Research Workflows - Towards reproducible science via detailed provenance tracking in Open Science Chain

Viswanath Nandigam
San Diego Supercomputer Center
University of California San Diego
La Jolla, California
vnandigam@ucsd.edu

Kai Lin
San Diego Supercomputer Center
University of California San Diego
La Jolla, California
klin@sdsc.edu

Manu Shantharam
San Diego Supercomputer Center
University of California San Diego
La Jolla, California
mshantharam@sdsc.edu

Scott Sakai
San Diego Supercomputer Center
University of California San Diego
La Jolla, California
ssakai@sdsc.edu

Subhashini Sivagnanam
San Diego Supercomputer Center
University of California San Diego
La Jolla, California
sivagnan@sdsc.edu

ABSTRACT

Scientific research has always struggled with problems related to reproducibility caused in part by low data sharing rates and lack of provenance. Credibility of the research hypothesis comes into question when results cannot be replicated. While the growing amount of data and widespread use of computational code in research has been pushing scientific breakthroughs, their references in scientific publications is insufficient from a reproducibility perspective.

The NSF funded Open Science Chain (OSC) is a cyberinfrastructure platform built using blockchain technologies that enables researchers to efficiently validate the authenticity of published data, track their provenance and view lineage. It does this by leveraging blockchain technology to securely store metadata and verification information about research data and track changes to that data in an auditable manner.

In this poster we introduce the concept of "research workflows", a tool that allows researchers to create a detailed workflow of their scientific experiment, linking specific data and computational code used in their published results in order to enable independent verification of the analysis. OSC research workflows will allow for detailed provenance tracking both within the OSC platform as well as external repositories like Github, thereby enabling transparency and fostering trust in the scientific process.

CCS CONCEPTS

• **Information systems** → **Data provenance**; *Integrity checking*.

KEYWORDS

Data Reproducibility, Data Provenance, Data Integrity, Blockchain

PEARC '20, July 26–30, 2020, Portland, OR, USA

ACM ISBN 978-1-4503-6689-2/20/07…$15.00
https://doi.org/10.1145/3311790.3399619

ACM Reference Format:
Viswanath Nandigam, Kai Lin, Manu Shantharam, Scott Sakai, and Subhashini Sivagnanam. 2020. Research Workflows - Towards reproducible science via detailed provenance tracking in Open Science Chain. In *Practice and Experience in Advanced Research Computing (PEARC '20), July 26–30, 2020, Portland, OR, USA.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3311790.3399619

1 INTRODUCTION

The credibility of research findings are compromised when results from the study cannot be replicated. Reproducibility is increasingly becoming a problem in multidisciplinary research with the emergence of big data, more complex computational methods and technologies [1]. According to recent study [3], many published Computer Science research hypotheses were not replicable within the context of the referenced computational code. Out of the 613 articles published in 13 top-tier systems research conference publications, the study found that only 25% of the results from the articles are replicable due to various reasons including modified data or unavailable data. Independent verification of scientific hypotheses is critical for the advancement of research, especially with the growing use of data and computation to achieve these results.

2 OPEN SCIENCE CHAIN

While there are several research data sharing repositories (e.g. figshare, Dryad Digital Repository) that focus on making research data available, they lack the ability to securely track lineage and authenticity in an immutable manner. The National Science Foundation funded Open Science Chain (OSC) is building a cyberinfrastructure platform, using consortium blockchain technologies, with the goal of enabling a broad set of researchers to efficiently share, verify and validate the authenticity of scientific data while preserving the provenance.

The OSC cyberinfrastructure comprises a consortium blockchain platform implemented using the open source Hyperledger Fabric framework, middleware services for interacting with the platform and a web portal that lowers the complexity barrier for researchers to utilize the system [10]. Researchers register verification information (eg. SHA256 hash) and additional metadata information about the data in the OSC blockchain. The actual data itself is stored

o-chain. Storing only the metadata and verification information of the dataset in the blockchain is more practical and efficient in an era of Big Data. It also enables researchers to share verification information for sensitive datasets that cannot be shared publicly due to privacy restrictions or licensing reasons. When updates are made to a dataset or data collection in OSC, all metadata changes including the SHA256 hashes for each and every file in that data collection are tracked in the blockchain, enabling users to view a detailed immutable history of that dataset over time.

