Reproducibility by Other Means: Transparent Research Objects with Science-Oriented Provenance

Timothy McPhillips Bertram Ludäscher
School of Information Sciences, University of Illinois at Urbana-Champaign
{tmcphill,ludaesch}@illinois.edu

1 INTRODUCTION

Publicly funded efforts around the world currently are underway to ensure that computational components of scientific research can be made "reproducible" or "replicable". The ongoing discourse about the perceived "reproducibility crisis in science" is just one illustration of the importance of these efforts. The energy invested in the wide-ranging debate over the precise meanings of the terms *reproducible*, *replicable*, *transparent*, etc, with respect to research results, processes, and settings is perhaps an even greater indication of both the significance of these efforts and the challenges they face. For while each effort aimed at facilitating reproducible computing in the sciences must clearly define its mission and apply the bulk of its resources to the specific problems it sets out to address, these efforts necessarily do so within the context of broader discussions about the nature, importance, and precise definitions of the qualities of science we wish to extend to computing over the longer time scale.

Within a particular effort it is useful to define terms such as *reproducible* operationally. For example, in the Whole Tale project we define a *Reproducible Tale* as one that *includes sufficient information for the Tale to be re-executed for the review and verification of results.* Adopting this definition allow us to focus our requirements analysis, system design, and software implementation efforts on the specific problems Whole Tale is funded to solve and the use cases we aim to support. Supporting publishers who request authors to include all new data, code, and workflows needed to reproduce computed artifacts supporting claims in a paper is one such use case targeted by Whole Tale. We anticipate that facilitating re-execution of code used to generate key products of a study will enable publishers to routinely confirm that provided data and code do in fact produce those results—thereby addressing a key dimension of the reproducibility challenge currently facing science.

At the same time, it is critical that efforts like Whole Tale contribute to a global vision of computational reproducibility in the sciences, and clearly situate its particular mission, use cases, and engineering deliverables in this context. For while the particular technical problems that Whole Tale and similar projects aim to address are particularly pressing, current efforts by no means represent the entire landscape of concepts, problems, and technical options that will require further discussion, clarification, and analysis if we are to meet the challenges of reproducibility now facing the sciences. In particular, current engineering efforts are unlikely to elevate the computational components of research to the level of reproducibility historically expected of studies in the pure natural sciences such as physics, chemistry and biology.

Consequently, we view the Whole Tale project—as currently chartered and funded—as just a step towards the kind of platforms, infrastructure, and standards needed to enable researchers using computing technology to routinely achieve the reproducibility long considered the essence of science as a whole. In support of this longer-term vision, we outline in this paper just a few of the issues we aim to investigate and discuss with the broader community over the next few years. We anticipate that future iterations of Whole Tale and its sibling efforts will be driven in part by the problem definitions and solution proposals we collectively develop between now and then.

2 OUTLINE OF PAPER

In the remainder of this paper we briefly discuss four topics we plan to investigate in the course of the Whole Tale project. In Section 3 we review the general notion of reproducibility in science, and in Section 4 highlight how digital computing in principle makes possible a completely new kind of reproducibility: *exact repeatability*. We emphasize that the notion of *transparency*—long a critical element of reproducibillity in the pure natural sciences—has a role to play even for those computational components of research where exact repeatability is feasible. In Section 5 we provide an overview of several dimensions of the terminological debate around reproducibility generally, and propose that a pluralistic approach to defining key terms is essential if a general concept of reproducibility is to be shared across disciplines. In Section 6 we summarize a number of limitations on exact repeatability in practice, and in Section 7 show how science-oriented provenance queries can mitigate such limitations by maintaining the transparency most essential to reproducibility in science. Throughout, we highlight the role that Research Objects can serve in supporting and maintaining reproducibility by encapsulating the information needed to rerun the computational steps in a study, and by enabling transparency via queries of provenance information packaged in the object.

