problem of data and metadata sharing and video generation. The system provides researchers with a cloud-based service that may be used via a web browser, through web services, and by use of application clients and command line tools. We also briefly discuss how researchers from several science communities are currently using *SeedMe*. Finally, we provide an overview of the upcoming next iteration of the *SeedMe* project, which is more capable, modular, and deployable. This new version will be available to the research community as an open source project that can be easily deployed on premise or in the cloud, and customized and extended for a particular project's needs.

Keywords
Data sharing; Collaboration; Visualization; Video encoding

1. INTRODUCTION
Research processes have become increasingly collaborative. This is especially true for processes that rely on High Performance Computing (HPC). In such environments the ability to share data, results, and context with the research team is critical for efficient collaboration.

Presently researchers rely on a variety of cumbersome methods to share information with each other. Some methods are specific to a particular HPC cluster while others may rely on downloading the results to a personal computer then sharing it via emails, FTP, Globus[1], social networking, or public sharing sites such as DropBox[2]. Furthermore, in context of visualization, the generation of videos from sequences of images is burdensome for researchers, as they need to find, install, and figure out the complexities of video codecs, bitrates, and compression settings for each target playback platform, from desktops to mobile phones. Busy users may default to older or simpler video codecs, or whatever is built in to their visualization tools, and unintentionally produce lower quality videos that can have visual artifacts that impact scientific assessment.

2. MOTIVATION
Remote computing and HPC users have a pressing need for a platform to share science data, including text files, images, image sequences, surfaces, plots, tables, document files, scripts, configuration files, and other binary data. Such data is often more complex, more interrelated, and in different file formats than are supported by consumer-oriented data sharing environments. To support this data well, and the HPC sources of this data, a science-oriented data sharing system is needed that can help transfer, store, organize, and secure content, that moves to and from HPC systems, desktops, and mobile devices. Once available within a science data sharing system, users need to be able to describe, discuss, reorganize, and share the data with each other. For visualization results, the system also needs to provide access to key processing tools, such as those for video encoding from image sequences.

In this work we discuss and describe the *SeedMe* platform, which addresses these needs. It provides a web based presentation system as well as programmatic interfaces to use this platform from remote computing, HPC, and mobile devices.

3. RELATED WORK
Web based collaboration methods have existed since the early days of the Web; they have morphed and matured over time. The simplest form of collaboration has been to provide direct access to the data and results on HPC resources. However, this requires team members to all have accounts on highly restricted and well-controlled systems. They need to know where to find files, how they are organized, and what is the context of data. After several iterations of an experiment, discovery of appropriate content can become onerous. Consumer-oriented file sharing services can be used, but support for large amounts of data, file transfers from HPC environments, science data file formats, visualization processing tools, secure access, and threaded discussion is

problematic or lacking. General-purpose Content Management Systems (CMS) such as (Drupal [3], Joomla [4], MediaWiki [5] and WordPress [6]), offer a rich way to organize, store and present content, but out of the box none of them support the data type and task breadth required for science data sharing. However, CMS systems do provide a solid foundation that can be extended to support the needs of science researchers. This is the approach used in this project, which is based upon the Drupal CMS.

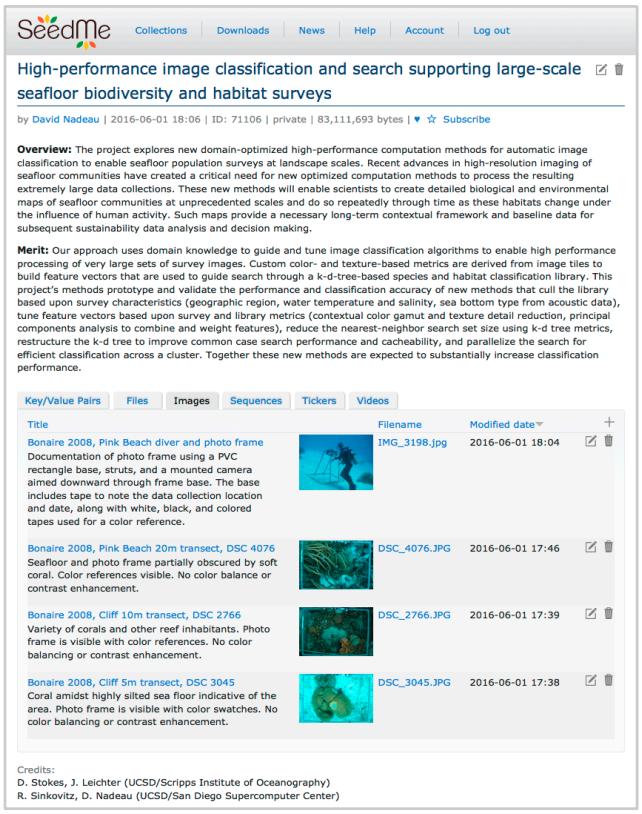

Figure 1. A snapshot of a *SeedMe* collection for a research project showing content in various tabs with image tab visible.

