Reliable Research: Towards Experimental Standards for Computer Science

Justin Zobel

Department of Computer Science, RMIT
GPO Box 2476V, Melbourne 3001, Australia
jz@cs.rmit.edu.au

Abstract. Scientists are expected to keep thorough records of their research, to provide experimental rigour, allow reproduction and verification, and inhibit fraud and negligence. However, standards for recording of experiments have not been widely adopted in computer science. In this paper I discuss why such record-keeping is believed to be appropriate, outline the roles of experimental records, and examine whether standards within our discipline should be improved. It seems clear that computer science researchers do not meet the standards expected by the wider scientific community; as a first step towards agreed standards I suggest an approach to record-keeping for computer science that involves only moderate change to current research practice.

1 Introduction

Researchers are expected to be ethical. Standards for behaviour are determined by accepted codes, both implicit and explicit, covering areas such as authorship, abuse of power, scientific fraud, objectivity, and human and animal experimentation. Many of these issues are generic; for example, standards for determining authorship and maintaining objectivity apply in every field of research. But for other issues the ethical norms differ between disciplines—a consequence of variations in methods and purposes of research, forms of publication, and so on.

In the sciences, a key element of ethical behaviour is the standards for conduct and recording of experiments. In medical trials of new drugs, for example, the records of the trials provide the evidence required to support publication, and the expectation that records be kept mitigates against both deliberate fraud and accidental distortion of results. At least in principle these records, often made publicly available, allow other researchers to check the results, to analyse them for further properties, and to reproduce the original experiments.

In computer science, however, there are no widely accepted standards for recording experiments or for making such records available. On the one hand, it can be argued that records are not necessarily as crucial as in other disciplines, since computing research can often be verified or recreated by different means.

Proceedings of the 21st Australasian Computer Science Conference, Perth, Australia, February 2–4 1997.

On the other hand, lack of generally-accepted standards can be interpreted as tacit approval of sloppy or fraudulent research, and allows careless or deliberately unethical researchers to publish claims that their experiments do not support.

In this paper I argue that standards for recording and reporting experiments in computer science do not meet the expectations of the broader research community (indeed, it is difficult to identify any standards for recording of computing experiments) and, measured against guidelines such as those published by the Australian Vice-Chancellor's Committee, practice in computer science is inadequate. Yet it not obvious what processes for record-keeping might be adopted in computer science, where methods and materials are, superficially at least, rather different to those of the other sciences.

The purpose of this paper is to explore these issues and initiate debate about experimental practice in computer science. I describe expected standards for recording experiments, considering ethical codes, practical guidelines, and motivations underlying these standards. Verification and reproduction are considered in the context of computer science, and their relationship to adequate experimental records. In the context of these expectations of good research I discuss record-keeping for computer science; current practice and community standards is then used to propose a pragmatic approach to recording of computing research.

2 Standards for experimentation

An experiment is an interaction with external properties of some subject, with the purpose of making some kind of measurement or observation of it [Radder, 1995]. Usually some kind of apparatus mediates between the scientist and the subject; apparatus can range in power from, say, a pair of binoculars to the Hubble telescope. In this broad, inclusive definition, any interaction with the subject can be viewed as an experiment, so long as the interaction has some kind of identifiable, measurable outcome.

In the broad research community, methods for conducting and recording experiments have developed over several centuries. The standards of this community can be judged from recent developments such as the activities of the Commission on Research Integrity in the USA [Frankel, 1995; Ryan, 1995], and from documents such as the joint NH&MRC/AVCC "Statement and Guidelines on Research Practice" [AVCC, 1997], which makes detailed recommendations on experimental records. It is worth quoting from these guidelines at length.

> 2.1 Data (including electronic data) must be recorded in a durable and appropriately referenced form ...
> 2.3 Data must be held for sufficient time to allow reference. For data that is published this may be for as long as interest and discussion persists following publication. It is recommended that the minimum period for retention is at least 5 years from the date of publication ...
> 2.4 Wherever possible, original data must be retained in the department or research unit in which they were generated. Individual researchers should be able to hold copies of the data for their own use. Retention solely by the individual researcher provides little protection to the researcher or the institution in the event of an allegation of falsification of data.

