Empir Software Eng (2012) 17:75–89
DOI 10.1007/s10664-011-9181-9

On the reproducibility of empirical software engineering studies based on data retrieved from development repositories

Jesús M. González-Barahona · Gregorio Robles

Published online: 18 October 2011

Editors: Martin Shepperd and Tim Menzies

Abstract Among empirical software engineering studies, those based on data retrieved from development repositories (such as those of source code management, issue tracking or communication systems) are specially suitable for reproduction. However their reproducibility status can vary a lot, from easy to almost impossible to reproduce. This paper explores which elements can be considered to characterize the reproducibility of a study in this area, and how they can be analyzed to better understand the type of reproduction studies they enable or obstruct. One of the main results of this exploration is the need of a systematic approach to asses the reproducibility of a study, due to the complexity of the processes usually involved, and the many details to be taken into account. To address this need, a methodology for assessing the reproducibility of studies is also presented and discussed, as a tool to help to raise awareness about research reproducibility in this field. The application of the methodology in practice has shown how, even for papers aimed to be reproducible, a systematic analysis raises important aspects that render reproduction difficult or impossible. We also show how, by identifying elements and attributes related to reproducibility, it can be better understood which kind of reproduction can be done for a specific study, given the description of datasets, methodologies and parameters it uses.

Keywords Repeatable results · Mining software repositories · Reproducibility

J. M. González-Barahona
Universidad Rey Juan Carlos, Mostoles, Spain

G. Robles (✉)
Universidad Rey Juan Carlos, Fuenlabrada, Spain
e-mail: grex@gsyc.urjc.es

1 Introduction

Reproducibility of experiments is one of the basic rules in the scientific method. Reviewers in scientific and engineering journals are well aware of this fact, and therefore check that publications include enough details. Usually, they focus on the description of the methodology used in experiments or studies, ensuring that other research teams can reproduce them with the same, similar, or completely different source data to verify, complement or extend the results. However, when software tools and complex collections of data are involved, the description of the methodology in a paper may not be enough to enable reproducibility. What is even more important, the lack of access to such software and data is certainly a barrier that discourages and makes reproduction more difficult.

This situation has been discussed during the last years in many different fields involving computational research (de Leeuw 2001; Donoho et al. 2009; Fomel and Claerbout 2009; Vandewalle et al. 2007; Koenker and Zeileis 2009; Hothorn and Leisch 2011). Recognizing the current state of affairs, reproducible research as such is promoted by several services such as the Reproducible Research Planet,[1] which advocates for the publication of "reproducible research compendiums", including not only the final paper but also the data and software tools needed to reproduce both the paper and the research study it presents (Gentleman and Lang 2007), and the Reproducible Statistical Computing Repository,[2] which facilitates the dissemination of reproducible statistical computations.

In the empirical software engineering community the issue of reproducibility has received increasing attention during the last years. There are many examples of this interest (Basili et al. 1999; Shull et al. 2004; Miller 2005; Vegas et al. 2006; Shull et al. 2008), that has lead to the establishment of specific workshops, of which the International Workshop on Replication in Empirical Software Engineering Research (RESER)[3] (Knutson et al. 2010) is probably the most prominent example. The idea of "reproducible research compendiums" found in the context of computational research appears also in empirical software engineering as "replication packages" (Vegas et al. 2006). At least in part as a result of this concern, some repositories have been established with datasets that can be used to facilitate the reproduction of studies: PROMISE (Boetticher et al. 2007), FLOSSMetrics (Herraiz et al. 2009), FLOSSMole repository of data about forges (Howison et al. 2006), Notre Dame SourceForge Research Repository,[4] and the Helix software evolution data set.[5]

Among empirical software engineering studies, those based on data retrieved from development repositories (such as those of source code management, issue tracking or communication systems) are specially suitable for reproduction. They are based on data that can be easily shared, and the analysis is in many cases performed with tools that can also either be shared, or described with great detail. Despite these facts, the reproducibility of many studies in this area is hindered by many factors,

[1] http://rrplanet.com

[2] http://freestatistics.org/

[3] http://sequoia.cs.byu.edu/?page=reser2011

[4] http://zerlot.cse.nd.edu/

[5] http://www.ict.swin.edu.au/research/projects/helix/

rendering them unreproducible or difficult to reproduce, even in part, due to lack of identification of the data or software tools used, as was found by one of the authors of this paper (Robles 2010). This situation led us to study the elements and attributes that are important for the reproducibility of this kind of studies. As a result, we present in this paper a methodology for assessing to which extent a study in this area is reproducible.

