AI Buzzwords Explained: Scientific Workflows

Daniel Garijo (Information Sciences Institute and Department of Computer Science, University of Southern California; dgarijo@isi.edu)

DOI: 10.1145/3054837.3054839

The reproducibility of scientific experiments is crucial for corroborating, consolidating and reusing new scientific discoveries. However, the constant pressure for publishing results (Fanelli, 2010) has removed reproducibility from the agenda of many researchers: in a recent survey published in Nature (with more than 1500 scientists) over 70% of the participants recognize to have failed to reproduce the work from another colleague at some point in time (Baker, 2016). Analyses from psychology and cancer biology show reproducibility rates below 40% and 10% respectively (Collaboration, 2015) (Begley & Lee, 2012). As a consequence, retractions of publications have occurred in the last years in several disciplines (Marcus & Oransky, 2014) (Rockoff, 2015), and the general public is now skeptical about scientific studies on topics like pesticides, depression drugs or flu pandemics (American, 2010).

Reproducing the results of a previous study can be a challenge, as even when the original datasets and end results are available, a significant investment in time may be required (Garijo et al., 2013). Fortunately, the community has started to pay attention to initiatives for preserving the data and software used in scientific publications (e.g., Zenodo,[1] Github,[2] etc.). In computational sciences, **scientific workflows** were proposed in the last decade as a means to address reproducibility. A scientific workflow defines the set of computational tasks and dependencies needed to carry out in silico experiments (Taylor, Deelman, Gannon, & Shields, 2006). Typically, scientific workflows are represented as directed graphs, where the nodes represent computational tasks and the edges represent their dependencies. Figure 1 shows an example with two workflows, one for text analytics on the left and another one for neuro-image analysis on the right.

[1] https://zenodo.org/

[2] http://github.com/

Scientific workflows have been used in many domains, including astronomy [10], brain image analysis (Dinov et al., 2009) and bioinformatics (Wolstencroft et al., 2013). Besides improving reproducibility, scientific workflows have also proved to be helpful in teaching new users to visualize the overall structure of a method, save time when reusing an existing method and debug or inspect and modularize scientific experiments (Goderis, 2008; Garijo et al., 2014).

There are many challenges associated to scientific workflows. During the last decade plenty of systems have been designed to efficiently represent and execute them in both local and distributed environments (e.g., (Wolstencroft et al., 2013; Gil et al., 2011; Deelman et al., 2004; Callahan et al., 2006; Ludscher et al., 2006; Filgueira et al., 2014; Giardine et al., 2005), etc.). Different approaches have focused in optimizing workflow execution (e.g., (Deelman et al., 2004)) and their results (e.g., (Holl, 2014)). Other works have addressed workflow reuse (Garijo et al., 2014), (Goderis, Sattler, Lord, & Goble, 2005), recommendation (Starlinger, Brancotte, Cohen-Boulakia, & Leser, 2014; Bergmann & Gil, 2014) and discovery (Goderis, 2008), (Bergmann & Gil, 2014), as building on previous findings is considered to be critical to push science forward. Here we overview those aspects of workflows related to reproducibility, i.e., workflow preservation, traceability of the results and workflow sharing.

There are two ways in which a workflow may be preserved. The first way is by documenting the method captured by the workflow itself, i.e., providing enough details on each of the tasks of the workflow for anyone to be able to understand their functionality (Garijo & Gil, 2011; Belhajjame et al., 2015). The rationale is simple: given the pace at which software and data evolve, it is difficult to ensure that within five, ten or twenty years the whole workflow will still be reusable. This