3 RESEARCH WORKFLOWS

While registering data with OSC can provide immutable provenance tracking for a digital object, scientific experiments typically involve a combination of algorithms or computational code, raw data and derived products. In several cases researchers make use of data and code owned and maintained by other providers and which is maintained and updated external to OSC. Computational code is increasingly used in data transformation, processing, integration and analysis including modelling or visualization. This is consistent across popular science gateways like the NSF funded OpenTopography [9] and the NSF/NIH funded Neuroscience gateway [5] and others where users are guided through a structured scientific workflow that involves applying one or more algorithms to a dataset to produce a derived product or result. In some cases, the original dataset can be updated (e.g. spatial dataset undergoes a projection change or data gets corrected) or the algorithm itself gets updated which can produce results that differ from the original scientific experiment.

Research workflows in OSC will provide researchers the ability to create a detailed workflow linking data and computational code metadata along with additional documentation detailing the experiment (e.g. dependencies, parameters) and store it in the OSC blockchain to enable independent reproducibility of their scientific experiments. OSC research workflows can help independent researchers reproduce the original scientific experiment by explicitly referring to specific instances or versions of the data or computation code used along with a detailed narrative to understand the experiment. Research workflows also have the ability to notify the original creator of the workflow if the referenced data or computational code undergoes changes or updates.

Computational code is typically maintained and distributed via multiple diverse platforms with versioning capability like the popular repository hosting service GitHub. Scientific datasets can be deposited for wider distribution in various funding agencies recommended domain specific data sharing repositories [6] [4]. For this reason, OSC research workflows tool was designed with the capability to link data and computational code from not just OSC but other external third party data and code repositories like GitHub, etc. We are in the process of enabling additional third party code and data repositories (e.g. figshare) integrations in the OSC research workflows.

3.1 Use Case

A researcher is working with pre and post event lidar data of 2016 Mw 7.0 Kumamoto, Japan, earthquake, which ruptured 40 km of the Futagawa-Hinagu Fault Zone on Kyushu Island available from OpenTopography [8] [7]. Both of these OpenTopography datasets are registered with the OSC platform. The researcher then proceeds to use a windowed implementation of the Iterative Closest Point (ICP) algorithm [2] to calculate the 3-D surface deformation produced by the earthquake using a code available via GitHub and uses the results of these analysis to produce a hypothesis. In order to ensure that the results of the analysis are transparent and more importantly reproducible, the researcher has the ability to create a scientific workflow in OSC for this experiment, linking the specific versions of the OpenTopography datasets used along with the location of the GitHub repository hosting the Iterative Closest Point algorithm code. The OSC scientific workflow will pull metadata for that GitHub repository and maintain a reference to that specific version of the code. The researcher also has the ability to provide additional details including spatial extents and other algorithm parameters used to enable others to replicate the experiment. A complete representation of this scientific workflow is shown in figure 1.