In this paper we consider reproducibility as the ability of a study to be reproduced, in whole or in part, by an independent research team. We consider partial reproducibility because being able of repeating specific steps of a study, while changing others, is the basis for accurate benchmarking of research methodologies and tools, and to detect intermediate steps subject to improvement. We say that a study "increases its reproducibility" if after some action either less effort is required to reproduce it, or it allows new types of reproduction studies. We consider as a "reproduction study" any study that reproduces in part or in whole another one, by reusing at least a part of its tools, data sources, datasets and parameters.

We will consider only studies based on data previously collected by some system supporting the software development process (such as a source code management system, an issue tracking system or a developers communication system). For them, the research work is based on the retrieval of data from those systems, hence the term "based on data retrieval" in the title of the paper. Although only a fraction of all empirical software engineering studies, they are gaining increasingly attention by the research community, probably because of the massive amount of data available in the repositories for those systems, and its relevance to understand the details of how software is developed.

Our approach is focused on characterizing elements according to how they impact on the reproduction of a study. The paper is written on the assumption that reproducibility is, in general, a desirable characteristic from the point of view of the research community, but this general benefit can conflict with the specific legitimate interests of a specific research team. This paper does not intend to enter into that conflict, which is addressed in some detail in Barr et al. (2010).

The structure of the paper is as follows. Next section introduces the elements that impact on the reproducibility of studies in our area of interest, and their main attributes. Section 3 uses the defined elements to characterize some kinds of reproducibility studies. A reproducibility assessment methodology for publications is then proposed, and later applied on two papers as case studies. In Section 5 we discuss our approach and some lessons learned. Finally, we provide some conclusions.

2 Elements of Studies with an Impact on Reproducibility

There is great variety in the characteristics of empirical software engineering studies based on data retrieval from development repositories. However, the analysis of the MSR papers (Robles 2010) and some other cases has led us to propose a general process model usable for most of them. Almost all studies in this field start by retrieving data from some system (or systems) related to software development. Those data are later cleaned, organized and maybe sampled. Then, some analysis is performed on this dataset (usually using some mathematical techniques), until the final results are produced (from which conclusions and research outcomes are

drawn). In this process, several elements of interest for reproducibility can be identified (see Fig. 1):

1. Data source. Where the "real world" data resides. It can be a repository (such as the source code management repository for a software project) or an object (such as the source code for a certain release of a software package or the electronic archive with the timesheets of the employees in a software company). For the purposes of this paper, a collection of data sources will be also considered as "the" data source.
2. Retrieval methodology. In most cases researchers cannot work directly with the data within the data source, and have therefore to retrieve it. This is usually implemented with software tools that automate the process.
3. Raw dataset. Data "as such", directly obtained from the data source by means of the retrieval methodology. These datasets can be stored in repositories, and be reused by other research teams for their studies.
4. Extraction methodology. Process, usually implemented (totally or in part) with software tools, of extracting, cleaning and storing the relevant data from the raw dataset.
5. Study parameters. In most cases, not all the data available in the data source or in the raw dataset is considered for the study: some parameters control which part actually is being analyzed. Those can be time periods, types of information, etc. They can be applied in the extraction phase, but also on the raw or even processed dataset.
6. Processed dataset. The application of the extraction methodology (and possibly the study parameters) to the raw dataset produces the processed dataset, which will be the input to the analysis methodology. For example, an SQL database or a CSV file ready to be imported in a spreadsheet.
7. Analysis methodology. Process, usually implemented (totally or in part) with software tools, of how the processed dataset is analyzed and studied to obtain the results dataset.
8. Results dataset. It is produced by applying the analysis methodology to the processed dataset and will be the basis for the research results and outcomes. Although they are data, they can be presented in the paper as graphs or in some other processed forms.

Each of these elements may have an impact on reproducibility, and are potentially subject to be reused in a new study.

