Tools for Computational and Data-Intensive Research: Lessons Learned From Science Gateways and Preservation Frameworks

Sandra Gesing*
*University of Notre Dame, sandra.gesing@nd.edu

Software is more and more recognized for its importance for research, evident in surveys such as {1] where over 90% of researchers answering that they use software for their research and over 65% expressing that they even could not do their research without software. Science gateways [2] are a subgroup of software in research focusing on providing end-to-end solutions with an easy-to-use interface while integrating complex research infrastructures such as distributed computing and data infrastructures and/or lab instruments. Science gateways are used by over 77% of XSEDE [3] users, for example, applying the research infrastructure via science gateways and not via commandline.

The Science Gateways Community Institute (SGCI) [4] offers services to create and sustain science gateways and to understand the science gateway eco-system. In the last decade quite a few science gateway frameworks have been developed to ease the development of science gateways and providing building blocks for diverse aspects such as connectors to distributed computing infrastructures, security packages, and workflow integration. Widely used solutions can be distinguished in three main areas: (1) complete frameworks, e.g., Galaxy [5], HUBzero [6], Globus Data Portal [7]; (2) RESTful APIs and support of multiple programming languages, e.g., Apache Airavata [8], the Agave platform [9], or reused interface implementations, e.g., CIPRES [10] with its RESTful API; (3) science gateways as a service with provision of hardware in the background such as SciGap (Science Gateway Platform as a Service) [11]. Successful approaches are characterized by being technology agnostic, using APIs and standard web technologies or delivering a complete solution for serving a community efficiently.

The uptake of containerization approaches such as Docker [12] or Singularity [13] allows for providing the full environment for a computational method - as full science gateway infrastructure or for submission of computational tools on the underlying infrastructure addressing the portability between different hardware architectures. Containerization addresses additionally a further aspect: reproducibility of science and preservation of software environments. The long-term aspect of preservation of software for over 15-20 years is probably not well addressed via containerization - immanent in dependencies on container versions, operating systems and existing research infrastructures - but delivers an intermediate solution. The need for preservation of software and data has led to quite a few projects, e.g., ReproZip [14], Code Ocean [15] and PresQT [16]. The first two aim at preserving the full execution environment of a computational method. The latter is a project to create RESTful web services that enhance the information about data and software with metadata, apply fixity checks to ensure that software has not been changed during a network transfer and add keywords to improve the discoverability of software and data. The concept is to be easily integrable so that researchers can keep working in their chosen computational environment and can receive additional features instead of having to switch to a different software..

A main lesson learned from science gateways and preservation projects is the switch from a system-centered view with expecting users to spend substantial effort and time in learning computing environments to a user-centric view to support research more effectively by considering usability, scalability and interoperability.

References

[1] Udit Nangia, Daniel S. Katz. Survey of National Postdoctoral Association. http://doi.org/10.5281/zenodo.843607

[2] Barker, M., Olabarriaga, S., Wilkins-Diehr, N., Gesing, S., Katz, D.S., Shahand, S., Henwood, S., Glatard, T., Jeffery, K., Corrie, B., Glaves, H., Wyborn, L., Hong, N.C., and Costa, A. (2019The Global Impact of Science Gateways, Virtual Research Environments and Virtual Laboratories. Future Generation Computer Systems, Volume 95, June 2019, Pages 240-248

[3] John Towns, Timothy Cockerill, Maytal Dahan, Ian Foster, Kelly Gaither, Andrew Grimshaw, Victor Hazlewood, Scott Lathrop, Dave Lifka, Gregory D. Peterson, Ralph Roskies, J. Ray Scott, Nancy Wilkins-Diehr, "XSEDE: Accelerating Scientific Discovery", Computing in Science & Engineering, vol.16, no. 5, pp. 62-74, Sept.-Oct. 2014, doi:10.1109/MCSE.2014.80