3 REPRODUCIBILITY IN SCIENCE

Modern science is founded on the expectation that the observations, experiments, and predictions that comprise scientific research be independently verifiable by others. This requirement, referred to as the *reproducibility* or *replicability* of science, applies not only to the products of research studies (*substances*, *results*, *conclusions*, *models*, *data products*, *predictions*), but also to the activities that ultimately give rise to these products (*methods*, *protocols*, *workflows*); the materials employed in these activities (*reagents*, *instruments*, *software*); and the conditions under which which these activities are carried out (*temperatures*, *instrument settings*, *software parameters*, *computing environments*). When sufficient details are available such that the research products and methods can be reviewed, interpreted, and evaluated by other researchers *without* repeating the work, a study is said to be *transparent.*

While it is true that studies attempting primarily to reproduce previous results are relatively rare in the pure natural sciences, even the most groundbreaking studies in these fields include components that explicitly or implicitly confirm the reproducibility of previously reported results and procedures. The expectation is that new studies will reliably produce meaningful results consistent with previous work only if the prior work on which they are based or otherwise relates to is reproducible. In this sense, the whole of basic research in the natural sciences can be seen as an ongoing, massively-parallel reproducibility study that also happens to produce a steady stream of new results. Exceptions to this pattern occur when studies appear to overturn well-established understandings of nature, violate the expectations of how research in a particular field is to be carried out, or otherwise cause controversy. In these cases direct attempts may be made to reproduce results by duplicating as carefully as possible the reported methods and conditions described in the controversial study.

Even when attempts are made specifically to confirm the reproducibility of particular studies or results, investigators in the natural sciences generally do not expect the processes and products of research to be duplicated exactly. The vast majority of quantitative observations made of real world phenomena using scientific instruments are associated with limited precision and other intrinsic uncertainties that must themselves be characterized and well understood for science based on them to be considered reproducible. It is a hallmark of trustworthy science that quantitative observations and claims are inseparable from these uncertainties in measurement and their propagation through data analysis.

Similarly, the materials and processes employed in the natural sciences generally are impossible to duplicate exactly. In a chemistry laboratory, the precise quantities of input reagents will vary, temperatures will differ, and heating or cooling rates will be unique for each run of a chemical synthesis, no matter how carefully these conditions are controlled; the yield and purity of the intended product necessarily will vary as well from run to run. A similar situation holds when measurements are made on samples using a scientific instrument. Different instruments of the same model will vary slightly and produce slightly different results even under identical conditions on identical samples. Generally, the original researchers are in the best position to assess how the minimum variation expected between runs of a synthesis (they have access to the same batch of reagents and the same equipment), or between repeated readings of an instrument on the same or equivalent samples (they can prepare multiple samples at the same time, and run these samples through the instrument one after the other). A researcher attempting to duplicate another's work can expect to see greater deviation from the reported results because the materials and conditions involved will necessarily differ to a greater degree.

This asymmetry between the original researcher and another repeating the work is reflected in the longstanding distinction between *reproducibility* and *replicability* in experimental biology. In Section 4 we will examine definitions of these terms jointly adopted by twenty-nine research societies in the biological sciences. For now we note that the notion of *replicates*, repeated measurements made to quantify experimental variability, is represented by a rich literature. This literature distinguishes between distinct modes of experimental replication. The term *technical replicates*, for example, refers to repeated measurements performed on the same sample. These are used to assess the variation intrinsic to the procedure, apparatus, and instrument employed. *Biological replicates* represent measurements made on different but equivalent samples. In practice both generally are performed by the original researcher under conditions otherwise held as constant as possible.

FOOTNOTE: Generating multiple gigabytes of raw data requiring intensive computational analysis for each replicate, Next-Generation Sequencing (NGS) represent just one sub-domain where the reproducibility terminologies in the natural sciences and in computing unavoidably collide.

4 COMPUTATIONAL REPEATABILITY

In contrast to expectations in the experimental natural sciences„ digital computing make it possible to repeat *exactly* certain computational aspects of research, even by *different* researchers using *different* computers. Indeed, it generally is expected that computational processes, the implementation of hardware and software enabling those processes, and the outputs of those processes all can be repeated exactly by others–at least in principle. This potential of exact repeatability is unquestionably of enormous value to any field of research employing computers, and certainly will contribute to the ability of researchers in every field to reproduce or build on others' work. At the same time, there is at least some risk of this new expectation of exact repeatability being conflated (consciously or unconsciously) with the longstanding understanding of reproducibility in the basic sciences. It is essential that the new concept be kept distinct.

