Computational Reproducibility: State-of-the-Art, Challenges, and Database Research Opportunities

Juliana Freire
NYU Poly
juliana.freire@nyu.edu

Philippe Bonnet
IT University of Copenhagen
phbo@itu.dk

Dennis Shasha
New York University
shasha@courant.nyu.edu

ABSTRACT

Computational experiments have become an integral part of the scientific method, but reproducing, archiving, and querying them is still a challenge. The first barrier to a wider adoption is the fact that it is hard both for authors to derive a compendium that encapsulates all the components needed to reproduce a result and for reviewers to verify the results. In this tutorial, we will present a series of guidelines and, through hands-on examples, review existing tools to help authors create of reproducible results. We will also outline open problems and new directions for database-related research having to do with querying computational experiments.

Categories and Subject Descriptors

D.0 [**Software**]: General; H.4 [**Information Systems Applications**]: Miscellaneous; H.2 [**Database Management**]: Database Applications

General Terms

Documentation, experimentation

Keywords

Computational reproducibily, provenance

1. COMPUTATIONAL EXPERIMENTS AND REPRODUCIBILITY

Important scientific results not only give insight but also lead to practical progress. The ability to test results is crucial for science to be self-correcting. In natural science, long tradition requires that results be reproducible by researchers the world over: an experiment done by laboratory L at time t is deemed to be reproducible if it can be repeated at a possibly different laboratory L' at some later time t'. For this to be the case, the description of the experiment must be sufficiently precise to allow repetition. When that fails, scandal ensues [11, 26, 27].

SIGMOD '12, May 20–24, 2012, Scottsdale, Arizona, USA.

Computational experiments support not only computer science research, but also natural science, social science and humanities research. For that reason at least, computational experiments should meet the same standards of reproducibility as natural science experiments. The lack of reproducibility for computational results currently reported in the literature has raised questions about their reliability [8] and has led to a widespread discussion of the importance of computational reproducibility. This has already resulted in a number of workshops, special issues, and software tools [17, 3, 18, 22, 30, 1, 13, 6]. Academic institutions such as ETH in Switzerland, funding agencies, conferences and journals have all pushed for authors to include reproducible results in their publications [10, 25, 5, 19, 20, 33, 4, 12].

Our goals for computational experiments follow the spirit of reproducibility for natural science experiments. A computational experiment that has been developed at time t on hardware/operating system s on data d is reproducible if it can be executed at time t' on system s' on data d' that is similar to (or potentially the same as) d. Such experiments are the basic building block for reproducible research papers, which, in addition to text, include data, specification of computational processes, software/code, as well as information about the environment used to derive the results.

There are many side benefits to producing and using reproducible experiments:

1. *Reproducible programs can be compared.* Different programs for the same purpose can be compared based on performance and quality on a common data set.

2. *Reproducible software and results are documented.* Newcomers to a field or to a research group can understand how to use and modify the software and its data inputs. This in turn allows others to more easily build upon existing research.

3. *Reproducible software is portable.* The software that drives the original experiment may be combined with other software in some target environment to result in an entirely new artifact.

4. *Reproducible experiments are cited.* Recent work has shown that papers including reproducible experiments have higher impact and visibility than those that don't [34] (though admittedly there may be confounding factors).

Within computer science, the database community has been a pioneer in the adoption of reproducible experiments [20, 19, 5]: SIGMOD has had a repeatability committee since 2008, and starting in 2012, VLDB will also provide repeatability evaluation. As reproducible computational experiments become ubiquitous and are

shared across different scientific domains, the database community is uniquely qualified to contribute to tools that play on them: these experiments can, after all, be viewed as data items that can be queried, modified, executed, and visualized [15].

Even though it is clear that reproducible computational experiments will make science better, there are many challenges that need to be addressed, including: how to create and package these experiments; how to query and combine experiments to make them useful to future researchers; and how to mine experimental repositories to find useful patterns. This tutorial will present a series of tools and will provide hands-on examples of how to create reproducible experiments. Besides suggesting guidelines to create the experiments, we will also discuss many of the pitfalls that should be avoided. We will conclude with a discussion on new research problems this initiative suggests.

