Reproducibility by Other Means: Transparent Research Objects

Timothy M. McPhillips*, Craig Willis*†, Michael R. Gryk*, Santiago Nuñez-Corrales‡, and Bertram Ludäscher*†

*School of Information Sciences, University of Illinois at Urbana-Champaign, Champaign IL
†National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign, Urbana IL
‡Illinois Informatics, University of Illinois at Urbana-Champaign, Champaign IL

Email: {tmcphill,willis8,gryk2,nunezco2,ludaesch}@illinois.edu

***Abstract*—Research Objects have the potential to significantly enhance the reproducibility of scientific research. One important way Research Objects can do this is by encapsulating the means for re-executing the computational components of studies, thus supporting the new form of reproducibility enabled by digital computing—exact repeatability. However, Research Objects also can make scientific research more reproducible by supporting transparency, a component of reproducibility orthogonal to re-executability. We describe here our vision for making Research Objects more transparent by providing means for disambiguating claims about reproducibility generally, and computational repeatability specifically. We show how support for science-oriented queries can enable researchers to assess the reproducibility of Research Objects and the individual methods and results they encapsulate.**

I. Introduction

Publicly funded efforts around the world currently are underway to ensure that computational components of scientific research can be made "reproducible" or "replicable" [12], [13], [20], [32]–[34], [41], [45], [46], [50]. The ongoing discourse about the perceived "reproducibility crisis in science" [8], [18], [21] 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. [9], [17], [19], [23], [25], [26], [29], [31], [43], 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. So 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 [12], [52] a *Reproducible Tale* is one that *includes sufficient information for the Tale to be re-executed for the review and verification of results*. Adopting this definition focuses requirements analysis, system design, and software implementation efforts on the specific problems Whole Tale is funded to solve and the use cases it aims 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. Facilitating re-execution of code used to generate key products of a study will enable publishers routinely to 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 their particular missions, use cases, and engineering deliverables in this context. For while the particular technical problems these 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. 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, the current generation of efforts to build reproducility platforms represents 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 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.

In the remainder of this paper we briefly discuss five topics we plan to investigate as part of this research agenda. In Section II we review the general notion of reproducibility in science, and in Section III 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 reproducibility in the pure natural sciences—has a role to play even for those computational components of research where exact repeatabil-

ity is feasible. In Section IV 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 V we summarize a number of limitations on exact repeatability in practice, and in Section VI 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 [10] can serve in supporting and maintaining reproducibility by encapsulating the information needed to rerun the computational steps in a study, by disambiguating claims about reproducibility, and by enabling transparency via queries of provenance information packaged in the object.

II. 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 these activities are carried out (*temperatures*, *instrument settings*, *software parameters*, *computing environments*) [17], [26]. 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* [17], [22].

While it is true that studies attempting primarily to reproduce previous results are relatively rare in the pure natural sciences [16], 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 relate 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 [35], violate the expectations of how research in a particular field is to be carried out, or otherwise cause controversy [16]. 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 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 can 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 III we will examine definitions of these terms jointly adopted by thirty 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 [48], [51]. 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.[1]

III. Computational Repeatability

In contrast to expectations in the experimental natural sciences, digital computing makes it possible to repeat *exactly* certain computational aspects of research, even by *different*

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

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 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

1) research procedures be repeatable by others in principle;
2) the means of repeating the work be subject to review and evaluation; and
3) 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 [40].

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*[2] 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.

IV. Terminology

What are some specific ways that Research Objects [10] can help make scientific research more transparent? Many of the objectives and current capabilities of Research Objects already can be seen as supporting transparency [39]. 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 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 *transparent*. A very simple yet important use case for Research Objects (ROs) could be the declaration of the senses in which the research study and results associated with the RO 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. ROs can help, not just be providing a place to make such declarations, but by preventing misunderstandings of what is meant by particular terms.

[2]In the geometric (not the statistical) sense of the word. Most scripts associated with a study, for example, likely contribute to both transparency and repeatability. But not infrequently a particular component of a study contributes only to one or the other of these dimensions. For example, an invocation of a web service that operates as a black box can be considered repeatable but not transparent.