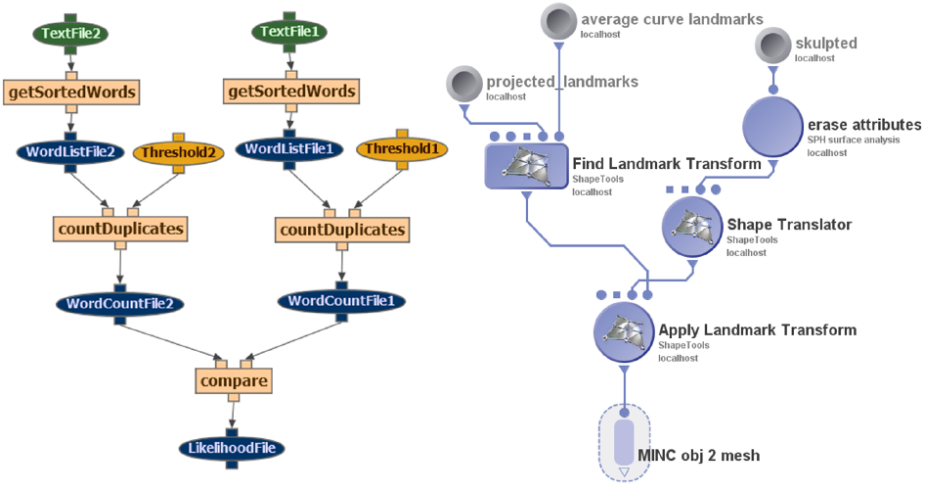


Figure 1: Two scientific workflows from two different workflow systems. The one on the left represents tasks as rectangles and data with ovals, while the one on the right represents task in blue and inputs in grey.

is common in domains where scientific workflows rely on external web services and evolving community-built datasets (e.g., the Protein Data Bank[3] in bioinformatics). New releases of software, changes to the existing APIs or new data discoveries may supersede existing resources, making them outdated and sometimes incompatible with the rest of the tasks in the workflow. Therefore, documentation approaches tend to contextualize, describe and generalize the functionality of every dataset and task used in the workflow. Documentation approaches are usually complemented with sample data, pointing to archived versions of the software to facilitate understanding the original method. Another key feature of these approaches includes documenting the provenance of the results of a workflow. The provenance of a result aims to capture its creation process, i.e., all the steps that contributed to its outcome, including the original datasets and intermediate data. A provenance record also attributes credit to the scientists responsible for producing the result. There is a standard model for provenance publishing on the web (Lebo et al., 2013), and related work has extended it to publish scientific workflow metadata[4] (Garijo & Gil, 2011; Belhajjame et al., 2015; Missier, Dey, Belhajjame, Cuevas-Vicenttn, & Ludscher, 2013). Once a workflow is documented, it may be included as part of a repository (Roure, Goble, & Stevens, 2009; Mates, Santos, Freire, & Silva, 2011; Belhajjame et al., 2013) for others to reuse.

The second way to preserve workflows is by capturing their functionality in containers (e.g., Docker[5]) or virtual machines. This way the workflow becomes a black box that performs the experiment functionality, including inputs, software and dependencies for execution. The challenge relies in the creation process of such containers. Approaches like (Chirigati, Shasha, & Freire, 2013) monitor the execution of the experiment to create a virtual machine, while approaches like (Santana-Prez & Prez-Hernndez, 2015) depend on the authors to document the infrastructure details for the workflow. Recent work has proposed a more flexible approach, capturing each of the steps of the workflow as an independent container (Qasha, Cala, & Watson, 2016). Finally notebooks[6] are gaining a lot of momentum as an alternative lightweight method to encapsulate and test script based experiments.

Scientific workflows have demonstrated to be useful to re-execute, reuse and share the methods and tasks commonly used in a community (Garijo et al., 2014). Workflows should be treated as first class citizens in cyberinfrastructure (Gil et al., 2007), since they provide the means of transparent and reproducible

[3]http://www.rcsb.org/pdb/home/home.do
[4]http://vcvcomputing.com/provone/provone.html
[5]https://www.docker.com/
[6]http://jupyter.org/

work. There are still open challenges in workflows, and venues like eScience[7] and Super Computing[8] discuss and publish new research every year.

References

American, S. (2010, October). In Science We Trust: Poll Results on How you Feel about Science. *Scientific American*.

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, *533*(7604). Retrieved from http://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970

Begley, C., & Lee, M. (2012, March). Drug development: Raise standards for preclinical cancer research. , *483*, 531–533. doi: 10.1038/483531a

Belhajjame, K., Zhao, J., Garijo, D., Gamble, M., Hettne, K., Palma, R., ... Goble, C. (2015). Using a suite of ontologies for preserving workflow-centric Research Objects. *Web Semantics: Science, Services and Agents on the World Wide Web*.