Moreover, while computational experiments and analyses may be exactly repeatable in principle, in practice the complexities of real-world hardware and software currently make computational repeatability challenging to achieve in practice except over very limited time scales. Because of the obvious value that exact repeatability brings when it is feasible, it is important that we work to expand the fraction of scenarios in which the computational components of research can be automatically repeated exactly over ranges of time and space relevant to scientific research and discourse. These efforts are particularly important for the research community to pursue, and for science funding agencies to support, because the computing industry generally does not have requirements for exact repeatability across significant spans of time.

However, we emphasize that the concept of exact repeatability is qualitatively different from the concept of reproducibility that underlies the natural sciences. In particular, scientific reproducibility is not simply a weaker form of computational repeatability. *Approximating or achieving computational repeatability does not automatically deliver scientific reproducibility.*

It is in a sense both bad and good news that exact computational repeatability is not tantamount to scientific reproducibility. The disappointing news, perhaps, is that it is possible to put much effort into achieving computational repeatability, exact where practical and inexact otherwise, without delivering the kind of reproducibility that is critical for producing trustworthy science. The good news is that scientifically meaningful reproducibility can be realized in cases (or over spans of time) where computational repeatability

is impractical due to the limitations of available technology or affordable resources. Thus, the older concept of reproducibility that permeates the basic natural sciences has a very useful role even where digital computing makes exact repeatability a theoretical possibility.

Researchers in the natural sciences are comfortable with the idea that it is not possible to exactly repeat all reported observations, procedures, and experimental results. They do not see this as a contradiction to their demand that science be reproducible. What the natural sciences actually do demand is that (a) research procedures be repeatable by others in principle; (b) the means of repeating the work be subject to review and evaluation; and (c) such review and evaluation be possible *without* actually repeating the work. To be perfectly clear about the third demand: in the natural sciences it is actually considered a *problem* if exact repetition of the steps taken in reported research is required either to evaluate the work or to reproduce results.

Consequently, it is not necessary to achieve or maintain perfect repeatability of the computational components of research for scientists to consider a study reproducible and therefore trustworthy. At the same time it is important that the standards, technologies, computational best-practices, and infrastructure we develop and advocate in fact support scientific reproducibility. It is not enough, in the long run, to pursue and support exact computational repeatability where we can, and to get as close as possible otherwise. Rather, computational repeatability is best seen as a dimension of research reproducibility *orthogonal* to the dimension of transparency. It is possible to achieve computational repeatability without providing research transparency—and vice versa. Moreover, exact repeatability is not an essential element of scientific reproducibility in the broadest sense of the term. Transparency arguably is.

5 TERMINOLOGY

What are some specific ways that Research Objects can help make scientific research more transparent? Many of the objectives and current capabilities of Research Objects already can be seen as supporting transparency. Examples of Research Objects support research reproducibility by enhancing transparency include...

In the remainder of this paper we propose that Research Objects can help in additional ways that not just enhance the transparency of research, but also ensure that transparency and other key elements of scientific reproducibility can be achieved, described, and shared meaningfully for all domains of research—including those that include both an experimental and computational elements.

The first way in which Research Objects can help is by helping researchers safely navigate the terminological quagmire surrounding the definitions of terms such as *reproducible*, *replicable*, and *transparency*. A very simple yet important use case for Research Objects could be the declaration of the senses in which the research study and results associated with the Object are in fact reproducible, replicable, computationally repeatable, and so on. Before extending or depending on others' works, methods, or results in their own studies, researchers reasonably want to know if that previous work is reproducible in various senses of the word. Research Objects can help, not just be providing a place to make such declarations, but by preventing misunderstandings of what is meant by particular terms.