The current debate over the meaning of key terms describing scientific reproducibility is motivated primarily by a desire to avoid just such confusion [9], [17], [19], [26], [29], [43], [47]. The recommendations from the Federation of American Societies for Experimental Biology[3] (FASEB) [22] 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 recent report from the National Academy of Sciences (NAS) Committee on Reproducibility and Replicability of Science [17] asserts that "the difficulties in assessing reproducibility and replicability are complicated by this absence of standard definitions for these terms."

The recommendations from these two organizations are representative 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[4] require from *replicability* a greater fidelity to the original study [22, p.3]:

> **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.

[3]FASEB is a federation of thirty distinct scientific societies representing over 130,000 researchers in the biological sciences [2].

[4]In accordance with the terminology around *replicates* described in Section II.

In contrast, the definitions in the report from the National Academy of Sciences reverses the relative fidelity implied by the terms 'reproducibility' and 'replicability' [17, p.4]:

> **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*. The reversal of the meanings of these terms between various research domains is well documented within the NAS report, which in turn depends on Barba's comprehensive study of the terms [9].

This aspect of the disagreement over terminologies is in a sense trivial[5], although the NAS likely is correct in asserting that the "different meanings and uses across science and engineering" has "led to confusion in collectively understanding problems in reproducibility and replicability." Far more notably, 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".[6]

An even more intriguing aspect of the NAS definitions [17] is that experiments not carried out entirely *in silico* apparently are left only with the term *replicability*. Satisfying the NAS definition of *reproducibility* requires "computational steps" and "code". 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 that the term *reproducible* only can be applied 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 non-computational methods, measurements, and results.

Note that by highlighting these aspects of the NAS definitions of reproducibility and replicability we are *not* arguing that the FASEB definitions of these terms are superior. In

[5]The NAS report points out that the words *reproducibility* and *replicability* are "interchangeable in everyday discourse." We note, however, that both the high-fidelity *replication* of DNA (in the *replisome* [49]) and the lower-fidelity *reproduction* of organisms *are* matters of everyday discourse for researchers who study these processes in nature or employ them in the lab. Furthermore, we observe a clear analogy between the exacting replication of DNA and careful replication of measurements and samples in the lab on the one hand; and on the other hand between 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.

[6]The NAS report section *Precision of Measurement* quotes a portion of the International Vocabulary of Metrology that twice employs the term *replicate measurement*.

particular, we do not propose that the latter definitions be adopted universally instead. On the contrary, we suggest that the differences in the content of these (and the many other) definitions of these two terms likely reflect specific, critical needs of the researchers and communities that adopt them.

Similarities and differences also appear in definitions and usages of the term *transparency*. According to FASEB [22], **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 [17] 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.

Again, the NAS usage of the term assumes that transparency is associated with digital artifacts. This could be of concern to those expecting experimental procedures to be transparent as well.

What is more important, however, is what the two concepts of transparency have in common. Both definitions imply that transparency is a desirable *component* of scientific reproducibility in the broader sense of the term. This shared insight suggests a role for Research Objects to play in the resolution of this terminological conundrum. In short, we propose that (1) ROs *provide users with vocabularies for asserting and querying the reproducibility of studies, results, and methods along multiple dimensions*; and (2) that we *enable users interacting with ROs to employ terminologies intuitive to researchers in their own communities*. 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 an RO might assert that the study is reproducible *sensu* Whole Tale. Another researcher filtering discovered ROs 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 RO satisfy the requirements of the latter term in conjunction with the implications of WT::reproducible.

The same approach could be applied to ROs published by communities adopting specific procedures for evaluating and verifying computational artifacts. For example, the American Journal of Political Science (AJPS) has adopted a policy of verifying quantitative research [14]. ACM SIGMOD has defined procedures for assessing database research reproducibility [5], [11]. ROs deemed reproducible through these community-defined workflows could be identified as AJPS::reproducible:v1.2[7] and SIGMOD::reproducible, respectively. The ACM additionally awards four different reproducibility badges [7] to publications with artifacts that meet particular criteria[8]. Given that a particular researcher is likely to depend on or extend results that have been assessed for reproducibility by different workflows (or different versions of those workflows), the ability to query the reproducibility of research products using the definitions and criteria of one's choosing is critical if these reproducibility assessment efforts are to have lasting value.