> 2.5 Data related to publications must be available for discussion with other researchers ...
> 2.10 When the data are obtained from limited access databases, or via a contractual arrangement, written indication of the location of the original data, or key information regarding the database from which it was collected, must be retained ... [AVCC, 1997]

(To understand these guidelines the terminology must be interpreted for the context of computing. In particular the term "data" is confusing: it is used by the broader scientific community to refer to the outcomes of experiments, but in computer science it refers also to the subject of experiments, that is, the stuff experimented on. The guidelines do not require scientists to retain the subject of their experiments—tissue samples or whatever—but their records of them. It is this sense in which "data" is used in this paper.)

The AVCC guidelines are not law, and it would be a mistake to treat them as such—they are written for the research community as a whole and some of the detail will not apply to all disciplines. However, they do set standards, and a discipline that chooses to deviate from such standards needs to have clearly reasoned motives for doing so.

The need for good records has been highlighted by recent cases of academic fraud, in particular in the biological and medical sciences. In the case of William McBride, for example, his claims about the morning sickness drug Bendictin—based on concocted results—were estimated to have cost the manufacturer around $100,000,000. In another case, that of Stephen Breuning, his recommendations led to unwarranted changes in drug regimes for mentally retarded patients—and two felony convictions. Other cases have ended researchers' careers or led to involved legal proceedings [Dingell, 1993]. In each case research data has been a primary element of the defence against accusations of fraud.

Data can be narrowly defined to be observed values of experimental variables, but in the laboratories of the traditional sciences researchers are expected to record a considerable mass of material. Wilson, for example, recommends: that records be kept in ink in bound notebooks and include observed values of the variables, a "rather complete description of the apparatus", notes of modifications to apparatus as they are made, statements of the purpose of each experiment and conclusions reached, and witnessing and dating in the event of future disputes such as patent rights; that "bad or unpromising experiments, even those deemed failures, should be fully recorded"; that all drawings, sketches, and notes should be retained, however rough; and that all paperwork should be initialled and dated [Wilson, Jr., 1952, pp. 131–133]. Porush makes similar recommendations, and adds that, as a device for recording experimental work, computers have the disadvantage that "it is too easy to manipulate, alter, and lose data and observations ... *Using a computer makes protecting the integrity of your data more difficult*" [italics in original] [Porush, 1995, p. 36]. There are many undergraduate introductions to laboratory method that offer similar guidelines.

Bogen and Woodward noted the distinction between *data* and *phenomena* [Radder, 1995]. Experimental subjects exhibit certain phenomena; in the context of experiments these phenomena result in observed data. One of the roles of the researcher is to explain that this data is evidence of the underlying phenomena, or to infer the phenomena from the data. Although this distinction is simplistic [Radder, 1995], it does provide a valuable insight into the aims of experimentation.

For example, an astronomer might plan a night of observations, collect a series of photographic exposures, and list the interesting features in the photographs. A chemist might propose a measurement of, say, the calcium content of some unknown substance, develop and describe an appropriate apparatus, record the quantities of chemicals added to the substance and the temperatures attained at each stage, and explain why the results (a certain weight of precipitate) show that a certain quantity of calcium was present and why it is unlikely that calcium was lost during the refinement. A psychologist might propose a test to measure fear of the dark in adults, develop an environment and quiz-based metrics, explain how the subjects will be isolated from experimenter bias, and record the behaviour of each subject as they sit the experiment. In each case the experimenter has gathered some data with an indirect relationship to the underlying phenomena.

The tradition of thorough record-keeping has a long history—the notes of scientists such as Newton and Babbage are still available today—and remains strong in many disciplines. Yet in computer science, within Australia at least, it is neither taught to undergraduates, required as part of the "method" of completing practical work, nor expected of researchers or postgraduate students. The central questions to be addressed in this paper are whether such record-keeping can be valuable in computer science and what form records might take; to answer these questions it is first necessary to consider the value of records in other disciplines.

3 Motivations for record-keeping

It seems clear that the scientific community regards good record-keeping as extremely important—it is seen as a key element of ethical scientific conduct. There are several reasons for valuing record-keeping, all of which relate to the need to have a clear testimony showing what took place and when.

A fundamental motivation for record-keeping is that it provides evidence of *precedent*, used for example to establish prior invention in patent disputes, or used in event of accusations of plagiarism. Another fundamental motivation is that for many kinds of research records are the only *evidence* resulting from an experiment; an investigation of, say, wildlife populations may be documented solely by handwritten observations taken in the field. In such experiments, which are effectively irreproducible, the records are the single source of data used for the basis of published research. Moreover, such records can be re-used in the future to draw new inferences. Because Newton's notes still exist we now know, for example, that in his work on optics he observed effects that were not explained until the advent of quantum mechanics—observations that a less careful scientist might have discarded as irrelevant curiosities.