2. TUTORIAL OUTLINE

Computational Reproducibility in Science: Overview. We will discuss the motivation for and importance of reproducible experiments, as well as their benefit to authors and to the scientific community in general. We will also review the key issues, social, legal [31], and operational [2] associated to the publication of reproducible experiments. Last but not least, we will provide examples of how different communities are approaching this problem (e.g., [9, 28, 33, 4]).

Creating and Publishing Reproducible Experiments. Although full reproducibility is the ideal, it may not be attainable in practice. Based on our experience with the SIGMOD Repeatability effort and users of the VisTrails reproducibility infrastructure [17, 16], we have identified the following criteria to characterize experiments with respect to the level or reproducibility:

1. *Depth:* How much of an experiment is made available or archived: (a) a set of figures associated to a manuscript (this is the default level today); (b) the script (or spreadsheet file) used to generate the figures included in the paper together with the appropriate data sets; (c) the raw data measured during the experiments together with the scripts used to obtained different levels of derived datasets (e.g., the raw data is IO response time measured for a period T and for various IO sizes, and the derived datasets are the average response time per IO size); (d) the set of experiments (system configuration and initialization, scripts, workload, measurement protocol) used to produce the raw data; (e) the software system as a white box (source, configuration files, build environment) or black box (executable) on which the experiments are performed.

2. *Portability:* Can the results be reproduced (a) on the original environment (basically the author of the experiment can replay it on his or her machine);[1] (b) on a similar environment (i.e., same OS but different machines), or (c) on a different environment (i.e., on a different OS or machine).

3. *Coverage:* How much of the experiments can be reproduced: (a) partial, or (b) full reproducibility.

In this part of the tutorial, we will present a series of guidelines that authors can follow during the design (and execution) of their experiments to make them reproducible. We will discuss potential pitfalls that can hamper the quality of reproducible experiments. While in some cases it may be possible to reproduce the whole process from data acquisition to the final plot in the paper, in others it may be possible (or desirable) to reproduce only *part* of the process. There are, however, simple workarounds that can be used to ensure that key parts of the experiments can be verified. For example, even though access to a third-party Web service may not be reproducible, by storing the results obtained from the Web service, not only can these results be examined, but the remaining components of the experiment can be reproduced. Similarly, results derived by complex processes (e.g., long-running simulations) can be cached and packaged together with the experiments.

Tools and Resources. We will give an overview of state-of-the-art tools and resources that are available to authors. We will address the following key questions. First, what are the tools that a research team should use to archive and distribute experiments at different levels of depth, portability and coverage? Second, what is the infrastructure that should be deployed to publish experiments for the community, and eventually to allow researchers to build on the experiments provided by others?

To demonstrate different tools that are essential in the creation of reproducible experiments, we will use real (hands-on) examples. For example:

- We will illustrate the usefulness of *version control systems* for creating and managing reproducible experiments as well as discuss their limitations.

- We will provide examples of the problems that ensue when portability is ignored, and show how to create portable experiments using virtual machines and packaging systems such as CDE.[2]

- To automate the repeatability process, it is necessary to capture (i) the computational process (i.e., workflow or implicit script) that was used to generate the figures in a paper and (ii) the provenance information for all components of the experiment. We will demonstrate tools such as Madagascar[3] and Vistrails [16][4], which provide support for these tasks.

- An important component of a research project is the publication of the results. We will give an overview of different approaches for including reproducible results in research articles (see e.g., [17, 22]) as well through interactive Web-based interfaces [29, 21].

Open Research Problems. Computational reproducibility efforts in many communities will soon give rise to experiment repositories that hold data and software. The database community can take the lead in showing how to advance science by providing new techniques and tools for designing the experiment repositories and exploring the information they hold. In this section, we will discuss open problems for database research related to the design, management and exploration of experiment repositories, including but not limited to:

- *Design and Management of Experiment Repositories.* Although there are already some repositories in use that cater to different aspects of reproducibility (see e.g., [24, 9, 28, 23, 7, 21]), they are still in their infancy, and there are many open questions about what their architecture should be, or even how they will be used. For example: will publishers such as

[1]This might be the preferred mode for proving the reproducibility of proprietary software: outsiders would be allowed to play with software and data sets but not download the software.