Capabilities for reasoning about definitions, verification workflows, and awarded badges in this way also would benefit emerging reproducibility platforms that target research with non-computational components. Biomolecular nuclear magnetic resonance (bioNMR) spectroscopy is computationally intensive, with studies requiring dozens of software tools for data processing and analysis. The NMRbox project [36] aims to foster reproducibility of such studies by providing researchers with access to Xubuntu virtual machines provisioned with more than 100 software tools used by this community. Each VM image is versioned, stored in its entirety, and made available long-term to make the software environments, computations, and end-to-end workflows employed in a particular study computationally reproducible. NMRbox uses PREMIS [3] to represent and store provenance metadata, and packages this metadata with processed data and results to support transparency both of the analysis and of the NMR experiment itself [28]. Queries probing the reproducibility of a particular result produced, archived, and shared using NMRbox necessarily would span claims made both about the (non-computational) experimental procedures employed, *and* the computational analysis performed on the resulting experimental data.

Going forward, we aim to explore the various terminologies surrounding reproducibility with the goal of identifying what might be considered the principal components or perhaps even the *dimensions* of scientific reproducibility[9]. We could then determine how various terms and definitions, put forward to meet the needs of particular research communities, can be seen as compositions of these shared components. This in turn would reveal how RO 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.

V. Maintaining repeatability

Another way Research Objects can contribute to scientific transparency is by clarifying claims about computational repeatability. Just as the overall scientific reproducibility of a study represented by a Research Object might be described precisely in terms of individual components required to satisfy

[7]The AJPS workflow is versioned. See for example https://ajps.org/wp-content/uploads/2019/01/ajps-quant-data-checklist-ver-1-2.pdf

[8]The current guidelines for awarding badges provide links to seven past versions of the guidelines that have been used since 2015.

[9]The idea that *scientific reproducibility* in general might comprise multiple, independent dimensions is analogous to the claim made by the PRIMAD model [47] that the information gained from a *reproducibility study* is related to—and meaningful only in terms of—the ways in which it varies from the original study along different dimensions: platform, research objective, implementation, etc.

particular (namespaced) definitions of *reproducible* or *replicable*, additional statements could be made about the various dimensions of computational reproducibility in particular.

There have long been discussions about packaging ROs for transparency and reproducibility [15], [24], [42], but less emphasis has been placed on *describing* characteristics of the published ROs. The aim would be for researchers publishing their work via ROs to be fully aware of the implications of the claims they make about the computations represented by the RO. Researchers discovering, evaluating, or using the Research Object for further research would be able to interpret these claims unambiguously. The implications for the possibilities of rerunning, reproducing, or exactly repeating the computations described in the RO under different conditions would be clear to all parties.

As discussed above, *exact repeatability* promises increasingly to be a powerful, new addition to the modern researcher's repertoire of reproducibility techniques. At the same time, there appears to be (possibly growing) confusion over what is actually possible in terms of computational reproducibility generally, exact repeatability specifically, and the conditions required to achieve them in practice.

The fundamental limitations computers impose on the exact replicability of program executions are well known. At the lowest level, finite precision arithmetic, differing word sizes between processors, the effects of round-off errors, and the implications of choosing between different mathematically equivalent orderings of operations all have the potential to impose limits on the replicability of scientific computations across different computing environments. Virtual machines and software containers cannot fully address issues at this level.

The fact that we can expect options such as full-processor emulation, either in software [44] or on customizable hardware, to provide better guarantees of exact computational repeatability under more circumstances over time reveals the true crux of the problem. What we can expect from computers in terms of reproducibility in general, and exact repeatability in particular, is changing quickly—and likely will continue to do so for the foreseeable future. In the case of hiding hardware differences, time is on our side—or can be if we happen to save the information actually needed to enable exact re-execution of our analyses in the future.