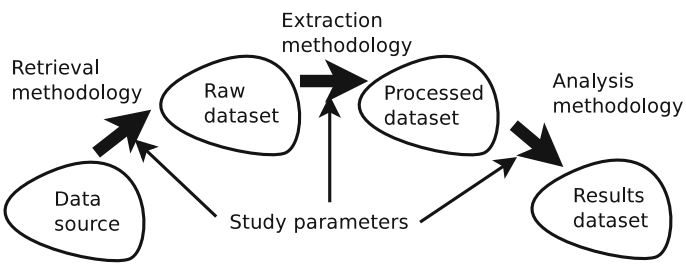

Fig. 1 Elements with an impact on reproducibility, organized according to their relationships during the research process

The identification of these elements in research papers is varied. Studying the papers analyzed in Robles (2010) we have gathered some details of how the participants in the Mining Software Repositories Workshop/Working Conference consider them. More than 60% of those papers were based on a public data source which was identified to a certain level. About one fifth of them identified publicly available tools for implementing the methodology, and an additional fifth mentioned those tools, although they were not identified. However, it was not common to differentiate between the three methodology steps we mention. Raw and processed datasets are mentioned in many of the papers, but only Germán (2004) was found to offer them publicly. Some others identified public datasets: Hayes et al. (2005) uses the MODIS dataset, from NASA, which is available from the PROMISE repository, while Panjer (2007) uses the 2007 MSR Challenge Dataset. Study parameters are also identified in some of the papers, but this is not a common case. Maybe Germán (2004) is again a good example, identifying parameters in sentences such as "in our experiments we have found that $\tau_{\max} = 45s$ and $\delta_{\max} = 600s$ are good values for these parameters". No paper was found to offer the results dataset.

The process model on which we identified the elements is quite similar to the KDD Process described in Fayyad et al. (1996), but adapted to our scope. We add one extra step at the beginning, by considering the retrieval of data from a repository as a part of the process, instead of starting with the raw dataset, as KDD does using the name "data" for it. Selection, which in KDD is performed only on this starting "data", is in many studies performed at other stages, so we have modeled it via the application of study parameters at any point. We have also merged KDD's "target", "preprocessed" and "transformed" data into "processed dataset" since the steps that produce them are usually mixed in the studies we have considered. Finally, we do not consider KDD's "interpretation/evaluation" step, since it is not relevant for reproducibility.

The level of detail with which all of the identified elements are described in a study, and their availability and characteristics may impact heavily on the reproducibility. To capture this impact, we identify several attributes of the elements with an impact on reproducibility:

- Identification: Where can the (original) element be obtained from?
- Description: How detailed is the published information about the element, including its internal organization and structure, and its semantics?
- Availability: How easy is it for a researcher to obtain the element, or have access to it?
- Persistence: How likely is the element to be available in the future?
- Flexibility. How flexible is the element, how easily can it be adapted to new environments?

These attributes are independent from each other, therefore showing different "dimensions" of how easy (or difficult) the reproduction of a study will be. For example, an element could be highly available (e.g. public), but badly identified (e.g. only with a generic name that makes it impossible to ensure which one it is exactly). Or an element could be unavailable, but stored in a well maintained, persistent private repository, ready for future use.

Values for the reproducibility attributes are inferred only from information published with the study. From this point of view it is irrelevant if an element is

Table 1 Attributes for each type of element

	Data source	Datasets	Parameters	Tools
Identification	X	X	X	X
Description	X	X	X	X
Availability	X	X		X
Persistence	X	X		X
Flexibility		X		X

perfectly identified in the internal, unpublished documents of the researchers, but not in the published study. In the same sense, it does not matter what the researchers consider about the different elements in the study if they are not specified in the paper. For example, tools are deemed as unavailable if authors do not mention explicitly how to obtain them (or the procedure to obtain them).

Not all reproducibility attributes can be applied to all reproducibility elements, and their exact interpretation may vary according to the type of element. Table 1 shows which attributes can be described for each type of element.

3 Types of Reproduction Studies

The presented elements can be used to characterize reproduction studies. We have 8 elements to consider, each of which could be either reused or new when reproducing, leading to many different kinds of reproduction. Diagrams based on Fig. 1 can be drawn to illustrate them, shadowing (in gray) the reused elements, and assuming that those not shadowed are new elements that differ from the original study.