Belhajjame, K., Zhao, J., Garijo, D., Garrido, A., Soiland-Reyes, S., Alper, P., & Corcho, O. (2013). A workflow PROV-corpus based on Taverna and Wings. In *Proceedings of the Joint EDBT/ICDT 2013 Workshops* (pp. 331–332). Genoa, Italy: ACM. Retrieved from http://doi.acm.org/10.1145/2457317.2457376 doi: 10.1145/2457317.2457376

Bergmann, R., & Gil, Y. (2014). Similarity assessment and efficient retrieval of semantic workflows. *Information Systems*, *40*, 115 – 127. Retrieved from http://www.sciencedirect.com/science/article/pii/S0306437912001020 doi: http://dx.doi.org/10.1016/j.is.2012.07.005

Callahan, S. P., Freire, J., Santos, E., Scheidegger, C. E., Silva, C. T., & Vo, H. T. (2006). Vistrails: Visualization meets data management. In *In ACM SIGMOD* (pp. 745–747). ACM Press.

Chirigati, F., Shasha, D., & Freire, J. (2013). ReproZip: Using Provenance to Support Computational Reproducibility. In *Proceedings of the 5th USENIX Workshop on the Theory and Practice of Provenance* (pp. 1:1–1:4). Lombard, Illinois: USENIX Association. Retrieved from http://dl.acm.org/citation.cfm?id=2482949.2482951

Collaboration, O. S. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251). doi: 10.1126/science.aac4716

Deelman, E., Blythe, J., Gil, Y., Kesselman, C., Mehta, G., Patil, S., ... Livny, M. (2004). Pegasus: Mapping Scientific Workflows onto the Grid. In M. Dikaiakos (Ed.), *Grid Computing* (Vol. 3165, pp. 11–20). Springer Berlin / Heidelberg.

Dinov, I. D., Horn, J. D. V., Lozev, K. M., Magsipoc, R., Petrosyan, P., Liu, Z., ... Toga, A. W. (2009). Efficient, Distributed and Interactive Neuroimaging Data Analysis Using the LONI Pipeline. In *Frontiers in Neuroinformatics* (Vol. 3). doi: 10.3389/neuro.11.022.2009

Fanelli, D. (2010). Do Pressures to Publish Increase Scientists Bias? An Empirical Support from US States Data. *PLoS ONE*, *5*(4). doi: 10.1371/journal.pone.0010271.

Filgueira, R., Atkinson, M., Bell, A., Main, I., Boon, S., Kilburn, C., & Meredith, P. (2014). eScience Gateway Stimulating Collaboration in Rock Physics and Volcanology. In *e-Science (e-Science), 2014 IEEE 10th International Conference on* (Vol. 1, pp. 187–195). IEEE.

Garijo, D., Corcho, O., Gil, Y., Braskie, M. N., Hibar, D., Hua, X., ... W.Toga, A. (2014). Workflow Reuse in Practice: A Study of Neuroimaging Pipeline Users. In *10th IEEE International Conference on eScience 2014*.

Garijo, D., & Gil, Y. (2011). A New Approach for Publishing Workflows: Abstractions, Standards, and Linked Data. In *Proceedings of the 6th workshop on Workflows in support of large-scale science* (pp. 47–56). Seattle: ACM. Retrieved from http://dl.acm.org/citation.cfm?id=2110504

Garijo, D., Kinnings, S., Xie, L., Xie, L., Zhang, Y., Bourne, P. E., &

[7] http://escience-2016.idies.jhu.edu/

[8] http://www.supercomp.org/

Gil, Y. (2013). Quantifying Reproducibility in Computational Biology: The Case of the Tuberculosis Drugome. *PLoS ONE*, *8*(11), e80278. Retrieved from http://dx.doi.org/10.1371%2Fjournal.pone.0080278 doi: 10.1371/journal.pone.0080278

Giardine, B., Riemer, C., Hardison, R. C., Burhans, R., Elnitski, L., Shah, P., ... Nekrutenko, A. (2005, October). Galaxy: a platform for interactive large-scale genome analysis. *Genome Research*, *15*(10), 1451–1455.

Gil, Y., Deelman, E., Ellisman, M., Fahringer, T., Fox, G., Gannon, D., ... Myers, J. (2007, December). Examining the Challenges of Scientific Workflows. *Computer*, *40*(12), 24–32. doi: 10.1109/MC.2007.421

Gil, Y., Ratnakar, V., Kim, J., Gonzlez-Calero, P. A., Groth, P. T., Moody, J., & Deelman, E. (2011). Wings: Intelligent Workflow-Based Design of Computational Experiments. *IEEE Intelligent Systems*, *26*(1), 62–72.