In many other cases, time works against repeatability [30]. A Dockerfile that today correctly produces the software environment in which computations were originally performed may not do so a year from now–if it builds at all. Due to the dependencies of most scientific software on packages not bundled with the language compiler or runtime (with these packages typically depending on other packages, and so on), the chances of rebuilding or rerunning software equivalent to that used to produce a result in a Research Object decreases rapidly with time. Fortunately, time also works for us in this dimension as well, as new ways of specifying software environments and archiving dependencies emerge. But again the issue arises—are we saving the right information to enable computational repeatability in the future?

What Research Objects can offer here is analogous to the proposed function of mediating between competing and contradictory definitions of reproducibility and replicability. Rather than trying to anticipate all future developments in the area of computational reproducibility, and representing computing environments, software dependencies, and machine information in a way that we hope will be usable by future technologies, we can take the pluralistic path here as well. We can characterize the various dimensions in which computing technology currently supports—or fails to support—exact repeatability; then create mappings from the specific capabilities of existing technologies (Docker [1], Singularity [6], Jupyter [4], etc) and software stacks (Binder [34], Whole Tale [52], etc) onto these dimensions. As new technologies that better (or differently) support computational reproducibility emerge or gain acceptance, the capabilities of these tools can be mapped as well; and the common, underlying model can be enhanced as needed.

The advantage of including such capabilities in Research Objects is that researchers could be made aware of the implications of the various technologies, programming environments, and specification standards they choose to use employ. It is easy to imagine a current-day researcher intending to enable others to reproduce their computational results by sharing any custom source code or scripts in a Git repository, along with the Dockerfile they used to create the computing environment in which they worked. While this is laudable, and almost certainly better than nothing, in many cases it is likely the researcher's expectations with regard to how these actions will ensure reproducibility will exceed what is actually the case. If instead the researcher composed their study as a Research Object, they could be prompted—by whatever software environment they are using to create the RO—for details about their precise expectations with regard to reproducibility. They may then find that a Dockerfile that does not specify the version of the base image, for example, is not sufficient to meet their expectations. When faced with the current limitations of available technology they may choose to archive the Docker image itself, or even a virtual machine image, while still being made aware of the limitations associated with these alternative approaches.

Researchers evaluating an RO similarly could probe its reproducibility capabilities. One might discover for example that an RO comes with a Dockerfile that currently references a non-existent base image. Or that it depends on a software package no longer available in the Ubuntu Apt Repository. The archived Docker image referenced in the RO might no longer be compatible with the latest version of Docker. These are all challenging issues to discover, debug, and remediate even for experts in these technologies. Making Research Objects *transparent* with respect to their actual reproducibility capabilities would be a step forward for making the computational components of scientific research reproducible and transparent.

VI. Reproducibility Queries

Research Objects long have been advocated as vehicles for sharing the provenance of scientific results and data products [10]. Here we list some ways we see extensions of our prior efforts in the area of science-oriented provenance queries enhancing reproducibility and transparency in Research Objects generally.

For provenance management systems, representations, and user interfaces to support scientific transparency, they must support science-oriented queries. Provenance must address 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 is key to enabling researchers to build on computed results reported in prior work with confidence.

For provenance to serve this function, however, it must be possible for researchers unversed in the detailed specifications of Research Objects and the PROV standard [27] to pose questions and receive answers meaningful for evaluating, using, and building on the processes and products of prior research. In [37] we provided a number of example queries about a run of a scientific workflow implemented in Python and marked up with YesWorkflow [38] annotations. Answers to these queries revealed the transitive dependencies of particular workflow outputs on the experimental samples, instrument settings, and intermediate data products. The queries were phrased in terms familiar to researchers in the example domain, and demonstrated how provenance queries can be used by scientists to answer scientific questions about research.

We plan to extend this approach to Research Objects that support the other capabilities proposed in this paper. Such queries would allow a researcher to determine not just that a study as a whole is FASEB::reproducible, for example, but also that a particular result is NAS::reproducible (which is a completely different thing). For studies that do not qualify as FASEB::reproducible as a whole, researchers could discover which results of the study are FASEB::replicable, and which are not. Where a particular computed result is not NAS::reproducible, they could pose a query that reveals what part of the method that produced the result is not repeatable. Answers to such queries could take available technology for reproducing computations into account. For example, a particular result might no longer be NAS::reproducible using the latest version of Docker.