Some of these kinds of studies, classified according to the major groups of methods for verifying findings presented in Gomez et al. (2010), are:

- Complete new study. All elements are new, produced by the research team performing the study. Hence, it cannot in fact be considered as a reproduction study, and is included in this list only for the sake of better understanding the diagrams. Figure 1 was an example of such a study.
- Procedural validation. This study is a complete reproduction of the original one, using all its methodologies and datasets. All elements are reused, hence all would be shadowed in gray. According to Gomez et al. (2010), this would "follow the same method" and "use the existing datasets". Starting from the same data source, and using the same retrieval methodology and study parameters, a raw dataset is produced. Then, using the same extraction methodology (again with the same parameters), a processed dataset is produced, on which the analysis methodology is applied to produce the results dataset. Raw, processed and results datasets are compared with the original ones. Therefore, the study is nothing else than a procedural validation with the same datasets. It could be used to asses that the original study was actually performed as described, and that the results are correct. Although from the research point of view this case does not produce new knowledge, some authors argue that it is well suited for education, or when researchers want to get better insight into the tools and processes used by other research teams (Robles and Germán 2010).
- New analysis based on the same processed dataset (Fig. 2). This could be a common case in which a research team starts with the processed dataset of

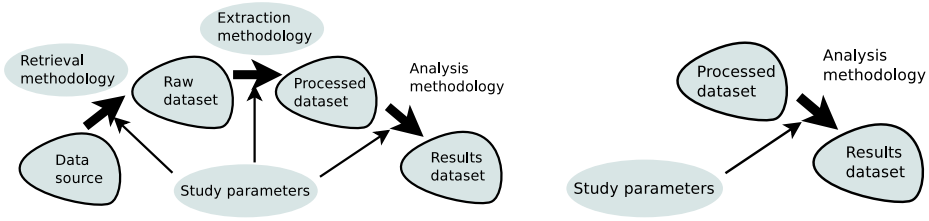

Fig. 2 New analysis based on the same processed dataset. Only the analysis methodology is new. In the left diagram, all the elements of the study are shown. In the right one, only those that would actually be used by the research team performing the reproduction study

the original study, trying to reproduce it by using a different methodology or a different set of tools. According to Gomez et al. (2010), this would "use a different method" and "use the existing datasets". Although all the other elements of the original study (except for the analysis methodology) are reused, the research team only deals really with four elements: the processed and results dataset, the study parameters (all reused) and the analysis methodology (new). In fact, the original results dataset can be reused or not: if it is, it can be compared with the new dataset produced by the new analysis methodology.

- New analysis based on the same raw dataset, with different parameters and methodologies. In this case, only the raw dataset (and indirectly the data source and the retrieval methodology) are reused. This would also fall in "use a different method" and "use the existing datasets", according to Gomez et al. (2010). In this case the research team benefits from an already collected raw dataset for performing a study that can be significantly different from the original. However, it could also look for the same type of results, but following a completely different path. Being able of reusing the results dataset would allow for easy comparison.
- New study reusing only the retrieval tools (Fig. 3). This is a common case, where the only reused element is the retrieval methodology, usually implemented by some retrieval tools (in part "follow the same method" (Gomez et al. 2010)) The data source, and the rest of the elements in the study are new. In fact, this is commonly not considered a reproduction study, but illustrates a common case of reuse of a single element.

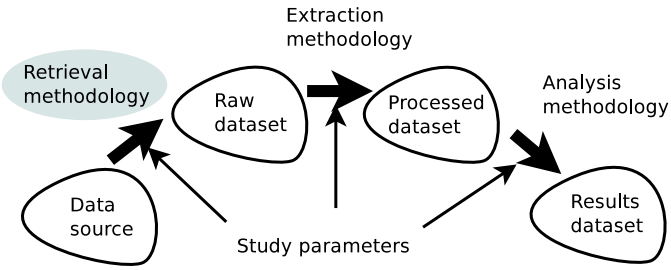

Fig. 3 New study reusing only the retrieval tools

Table 2 Example of assessment and tags for the elements of a certain study

Element	Assessment	Tag
Data source	Usable	U
Retrieval methodology	Not usable	N
Raw dataset	Usable with some difficulty	D
Extraction methodology	Usable	U
	Likely available in future	+
	Flexible	*
Study parameters	Not usable	N
Processed dataset	Not usable	N
Analysis methodology	Not usable	N
Results dataset	Usable	U
	Flexible	*

4 Reproducibility Assessment

The proposed assessment characterizes a paper according to its reproducibility, also specifying the type of reproduction studies that are feasible. To be useful, the assessment methodology should produce information on the reproducibility of each element of the study so that authors, reviewers or prospective reproducers can understand its general reproducibility status. For that, we propose a table constructed by assigning simple tags to each attribute of each element, according to its reproducibility status. Tags are one character (N: not usable for reproduction, D: usable for reproduction with some difficulty, U: usable for reproduction, "−": irrelevant or nonexistent) maybe followed by some signs ("+" for indicating that availability is foreseeable in the future, "*" for indicating flexibility).