[2]http://www.stanford.edu/ pgbovine/cde.html

[3]http://www.reproducibility.org/wiki/Tutorial

[4]http://vistrails.org

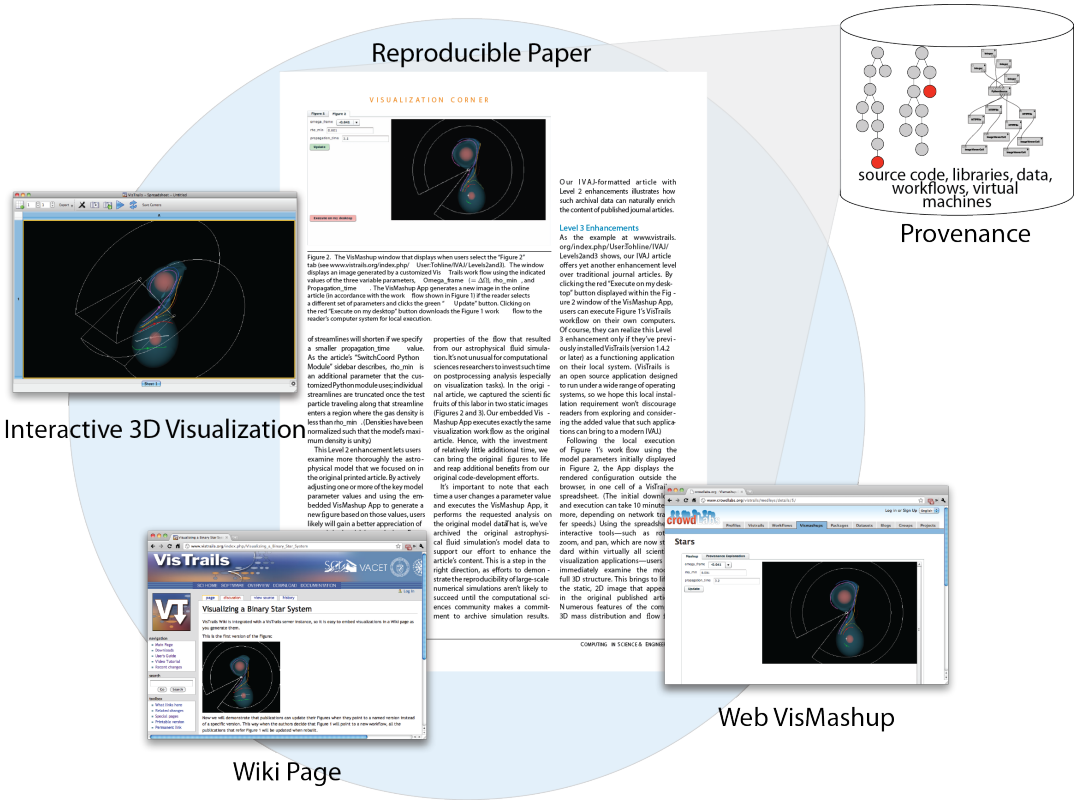

Figure 1: A reproducible paper [32]. A result reported in a reproducible paper comes with a deep caption that includes the components required to reproduce that result, including, for example, the specification of the computational process used, the underlying source code and libraries, raw data, and the environment. Results can be published using different media, including PDF, Word, Wiki, HTML pages, and they can be both static and interactive.

ACM host a mega-repository? Or will authors publish experiments on their own Web site and *experiment search engines* be developed to locate and index them? How will usage be controlled and tracked? In addition, since these repositories will hold diverse data—source code, binary files, raw data, structured workflows, provenance—an important question is how to efficiently store, index and connect this information.