VII. Conclusion

By combining capabilities for querying transitive data dependencies with approaches we propose in this paper for precisely characterizing the reproducibility of computed results and data products, we expect to bring the computational components of research studies represented by Research Objects closer to the level of reproducibility that characterizes the natural sciences. Such Research Objects would make clear what studies are reproducible and what methods underlying reproducible results can be used in future studies. Research Objects that are transparent in this sense would allow researchers to build on others' work with greater confidence—without actually having to rerun another researcher's study. This is reproducibility by other means.

Acknowledgments

The authors would like to thank Victoria Stodden, Matthew B. Jones, Kacper Kowalik, Ana Trisovic, and members of the Whole Tale team for useful discussion. This work was supported in part by National Science Foundation Awards OAC-1541450 and SMA-1637155.

References

[1] Docker documentation. https://docs.docker.com/, 2019.
[2] Federation of American Societies for Experimental Biology. https://faseb.org/, 2019.
[3] PREMIS data dictionary for preservation metadata. https://www.loc.gov/standards/premis/, 2019.
[4] Project Jupyter. https://jupyter.org/, 2019.
[5] SIGMOD DB research reproducibility. http://db-reproducibility.seas.harvard.edu, 2019.
[6] Singularity. https://singularity.lbl.gov/, 2019.
[7] ACM. Artifact Review and Badging, Apr. 2018. https://www.acm.org/publications/policies/artifact-review-badging.
[8] M. Baker. Is there a reproducibility crisis? 1,500 scientists lift the lid on reproducibility. *Nature*, 533:452–454, 2016.
[9] L. A. Barba. Terminologies for Reproducible Research. *arXiv:1802.03311 [cs]*, Feb. 2018.
[10] S. Bechhofer, I. Buchan, D. De Roure, P. Missier, J. Ainsworth, J. Bhagat, P. Couch, D. Cruickshank, M. Delderfield, I. Dunlop, M. Gamble, D. Michaelides, S. Owen, D. Newman, S. Sufi, and C. Goble. Why linked data is not enough for scientists. *Future Generation Computer Systems*, 29(2):599–611, Feb. 2013.
[11] 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 Rec.*, 40(2):45–48, Sept. 2011.
[12] A. Brinckman, K. Chard, N. Gaffney, M. Hategan, M. B. Jones, K. Kowalik, S. Kulasekaran, B. Ludäscher, B. D. Mecum, J. Nabrzyski, V. Stodden, I. J. Taylor, M. J. Turk, and K. Turner. Computing Environments for Reproducibility: Capturing the "Whole Tale". *FGCS*, 94:854–867, 2019.
[13] K. Chard, N. Gaffney, M. B. Jones, K. Kowalik, B. Ludäscher, J. Nabrzyski, V. Stodden, I. Taylor, M. J. Turk, and C. Willis. Implementing computational reproducibility in the whole tale environment. In *Proceedings of the 2nd International Workshop on Practical Reproducible Evaluation of Computer Systems*, P-RECS '19, pages 17–22, New York, NY, USA, 2019. ACM.
[14] T.-M. Christian, S. Lafferty-Hess, W. G. Jacoby, and T. Carsey. Operationalizing the replication standard. *IJDC*, 13(1):114–124, 2018.
[15] J. F. Claerbout and M. Karrenbach. Electronic documents give reproducible research a new meaning. In *SEG Technical Program Expanded Abstracts 1992*, pages 601–604. Society of Exploration Geophysicists, 1992. doi:10.1190/1.1822162.
[16] H. Collins. *Changing Order: Replication and Induction in Scientific Practice*. Chicago Press, 1985.
[17] Committee on Reproducibility and Replicability in Science. *Reproducibility and Replicability in Science*. The National Academies Press, Washington, DC, 2019.
[18] D. L. Donoho, A. Maleki, I. U. Rahman, M. Shahram, and V. Stodden. Reproducible Research in Computational Harmonic Analysis. *Computing in Science & Engineering*, 11(1):8–18, Jan. 2009.
[19] D. C. Drummond. Replicability is not Reproducibility: Nor is it Good Science. In *Proc. Evaluation Methods for Machine Learning Workshop at the 26th ICML*, Montreal, June 2009. National Research Council of Canada.