Goderis, A. (2008). *Workflow re-use and discovery in Bioinformatics* (Unpublished doctoral dissertation). School of Computer Science, The University of Manchester.

Goderis, A., Sattler, U., Lord, P., & Goble, C. (2005). Seven Bottlenecks to Workflow Reuse and Repurposing. In *The Semantic Web ISWC 2005* (Vol. 3729, pp. 323–337). Springer Berlin Heidelberg. Retrieved from http://dx.doi.org/10.1007/11574620_25

Holl, S. (2014). *Automated Optimization Methods for Scientific Workflows in e-Science Infrastructures* (Doctoral dissertation, University of Bonn). Retrieved from http://hss.ulb.uni-bonn.de/2014/3508/3508.htm

Lebo, T., McGuiness, D., Belhajjame, K., Cheney, J., Corsar, D., Garijo, D., ... Zhao, J. (2013, April). *The PROV ontology, W3c Recommendation* (Tech. Rep.). WWW Consortium. Retrieved from http://www.w3.org/TR/2013/REC-prov-o-20130430/

Ludscher, B., Altintas, I., Berkley, C., Higgins, D., Jaeger, E., Jones, M., ... Zhao, Y. (2006). Scientific workflow management and the Kepler system. *Concurrency and Computation: Practice and Experience*, *18*(10), 1039–1065.

Marcus, A., & Oransky, I. (2014, December). Top retractions of 2014. *The Scientist*.

Mates, P., Santos, E., Freire, J., & Silva, C. T. (2011). CrowdLabs: Social Analysis and Visualization for the Sciences. In *23rd International Conference on Scientific and Statistical Database Management (SSDBM)* (pp. 555–564). Springer.

Missier, P., Dey, S., Belhajjame, K., Cuevas-Vicenttn, V., & Ludscher, B. (2013). D-PROV: Extending the PROV Provenance Model with Workflow Structure. In *Proceedings of the 5th USENIX Workshop on the Theory and Practice of Provenance* (pp. 9:1–9:7). Lombard, Illinois: USENIX Association. Retrieved from http://dl.acm.org/citation.cfm?id=2482949.2482961

Qasha, W., Cala, J., & Watson, P. (2016). A Framework for Scientific Workflow Reproducibility in the Cloud. In *IEEE 12th International Conference on eScience.*

Rockoff, J. D. (2015, September). Amgen Finds Data Falsified in Obesity-Diabetes Study Featuring Grizzly Bears. *The Wall Street Journal*. Retrieved from http://www.wsj.com/articles/amgen-finds-data-falsified-in-obesity-diabetes-study-featuring-grizzly-bears-1441123200

Roure, D. D., Goble, C. A., & Stevens, R. (2009). The design and realisation of the myExperiment Virtual Research Environment for social sharing of workflows. *Future Generation Comp. Syst.*, *25*(5), 561–567.

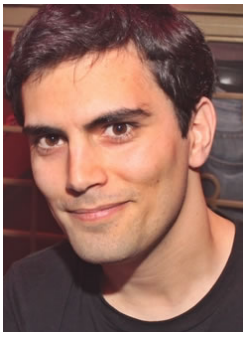
Santana-Prez, I., & Prez-Hernndez, M. (2015). Towards Reproducibility in Scientific Workflows: An Infrastructure-Based Approach. *Scientific Programming*, *2015*, 11. doi: 10.1155/2015/243180

Starlinger, J., Brancotte, B., Cohen-Boulakia, S., & Leser, U. (2014). Similarity Search for Scientific Workflows. *Proceedings of the VLDB Endowment*, *7*(12), 1143–1154.

Taylor, I. J., Deelman, E., Gannon, D. B., & Shields, M. (2006). *Workflows for e-Science: Scientific Workflows for Grids*. Secaucus, NJ, USA: Springer-

Verlag New York, Inc.

Wolstencroft, K., Haines, R., Fellows, D., Williams, A., Withers, D., Owen, S., ... Goble, C. (2013). The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud. *Nucleic Acids Research*.



Daniel Garijo is a postdoctoral researcher at the Information Sciences Institute of the University of Southern California. His research activities focus on e-Science and the Semantic web, specifically on how to increase the understandability of scientific workflows using provenance, metadata, intermediate results and Linked Data.