Table 2 shows an example of an assessment of a study. In this case, only the data source, the tools for the extraction methodology and the results dataset are completely usable for a reproduction study. The tools implementing the extraction methodology are stored in a way that suggests future availability (for example, they are hosted as a project in a major forge), and are flexible (for example, they include source code). The results dataset has also been evaluated to be flexible (for example, it is an SQL file, which can be easily converted to other formats). The raw dataset is usable with some difficulty (e.g., is not public but can be obtained through a well defined procedure).

Table 3 Reproducibility assessment for the first paper

	Ident.	Description	Availability	Persistence	Flexibility	Assessment
Data source	Partial	Detailed	Public	Likely	–	D+
Retrieval meth.	Partial	Source code	Public	Likely	Complete	D+*
Raw dataset	No	No	No	N/A	N/A	N
Extraction meth.	Partial	Source code	Public	Likely	Complete	D+*
Parameters	Complete	Complete	–	–	–	U
Processed dataset	No	No	No	N/A	N/A	N
Analysis meth.	No	Textual	No	N/A	N/A	N
Results dataset	No	No	No	N/A	N/A	N

Table 4 Reproducibility assessment for the second paper

	Ident.	Description	Availability	Persistence	Flexibility	Assessment
Data source	Partial	Partial	Public	Likely	–	D+
Retrieval meth.	No	Textual	No	N/A	N/A	N
Raw dataset	Partial	Partial	Public	Likely	Complete	D+*
Extraction meth.	Partial	Source code Textual	No	N/A	N/A	N
Parameters	Partial	Complete	–	–	–	D
Processed dataset	Partial	Textual	Partial	Unknown	No	D
Analysis meth.	No	Textual	No	N/A	N/A	N
Results dataset	No	No	No	N/A	N/A	N

Given this assessment table, it is quick to analyze if a certain kind of reproduction study is easy, difficult or impossible: it is enough to consider the rows in the table corresponding to the elements relevant for it (according to Section 3).

Let us now present the reproducibility assessment for real cases: two journal papers in the field. Both have been authored (among others) by ourselves, but have been evaluated with respect to reproducibility only on the basis of what is available in the published paper. This same process has been done on many other papers, but we do not intend, at this point, to present a detailed massive assessment study, only to illustrate on the application of the proposed methodology (an advance of the results was presented in Robles 2010).

The first paper (Robles et al. 2006) analyzes a well known version management repository for files of different types (source code, build, translations, documentation, etc.). The corresponding assessment is presented in Table 3. A quick inspection raises two very simple actions that would have enhanced significantly the reproducibility of the study: the proper identification of the data source and of the tool used in the retrieval and extraction methodologies.

The second paper (González-Barahona et al. 2009) is a longitudinal study of several characteristics (size, dependencies, etc.) of Debian, a large Linux-based distribution. Again, we have identified the elements with impact on reproducibility, and their attributes, obtaining the assessment presented in Table 4. The inspection of that table permits the identification of actions that would enhance the reproducibility of the paper: all data sources could have been identified and properly cited, the processed dataset (which was published, but in an inconvenient way for reproduction) could have been offered in a more convenient format, and tools and datasets regarding dependencies could have been made public. Considering the details of the study, those actions ware indeed very easy to perform.

5 Discussion

The identification of the elements impacting on the reproducibility may seem trivial or even arbitrary at first sight. However, after the assessment of many studies, we have reasons to believe it is fundamental for an ordered and formal consideration of reproducibility. Elements have to be clearly identified so that they can be later

examined. Of course, some other sets of elements could be identified, but the one proposed in this paper has proved to permit easy identification in almost all the papers we have considered. Just to discuss one aspect that led to our list of elements, the differences between a "data source" and a "raw dataset" may seem negligible. However, we have found it to have a clear impact in many cases, since it provides a useful barrier between the "real world" and the "research world". Data sources, being out of the control of the researcher, can change over time, and even disappear. Raw datasets can be made immutable and persistent, and can be maintained in research facilities. For example, a CVS or Subversion repository run by a development project or company can be modified in unintended ways (eg, to remove all references to some code causing intellectual property problems), which will render exact reproduction impossible, something that will not happen with properly preserved datasets.

