6 Working groups

6.1 PRIMAD – Information gained by different types of reproducibility

Andreas Rauber (TU Wien, AT), Vanessa Braganholo (Fluminense Federal University, BR), Jens Dittrich (Universität des Saarlandes, DE), Nicola Ferro (University of Padova, IT), Juliana Freire (New York University, US), Norbert Fuhr (Universität Duisburg-Essen, DE), Daniel Garijo (Technical University of Madrid, ES), Carole Goble (University of Manchester, GB), Kalervo Järvelin (University of Tampere, FI), Bertram Ludäscher (University of Illinois at Urbana-Champaign, US), Benno Stein (Bauhaus-Universität Weimar, DE), and Rainer Stotzka (KIT – Karlsruher Institut für Technologie, DE)

6.1.1 What is Reproducibility

What is "reproducibility" anyways? And how is it different from "repeatability", "replicability", or any of the other r-words? There are already a number of attempts at defining and sorting out these different notions. De Roure [1] lists 21 different r-words grouped into 6 categories, stating that reproducibility means reusing a research object with a change to some circumstances, inputs, resources or components in order to see if the same results are achieved independent of those changes. Often these notions are context-sensitive (e.g., validation vs verification have rather precise and very different meanings in different communities.

As an alternative approach to sort out terminological confusions, we attempted to look at a different perspective. When trying to reproduce a study, what are the things that are kept the same (e.g., the overall method or algorithm) and what is changed (e.g., the input data or implementation language, etc.)? More importantly, while changing these things, what information is gained by successfully reproducing (or failing to reproduce) a study?

6.1.2 The PRIMAD Model

As a starting point, we defined a preliminary list of "variables" that could potentially be changed:

- (R) or (O) Research Objectives / Goals
- (M) Methods / Algorithms
- (I) Implementation / Code / Source-Code
- (P) Platform / Execution Environment / Context
- (A) Actors / Persons
- (D) Data (input data and parameter values)

This spells: OMIPAD. Rearranging the letters that we use to represent the several aspects that can be changed, it can be remembered as PRIMAD: (P)latform, (R)esearch Goal, (I)mplementation, (M)ethod, (A)ctor, (D)ata (both input and parameter data), which allows us to ask: What variables have you "primed" in your reproducibility study?

As a concrete example of the meaning of these variables, let's assume our (R)esearch objective is to sort a data set. We could use Quick Sort as the sorting (M)ethod (algorithm), which could be (I)mplemented as a script in Python and run over a Python 2.7 compiler on an iMac running MacOS 10 (and this would be the execution (P)latform). We could run this

over a specific (D)ataset (data.csv) using 0 as the pivot parameter. The (A)ctor, in this case, is the researcher that is executing the sorting. Summarizing:

- Research goal: sorting the input
- Method: quick sort
- Implementation: script in Python
- Platform: Python 2.7, MacOS, iMac, etc
- Input Data: the data that is to be sorted
- Parameter: the position of the pivot
- Actor: user that is executing the experiment

As a more concrete example, we can take Tandy Warnow's statistically binning paper and the controversy around it[22]. In this case, the controversy was that her initial approach (we will call it method M, proposed by team T1) was claimed (by team T2) to be non-reproducible. More specifically, team T2 implemented method M and could not reproduce the original results obtained by team T1. So, in this example, we have the following scenario:

- Research Objective: Improve state-of-the-art in phylogenetic tree construction
- Method: Statistical binning (supposedly M = M', but one side is arguing that M != M')
- Implementation: two available, by team T1 and by the "opposing" team T2
- Platform: various (we suppose)
- (input) Data: different datasets – some arguments were made about the suitability here as well, since apparently team T2 did not respect some premises of how the input data should be organized.

To describe this reproducibility study in terms of these variables, only the research objective R and the method M are fixed; everything else is varied (team T1 actually argues that the implementation I2 isn't of the method M, but of another method M'). To represent what changed, we use primed variables.

In this case, T1 argues: P'RI'M'A'D', while T2 argues P'RI'MA'D' (variables with apostrophe were changed, and non-apostrophe variables were kept the same). Thus, both teams actually disagree on whether M = M' or not!

6.1.3 Gains from different types of reproducibility

Reproducibility in its various forms, however, is never a goal in itself. We do it in order to gain something. By changing some (or several) of these variables, we gain different kind of knowledge. For example, if one keeps R, M, and I fixed, but varies the platform $P \longrightarrow P'$, then the reproducibility study tests the portability, stability, or platform-independence of the experiment.