The current debate over the meaning of key terms describing scientific reproducibility are motivated primarily by a desire to avoid just such confusion. The recommendations from the Federation of American Societies for Experimental Biology (FASEB) cite "lack of uniform definitions to describe the problem" as one of the top three factors that "impede the ability to reproduce experimental results." The report National Academy of Sciences Committee on Reproducibility and Replicability of Science states that "the difficulties in assessing reproducibility and replicability are complicated by this absence of standard definitions for these terms."

The recommendations of these two organizations are representatives of numerous recent studies, papers, and proposed definitions intended to enhance reproducibility by providing a uniform terminology for describing it. The FASEB recommendations originate in one domain of science while the NAS definitions explicitly "are intended to apply across all fields of science." Given the interdisciplinary character of modern research—and in particular the ubiquity of computing in science—it is hard to argue against attempts to facilitate communication about reproducibility across science as a whole.

What can be surprising to researchers new to this debate is how many ways the proposed definitions can differ. First, there is disagreement over which term, *reproducibility* or *replicability*, indicates a greater adherence to the procedures, material, and methods employed in the original research. The FASEB definitions (in accordance with the terminology around *replicates* described in section 4) require from *replicability* a greater fidelity to the original study:

Replicability: the ability to duplicate (i.e., repeat) a prior result using the same source materials and methodologies. This term should only be used when referring to repeating the results of a specific experiment rather than an entire study.

Reproducibility: the ability to achieve similar or nearly identical results using comparable materials and methodologies. This term may be used when specific findings from a study are obtained by an independent group of researchers

According to FASEB, *replicability* indicates a higher degree of fidelity than does *reproducibility*, both with respect to the prior result to be confirmed, and to the materials and methodologies employed. Replicability also appears likely more feasible for the original researchers (they presumably have access to the "same source materials" and are in the best position to use the same methodologies), whereas reproducibility is feasible for "an independent group of researchers". Both definitions may be applied to experimental results, but neither definition precludes application to in silico experiments or to the computational elements of laboratory studies.

In contrast, the definitions in the recent report from the National Academy of Sciences reverses the relative fidelity implied by the terms 'reproducibility' and 'replicability':

Reproducibility is obtaining consistent results using the same input data, computational steps, methods, and code, and conditions of analysis.

Replicability is obtaining consistent results across studies aimed at answering the same scientific question, each of which has obtained its own data.

The NAS definition of replicability is most similar to the FASEB definition of reproducibility. This reversal of the meanings of these terms between various research domains is well documented. This aspect of the disagreement over terminologies is in a sense trivial, although the NAS rightly states that the "different meanings and uses across science and engineering" has "led to confusion in collectively understanding problems in reproducibility and replicability." It is interesting that the NAS report does not suggest new terms for referring to the *technical replicates* and *biological replicates* so important in experimental biology, should biologists adopt the recommendation of restricting *replication* to "obtaining consistent results across studies". FOOTNOTE: The NAS report section "Precision of Measurement" quotes a portion of the International Vocabulary of Metrology that twice employs the term *replicate measurement*. What might come as news to biologists is the assertion that the NAS "committee adopted specific definitions" of reproducibility and replicability, "which are otherwise interchangeable in everyday discourse." Not only is the high-fidelity *replication* of DNA (in the *replisome*) and the lower fidelity *reproduction* of organisms matters for everyday discourse for the many biologists study these processes in nature or employ them in the lab, it is easy to see an analogy between replication of DNA and careful replication of measurements and samples in the lab on the one hand, and on the other the reproduction of organisms where variation is encouraged in nature (for example through sex) and the reproduction of scientific results across studies where, again, some variation is both expected and desirable.

A far more intriguing aspect of the NAS definitions is that experiments not carried out entirely in silico apparently are left with only the term *replicability*. Satisfying the definition of reproducibility requires "computational steps" and "code", and the report goes on to clarify that reproducibility "is synonymous with computational reproducibility," and "the terms are used interchangeably in this report." Indeed the executive summary of the report states not only that "We define reproducibility to mean computational reproducibility" but also that "the committee adopted definitions that are intended to apply across all fields of science." The clear implication is the term *reproducible* only can be applied only to the computational components of research. Because this term is analogous to *replicable* as defined by FASEB, the NAS definitions do not provide a vocabulary that would enable experimentalists to report the intrinsic repeatability of their own methods, measurements, and results.