- *Querying and Searching Experiments.* Because reproducible experiments have several levels of data and meta-data, they can enable researchers to identity related work in new ways. They can formulate queries over the method of an experiment (e.g., find experiments that perform data cleaning using Algorithm A followed by integration using System B); over the raw data (e.g., find all experiments that operate over a dataset with salinity information from the Columbia River); that straddle raw data and method (e.g., find a workflow that generates a volume rendering of salinity data); etc. Because of the inherent heterogeneity of the data, which include structured and unstructured data contributed by many different users, there are new challenges for querying. A flexible query system (and language) is needed that is able to cross the boundaries between structured and unstructured data perhaps in a similar way to querying on dataspace systems [14]. In addition, intuitive and visual interfaces are needed that can be used by a broad range of users who do not necessarily have computer science expertise, and that allow them to iteratively explore the information in the repository.

- *Mining Experiments* The availability of large collections of experiments opens up new opportunities for knowledge discovery. A notable example is the ability to measure the impact of a specific piece of work or of a given area. For example, given an algorithm A, count other papers that use A (directly or indirectly); or given a set of algorithms in a given area, count their use in other areas. Many other questions and meta-questions can potentially be answered, for example: what are the characteristics of experiments that have had high impact? what are the most common workflows used to solve a given problem? The nature of the data creates new challenges for mining.

3. ABOUT THE PRESENTERS

Juliana Freire is a Professor of Computer Science at NYU Poly. Her research interests include Web mining and crawling, large-scale information integration, information visualization, and scientific data management. She is a co-creator of VisTrails, an open-source data analysis and visualization system that supports the creation and publication of reproducible results (www.vistrails.org). Since 2010, she has been working with the repeatability initiative of SIGMOD.

Philippe Bonnet is an associate professor at IT University of Copenhagen. He is an experimental computer scientist; his research interests include flash-based database systems, sensor data management and computational repeatability. Philippe currently serves as chair for the SIGMOD and VLDB reproducibility committees.

Dennis Shasha is a professor of computer science at New York University where he works with biologists on pattern discovery for network inference; with physicists and financial people on algorithms for time series; on database applications in untrusted environments; on database tuning; and on computational reproducibility. He has been working with the repeatability and workability initiative of SIGMOD since 2008.

4. ACKNOWLEDGMENTS

This work was partially supported by the the Department of Energy and National Science Foundation awards IIS-1139832, IIS-1142013, IIS-0905385, IIS-1050388, IOS-0922738, MCB-0929339, and NIH 2R01GM032877-25A1. That support is greatly appreciated.

5. REFERENCES

[1] ICIAM Workshop on Reproducible Research: Tools and Strategies for Scientific Computing. http://www.mitacs.ca/events/index.php?option=com_content&view=article&id=214&Itemid=230&lang=en, 2011.

[2] B. Bauer, J. Gukelberger, B. Surer, and M. Troyer. Publishing provenance-rich scientific papers. In *USENIX Workshop on the Theory and Practice of Provenance*, 2011.

[3] Beyond the PDF Workshop. https://sites.google.com/site/beyondthepdf, 2011.

[4] Biostistics: Information for authors. http://www.oxfordjournals.org/our_journals/biosts/for_authors/msprep_submission.html.

[5] P. Bonnet, S. Manegold, M. Bjørling, W. Cao, J. Gonzalez, J. Granados, N. Hall, S. Idreos, M. Ivanova, R. Johnson, D. Koop, T. Kraska, R. Müller, D. Olteanu, P. Papotti, C. Reilly, D. Tsirogiannis, C. Yu, J. Freire, and D. Shasha. Repeatability and workability evaluation of sigmod 2011. *SIGMOD Record*, 40(2):45–48, 2011.

[6] K. M. Chandy, J. Kiniry, A. Rifkin, and D. Zimmerman. Webs of Archived Distributed Computations for Asynchronous Collaboration. *The Journal of Supercomputing*, 11(2):101–118, Oct. 1997.

[7] CrowdLabs. http://www.crowdlabs.org.

[8] D. Donoho, A. Maleki, I. Rahman, M. Shahram, and V. Stodden. Reproducible research in computational harmonic analysis. *Computing in Science & Engineering*, 11(1):8–18, Jan.-Feb. 2009.

[9] Dryad. http://datadryad.org.

[10] Guidelines for Research Integrity and Good Scientific Practice at ETH Zurich. http://www.vpf.ethz.ch/services/researchethics/Broschure.

