The Journal of Open Source Software

Science Capsule - Capturing the Data Life Cycle

Devarshi Ghoshal[1], **Ludovico Bianchi**[1], **Abdelilah Essiari**[1], **Michael Beach**[1, 2], **Drew Paine**[1], **and Lavanya Ramakrishnan**[1]

1 Lawrence Berkeley National Lab **2** University of Washington

DOI: 10.21105/joss.02484

Software

- Review
- Repository
- Archive

Editor: Amy Roberts

Reviewers:

- @cmbiwer

Submitted: 28 June 2020
Published: 17 July 2020

Summary

The volume and variety of data that is generated at DOE Office of Science user facilities is at risk of not being usable due to the complexity of data and associated processing. It is necessary that we develop appropriate tools and technologies to capture, preserve, share, reproduce, and make optimal use of the data and workflows.

Science Capsule captures, organizes, and manages the end-to-end scientific process to facilitate capture of provenance, easy sharing of data and workflows across domains and ensuring reproducibility. It captures the artifacts of a scientific workflow process that have not been captured in traditional computational workflows, such as, the pre-processing scripts, lab notebooks, metadata, and provenance. These artifacts provide critical information about the process that is needed for data analyses, reproducibility, sharing, and reuse.

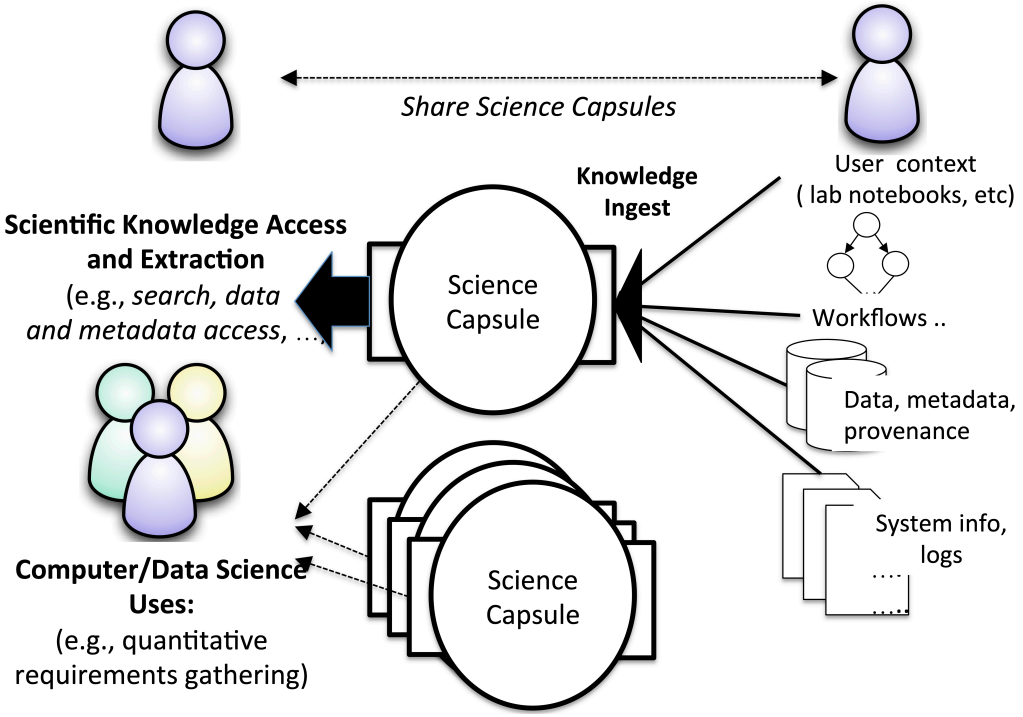

Figure 1: Science Capsule abstraction collects necessary information for reproducible science.

Event Capture

Different monitoring tools can be used in Science Capsule to capture events corresponding to scientific workflow management. Currently, Science Capsule captures events from inotify

(Fisher (2017)), Linux's strace utility and Python's watchdog module depending upon the underlying platform and level of granularity of workflow events. Additionally, users can also configure and select tools for capturing workflow events based on their requirements.

CLI

Science Capsule is developed in Python and works on Linux, MacOS and Windows platforms. It also provides the necessary tools for containerizing and sharing scientific workflows using Docker. In order to setup and use a Science Capsule environment, users can use the command-line interface (CLI). A Science Capsule environment can be configured and instantiated using the following commands:

```
export SC_CONFIG_DIR=<config-directory>

sc bootstrap "$SC_CONFIG_DIR" --monitored-dir <mydir> --event-sources inotify,stra

sc services start all
```

All activities in `<mydir>` are now observed and captured by Science Capsule. Users can add multiple directories using the `--monitored-dir` option. `inotify` captures all filesystem events, whereas all process events are capturedd through `strace`. To monitor the status of event capture, users can use the following command:

```
sc services tail -f capture
```

A Web UI displays a timeline of all the events corresponding to a workflow execution.

Key Features

Science Capsule enables researchers to ensure reproducibility by providing:

- Automatic capture of user workflow events
- A customizable and extendable interface for event monitoring and capturing from multiple sources and at different granularities
- Capability to create, save, share, and extend workflows using containers
- Allows for easy migration and execution of workflows across multiple platforms (Mac, Windows and Linux)
- A web based user interface for viewing timelines of workflow events and annotating with additional information including digital notes, scanned images, and other attachments

Acknowledgements

This work is supported by the U.S. Department of Energy, Office of Science and Office of Advanced Scientific Computing Research (ASCR) under Contract No. DE-AC02-05CH11231.

References

Fisher, C. (2017). Linux filesystem events with inotify. *Linux Journal*, *2017*(280), 2.