Intriguing similarities and differences also appear in definitions of transparency. According to FASEB, transparency is

The reporting of experimental materials and methods in a manner that provides enough information for others to independently assess and/or reproduce experimental findings

While the NAS report states:

When a researcher transparently reports a study and makes available the underlying digital artifacts, such as data and code, the results should be computationally reproducible.

According to the latter definition, transparency, like reproducibility, requires digital artifacts which could be of concern to those expecting experimental procedures to be transparent. However both definitions imply that transparency is a necessary *component* of reproducibility, a position that suggests a role for Research Objects to play in the resolving this terminological conundrum.

In short, we propose that users of Research Objects be provided with a vocabulary for asserting and querying the reproducibility of studies, results, and methods along multiple dimensions. Namespaces would support multiple definitions of terms without conflict. Synonym relationships and other mappings between the vocabularies would enable reasoning about reproducibility and support assertions and queries phrased using terminologies selected by the user. For example, a researcher publishing a research object might assert that the study is reproducible *sensu* Whole Tale. Another researcher filtering discovered Research Objects by the property NAS::reproducible would find this study either if WT::reproducible had been found to imply NAS::reproducible generally, or if other assertions made by the author about the Object satisfy the requirements of the latter term in conjunction with the implications of WT::reproducible.

Going forward, the Whole Tale project aims to explore the various terminologies for reproducibility with the goal of identifying what might be considered to be the principle components of reproducibility in science as a whole. As a community we could then determine how various terms and definitions can be seen as compositions of these components. This in turn would reveal how Research Object infrastructure should reason about these terms, and how claims made in terms of one set of definitions could be converted to claims using another set of definitions.

6 COMPUTATIONAL CHALLENGES

The fundamental limitations computers impose on replicability of program executions are well known.

Finite precision arithmetic, different word sizes on different processors, round-off errors, etc, impose limits on scientific computations and their replicability across different computing environments. Virtual machines and containers do not address these issues. Full emulation is required to run the same binary in identical fashion on a different processor. This is typically slow. These limitations are even more challenging to manage reproducibly because programs typically are compiled, meaning that the exact sequence of machine instructions executed even by a single processor cannot generally be controlled. A different compiler, or a newer version of the same compiler will yield a different sequence of machine instructions.

Replicating the outputs of a program is far from straightforward

Observing that a program or set of programs can be executed again is not sufficient to conclude that the underlying computation was replicated. The outputs of the programs must be checked for equivalence. Because of the expected variation in run time behavior of programs due to the issues above, checking that outputs of a program run are equivalent to the outputs of a run of that program is not always as simple as comparing the outputs for bitwise identity. Robustly checking for equivalence of output generally must be confirmed in some way other than comparing files at the bit level. Footnote: The excellent practice of including accurate provenance and other meaningful metadata in data file headers makes it even more unlikley that outputs from different runs will be bitwise-identical.

Replicating just the software running the program is challenging in practice

How can we ensure that the stream of instructions sent to the processor for two executions is identical? Even holding the computer hardware and compiler version constant, programs depend on language libraries, OS libraries, and system calls. Much scientific software also depends directly and indirectly on large numbers of 3rd-party libraries. Only direct dependencies can be controlled reliably at build-time. And many dependencies are via shared libraries that can change between executions of the exact same executable–no recompile is needed to get a new effective executable. Footnote: Fans of the Go programming language are bringing back the static executable. Recompiling or even just rerunning the "same" program a week later can result in a completely different effective instruction stream.