Figure 1 shows an attempt to categorize and label the various types of reproducibility and to summarize the gain they bring to a computational experiment. The precise terminology to use is still subject to further debate and no final agreement could be reached, specifically with respect to the labels and the mapping to the terminology found in the literature to describe different types of reproducibility. This may be partially due to the fact that many of the terms used describe repeatability settings refer to combinations of the above, e.g. to differentiate between obtaining a certain level of repeatability within the same lab or by an external lab. But even independent of this combinatorial issues the exact terminology proves to be difficult to agree upon already within a computing setting, not to mention beyond this domain.

[22] https://youtu.be/-0jd0x7Kg90?list=PLO8UWE9gZTlAgHZPaxQbpUNY0T26zeL_f

Label	Data		Platform / Stack	Implementation	Method	Research Objective	Actor	Gain
	Parameters	Raw Data						
Repeat	-	-	-	-	-	-		Determinism
Param. Sweep	x	-	-	-	-	-		Robustness / Sensitivity
Generalize	(x)	x	-	-	-	-		Applicability across different settings
Port	-	-	x	-	-	-		Portability across platforms, flexibility
Re-code	-	-	(x)	x	-	-		Correctness of implementation, flexibility, adoption, efficiency
Validate	(x)	(x)	(x)	(x)	x	-		Correctness of hypothesis, validation via different approach
Re-use	-	-	-	-	-	x		Apply code in different settings, Re-purpose
Independent *x* *(orthogonal)*							x	Sufficiency of information, independent verification

Figure 1 PRIMAD Model: Categorizing the various types of reproducibility by varying the (P)latform, (R)esearch Objective, (I)mplementation, (M)ethod, (A)ctor and (D)ata, analyzing the gain they bring to computational experiments. x denotes the variable primed i.e. changed, (x) a variable that may need to be changed as a consequence, whereas – denotes no change.

We now elaborate on the various aspects that can be changed, and how we could "label" reproducibility studies that use such combinations of changes.

1. ϵ equals to not changing anything, simply repeating an earlier experiment within the same computational environment, using the same code and data, allows to verify that the computed results are deterministically consistent. **Suggested label:** [**repeat**]
2. **Data → Parameters:** changing the parameter settings (e.g. parameter sweep, 10-fold cross-validation, etc.) allows to determine the: robustness/sensitivity of an experiment wrt. the specific parameters. Suggested label: [**rerun: robustness check, parameter sweep**]
3. **Data → Raw (Input) data:** changing the raw data processed by an experiment allows to verify how far the statements made hold across a larger part of the input space. Depending on the degree of similarity/difference in the input data, statements on the generality can be made. It also allows to evaluate whether the data originally used is representative/comparable for a given domain. **Suggested label:** [**rerun: check generality**]
4. **Platform:** changing the execution platform (i.e. the context, execution environment, including the software and hardware stack, i.e. a Java virtual Machine, running on a specific version of some operating system, within some hypervisor, running on specific HW) allows to test the platform independence/portability of an experiment. It may gain wider adoption or higher stability by being runnable on a wider range of platforms. **Suggested label:** [**port**]
5. **Implementation:** changing the implementation allows to verify the correctness of the previous implementation. It may also gain you higher efficiency, provide broader set of execution platforms, leading to higher adoption in different communities. Note that changing the implementation may incur a change of the execution platform. **Suggested label:** [**re-code**]

6. **Method:** changing the method allows to validate the correctness of a hypothesis using a different methodological approach. This provides a method-independent verification, or may provide a more efficient method to support the claims made. Note that a change in the method by definition will incur a change in the implementation, and possibly also of the execution platform. **Suggested label:** [**validate**]
7. **Research Objective:** changing the research objective (hypothesis) basically constitutes a re-purposing / re-use of an earlier experiment, allowing science to progress faster, opening new avenues for research. It requires trustworthy results/components to offer a solid basis. **Suggested label:** [**repurpose**]
8. **Actor:** changing the actor is orthogonal to all changes discussed above. It allows both independent verification of the characteristics, and also determines whether the information provided is sufficient to achieve such independent verification. **Suggested label:** [**experimenter-independent <activity>**]

Consistency: success or failure of a reproducibility study has to be evaluated wrt. the consistency of the outcomes. The criterion to apply thus is not whether the outcomes of priming any of the above variables leads to identical results, but whether results are consistent with the previous ones. Depending on the setting, this may require identity of results, but may also be lessened to consistency within certain error bounds or allow differences that are not statistically significant.