There are several further important motivations: rigour, elucidation, reproduction, and verification. Record-keeping provides *rigour* because the act of record-keeping, often to some kind of predetermined template, requires that the researcher proceed with a certain degree of care. For example, scientists usually expect particular experimental outcomes, and quite naturally will investigate if something unexpected occurs, by, say, checking parameter settings. If good records are kept it should quickly be obvious if the scientist is failing to check expected outcomes in the same way.

If records include discussion of the purposes and outcomes of experiments they are likely to be the first written *elucidation* of the intent of the experiments. Such writing forces the researcher to state thoughts clearly and to clarify vague ideas, and provides an excellent resource when, later on, the researcher is assembling material for publication. In this model, notes are not just good ethical practice, they are good research practice.

Records can form the basis of *reproduction* of experiments, because they should provide a full statement of what was actually done. They are also the basis of *verification*—checking whether the experiments were conducted and analysed with appropriate care, or indeed whether they were conducted at all; whether the claims are justified by the results; and whether the published results are a fair reflection of the experimental outcomes.

Exact reproduction is often difficult and sometimes impossible; indeed, it has been said that experiments are never really reproduced [Radder, 1995]. However, reproduction should be thought of as gathering new data about the same phenomena; the data can differ yet remain consistent with being good evidence of the properties in question. Considering the scenarios discussed above, for example, there is little chance of reproducing the exact behaviour of the psychologist's sample; different subjects will constitute a different sample of the population and even the original subjects are unlikely to record exactly the same answers to a quiz about fear. In contrast, it may well be possible to reproduce the chemist's work. Technique and experience are important to the outcomes of such experiments [Gower, 1997], but given adequate explanation a competent chemist should be able to get a similar outcome—assuming of course that the original material is still available. The astronomer's work may be easy to reproduce, so long as a similar telescope is available; or may be impossible to reproduce, if the objects or the position of the earth in the universe have changed too drastically.

In each of these cases good records should provide a reliable basis for verification. Note that verification too is never exact—a given set of records will provide a measure of certainty that the work was conducted as described and with adequate care, but is not an absolute arbiter.

Records are also valuable for detecting misconduct; if record-keeping is expected, a scientist must plan to be unethical if research results are to be forged. A widely used definition of misconduct is "fabrication, falsification, plagiarism, or other practices that seriously deviate from those that are commonly accepted within the scientific community for proposing, conducting, or reporting research" [AVCC, 1997; Ryan, 1995]. (The "FFP" part of this definition is uncontroversial, but some writers have objected to the "other practices" component on the grounds that it is too sweeping [NAS, 1996]. However, the Commission for Research Integrity has recommended retention of the wider definition [Ryan, 1995].) A particular issue of relevance to record-keeping is

> *Misrepresentation*: A researcher or reviewer shall not with intent to deceive, or in reckless disregard for the truth, state or present a material or significant falsehood; or omit a fact so that what is stated or presented as a whole states or presents a material or significant falsehood [AVCC, 1997].

Similar positions are taken by other major research bodies [NAS, 1992]. Records do not prevent misconduct, but they do inhibit it—casual misrepresentation of outcomes, for example, is a good deal less trivial if records are made available. Thus, while it is difficult to prevent fraud, that does not mean that we should simply do nothing and thus condone it. The more thorough that records are expected to be, the harder it is to falsify them. With low standards for records, a researcher can decide on the spur of the moment to report a falsehood; with better standards, reporting a falsehood requires considerable effort or risk.

It has been argued that serious misconduct is rare, but that it may become more common in future [Goodstein, 1995]; while in medical research there is anecdotal evidence that strongly suggests that fraud may not be rare at all [Dingell, 1993]. Moreover, fraud can consist of small-scale distortions as well as outright concoction of results. For example, removal of outliers from sets of results is often justifiable, but should not be routine. In experiments with algorithms many factors can be considered: not just resource requirements such as disk traffic, network traffic, memory, and time, but, in the case of heuristics, the algorithms' effectiveness. Unethical researchers might emphasise factors that favour a new algorithm and neglect other factors. Likewise a scientist might search for an environment in which an algorithm does well—exploring different data sets, say, or combinations of cache size and buffering strategy—and report only the cases in which the algorithm is successful. That is, results based on samples from a large population (of parameter values, say) can be distorted by only publishing the outcomes of the most promising trials.