Even reproducing computing environments is hard

Containers and their discontents Footnote: By 'discontents' we do not mean that we object to the use of containers, but rather than we are not content with container technology alone. There currently is much enthusiasm around containers as a means of reproducing computing environments and making computational science replicable. Whole Tale is one of several projects leveraging the capabilities of containers for this purpose. Others include Binder and Code Ocean. In Whole Tale it is recognized that containers alone cannot satisfy researcher's needs for sharing their computing environments and computations. Rather, container technology such as Docker provide an invaluable tool for the reproducible science software stack architect. A major motivation for funding (and continuing to fund) projects like Whole Tale is that the containers on their own are insufficient as means to making computational science reproducible, and it is not practical for individual researchers and groups to use containers and other technology to actually achieve scientifically meaningful reproducibility over periods longer than the publication-cycle time scale.

What containers do not do Despite what sound like suggestions to the contrary in the literature, container technology such as Docker do not ensure computational replicability, and do not on their own solve any of the problems of computational replicability described above. What containers do provide a very convenient means for executing customized computing environments on behalf of researchers, withot having to run an entire virtual machine for each environment. In common with virtual machine technology, containers do not abstract or hide the underlying hardware architecture of the computer on which they run. They do not abstract the underlying operating system, but simply use the Linux kernel on the host. Kernel parameter settings on the host apply to all containers running on the host (reference famous blog post on the topic "Containers Do Not Contain"). Rebuilding an image from its Dockerfile specification is not guaranteed to yield the same image. It general it will not. Container images, once they are built, are not guaranteed to run on future releases of the container host. They also do not ensure that computations run within the container will replicable in the feature.

What containers are for What containers are good for is precisely what they were to do for the computing industry: enable developers to write and test code in a computing environment of a developer's choosing that can then be replicated on a very short time scale (hours or days) in staging and production environments. Containers also are good at managing conflicts in dependencies between different components of a multiprocess software architecture. Using containers to 'contain' dependencies in this way is most effective when an application can be split across multiple containers running in concert. The model of one container, one computing environment does not lend itself to dependency isolation.

The problem of time and dependencies An emerging threat to reproducibility of computational science is the spread misconception that sharing the definition of a container image, e.g. by including a Dockerfile in the Git repo for the project, is a guarantee that others (or even the original researcher) will be able to recreate the corresponding image and computing environment it represents. Researchers making this assumption may be less likely to preserve all of the information actually required to reproduce their computations. A major reason a Dockerfile is not enough is that the implicit dependencies of the built environment are constantly changing. This is well known to anyone working directly with Docker, or other container technologies. Here we will give a single example of the implications of this issue for reproducible science.

What is reproducibility really for?

Achieving meaningful replicability even for the computational parts of research is very challenging. But this is no reason to give up hope. Replicability is a means to an end–justification of scientific results–and Research Objects can help us achieve that end by other means. What is most exciting about Research Objects is that they can achieve this end despite the difficulty of computational replicability. And the primary means by which Research Objects can due this is by providing transparency.

7 TRANSPARENCY

In this paper we argue that the dimension of reproducibility most ripe for the contributions of computer science is research *transparency*, in particular through the modeling, recording, and querying of the provenance of research artifacts. In alignment with researchers in the natural sciences who recognize transparency as crucial, we are confident that provenance management has much to contribute to scientific reproducibility, even when it does not specifically enable exact repeatability of the computations they describe.

For provenance management systems, representations, and user interfaces to support reproducibility via transparency, however, they must support science-oriented queries. Provenance must be able to answer questions about the *science* that was performed–not just the sequence, dependencies, and flow of data through computational steps. The answers to these questions must enable others to evaluate the scientific quality of the work, and to learn what is necessary to reproduce the results without actually repeating every step taken in the original work. Provenance must enable researchers to build on the results and processes reported in prior work with confidence.

Finally, it must be possible for researchers unversed in the detailed specifications of Research Objects and the PROV standard to pose questions and receive answers meaningful for evaluating, using, and building on the processes and products of prior research. We suggest that Research Objects and related approaches are the

ideal vehicle for storing, sharing and making provenance queryable in this way. Research Objects thus can support scientific reproducibility even in the face of the many practical challenges to computational repeatability.

REFERENCES