The proposed methodology tries to be as much objective as possible. However, many details have to be addressed, resulting in certain degree of indeterminism:

- Although the general process for performing a study in the field is usually the one described in Section 2, in real cases there may be some deviations. For example, the difference between data source and raw dataset is not always clear. Consider the case of a Subversion repository cloned using *rsync*; can that clone be considered as a data source, or is it a raw dataset? The same can be said of methodologies; in many cases, the same tool can be used for several methodologies. For example, CVSAnalY is used to support the retrieval, extraction and partially the analysis methodologies. In many cases the barriers between retrieval, extraction, and analysis are fuzzy. However, the identification of the eight elements we use have proved useful in the papers we have analyzed, even after taking into account these deviations.
- Real studies usually have not just one data source, or one raw, processed or results dataset. On the contrary, it is usual that they use many of them. In this paper we have considered that all data sources used in a study, combined, compose the "data source for the study", and have assessed it as a whole. The same has been done for datasets and methodologies. This has been a conscious decision: considering all those as separate elements would have meant adding more complexity, requiring more tables and diagrams, without a significant increase in obtaining a better model. To assess one element which is really a combination of many elements, we have tried to answer the question: is the combined element reusable in a reproduction study? In general, this means that if a single, maybe of minor importance, "subelement" has a very bad impact on reproducibility (because it is unavailable, for example), the whole composed element is also considered to have a bad impact. This can mask situations where "most of" the elements have a good impact on reproducibility, assessing those papers as "difficult or impossible to reproduce".
- A common reason for the multiplicity of elements in a study is that a paper is really presenting not one, but several interrelated studies, some of which can be radically more reproducible than others. Therefore, another way of addressing the problem mentioned in the previous item is to separate those different studies, which usually are based on a single data source and single datasets, and assess

them separately. Up to now we have considered that each paper corresponds to a single study, but this idea of identification of the "inner" studies seems worth exploring.

As shown with both case studies, the methodology has proved useful for helping authors to detect why a study is difficult to reproduce, in many cases leading to simple solutions such as the publication of a certain dataset, or the clear identification of a certain tool. This could also be used by reviewers, suggesting authors simple ways of improving reproducibility.

Following this line of reasoning to its end, a program committee could ask authors to produce a detailed reproducibility assessment of their paper, which would be subject to the review process along with it. This way it would be easier to decide whether the study is "reproducible enough" for a certain standard, having also the chance of arguing over the assessment itself as part of the review. In addition, authors would be confronted to the limits of the reproducibility of their study; this could lead to considering this issue when designing the research study, and when deciding the elements to publish.

The proposed methodology can be applied with independence of the specific type of reproduction study to be performed. Usually the whole study is not reproduced: some elements are reused from the original study, and some others are either modified or build from scratch. For example, a reproduction study may be done to test a new specific analysis methodology, maintaining all other elements, which would allow for easy comparison of the results. In other cases, the reproduction is as much complete as possible, with the aim of validating the original research. And there are still many other reasons and cases. Therefore, we have not focused on "grading" how reproducible a study is, but on providing the means for assessing how difficult a specific kind of reproduction study is going to be, providing a multidimensional approach to the problem of reproducibility.

As was mentioned in the introduction, an interesting question is how to obtain high standards of reproducibility. It turns out that, when the different elements in the study are clearly defined and are publicly available, still some concerns about reproducibility exist. For reproducing the study, the elements need to available at the moment it is being reproduced, maybe some years later. In the interim period, elements that were publicly available may be no longer, or the software environments on which researchers work have changed, which could make both programs and data formats obsolete, rendering them unusable even if available. This is the main reason why the attributes of persistence and flexibility have been introduced: they try to predict if a given element will still be accessible/usable in the future.

This leads to another interesting consideration: if the study is to be reproduced in the future, storing all its elements in a public archive intended specifically to preserve them for long periods of time will help a lot. That will save researchers a lot of time in finding and evaluating availability of elements, and will make it easy to detect elements suitable for future reuse. This would also facilitate benchmarking, and the establishment of standard datasets and tools which would help in the comparison of methodologies and research results. Currently, some of those archives exist, but they are more oriented towards storing reusable datasets than to put in a single, persistent place all the elements of a study. There is, therefore, some

room for improvement in this area, which initiatives such as the "reproducible research compendiums" (Gentleman and Lang 2007) are trying to fill. "Replication packages" are opening this path in the empirical software engineering research field (Vegas et al. 2006).