What is not clear is the impact of such fraud. First, much published research is false. It has been said of physics that 90% of published results are wrong [Ziman, 1968], which implies that fraudulent results may have little effect. However it would be surprising if typical work in physics was false because it was sloppy; falsehood arises because the data yielded by measurement of natural phenomena is often ambiguous. Moreover, theoretical papers are concerned with models of the natural world rather than with the (unknowable) natural world itself, and in this context "false" often means "later shown to be incorrect". That is, there may be many competing theories explaining the same data, some of which will be falsified by new data or inability to make successful predictions. This situation is in contrast to computing, where both the causes and effects of the phenomena are artifacts, and only a fraction of research follows the model-and-test framework.

Second, the effect of wrong or fraudulent research can be hard to discern. This aspect of fraud is considered further below, but the fact that it may have little impact does not imply that it can be neglected. We cannot condone poor research: it allows amoral researchers to build careers on insubstantial work; it removes the incentive for scientists to check their results because the extra work is not necessary for publication; and it erodes trust in published work, and in science in general. Mechanisms for prevention or containment of poor work add to the reputation of science as well as to the quality of published research.

4 Experiments in computer science

Consideration of the kind of records that should be kept depends on the nature of the research. Records of human drug trials, for example, clearly require ongoing, detailed documentation of the individuals involved—without such records verification is impossible. For experiments using highly standardised apparatus, however, sketchier records may be appropriate; only a little information may be needed to describe results based on inspection of publicly-available satellite images, since reproduction and verification are likely to be straightforward.

The kinds of records needed for computer science, then, depend on the kind of research that is done. Computer science is a broad discipline covering many different kinds of research activity. The proceedings of the 1997 Australasian Computer Science Conference [Ramamohanarao and Zobel, 1997] (which represents a wider range of kinds of research activity than most conferences) includes, for example, several instances of experiments comparing different algorithms, such as grep methods and compiler optimisations; an experiment demonstrating the convergence of a method for eigenvalue computation; demonstration of learning in a neural network; evaluation of a browsing interface; evaluation of schedulers; and comparative measurements of system performance. (A surprising number of the papers in the conference have no experimental content, and moreover no concrete evaluation of the proposed methods. It is an open question as to whether this reflects low scientific standards in computing, or whether some research in computing should be judged by the ideas rather than the evaluation.)

This paper is not the place to attempt an exhaustive categorisation of research in computer science, even supposing that such a thing was useful or feasible, but several kinds of research can be identified. These categories are intended to be illustrative rather than precise; they overlap and by no means include all research in computing.

- Evaluating whether an algorithm (or more generally a system) behaves as predicted. Behaviour might involve resource requirements or correctness, for example.
- Comparing algorithms with regard to particular properties.
- Identifying appropriate parameters or typical resource requirements for an algorithm.
- Demonstrating that a concept is feasible in practice.
- Testing of human factors, such as reaction to an interface or ability of a retrieval system to identify relevant information.

These activities do not all require the same kinds of records. Human factors experiments, for example, are little different from many experiments in psychology, and presumably the same kinds of standards apply: there should be careful records of subjects and responses, together with descriptions of the apparatus and experimental environment. In the cases involving algorithms, the apparatus in some sense documents itself—the code embodies a great deal of the matter of the experiment. While other records are likely to be required, to note, say, version numbers and parameters, these records need not be nearly so comprehensive as for human trials. (Kinds of records for computer science are considered further below.)

The outcomes of experiments in computer science can nonetheless be as researcher-dependent as experiments in any other discipline. To test an algorithm requires an implementation of reasonable quality; an issue of particular significance when algorithms are being compared, because performance may relate to the standard of code rather than the embodied concepts. Also, the existence of code does not obviate the need for records. Experimenter skill is important—obtaining reliable disk timings on a desktop UNIX machine, for example, can require knowledge of the architecture of both hardware and operating system and the ability to "cold start" by flushing caches. Choice of data and of underlying system is also important. The relative performance of two algorithms may vary from architecture to architecture, and from data set to data set. The researcher needs to make appropriate choices, and to know why they are appropriate. That is, the ability to construct a good experimental design can have a significant impact on outcomes; and experimental designs must be recorded.