[20] B. T. Essawy, J. L. Goodall, W. Zell, D. Voce, M. M. Morsy, J. Sadler, Z. Yuan, and T. Malik. Integrating scientific cyberinfrastructures to improve reproducibility in computational hydrology: Example for HydroShare and GeoTrust. *Environmental Modelling & Software*, 105:217–229, July 2018.

[21] D. Fanelli. Opinion: Is science really facing a reproducibility crisis, and do we need it to? *Proceedings of the National Academy of Sciences*, 115(11):2628–2631, 2018.

[22] FASEB. Enhancing Research Reproducibility: Recommendations from the Federation of American Societies for Experimental Biology. Technical report, 2016.

[23] J. Freire, N. Fuhr, and A. Rauber. Reproducibility of Data-Oriented Experiments in e-Science (Dagstuhl Seminar 16041). *Dagstuhl Reports*, 6(1):108–159, 2016. doi:10.4230/DagRep.6.1.108.

[24] R. Gentleman and D. Temple Lang. Statistical analyses and reproducible research. *Journal of Computational and Graphical Statistics*, 16(1):1–23, 2007.

[25] C. Goble. What is Reproducibility? The R* Brouhaha, Apr. 2016. Symposium Reproducibility, Sustainability and Preservation, Alan Turing Institute.

[26] S. N. Goodman, D. Fanelli, and J. P. A. Ioannidis. What does research reproducibility mean? *Science Translational Medicine*, 8(341), June 2016.

[27] P. Groth and L. Moreau. PROV-Overview. An Overview of the PROV Family of Documents, Apr. 2013. http://www.w3.org/TR/2013/NOTE-prov-overview-20130430/.

[28] D. Heintz and M. R. Gryk. Curating Scientific Workflows for Biomolecular Nuclear Magnetic Resonance Spectroscopy. *International Journal of Digital Curation*, 13(1):286–293, Apr. 2019.

[29] M. A. Heroux, L. A. Barba, M. Parashar, V. Stodden, and M. Taufer. Toward a Compatible Reproducibility Taxonomy for Computational and Computing Sciences. Technical Report SAND2018-11186, Sandia National Laboratories, Albuquerque, New Mexico, Oct. 2018.

[30] K. Hinsen. Dealing With Software Collapse. *Computing in Science Engineering*, 21(3):104–108, May 2019.

[31] J. P. A. Ioannidis. The Reproducibility Wars: Successful, Unsuccessful, Uninterpretable, Exact, Conceptual, Triangulated, Contested Replication. *Clinical Chemistry*, 63(5):943–945, May 2017.

[32] I. Jimenez, A. Arpaci-Dusseau, R. Arpaci-Dusseau, J. Lofstead, C. Maltzahn, K. Mohror, and R. Ricci. PopperCI: Automated reproducibility validation. In *2017 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pages 450–455, Atlanta, GA, May 2017. IEEE.

[33] I. Jimenez, M. Sevilla, N. Watkins, C. Maltzahn, J. Lofstead, K. Mohror, A. Arpaci-Dusseau, and R. Arpaci-Dusseau. The Popper Convention: Making Reproducible Systems Evaluation Practical. In *2017 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, pages 1561–1570, Orlando / Buena Vista, FL, USA, May 2017. IEEE.

[34] Jupyter-Project. Binder 2.0 - Reproducible, Interactive, Sharable Environments for Science at Scale. 17th Python in Science Conference, 2018.

[35] T. Kuhn. *The Structure of Scientific Revolutions*. University of Chicago Press, 1962.