We have also noticed the importance of sticking to what is publicly available about a study when assessing its reproducibility. In many cases, the research team could have, internally, the elements needed for reproduction, and could provide them to third parties, even if that is not stated in the paper. However, only available information can be assessed, in the same way that a reviewer can only review what is in a paper. Therefore the proposed methodology uses as input only the information mentioned in the paper. Of course, authors can include references to companion documents or archives with further information, but even in that case, we consider that the availability of at least the central elements of the study should be clearly mentioned in the paper for being taken into account when assessing reproducibility. This is mainly a matter of cost-effectiveness and certainty. If assessors have to dig exhaustively in many different places to assess, the process becomes forbiddingly expensive, and a research task in itself. On the other hand, authors are very well situated to provide this information, prominently, in their paper if they have the right incentives (for instance, this being considered a positive aspect by reviewers).

6 Conclusions and Further Research

For research studies, reproducibility is a desirable property. However, analyzing if a given study is reproducible, or to which extent it can be reproduced, is a complex task. In this paper, we have focused on the field of empirical software engineering studies based on data retrieved from development repositories with the goal of improving our understanding on the factors that impact their reproducibility. In particular, we have discussed the factors affecting reproducibility in this specific field, and have proposed a methodology for assessing the reproducibility of a study.

Analyzing the usual process followed in the studies in this area, from the data source to the results dataset, we have identified eight elements of potential interest for reproduction studies. For each of them we have identified several attributes that will determine how feasible their reuse in a reproduction study is, and have studied how they affect each element. We have used them for characterizing reproduction studies according to which ones are being reused. This characterization of studies and types of reproduction has allowed us to provide a simple method for understanding if a type of reproduction study can be performed: it is just a matter of comparing the attributes of the elements in the original study that should be reused in the reproduction one.

We have also discussed two specific case examples: two papers, intended in principle to be reasonably reproducible, but which due to some details have rendered unreproducible, or difficult to reproduce, for many types of reproduction studies. This discussion sheds some light not only on the details of the application of the proposed methodology, but also on the complexities of doing the reproducibility assessment, and on how the methodology deals with them to produce sensible results.

We expect that this paper opens a bit more the field to further research on how

to improve reproducibility and how to detect factors that hinder it. The proposed methodology, as such, is expected to be of interest for:

- authors, who get a method to check the reproducibility of their studies, detecting the main barriers for a reproducible study,
- reviewers, who can also check whether the different elements in the study allow for an easy reproduction, and
- researchers performing reproduction of studies, who have now a method to quickly detect the problems they are going to face with a specific study.

Future work should be devoted to a more complete and formal application of the methodology to as much literature in the field as possible, to validate it, as well as to better understand the general situation of reproducibility in this research area.

Acknowledgements We thank the anonymous reviewers for their helpful comments and suggestions, and the researchers in this field for providing the matter of this study. This work has been partially funded by the European Commission, under the ALERT project (ICT-258098).

References

Barr ET, Bird C, Hyatt E, Menzies T, Robles G (2010) On the shoulders of giants. In: FoSER, pp 23–28

Basili VR, Shull F, Lanubile F (1999) Building knowledge through families of experiments. IEEE Trans Softw Eng 25(4):456–473

Boetticher G, Menzies T, Ostrand T (2007) PROMISE repository of empirical software engineering data. Department of Computer Science, West Virginia University. http://promisedata.org/

de Leeuw J (2001) Reproducible research. The bottom line. Technical report, UC Los Angeles: Department of Statistics, UCLA. http://escholarship.org/uc/item/9050x4r4

Donoho DL, Maleki A, Rahman IU, Shahram M, Stodden V (2009) Reproducible research in computational harmonic analysis. Comput Sci Eng 11:8–18

Fayyad U, Piatetsky-Shapiro G, Smyth P (1996) Knowledge discovery and data mining: towards a unifying framework. In: Proceedings of the 2nd international conference on knowledge discovery and data mining, KDD-96, Portland (Oregon, USA). AAAI Press, Menlo Park, pp 82–88

Fomel S, Claerbout JF (2009) Guest editors' introduction: reproducible research. Comput Sci Eng 11:5–7