Transparency: Another dimension to be considered is transparency. It denotes the ability to look into all necessary components to be able to understand the path from the hypothesis to the results. While many of the changes above can be performed on a black-box level (repeating a run using binary code, performing the repeatability evaluation on a virtual machine provided by the original authors) it does not allow to make qualified inspections on the internal functioning on the respective levels. Thus, the degree of transparency should be used as a measure for the degree of inspection possible.

6.1.4 Variations on PRIMAD

After analyzing the various aspects that can be changed, we realized that using just one letter to represent both input data and parameters may not be enough. We are also aware of the fact that the differences between these attributes may not always be very clear-cut, as e.g. the fuzzy distinction between parameter and data to be supplied to an algorithm, or the boundary between an implementation and the execution platform becoming less clear-cut via the use of static or dynamically linked libraries. Yet, we find that the current set of variables helps in distinguishing core concepts and challenges to repeatable experiments relying on computation. Thus, we tried to identify possible letters we could use to represent each of the aspects we discussed:

- (O,R,G) Research Objectives / Goals
- (M,A) Methods / Algorithms
- (I,C,S) Implementation / Code / Source-Code
- (E,C) Platform / Execution Environment / Context
- (D,I,R) Input Data (“raw” data)
- (P) Parameter values
- (A) Actors / Persons

In the future, we may define a new acronym using these letters to better represent all the possible variations. Some possibilities are APDEIMO, PDEIMOA, AOMIEDP, OMIEDPA,

OMIEPAD. We may also need a deeper analysis of the various attributes and their changes, seeing in how far these can be mapped to, first of all, the different definitions of types of reproducibility being used in different communities. Furthermore, with most scientific work today spanning several disciplines and crossing methodological boundaries we need to investigate, in how far the concept of fixing and changing various attributes can be applied in more general settings. However, while the precise labels being used may change, we have the feeling that having a precise definition and understanding of the attributes that are fixed or changed is essential to define the various types of reproducibility studies and, specifically, to understand the benefit we gain from them. Reproducibility is not a means to its own end. While showing deterministic results by simply repeating a computation without changing anything may already be an exciting fact in some settings we very likely will want to go beyond such basic settings of reproducibility studies, gaining deeper insights into scientific work and establishing trust in results, methods and tools for the benefit of science.

References

1 De Roure, D., (2014). The future of scholarly communications. Insights. 27(3), pp. 233–238.

6.2 Reproducibility Tools and Services

Tanu Malik (University of Chicago, US), Vanessa Braganholo (Fluminense Federal University, BR), Fernando Chirigati (NYU Tandon School of Engineering, US), Rudolf Mayer (SBA Research – Wien, AT), and Raul A. Palma de Leon (Poznan Supercomputing and Networking Center, PL)

Sharing code and data increase reproducibility, but such sharing may not reflect the overall method, which is typically published in research papers. The current format of research papers (text-based) does not link code and data at finer granularity, the page-limit restricts detailed description of analyses and/or reporting of negative results, and authors have little motivation to describe in detail on a companion website. The consequence is built-up of scientific bias, which can be hard to break, given long cycles of publishing and funding.

Consequently, there is a critical need for reproducibility tools that, along with the changing culture of reproducibility, can also help researchers achieve the desired state of reproducibility in an efficient manner. However, before developing and/or applying a tool-suite to solve a reproducibility problem, several issues at hand must be understood. These range from:

1. **Precise identification of gaps in the research lifecycle.** A precise identification of gap in the research life-cycles is needed to understand which tool is applicable for solving the problem. Three gaps are often identified in the research lifecycle. The first one is related to the lack of motivation from researchers to apply reproducibility on their research. Better methods to incentivize reproducibility are needed, e.g.: having regulations and funding agencies to "force" the practice of reproducible research.
   A second possible gap is due to the poor linking between computational assets and text-based research outputs: there is rarely a connection between computational artifacts (research material, data, samples, software, models, methods, etc.) and the published results (paper and review process). This gap is very much discipline specific: some disciplines have developed standards on how to handle these artefacts and document the