[36] M. Maciejewski, A. Schuyler, M. Gryk, I. Moraru, P. Romero, E. Ulrich, H. Eghbalnia, M. Livny, F. Delaglio, and J. Hoch. NMRbox: A Resource for Biomolecular NMR Computation. *Biophys J*, 112(8):1529–1534, 2017.

[37] T. McPhillips, S. Bowers, K. Belhajjame, and B. Ludäscher. Retrospective provenance without a runtime provenance recorder. In *Proceedings of the 7th USENIX Conference on Theory and Practice of Provenance*, 2015.

[38] T. McPhillips, T. Song, T. Kolisnik, S. Aulenbach, K. Belhajjame, R. K. Bocinsky, Y. Cao, J. Cheney, F. Chirigati, S. Dey, J. Freire, C. Jones, J. Hanken, K. W. Kintigh, T. A. Kohler, D. Koop, J. A. Macklin, P. Missier, M. Schildhauer, C. Schwalm, Y. Wei, M. Bieda, and B. Ludäscher. YesWorkflow: A User-Oriented, Language-Independent Tool for Recovering Workflow Information from Scripts. *Intl J. of Digital Curation*, 10:298–313, 2015.

[39] B. Mecum, M. B. Jones, D. Vieglais, and C. Willis. Preserving Reproducibility: Provenance and Executable Containers in DataONE Data Packages. In *2018 IEEE 14th International Conference on E-Science (e-Science)*, pages 45–49, Oct. 2018.

[40] M. Miłkowski, W. M. Hensel, and M. Hohol. Replicability or reproducibility? On the replication crisis in computational neuroscience and sharing only relevant detail. *Journal of Computational Neuroscience*, 45(3):163–172, Dec. 2018.

[41] L. Oliveira, D. Wilkinson, D. Mossé, and B. Childers. Supporting Long-term Reproducible Software Execution. In *Proceedings of the First International Workshop on Practical Reproducible Evaluation of Computer Systems - P-RECS'18*, pages 1–6, Tempe, AZ, USA, 2018. ACM Press.

[42] R. D. Peng. Reproducible research in computational science. *Science*, 334(6060):1226–1227, Dec 2011.

[43] H. E. Plesser. Reproducibility vs. Replicability: A Brief History of a Confused Terminology. *Frontiers in Neuroinformatics*, 11, 2018.

[44] QEMU: the fast! processor emulator. https://www.qemu.org/, 2019.

[45] R. Rampin, F. Chirigati, D. Shasha, J. Freire, and V. Steeves. ReproZip: The Reproducibility Packer. *The Journal of Open Source Software*, 1(8):107, Dec. 2016.

[46] R. Rampin, F. Chirigati, V. Steeves, and J. Freire. Reproserver: Making reproducibility easier and less intensive. *CoRR*, abs/1808.01406, 2018.

[47] A. Rauber, V. Braganholo, J. Dittrich, N. Ferro, J. Freire, N. Fuhr, D. Garijo, C. Goble, K. Järvelin, B. Ludäscher, B. Stein, and R. Stotzka. PRIMAD: Information gained by different types of reproducibility. In Freire et al. [23], pages 128–132. doi:10.4230/DagRep.6.1.108.

[48] K. Robasky, N. E. Lewis, and G. M. Church. The Role of Replicates for Error Mitigation in Next-Generation Sequencing. *Nature reviews. Genetics*, 15(1):56–62, Jan. 2014.

[49] L. M. Spenkelink, J. S. Lewis, S. Jergic, Z.-Q. Xu, A. Robinson, N. E. Dixon, and A. M. van Oijen. Recycling of single-stranded DNA-binding protein by the bacterial replisome. *Nucleic Acids Research*, 47(8):4111–4123, May 2019.

[50] D. H. T. That, G. Fils, Z. Yuan, and T. Malik. Sciunits: Reusable Research Objects. In *2017 IEEE 13th International Conference on e-Science (e-Science)*, pages 374–383, Oct. 2017.

[51] D. L. Vaux, F. Fidler, and G. Cumming. Replicates and repeats—what is the difference and is it significant? *EMBO Reports*, 13(4):291–296, Apr. 2012.

[52] Whole Tale project. https://wholetale.org/, 2019.