Gentleman R, Lang DT (2007) Statistical analyses and reproducible research. J Comput Graph Stat 16(1):1–23

Germán DM (2004) Mining CVS repositories, the softChange experience. In: Proceedings of the international workshop on mining software repositories, Edinburghh, UK

Gomez OS, Juristo N, Vegas S (2010) Replication, reproduction and re-analysis: three ways for verifying experimental findings. In: Proceedings of the 1st international workshop on replication in empirical software engineering research (RESER 2010), Cape Town, South Africa

González-Barahona JM, Robles G, Michlmayr M, Amor JJ, Germán DM (2009) Macro-level software evolution: a case study of a large software compilation. Empir Software Eng 14(3):262–285

Hayes JH, Dekhtyar A, Sundaram S (2005) Text mining for software engineering: how analyst feedback impacts final results. In: Proceedings of the second international workshop on mining software repositories, St. Louis, USA

Herraiz I, Izquierdo-Cortazar D, Rivas-Hernández F (2009) FLOSSMetrics: Free/Libre/Open Source Software Metrics. In: CSMR, pp 281–284

Hothorn T, Leisch F (2011) Case studies in reproducibility. Brief Bioinform

Howison J, Conklin M, Crowston K (2006) FLOSSmole: a collaborative repository for FLOSS research data and analyses. IJITWE 1(3):17–26

Knutson CD, Krein JL , Prechelt L, Juristo N (2010) Report from the 1st international workshop on replication in empirical software engineering research (RESER 2010). SIGSOFT Softw Eng Notes 35:42–44

Koenker R, Zeileis A (2009) On reproducible econometric research. J Appl Econ 24(5):833–847

Miller J (2005) Replicating software engineering experiments: a poisoned chalice or the holy grail. Inf Softw Technol 47:233–244

Panjer LD (2007) Predicting eclipse bug lifetimes. In: Proceedings of the fourth international workshop on mining software repositories, MSR '07, p 29

Robles G (2010) Replicating MSR: A study of the potential replicability of papers published in the mining software repositories proceedings. In: 2010 7th IEEE working conference on mining software repositories (MSR), pp 171–180

Robles G, Germán DM (2010) Beyond replication: an example of the potential benefits of replicability in the mining of software repositories community. In: Proceedings of the 1st international workshop on replication in empirical software engineering sesearch (RESER 2010)

Robles G, González-Barahona JM, Merelo-Guervós JJ (2006) Beyond source code: the importance of other artifacts in software development (a case study). J Syst Softw 79(9):1233–1248

Shull F, Mendonça MG, Basili VR, Carver J, Maldonado JC, Fabbri SCPF, Travassos GH, de Oliveira MCF (2004) Knowledge-sharing issues in experimental software engineering. Empir Software Eng 9(1–2):111–137

Shull FJ, Carver JC, Vegas S, Juristo N (2008) The role of replications in empirical software engineering. Empir Software Eng 13(2):211–218

Vandewalle P, Barrenexea G, Jovanovic I, Ridolfi A, Vetterli M (2007) Experiences with reproducible research in various facets of signal processing research. In: Proceedings of the international conference on acoustics, speech and signal processing. ICASSP 2007, vol 4, pp IV-1253–IV-1256

Vegas S, Juristo N, Moreno A, Solari M, Letelier P (2006) Analysis of the influence of communication between researchers on experiment replication. In: ISESE '06: Proceedings of the 2006 ACM/IEEE international symposium on empirical software engineering, pp 28–37

Jesús M. González-Barahona teaches and researches at Universidad Rey Juan Carlos, Mostoles (Spain). His research interests include libre software engineering, and in particular quantitative measures of libre software development and distributed tools for collaboration in libre software projects. In this area, he has published several papers, and is participating in some international research projects (more info at http://libresoft.urjc.es). He is also one of the promoters of the idea of a European masters program on libre software, and has specific interest in education relating to that area.

Gregorio Robles is Associate Professor at the Universidad Rey Juan Carlos, which has several campi distributed in the region of Madrid (Spain). He earned his PhD in 2006 and currently has mainly teaching duties in the field of computer networks, although he teaches in a master on libre software as well introductory computer courses to journalism students. Regarding research, he works in two fields: a) technology enhanced learning, which means that he tries to use technology to improve the learning processes, and b) software engineering research on Free/Libre/Open Source Software systems with special focus on empirical issues.