[11] ETH Zurich's head of research resigns. http://www.ethlife.ethz.ch/archive_articles/090921_Peter_Chen_Ruecktritt_MM/index_EN.

[12] The executable paper grand challenge, 2011. http://www.executablepapers.com.

[13] S. Fomel and J. Claerbout. Guest editors' introduction: Reproducible research. *Computing in Science Engineering*, 11(1):5 –7, 2009.

[14] M. J. Franklin, A. Y. Halevy, and D. Maier. From databases to dataspaces: a new abstraction for information management. *SIGMOD Record*, 34(4):27–33, 2005.

[15] J. Freire, P. Bonnet, and D. Shasha. Exploring the coming repositories of reproducible experiments: Challenges and opportunities. *PVLDB*, 4(12):1494–1497, 2011.

[16] J. Freire and C. Silva. Making computations and publications reproducible with vistrail. *Computing in Science Engineering*, 2012.

[17] D. Koop, E. Santos, P. Mates, H. T. Vo, P. Bonnet, B. Bauer, B. Surer, M. Troyer, D. N. Williams, J. E. Tohline, J. Freire, and C. T. Silva. A provenance-based infrastructure to support the life cycle of executable papers. *Procedia Computer Science*, 4:648 – 657, 2011. Proceedings of the International Conference on Computational Science, ICCS 2011.

[18] Madagascar. http://www.reproducibility.org/wiki/Main_Page.

[19] S. Manegold, I. Manolescu, L. Afanasiev, J. Feng, G. Gou, M. Hadjieleftheriou, S. Harizopoulos, P. Kalnis, K. Karanasos, D. Laurent, M. Lupu, N. Onose, C. Ré, V. Sans, P. Senellart, T. Wu, and D. Shasha. Repeatability & workability evaluation of SIGMOD 2009. *SIGMOD Record*, 38(3):40–43, 2009.

[20] I. Manolescu, L. Afanasiev, A. Arion, J. Dittrich, S. Manegold, N. Polyzotis, K. Schnaitter, P. Senellart, S. Zoupanos, and D. Shasha. The repeatability experiment of SIGMOD 2008. *SIGMOD Record*, 37(1):39–45, 2008.

[21] P. Mates, E. Santos, J. Freire, and C. Silva. Crowdlabs: Social analysis and visualization for the sciences. In *Proceedings of SSDBM*, pages 555–564, 2011.

[22] J. Mesirov. Accessible reproducible research. *Science*, 327(5964):415–416, 2010.

[23] myExperiment. http://www.myexperiment.org.

[24] nanoHub. http://nanohub.org.

[25] Nsf task force on data policies. http://www.nsf.gov/nsb/committees/tskforce_dp.jsp.

[26] Nobel laureate retracts two papers unrelated to her prize. http://www.nytimes.com/2010/09/24/science/24retraction.html?_r=1&emc=eta1, September 2010.

[27] It's science, but not necessarily right. http://www.nytimes.com/2011/06/26/opinion/sunday/26ideas.html?_r=2, June 2011.

[28] The openscience project. http://www.openscience.org.

[29] E. Santos, L. Lins, J. Ahrens, J. Freire, and C. T. Silva. Vismashup: Streamlining the creation of custom visualization applications. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1539–1546, 2009.

[30] SIAM Mini-symposium on Verifiable, Reproducible Research and Computational Science. http://meetings.siam.org/sess/dsp_programsess.cfm?SESSIONCODE=11845.

[31] V. Stodden. The legal framework for reproducible scientific research: Licensing and copyright. *Computing in Science and Engineering*, 11(1):35–40, Jan-Feb 2009.

[32] J. Tohline and E. Santos. Visualizing a journal that serves the computational sciences community. *Computing in Science Engineering*, 12(3):78 –81, may-june 2010.

[33] IEEE Transactions on Signal Processing. http://www.signalprocessingsociety.org/publications/periodicals/tsp.

[34] P. Vandewalle, J. Kovacevic, and M. Vetterli. Reproducible research in signal processing - what, why, and how. *IEEE Signal Processing Magazine*, 26(3):37–47, May 2009.