Publishing provenance-rich scientific papers

Bela Bauer, Jan Gukelberger, Brigitte Surer, Matthias Troyer
Institute for Theoretical Physics, ETH Zurich, CH-8093 Zurich

Abstract

Complete documentation and reproducibility of results are important goals for scientific publications. Standard scientific papers, however, usually contain only final results and document only parameters and processing steps that the authors considered important enough. By recording the complete provenance history of the data leading to a publication one can overcome this limitation and allow reproducibility for reviewers, publishers and readers of scientific publications. While the process of capturing provenance information is a growing research subject, here we discuss usually overlooked challenges involved in publishing provenance-complete papers. We report on our experience in preparing and publishing two specific *executable papers* where we used the VisTrails workflow system to embed full provenance information of the paper and discuss open challenges and issues we encountered.

1 Introduction

The ETH Zurich Research Ethics Guidelines demand that *[...] all steps in the treatment of primary data must be documented [...] in such a way as to ensure that the results obtained from the primary data can be reproduced completely. Primary data must be filed and safeguarded in such a way as to ensure that they can be securely retrieved for later use of verification [...].*

Achieving these goals in scientific publications presents serious challenges both to authors and publishers and affects all steps in the scientific process: from the early stages of simulation and data analysis, to preparing figures and tables, and finally the production process of a scientific journal and the long-term archival of the paper, data and provenance information. For the authors, recording complete provenance information for all steps of obtaining raw data and their analysis is a difficult task due to the exploratory nature of scientific work. If a significant amount of work from the user is required, it is likely that provenance information will be incomplete. It is thus imperative to automate the recording and publishing of provenance information in a way that uses existing tools that a scientist may have. For automatic recording of provenance information the VisTrails [1] workflow system was found to be suitable for parts of the process of running simulations and analyzing results. A provenance-enabled workflow system enables the creation of an executable paper [5]. This allows readers to follow and reproduce the process that lead to a publication. Publishers show interest in such multi-layered publications containing more than just the printed text. For example, the "executable paper grand challenge" was initiated by Elsevier [2]. Nevertheless, challenges arise during the publication process that have not been adequately addressed yet by any system. Here we discuss

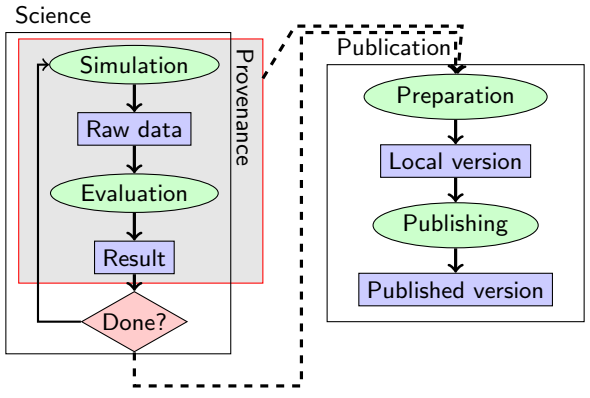

Figure 1: The scientific process. We broadly distinguish two phases: the scientific process of obtaining results, and the publishing process of creating a manuscript and publishing it in a peer-reviewed journal. Provenance information should be recorded for the first part of the process and then embedded in the final paper, which may require significant changes to the publishing process.

some challenges we encountered and our approach to solve some of these issues. Two papers with two different publishers are described: Ref. [3] is in the last stages of production and Ref. [4] is in preparation. We address the whole scientific process, illustrated in Fig. 1 including the preparation of a manuscript and the completion of the published paper with full provenance information (c.f. Fig. 2). We focus primarily on tools and practices found to be useful in real scientific projects and identify some remaining challenges.

2 Recording provenance during the scientific process

Recording the full provenance information during the exploratory scientific process requires an automated approach. In the following, we discuss the specific tools we used to capture this information. A workflow system is an environment where users can construct complex computational tasks from a restricted set of modules, which perform well-defined small tasks. After setup it is run in a well-defined execution environment permitting observation of its behavior and recording the provenance information. Such a system can conserve all modification stages of a workflow to construct the full provenance history. For the two publications discussed here we have used the VisTrails workflow system [1] and the ALPS libraries [3].

Vistrails: The VisTrails system is an open-source provenance-enabled workflow system based on the portable open-source tools Python and Qt. It is specifically designed for the exploratory nature of scientific work and is easily integrated with existing tools. VisTrails comes with a variety of modules which aid the user in data handling, visualization and enables users to construct their own VisTrails modules. VisTrails implements an efficient caching mechanism for intermediate results whereupon changing a module parameter or the workflow structure only the parts depending on changed data need to be recomputed. With the VisTrails *Persistence* package, this caching scheme is extended such that data can be stored in a central repository [6].

The ALPS (Algorithms and Libraries for Physics Simulations) [3] project has open-source basic libraries for simulations of classical and quantum lattice problems and applications containing widely-used algorithms. Integration of many different libraries and algorithms is achieved by using common data formats for both input and output files. VisTrails modules exist that expose much of the functionality for setting up and running simulations as well as analyzing data. ALPS also comes with a set of libraries which ease recording of provenance information in the output files of large-scale simulations.