[4] Gesing, S., Wilkins-Diehr, N., Dahan, M., Lawrence, K., Zentner, M., Pierce, M., Hayden, L.B., and Marru, S. (2017) Science Gateways: The Long Road to the Birth of an Institute. Proc. of HICSS-50 (50th Hawaii International Conference on System Sciences), 4-7 January 2017, Hilton Waikoloa, HI, USA, http://hdl.handle.net/10125/41919

[5] Enis Afgan, Dannon Baker, Bérénice Batut, Marius van den Beek, Dave Bouvier, Martin Čech, John Chilton, Dave Clements, Nate Coraor, Björn Grüning, Aysam Guerler, Jennifer Hillman-Jackson, Vahid Jalili, Helena Rasche, Nicola Soranzo, Jeremy Goecks, James Taylor, Anton Nekrutenko, and Daniel Blankenberg. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update, Nucleic Acids Research, Volume 46, Issue W1, 2 July 2018, Pages W537–W544, doi:10.1093/nar/gky379

[6] McLennan, Michael, and Rick Kennell. HUBzero: A Platform for Dissemination and Collaboration in Computational Science and Engineering. Computing in Science & Engineering 12, no. 2 (2010): 48-53.

[7] Chard K, Dart E, Foster I, Shifflett D, Tuecke S, Williams J. 2018. The Modern Research Data Portal: a design pattern for networked, data-intensive science. PeerJ Computer Science 4:e144 https://doi.org/10.7717/peerj-cs.144

[8] Suresh Marru, Marlon Pierce, Sudhakar Pamidighantam, and Chathuri Wimalasena. 2015. Apache Airavata as a Laboratory: Architecture and Case Study for Component-Based Gateway Middleware. In Proceedings of the 1st Workshop on The Science of Cyberinfrastructure: Research, Experience, Applications and Models (SCREAM '15). ACM, New York, NY, USA, 19-26. DOI: https://doi.org/10.1145/2753524.2753529

[9] Rion Dooley, Steven R. Brandt, and John Fonner. 2018. The Agave Platform: An Open, Science-as-a-Service Platform for Digital Science. In Proceedings of the Practice and Experience on Advanced Research Computing (PEARC '18). ACM, New York, NY, USA, Article 28, 8 pages. DOI: https://doi.org/10.1145/3219104.3219129

[10] Miller, M.A., Pfeiffer, W., and Schwartz, T. (2010) "Creating the CIPRES Science Gateway for inference of large phylogenetic trees" in Proceedings of the Gateway Computing Environments Workshop (GCE), 14 Nov. 2010, New Orleans, LA pp 1 - 8.

[11] Marlon Pierce, Suresh Marru, Eroma Abeysinghe, Sudhakar Pamidighantam, Marcus Christie, and Dimuthu Wannipurage. 2018. Supporting Science Gateways Using Apache Airavata and SciGaP Services. In Proceedings of the Practice and Experience on Advanced Research Computing (PEARC '18). ACM, New York, NY, USA, Article 99, 4 pages. DOI: https://doi.org/10.1145/3219104.3229240

[12] Dirk Merkel. 2014. Docker: lightweight Linux containers for consistent development and deployment. Linux J. 2014, 239, pages.

[13] Kurtzer GM, Sochat V, Bauer MW (2017) Singularity: Scientific containers for mobility of compute. PLoS ONE 12(5): e0177459. https://doi.org/10.1371/journal.pone.0177459

[14] Fernando Chirigati, Dennis Shasha, and Juliana Freire. 2013. ReproZip: using provenance to support computational reproducibility. In Proceedings of the 5th USENIX Workshop on the Theory and Practice of Provenance (TaPP '13). USENIX Association, Berkeley, CA, USA, Article 1, 4 pages.

[15] https://codeocean.com/about

[16] Meyers, N.K., Gesing, S., Wang, Z. and Johnson, R. (2018) Tools and RESTful Services to Improve Preservation and Re-use of Research Data & Software. AGU Fall Meeting 2018, December 10-14, 2018, Washington, DC, USA