Moreover, it can be hard to replicate results, particularly for experiments used to demonstrate properties such as small improvements between one algorithm and another, and precise reproduction of results is impossible when the underlying technology is rapidly developing. It is because of these kinds of factors that the distinction between data and phenomena is important. Different researchers will test something in different ways on different systems, and will probably observe different outcomes, but these outcomes should illustrate the same underlying phenomena.

Another problem with relying on code as records is that evaluation of an algorithm also involves other factors, such as data sets and parameter values. Without ongoing recording of details it is impossible to know whether the reported runs fairly reflect experimental outcomes, or have been selected, possibly with bias, from a much larger population of runs. In some ways, curiously, experiments in computer science can be rather like experiments in medicine. A particular algorithm (treatment) worked for particular data (patient), but knowing that it worked is not sufficient; it should be shown to work for a variety of data (people), and if it is to be trusted it is important to understand why it worked. If in some context it does not work (the patient died), that too is important information that must be recorded.

And, as in medicine, erroneous research results in computing can matter. Encouraging publication of sloppy or fraudulent work is poor because it rewards bad science and erodes confidence in research. But additionally bad research in computing can cause harm. This issue is perhaps illustrated best by analogy. False experimental results in medicine can have an immediately obvious consequence, such as death, but in other cases may be inconspicuous; a drug that in a small percentage of people leads over years to a particular illness will be only one of the possible causes of that illness, and thorough investigation is required to reveal the drug as the cause. Likewise, if engineering studies apparently showed that a new material has a certain strength, but in practice it sometimes fails catastrophically, the numerous possible causes can only be eliminated as instances of failure accrue. In computing, a false claim might relate to the resource requirements of an algorithm, in which case the harm is insidious: a small percentage increase in costs, say, for businesses using the algorithm. Or a claim might relate to the reliability of a process scheduler, where occasional errors could easily be attributed to software unreliability rather than a fundamental flaw in the underlying method. A compiler optimisation that occasionally results in wrong code could be extremely difficult to identify given the number of factors that affect reliability of a large software system. The consequences of false results might often be trivial, but much of the research in computing is concerned with practice and utility, and computers are involved in every part of our lives. Erroneous research could well lead to events where—just as with medicine or structural engineering—a scientist is held responsible for some disaster.

A failure could be the result of incorrect research, regrettable but unforeseeable. If the research has been conducted with adequate care the scientist has a reasonable defence against liability. But the onus is on scientists to take reasonable care, and to adequately document their work [AVCC, 1997].

5 Practice in computer science

One of my original motivations for investigating standards in computer science research was pragmatic: in teaching of research methods, what practices should be recommended to postgraduate and honours students? My own experience had exposed me to neither recommended standards for conduct of research nor consistent practice amongst fellow scientists. Indeed, although a good many researchers in computer science do keep some record of their work, many others appear not to, and furthermore it seems that failure to keep such records is not seen as particularly poor practice. To take an extreme case, it would be unusual to condemn a paper even if it was revealed that all the code used in experiments had been deliberately discarded.

More typically, it is not uncommon to find that the only code kept by a researcher is the most recent version. (Perhaps this is due to the training of many computer scientists as programmers, for whom different priorities apply: old versions, for example, are usually of limited value, and, in marked contrast to a scientist, a programmer whose code is working is unlikely to investigate *why* it is working. An interesting question is why computer scientists do not as undergraduates receive the kind of training in scientific method that is compulsory in other disciplines. Part of the answer might be that other undergraduates will use such methods as either professionals or researchers; which begs the question of whether professional computer scientists too would benefit from training in method.) Many experimental computer scientists can empathise with comments such as “we still have the code, but are not sure how to use it”, “we’re not sure which version we used”, or even “we can’t find the code anymore”. It would be in only a fraction of cases that a researcher would be able to find the original output from a run reported in a paper. In some disciplines, work is not published unless the supporting data is publicly available; in contrast, in computer science many researchers strongly resist publication of code or data. These observations are anecdotal, but, I believe, not unfair.

(Some departments and research institutes do of course encourage good practice. It is not that low standards are universal, but that there are no generally-accepted guidelines for conduct of research in computing.)

It might be argued that records are kept, in effect, by the automatic mechanisms of dumps and backups, but these mechanisms are not by themselves adequate: they store information indiscriminately and, once stored, it is effectively inaccessible. Even where versions are explicitly kept it can be difficult to determine which was used for the results in a particular paper.