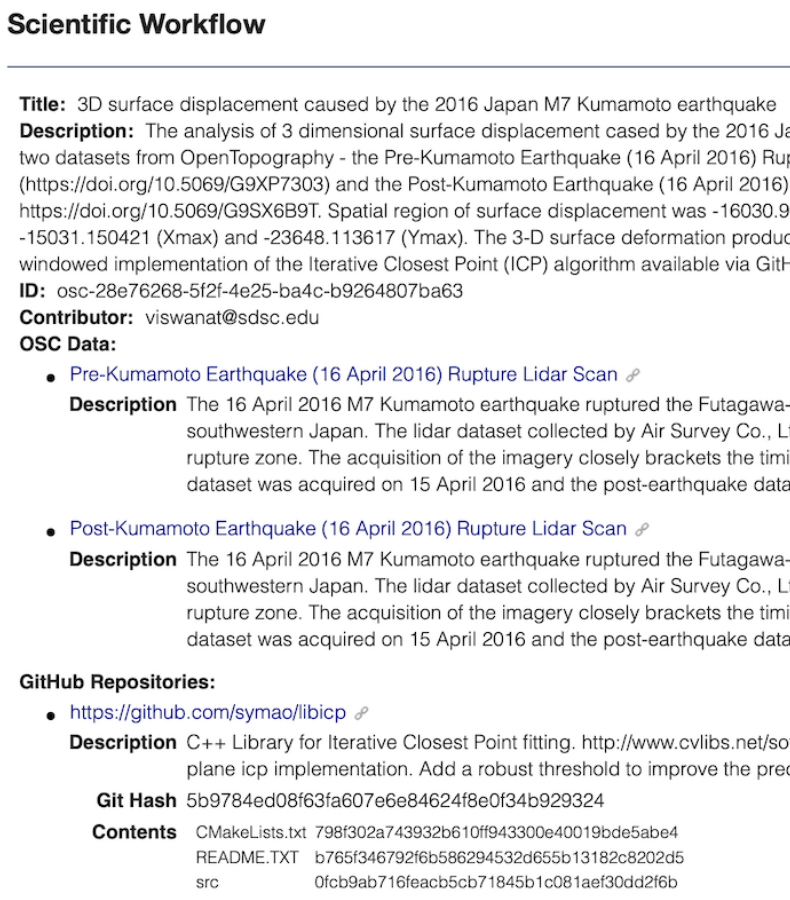

Figure 1: Sample research workflow for 3D surface displacement modelling caused by the 2016 Japan M7 Kumamoto earthquake

Additionally if there is an update to any of the two datasets or there is a new push to the ICP GitHub repository, the researcher who created the research workflow will be notified. The researcher can optionally update the workflow and all changes are securely tracked in the OSC blockchain, thereby providing a provenance trail to the workflow itself. Other researchers reviewing this workflow will also be made aware if any digital entity referenced in the scientific workflow has newer versions or have been updated.

4 SUMMARY

Preserving detailed linkages to specific instances of data and algorithms in scientific experiments especially those used in publications is vital to ensure reproducibility and integrity of the hypothesis. The research workflows capability in the Open Science Chain platform provides an efficient and secure solution for researchers to enable reproducibility and transparency to their scientific experiments.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. 1840218.

REFERENCES

[1] David H Bailey, Jonathan M Borwein, and Victoria Stodden. 2016. Facilitating reproducibility in scientific computing: Principles and practice. *Reproducibility: Principles, Problems, Practices, John Wiley and Sons, New York, to appear* (2016).
[2] Paul J Besl and Neil D McKay. 1992. Method for registration of 3-D shapes. In *Sensor fusion IV: control paradigms and data structures*, Vol. 1611. International Society for Optics and Photonics, 586–606.
[3] Gina Moraila, Akash Shankaran, Zuoming Shi, and Alex M Warren. 2014. *Measuring reproducibility in computer systems research.* Technical Report. Technical report, University of Arizona.
[4] Nature. 2020. *Recommended Data Repositories.* https://www.nature.com/sdata/policies/repositories
[5] nsgportal.org. 2020. *Neuroscience Gateway.* https://www.nsgportal.org/
[6] National Library of Medicine. 2020. *Open Domain-Specific Data Sharing Repositories.* https://www.nlm.nih.gov/NIHbmic/domain_specific_repositories.html
[7] OpenTopography. 2020. *Post-Kumamoto Earthquake (16 April 2016) Rupture Lidar Scan.* https://doi.org/10.5069/G9SX6B9T
[8] OpenTopography. 2020. *Pre-Kumamoto Earthquake (16 April 2016) Rupture Lidar Scan.* https://doi.org/10.5069/G9XP7303
[9] OpenTopography.org. 2020. *OpenTopography.* https://www.opentopography.org
[10] Subhashini Sivagnanam, Viswanath Nandigam, and Kai Lin. 2019. Introducing the Open Science Chain: Protecting Integrity and Provenance of Research Data. In *Proceedings of the Practice and Experience in Advanced Research Computing on Rise of the Machines (learning).* 1–5.