Good Record Keeping for Conducting Research Ethically Correct
Research Ethics

Cong Peng

peng.cong@bth.se

Introduction

Academic publication is the major way for sharing and communicating scientific research. But a research project is more than just a publication, the actual research activities and its records compose a research project together with its publication.

Records of research usually include planning and protocol descriptions, data manipulations, analysis procedures, and way more [1]. In the discipline of Computer Science, research records include more content like program code, experiment data, and logs produced by program. Here we consider the research record as all the data, notes, experiment records, proposals and other records that involved in the research but may not included in the published paper.

Research record keeping is not just simply storing the data or notes after publication, it also includes careful recording, clear documentation and proper management of those records during and after the research activities. It is not only for the sake of researchers' own planning, managing and retrospecting. The research records are also important sources for replicating research [2], checking validity, proving intellectual property [1], and in some cases for preventing and checking scientific misconduct [3].

As an important part of either before or after publishing research results, it is clear that good research record keeping is a key element for conducting not only rigorous but also ethical scientific research [3]. However, given the importance of good research record keeping to research ethics, the reality of the research performed in computer science in regarding to research record keeping are not optimistic. And there are no widely accepted standards [1] or commonly used guidelines for neither recording nor making the records accessible.

In this report, we discuss how can good research record keeping ensure and promote research in computer science being conducted ethically correct, and how it could be performed. Of course it is not highly necessary to require detailed research records for some categories of research studies that its publication can tell almost everything. So what we discuss in this report only applies to those research studies that the research records matter, such as research involves experiment, simulation, survey and interview. For the fact that research ethics covers many aspects, it is not possible to discuss all of them in a

short report. So here we cover only a few aspects, mostly related to research misconduct and reproducibility. Since there is no golden rule for research ethics that can distinguish between right and wrong, good practices are discussed together with dilemmas.

Research Records to Keep

Before the discussion of how research record keeping affects research being conducted ethically, it is better to define, at least in the scope of this report, what research record keeping is, what does it mean to be a record, and what involves in the activity of research record keeping.

As stated, record keeping is not just "keeping" as its literal meaning. It is meaningless to keep them if the content were not recorded in a proper manner. Which makes the research record keeping comprise the activities of recording, documentation and keeping.

As a broad discipline as computer science, it covers various categories of research activities. So the records that will be generated and to be kept are different with each research activity [3]. Here below are four types of research records that will be generated by different categories of research studies that generally appear, which are mostly refer to the categorization by J. Zobel [3], A.Schreier et al. [1] and C.Collberg et al. [4].

- Notes, intermediate study records such as descriptions of progressive ideas, experiment records, research proposals, progress reports.

- Data, such as source data for experiments, original questionnaire responses, complete interview transcripts.

- Code, programs of the experiment, prototype, and those involved in the research. With documentation, running environment.

- Logs, produced by researchers and programs.

Of course not the all research records in Computer Science research can be categorized according to this, and it also overlaps between types.

Research Keeping matters for Research Ethics

As agreed by the wide scientific community, research record keeping affects many aspects for conducting research ethically correct [1, 2, 3].

Reproducibility is essential for research [5]. However, reproducing other researchers' study in many cases could be difficult. One of the big obstacles is the communication of how to perform the study. Taking re-running the code of an experiment as an example, it will not be simple as getting the code and running it. Without corresponding document, bunch of errors always come along when it is lucky, it even cannot be executed in worse cases, not mention getting the approximate results to the ones in publication. Because the appliances for the experiment, the program in this case, are usually developed by researchers specifically for the study in a temporary usage concern, many specific details were not considered.

It would be totally different if the corresponding research record, the document for the code in this case, could be provided, with clear and complete descriptions like running environment, experiment procedures, logs of phenomenon when the researcher doing experiment.

Provide support for research reproduction is one of the essential research ethics. Good record keeping is the main basis of the support. It would be very supportive if researchers can keep in mind how they can support their fellow researchers for reproducing by making the research record well, just as they think how they can report the experiment when doing for their papers.

But without making the research records accessible to fellow researchers, it is in vain no matter how good the record is. As some 404 pages returned to me when I was trying to open the links of code snippets provided in the paper, and the anecdotes experienced by Collberg et al. [5], it is not just a waste of time for the research progress, but also raised a doubt on the results. For some (or we can say many) categories of studies, we can hardly be sure whether the results are valid without checking the code, data and so on.

And for some types of research methods like survey, it is somehow easy to forge or falsify the data, especially for those anonymously recorded. With publicly accessible research records, misconduct behaviours like fabrication and falsification could be checked by fellow researchers.