4. SEEDME PLATFORM

SeedMe is an acronym for *Stream Encode Explore and Disseminate My Experiments* [7, 8]. The acronym reflects on seeding and sharing of research content for efficient collaboration. The original focus of the project was to automate web-based video encoding, transcoding, streaming, and secure sharing for visualization animations, and especially those generated on HPC resources. Existing visualization tools rarely offer suitable built-in video encoding to handle the variety of settings appropriate for good playback on desktops, tablets, and mobile devices.

The *SeedMe* platform has since expanded beyond video encoding to handle general science data sharing needs, including the following key features:

- Organize any number of files into collections (see figure 1)
- Add descriptions for files and collections (see figure 1)
- Add metadata tables in form of key-value pairs (see figure 2)
- Support time-stamped messages to report computation progress (see figure 3)
- Provide video encoding, transcoding, and streaming (see figure 4 & 5)
- Support access control to restrict access to the content
- Enable threaded discussion of specific content
- Provide secure content upload and download
- Federate single sign-on from 2,266 institutions
- Support multiple clients to post/transfer and retrieve content

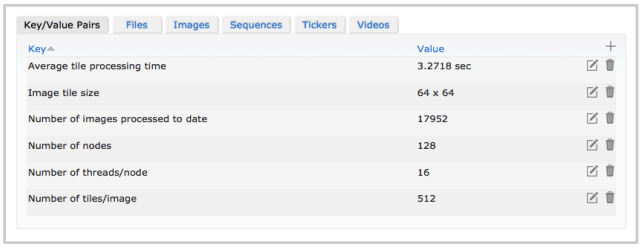

Figure 2. Key value pairs showing computation parameters.

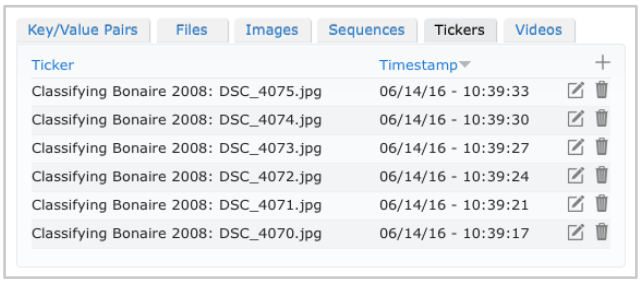

Figure 3. Tickers showing computation runtime progress.

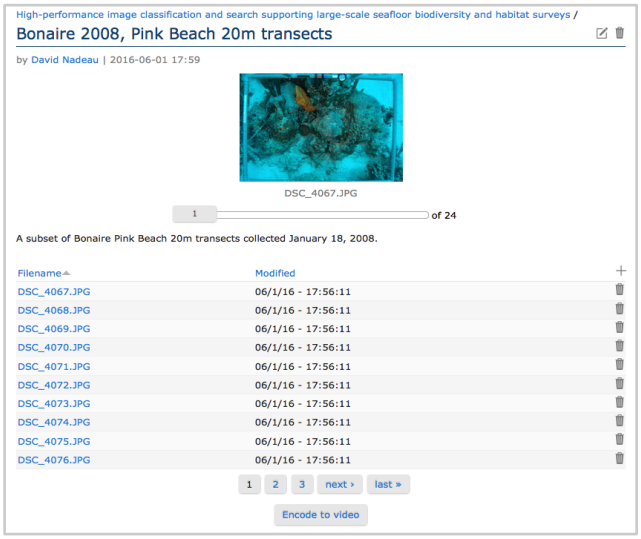

Figure 4. An image sequence ready for video encoding.

```
seedme.py -auth_path ~/seedme.txt \
          -title "My collection"  \
          -sequence_path  "images*.png" \
          -sequence_title "Animation" \
          -sequence_frame_rate 10 \
          -sequence_encode
```

Figure 5. A *SeedMe* command line to create a collection, post an image sequence then encode it to video.