Large-scale simulations: Reproducibility requirements are much harder to meet for large-scale computations performed on supercomputers. Repeating computations is often neither feasible nor desirable and we therefore treat such numerical simulations on the same footing as experiments. We record and document a limited set of information allowing an expert in the field to reproduce the results. In particular, we record all the information which the researchers need in order to reproduce the simulations as meta data in the results. In our experience, the most important information to store includes the complete input data, including random seeds, the exact source code used, build options (compiler and library versions, build flags) and how the code was run, i.e. on what machine, when, by whom, with what command line options. While recording such information can seem like a daunting task to a beginning researcher, this process can be greatly simplified by developing libraries and build tools which automate the process of collecting such information and adding it to the result files – as we have done in the ALPS project.

One missing ingredient is keeping track of the state (software and hardware) of a supercomputer system at a given time. This should not be the task of individual researchers but rather supercomputer centers. Similarly, tools and operating system support need to be developed to automatically capture the relevant information on individual workstations.

3 The publishing process

After identifying the information needed for full reproducibility of results and presenting ways to facilitate its recording, we now turn to the topic of publishing a manuscript containing full simulation results and provenance information. The full input and output data of experiments and simulations which cannot be easily repeated need to be publicly and permanently available. This issue concerns data formats and storage locations.

Data formats: Publishing raw data is only useful if done in an open and well-documented format accessible to readers of the paper. Formats need to be stable for a long time and should not be specific to software versions, which may be abandoned. Without open community formats, definitions and documentation are necessary first steps. In condensed matter physics, we have defined open data formats based on XML and HDF5. Both are widely used file formats and supported by a large community with open source codes which will be maintained for a long time. The schemas used to store data are documented by the ALPS project and are accessible to any reader.

Archiving data: While some communities maintain public archives for large datasets, especially in observa-

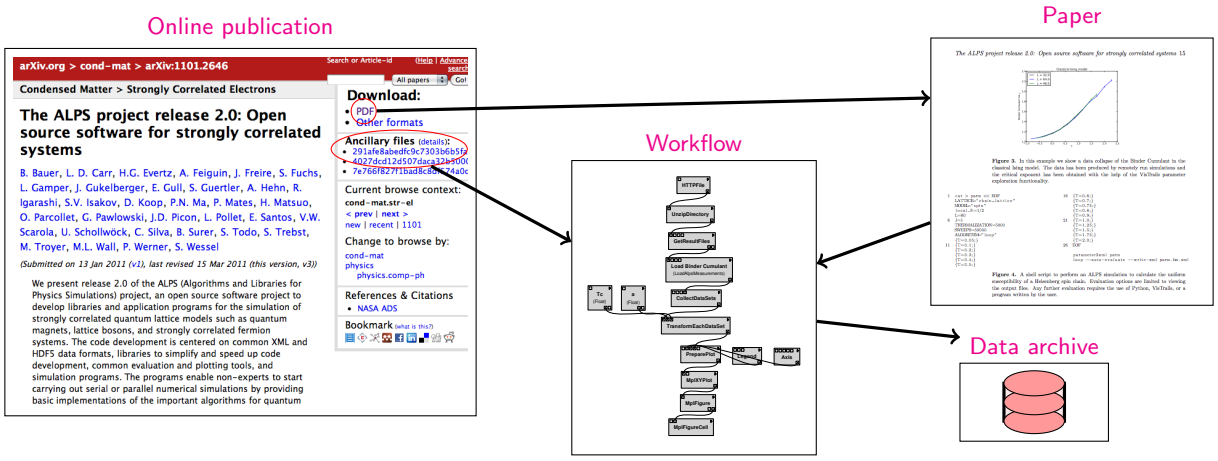

Figure 2: In a provenance-rich paper, the online publication (here on the arXiv preprint archive) will contain the paper and auxiliary files, for example the full workflow that leads from raw data to a certain figure. The paper itself will contain links to the workflow files stored by the publisher. These worflows themselves might refer to simulation results in a data archive, which can be maintained either on an institutional level or also by the publisher.

tional sciences, other fields have no established common archives for experimental results or numerical simulations. Thus researchers have to find their own archiving solution. Such archives could be maintained by individual research groups, their institutions, or by publishers. Of these options, single research groups are the least desirable as the lifetime of a scientific publication will usually exceed the professional career of its authors. Many publishers, including the American Physical Society and the Institute of Physics, permitted authors to publish a limited amount of auxiliary files along with their papers. Only recently, publishers have started lifting size limitations and aim to publish the full data sets together with the paper.

3.1 Workflows and codes

The optimum for reproducibility is publishing not only all the data but also workflows used to obtain the published results. Here we use the term workflows in a broad sense not limited to workflow files of a specific workflow system, but also to scripts or codes used to produce the results. Such an approach opens the possibility for executable papers [5] which enable readers to inspect and redo the published results.

Proprietary codes: Completely public codes are not a standard practice in many fields nor are they required by the ethics guidelines cited in the introduction. Computational results are often obtained using proprietary software that authors do not want publicly distributed. Not publishing the code, however, shifts the archival burden to authors. Authors might distribute source codes upon request assuring that the provenance information in the paper allows rerunning the simulations.