On the other hand, the requirement of making research records publicly accessible with details can somehow prevent potential misconduct. Because we know that forged content is vulnerable, researchers with malicious intentions will consider more before deciding to forge.

The similar effects apply to plagiarism, which might be the mostly appeared misconduct behaviour [6]. Though the plagiarism is revealed by chance if it was elaborately conducted, the stolen text excerpts, ideas, data, etc. could still be discovered from the research records since the circle is relatively small for the same topic.

Consider that the publication is a report of research's final results with a limited length with a clear scope, it cannot include all those works that helped to formulate and conduct the research. So the research record can help to give credits to and also share with fellow researchers those works that not mentioned in the publication.

Despite good research record keeping itself is part of the research ethics, and affects many aspects for conducting an ethical research, the reality did not reflect the same importance it should have.

A. Schreier et al. conducted a survey of 96 universities, in which over half of the officials reported they had been hampered in inquires or investigations by inadequate research records [1]. And 27% of 3,247 respondents admitted to inadequate research record keeping according to the survey by American National Institutes of Health [7].

Now let us take a look at the facts in the discipline of Computer Science, C. Collberg et al. examined 402 papers different ACM conferences and journals that should have code as research record to back the results [5]. This is only a check for repeating the part of running the code used in research, which is far from reproducing the work. However, only 85 codes could be found by the links provided in the paper, codes of 176 papers could not

be accessed even through email requests to the authors, either negative responses or no response within two months. Finally, only about half of the papers' (217 of 402) codes were able to be built, either easily or with difficulties.

As the importance of good research record keeping, it is surprising that large a proportion of research studies, rather than exceptional cases, without decent research record keeping. We cannot say that most of the researchers are unethical on conducting their research studies, there should be some reasons hinder research record keeping being better.

Here are some reasons concluded from the works by C. Collberg et al. [5] and B.Martinson et al. [7]:

- Difficult to maintain the host of research records, due to server downtime, operation expertise, since researchers were not trained on this.

- Software licensing constraints or commercial consideration.

- Complicated dependencies on specific versions of external software or types of hardware.

- Domain expertise to build and operate programs.

- There are no commonly used guidelines.

- It requires much effort to document for details.

- Lacking of sufficient formal training or guidance for new researchers and students.

Good Practices and Dilemmas

Though there are reasons, it is not saying that the standards in computer science are low, and OK to continue being low. As mentioned, the reasons caused these facts are multi-fold. One of the main reason is the lacking of generally accepted guidelines. But it could not and also should not have one commonly applied guideline due to the computer science being a discipline include various categories of research activities.

By referring to the works by J. Zobel [3], D.Wright [2] and the report of Good Research Practice [6], here we can discuss some probable good practices that can be applied in certain category of research activity, and some might be applied as general practices.

- Retention period:
  - As long as possible
- Code:
  - Keep and make accessible at least the exact code that was used to produce the results in the published paper
  - Provide comments at least for the essential parts
  - Document for building and running processes
  - Document for execution environment, ideally packed with environment like a docker file or virtual machine image

- Notebook
  - Guidebook for performing experiments
  - Records of experiments performed, with dates, intent, experiment setups, anticipate results, actual results, experiment equipment (software and hardware), problems occurred, interpretation, and other special notes for the experiment. Preferably include all the performed experiments rather than only the ones with satisfied results, it prevents bias for selected rounds
  - Notes on description of ideas, progressive thoughts of research, notable references

- Data:
  - Codebook for data records such as questionnaire responses, raw result data and input data set. Provide general information, indicate variables, measurements, and other explanatory information [8].
  - Keep the data with anonymity and confidentiality concerns

- Logs:
  - Raw log output of programs
  - Complete transcripts of interview, original questionnaire responses
  - Preferably all the records of outputted logs rather than only the succeed rounds

- Versioning of records:
  - Should keep versions when made major changes
  - Versioning together with code, notebook, logs
  - Preferably keep all major versions

While talking about practice, it always comes with trade-offs and dilemmas. It depends on the real situation when the researchers performing research activities, but by keeping in mind of ethically supporting others and the future self to understand and to reproduce the research.

For the retention period, it should be ideally as long as possible, but it dependents on the nature of the research. For example in Health and Medical research, the Australian Research Council and Universities Australia [9] recommended a minimum 5 years retention period of research data from the date of publication in general, and permanent retention for example if the work has community value.