5. HOW SEEDME WORKS

SeedMe organizes content into "collection" containers. Each one may have a title, description, Metadata, computation progress messages as tickers, files, images, image sequences, and videos (see figure 1). Visualization and data generated by users using their preferred tools can be uploaded, viewed and downloaded on *SeedMe* platform in multiple ways as shown in figure 6. For example, a user needs to sign-in to create a collection, then upload a set of images under the sequence category, and finally trigger generation of video. These steps may be performed via a web browser or via the command line using the Python module as shown in figure 5.

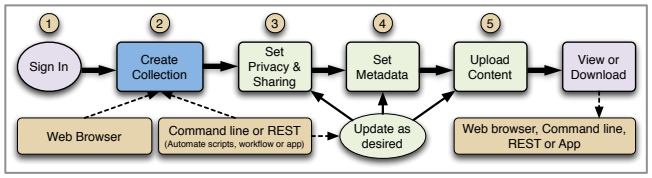

Figure 6: Illustration of how *SeedMe* works

6. SEEDME TECHNOLOGY

SeedMe is implemented as a web front-end to an integrated suite of back-end open source software that includes:

- *Apache* [9] web server
- *Drupal* [3] content management system
- *MySQL* [10] database
- *Gearman* [11] job scheduler
- *FFmpeg* [12] video encoding tool

Back-end software is integrated using custom and contributed *Drupal* modules as well as glueware written in PHP. *SeedMe* clients to interact with web services are written in JAVA and Python. The site runs on a cluster of primary and backup Linux servers and Mac minis for video encoding and large storage pool for storing files with backup.

7. DISCUSSION & CONCLUSIONS

SeedMe is a scientific data sharing and collaboration service that exists at a mid-point between consumer file sharing services and social networks. It supports remote browser-based access to shared files and collections in a way that is loosely similar to file hosting services like DropBox. At the same time, *SeedMe* provides collaboration and discussion features that are loosely similar to social networks. But unlike both of these, *SeedMe* provides metadata management, video encoding, and command-line and third party tools to programmatically post, query, and retrieve data using HPC job scripts.

SeedMe's user base spans a broad array of disciplines and its unique feature set is in active use by researchers as a tool to share computed and experimental data among distributed collaborators, track the status of long-running computations, and encode visualization images into videos used to quickly review computation job results. The platform has also been integrated into scientific applications such as the Kepler workflow system [13], and the Vapor [14], and VisIt [15] visualization tools that allow the users to seamlessly interact with *SeedMe* from the application. The platform has over 600 user registrations and over 140,000 content items that are continually growing (see figure 7). The usage of the platform is in bursts and the bulk of the content is not made public. This is expected as content is usually unpublished work in progress. Typically users generate visualization images and documents and share them with their collaborators using *SeedMe*. Overall the feedback from users has been very positive. Few example use cases for *SeedMe* are as follows

- Visualization video encoding and sharing from diverse scientific disciplines including climate, astrophysics, geoscience and many others. [16, 17]
- Visualization repository for Vortex induced vibration simulation website [18] that is a free online educational environment for high school and college level students to learn about the physical phenomena known as Vortex Shedding and Vortex-Induced Vibration
- Data sharing for experimental data [19]
- Progress monitoring by posting ticker messages from HPC resources. (see figure 5)

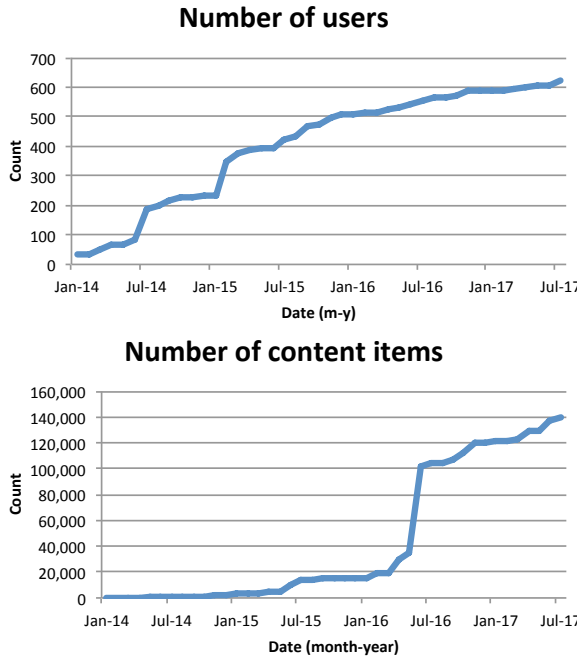

Figure 7. Number of *SeedMe* user registrations (top) and number of content items posted (bottom) by *SeedMe* users over time.