Adapting workflows to published versions of the data: An additional complication arises when the workflows refer to *local versions* of input data and have to be changed to the actual location of the *published versions* of the data. To address this issue it will be very important to avoid inconvenient and error-prone manual changes to workflows and to provide tools that can automatically provide these transformations in a safe and transparent way.

3.2 Publishing provenance-rich papers

The actual scientific paper is usually a PDF document downloadable from the website of the journal. We intend to provide the reader with the full provenance information by publishing the workflows and data along with the PDF document. The linking of the different components of such a multi-layered publication can be achieved by enhancing figures with "deep captions" – links from the figure to the provenance information showing how it was produced – and referencing the archival location of the data from the workflows. An overview of such a paper is shown in Fig. 2. Keeping the links intact and publishing such a multi-layered paper challenges the current publishing processes.

Preparing the paper: In the preparation of our papers we profit from a VisTrails LaTeX plugin which directly connects a figure to the workflow that generated it. This plugin is based on a VisTrails server infrastructure: the LaTeX source contains a command that takes the address of a VisTrails server, the workflow number and version used to create the figure. When compiling the

file, the figure is retrieved from the VisTrails server and computations are performed on the server. Due to the use of caching mechanisms the compilation of the LaTeX file does not take more time than usual. Clicking on the figure retrieves the workflow from the server, allowing the reader to inspect details of the workflow and its provenance, and even to rerun the workflow.

Publishing the paper: For submission of the final publication we need to create a self-contained bundle with the paper itself, corresponding workflows, and, if there is no other permanent public archive, all data. The main challenge is updating the links from figures to provenance information and data. For this purpose the permanent location of the files should be known, which seems like a catch-22 situation since in standard publishing procedure, paper identifiers become available only at the last stage of publication. In the following, we discuss how this restriction can be overcome in some situations, and what the open problems are.

An arXiv preprint: A first example is publishing on the arXiv preprint server. Here, permanent URLs to so-called *ancillary* files are available, but the paper identifier, and hence the exact URL, becomes available to the author only after the paper has been published. Since authors can replace their published arXiv preprints by uploading a new version keeping the same arXiv identifier the original manuscript can be replaced by a second version whose links point to the correct location on the arXiv server. While this approach is practicable yet inconvenient for the authors, the situation is even more difficult for publications in peer reviewed journals.

A journal publication: Publishing in a journal does not allow replacements of the paper or workflows after publication. After all, the published journal paper is supposed to be immutable and setting correct links currently requires a cumbersome manual intervention by the publisher leading to publication complications and delays.

An important issue is the stability of the links. While some publishers, such as the Institute of Physics, guarantee stable URLs for papers and auxiliary files, others, like the American Physical Society, prefer to use Digital Object Identifiers (DOIs) for this purpose. Tools to link to auxiliary files relative to a base DOI of the paper are currently lacking, as is the support for accessing data via DOIs or relative to resolved URLs corresponding to a DOI in all workflow systems known to us.

To conclude, publishing a paper first on a preprint server and then in a journal requires us to go through several stages in each of which we have to manually update not only the links from the main document to all workflows but also the locations of the raw data files. At the time of writing this article, we are working with the Institute of Physics to finalize a first publication that should follow these guidelines, and are in discussion with the editors at the American Physical Society on how to publish a manuscript in their journal.

4 Summary and Acknowledgements

In this experience report we have demonstrated how to publish provenance-rich papers and presented an overview over the complete scientific process from the preparation of input parameters for large-scale simulations to a published paper in a peer-reviewed journal. We pointed out the need for improvements in processes and tools and outlined how some of these challenges can be addressed by the developers and publishers.

We thank the VisTrails team, and especially D. Koop, P. Mates, E. Santos, J. Freire and C. Silva for continuing help and development of VisTrails. We also thank the staff of the Institute of Physics and Physical Review for taking on the challenge and agreeing to help with publishing our provenance-rich papers. We especially acknowledge discussions with T. Smith and J.W. Taylor. This work was supported by an ETH-internal grant, the Swiss HP2C initiative and a grant from the Army Research Office with funding from the DARPA OLE program.

References

[1] http://www.vistrails.org/.

[2] http://www.executablepapers.com/.

[3] BAUER, B., ET AL. The ALPS project release 2.0: Open source software for strongly correlated systems. *Accepted in Journal of Statistical Mechanics: Theory and Experiment* (2011). arXiv:1101.2646.

[4] FREEDMAN, M. H., GUKELBERGER, J., TREBST, S., TROYER, M., AND WANG, Z. Galois Conjugates of Topological Phases. In preparation.

[5] KOOP, D., ET AL. A Provenance-Based Infrastructure for Creating Executable Papers. *Journal of Computational Science*. Submitted.

[6] KOOP, D., ET AL. Bridging workflow and data provenance using strong links. In *Scientific and Statistical Database Management*, M. Gertz and B. Ludäscher, Eds., vol. 6187 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, 2010, pp. 397–415.