The current lack of standards is, I believe, not acceptable. It is inconsistent with the expectations of the wider scientific community, in breach of published guidelines, and encourages publication of poor research. Some record-keeping is essential; it is the form of the records that is debatable.

Record-keeping practices should be designed to meet the needs listed earlier. They should be reasonable (not an intolerable bureaucratic burden, since we wish to restrain the wicked rather than hamper the good) and widely accepted, so that record-keeping of a lower standard can be condemned by the community rather than by a controversial rule-book; guidelines must have consensual support if they are to have any authority. They should also be appropriate to the research activity; different kinds of work will require different records. Thus, for example, the conventional method of record-keeping with written notebooks is often not a good solution. Copying out results, or printing them and pasting them into a book, is ill-fitted to the normal routine of research in computing. Nor do notebooks lend themselves to central record-keeping.

It is therefore impossible, and certainly inappropriate, to prescribe a fixed rigorous standard for record-keeping. What can be established, however, are: understanding of the needs that records must fill; acceptance that the records themselves should be verifiable; and examples of "normal" record-keeping practice that researchers can fit to their work. The core principle of record-keeping should, arguably, be the creation of adequate *corroborating testimonies*. If one kind of record by itself completely documents the research, then no other records are required; but in the more typical case several kinds of record support each other. In some research, the best evidence is a carefully-maintained notebook with dated, witnessed pages; in other research, different evidence is appropriate.

In my view, published results in computer science should usually be based on three separate, mutually supporting elements—notebooks, code, and logs.

Notebooks can be used to record dates; daily notes; names and locations of code, scripts, input, and other files; important references and web addresses; minutes of discussions; bug reports; locations and identifying marks of paper records; experimental parameters; and intent, outcomes, and interpretation of experiments.

Such notes, well-maintained, can provide a "guidebook" to the experiments. They should contain descriptions of ideas and show the progress of the research. Simple activities should not involve time-consuming record-keeping, but it is reasonable to expect a few lines to describe the intent of a particular run and its outcomes, and rather more description of more substantial activities. However, considered as a fraction of the total research program (conducting the experiments, analysing the outcomes, and writing a paper), the additional effort involved in maintaining a notebook should be small.

Code is obviously required if the experiments are to be run again, and at an absolute minimum researchers should preserve the exact code used to yield any published results, and if possible the exact input. Although the code may not work at a later date, today's systems are by historical standards stable so that such problems are becoming less common, and moreover discussions of past research often centre on the details of an implementation rather than the output it produced.

What is less clear is how many versions of the code should be retained; it is not uncommon for a researcher to change code in trivial ways dozens of times in a day. What is important in such cases is for the notebook to discuss the kinds of changes that were made and why; if the changes are small enough to be quickly made by a competent programmer, and are documented in notebooks, there is no need to keep every distinct version. But major versions could be kept—the versions that embody a significant change to the apparatus of the experiment—with documentation indicating which version was used for which experiment.

Logs should be complete transcripts of the output of each experiment. By this I do not mean the vast reams of numbers generated by some experimental software (although it may be valuable to keep these numbers), but the data as reduced by some process for human consumption—whether a summary table, list of averaged values, or whatever.

As discussed earlier, it is inappropriate to keep only the output from some of the runs [Wilson, Jr., 1952], if only because the process of selecting which runs to keep presents the possibility of investigative bias. Any selection should take place prior to a run, not afterwards. For the same reason, unsuccessful variations of the code should be kept. There are even good reasons for keeping versions with known bugs—for example, prior to their detection these bugs may have been a factor in experimental results.

There are several reasons why record-keeping based on these elements is appropriate, and perhaps a bare minimum. First, it is not particularly onerous; while keeping logs, for example, may require a little work as the experiments are conducted, they can save work later on. Second, this material is useful to write-up, and greatly simplifies the typical "wake-up" stage that occurs when the research is revisited after a break. Third, these records meet the aims for record-keeping: reproduction, verification, rigour, and so on. The dates in the notebooks will match the creation or modify dates of the code, and if there is a centralised dump mechanism these dates will match those of the dump tapes. The notebooks allow location of material on dump tapes, and the tapes provide a central, relatively trusted repository. Without centralised dumps, more traditional methods are required for verification, such as researchers providing materials to be centrally filed. In many environments mechanisms such as witnessing of notebooks is probably not required: sufficient corroboration of dates is available automatically.