8. UPCOMING SEEDME PLATFORM

While *SeedMe* has proven to be a popular science collaboration tool, the current project design has limitations that we are addressing in its next iteration [18]. This new version expands collections to support arbitrary directory trees of files, metadata, and threaded discussions so that there is greater flexibility for a hierarchical organization of content. The next generation of *SeedMe* also integrates data viewing features that automatically build visualization plots and diagrams from data posted in a variety of well-known data formats such as CSV, JSON, and HTML. All features are being developed as modular building blocks that can be separately used in other *Drupal*-based sites or some even in non-*Drupal* projects.

9. ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. ACI-1235505 and ACI-1443083.

10. REFERENCES

[1] I. Foster. Globus Toolkit Version 4: Software for Service-Oriented Systems. IFIP International Conference on Network and Parallel Computing, Springer-Verlag LNCS 3779, pp 2-13, 2006.

[2] Dropbox. 2017. *Dropbox*. Retrieved Jun 15, 2017 from http://dropbox.com/

[3] Drupal. 2017. *Drupal – Open Source CMS*. Retrieved Jun 15, 2017 from http://drupal.org/

[4] Joomla. 2017. *Joomla – The CMS Trusted By Millions for their Websites*. Retrieved Jun 15, 2017 from https://www.joomla.org/

[5] MediaWiki, 2017. *MediaWiki*. Retrieved Jun 15, 2017 from https://www.mediawiki.org/

[6] WordPress. 2017. *WordPress – Create your site today*. Retrieved Jun 15, 2017 from https://wordpress.com.

[7] Chourasia, A. Wong-Barnum, M. Dmitry, M. Nadeau, D.R. and Norman, M.L. 2016. 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

[8] SeedMe. 2017. SeedMe (Stream Encode, Explore and Disseminate My Experiments). Retrieved Jun 15, 2017 from https://www.seedme.org/

[9] Apache. 2017. *The Apache HTTP Server Project*. Retrieved Jun 15, 2017 from http://httpd.apache.org/

[10] MySQL. 2017. *MySQL*. Retrieved Jun 15, 2017 from http://www.mysql.com/

[11] Gearman. 2017. *Gearman Job Server*. Retrieved Jun 15, 2017 from http://gearman.org/

[12] FFmpeg. 2017. *FFmpeg*. Retrieved Jun 15, 2017 from http://ffmpeg.org/

[13] Altintas, I. Berkley, C. Jaeger, E. Jones, M. Ludascher, B. and Mock, S. 2004. *Kepler: an extensible system for design and execution of scientific workflows*. Scientific and Statistical Database Management, 2004. Proceedings. 16th International Conference on. Pages: 423-424.

[14] Clyne, J., Mininni, P., Norton, A., and Rast, M. "Interactive desktop analysis of high resolution simulations: application to turbulent plume dynamics and current sheet formation", New Journal of Physics 9 (2007) 301.

[15] VisIt. 2017. *VisIt*. Retrieved Jun 15, 2017 from https://wci.llnl.gov/simulation/computer-codes/visit

[16] James Schiavone. WRF Users' Workshop 2016 Extended Abstract Visualizations | SeedMe.org - Scientific data sharing made easy. 2017. Retrieved Jul 28, 2017 from https://www.seedme.org/node/70880

[17] Amit Chourasia. Visualization of Geodynamo Run-3 | SeedMe.org - Scientific data sharing made easy. 2017. Retrieved Jul 28, 2017 from https://www.seedme.org/node/35865

[18] Vortex Shedding 0.1. 2017. Retrieved Jul 28, 2017 from http://js-172-83.jetstream-cloud.org/vortexshedding/

[19] Istvan Ladunga. ExpressOrtho | SeedMe.org - Scientific data sharing made easy. 2017. Retrieved Jul 28, 2017 from https://www.seedme.org/node/162052

[20] SeedMe Science. 2017. Retrieved Jun 15, 2017 from https://dibbs.seedme.org