Although computer science is a fast evolving discipline, many research projects that published 5 or more years ago are still valuable and have high impacts. It requires time for fellow researchers to look into the research and its record to reproduce the study and check the misconduct of the research. If the record is kept just in a short time, then the reproduction and misconduct checking is unrealizable, and there is no strict rule to require the researcher to provide the records. So a retention period of 5 to 10 years seems applicable. And if the research is considered important, then keep it longer. As the tools gets handy, host the records online, if they are not confidential, provides easy accessibility.

And once the record is hosted online, it will not cost too much effort to maintain its accessibility.

Usually, research should be made as open as possible to openly account for methods, processes and results. However, for research involves sensitive data, e.g. health data mining, the researchers have to make a good balance between confidentiality and openness, and openness may compromise in this case. This dilemma also exists in those research projects restricted by certain academia-industry collaboration agreements or laws in areas where the research performed.

It is suggested to pack the code with its execution environment. This might be argued too much for researchers to do. But packing with environment can significantly reduce the steps for fellow researchers to replicate the work, and it can also reduce much documentation work.

The code execution problem could be mostly solved by good documentation and packing with execution environment. But the real world environment is much harder to deal with. For cases that the code would be no longer valid due to changes of real world environment, like code needs to instantly scrape data from the Web, then the particular results could never be reproduced or even the code could not get the data due to changes from service provider. Then in cases like this, state clearly the reasons and provide a description of how the code should work are necessary, and also keep the exact code at the same time.

There are web based services such as RunMyCode[1], ResearchCompendia[2] and CARMEN[3] (Code, Analysis, Repository and Modelling for e-Neuroscience) to allow researchers to publish the code, data and other records related to their papers, and can even execute the code remotely. Although currently these services are either unavailable from time to time, or not actively being used, it is a good direction to reduce the effort for researchers to keep their research records. Institutional level cooperation with third-party services may achieve a better outcome. It's not only for a better financial support for these services, but also for having some kind of regulations or guidelines since there are similar requirements of record keeping within a department or faculty.

The very detailed record on experiments can be time consuming, and looks unnecessary to researcher. But it provides the groundwork for fellow researchers to reproduce, refute and refine the research, and it also protects the researcher from the suspect of fraud or misconduct in some cases [2].

Conclusion

Research is expected to be ethical, but researchers are humans, humans make mistake. The right or wrong of the research being conducted lies behind the publication and its records, not told by the researchers themselves. Unethical research need to be prevented and monitored under certain mechanism. A requirement of good record keeping is one of the ways to help preventing and monitoring intentional and unintentional unethical research.

---

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

[2]http://researchcompendia.org

[3]http://www.carmen.org.uk

The current lacking of common standards for research record keeping in computer science is inconsistent with the wider scientific community, and somehow fostered sloppy and unethical research [3]. Given the fact that computer science covers various categories of research activities, and the research ethics itself is also not static [6], researchers need to find a reasonable balance among various interests. It is the researcher's responsibility to keep her/his research records in a good manner with supportive, rigorous and open.

As an extremely inexperienced researcher, there must be many factors that are not considered or considered wrong, the content of this report is really restricted by my experience.

References

[1] A. A. Schreier, K. Wilson, and D. Resnik, "Academic research record-keeping: best practices for individuals, group leaders, and institutions.," *Academic medicine : journal of the Association of American Medical Colleges*, vol. 81, pp. 42–7, jan 2006.

[2] D. Wright, "Research ethics and computer science: an unconsummated marriage," *Proceedings of the 24th annual ACM international conference on Design of communication*, pp. 196–201, 2006.

[3] J. Zobel, "Reliable Research: Towards Experimental Standards for Computer Science," in *Proceedings of the Australasian Computer Science Conference*, (Perth, Western Australia), pp. 217–299, Springer-Verlag, 1998.

[4] "42 CFR 93.224 - Research record. | US Law | LII / Legal Information Institute."

[5] C. Collberg and T. A. Proebsting, "Repeatability in computer systems research," *Communications of the ACM*, vol. 59, no. 3, pp. 62–69, 2016.

[6] The Scientific Research Council's Expert Group on Ethics, *Good Research Practice.* 2011.

[7] B. C. Martinson, M. S. Anderson, and R. de Vries, "Scientists behaving badly," *Nature*, vol. 435, pp. 737–738, jun 2005.

[8] National Addiction & HIV Data Archive Program, "What is a codebook?," 2006.

[9] the National Health and Medical Research Council, the Australian Research Council, and Universities Australia, *Australian Code for the Responsible Conduct of Research.* 2007.