(Dates and dump tapes can of course be forged or tampered with, but in an environment of shared machines such tampering is not straightforward. If stricter anti-fraud mechanisms are required then it may be necessary to resort to techniques such as verifiable timestamps or secure signatures, but at some effort and possibly with little gain. Some judgement must be made as to when an approach to record-keeping is sufficiently secure.)

Notebooks can be maintained electronically, but are trustworthy only where there are external mechanisms for verifying them, such as timestamps or centralised dumps. For example, the progressive versions of a well-maintained "e-notebook" (as captured at intervals on a dump tape or stored in a source control system) should show the same sort of development as a written notebook, with careful dating of every entry and material added but never changed or erased. Note that electronic versions of written documents need not be a single file; an e-notebook could for example be a more accessible structure such as a hierarchy of web pages with an entry per day.

A natural extension of keeping records online is that they can be made publicly available, by request or via mechanisms such as ftp or the web. In my opinion good reasons are needed to not make code available—publication of code shows that researchers have high confidence in their results, and, for the community as a whole, reimplementing algorithms is a waste of resources. The principle of publishing code is accepted, for example, by the ACM Journal of Experimental Algorithmics, where code and data are required to substantiate results.

6 Summary

Standards for recording research are not high in computer science. Both the published national standards for scientists and practice in other disciplines suggest that we should be taking more care to capture the day-to-day progress of our experiments. As a discipline we need to develop agreed standards and practices for conducting and recording our research. Such record-keeping need not be a burden: it encourages experimental rigour and can reduce the effort of producing finished research.

To provide a basis for debate about standards I have outlined the issues that standards must address. Computer scientists sometimes make the error of urging a technical solution before the requirements of the problem are well understood. In this paper these requirements have been used to argue that records should include all major versions of code, output of runs, and notebooks recording the progress of the experiments.

At conferences such as ACSC I have often heard researchers debate the question of whether computer science is a science—a question which as it stands is probably meaningless, as it depends on highly individual interpretations of what "science" is. However, given that "much suggests that the paradigms of modern science are appropriate for computer science" [Stewart, 1995], a more pertinent question is whether computer science research adheres to scientific standards. Too often, the answer to that question is “no”. Adoption of better experimental practice will help to change that answer.

Acknowledgements

Thanks to Tim Bell, Lawrence Cavedon, Vic Cieselski, Philip Dart, Michael Fuller, James Harland, Andrew McDonald, Alistair Moffat, Lin Padgham, Zoltan Somogyi, and Hugh Williams, for their comments and opinions, some of which are reflected in this paper.

References

AVCC (1997). Joint National Health & Medical Research Council/Australian Vice-Chancellor’s Committee statement and guidelines on research practice. Available at http://www.avcc.edu.au.

Dingell, J.D. (1993). Shattuck lecture—Misconduct in medical research. *The New England Journal of Medicine*, 328:1610–1615.

Frankel, M.S., editor (1995). *Professional Ethics Report*, volume VIII(2). American Association for the Advancement of Science.

Goodstein, D. (1995). Conduct and misconduct in science. California Institute of Technology. Public address, available at http://www.caltech.edu/˜goodstein.

Gower, B. (1997). *Scientific methods: an historical and philosophical introduction.* Routledge, London.

NAS (1992). Responsible science: Ensuring the integrity of the research process. National Academy of Sciences, Washington, D.C., Panel on Scientific Responsibility and the Conduct of Research.

NAS (1996). Open letter to Dr. William Raub. National Academy of Sciences, Washington, D.C.

Porush, D. (1995). *A Short Guide to Writing About Science.* Harper-Collins, New York.

Radder, H. (1995). Experimenting in the natural sciences. In Buchwald, J.Z., editor, *Scientific practice: theories and stories of doing physics.* University of Chicago Press, Chicago.

Ramamohanarao, K. and Zobel, J., editors (1997). *Proc. Twentieth Australasian Computer Science Conference*, Sydney, Australia. Australian Computer Science Association.

Ryan, K.J. (1995). Integrity and misconduct in research: report of the Commission on Research Integrity.

Stewart, N.F. (1995). Science and computer science. *Computing Surveys*, 27(1):39–41.

Wilson, Jr., E.B. (1952). *An Introduction to Scientific Research.* Dover, New York.

Ziman, J. (1968). *Public Knowledge.* Cambridge University Press, Cambridge.