Querying XML: benchmarks and recursion
Afanasiev, L.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
One important aspect of experimental studies is their repeatability. The aim of this chapter is to determine and improve the repeatability of the experimental studies in the database field. We report on a repeatability review that was conducted for the research articles submitted to the conference of the Special Interest Group On Management of Data (SIGMOD) of the Association For Computing Machinery (ACM) in 2008 [ACM, 2008].

From the repeatability review, we learn a number of lessons about how to describe experimental studies in order to ensure their repeatability; these lessons are implemented in XCheck, a software tool for executing XML query benchmarks. We present XCheck in Chapter 5.

This chapter is based on work previously published in [Manolescu et al., 2008a].

4.1 Introduction

Repeatability is a property that allows experimental studies to be repeated and their results reproduced. This property is at the basis of the experimental sciences and it has two functions: (i) it guarantees the consistency of experimental results and thus, the soundness of the observations and conclusions drawn; and (ii) it makes experimental results available for comparison with other research, which facilitates scientific proliferation. A more precise definition of repeatability follows. For experimental studies conducted in computer science repeatability is equally important [Jain, 1991].

Performance of database management systems (DBMSs) is a key research topic in the database community, while experimental studies are the main performance evaluation method. Nevertheless, in the last decade or more, the field has suffered from a lack of standard methodology for conducting and reporting experimental studies [Manolescu and Manegold, 2007, 2008]. In particular, the
repeatability of experimental studies is not ensured. We believe that this affects the quality and progress of database research.

This leads us to the following question:

**4.1. Question.** How to ensure the repeatability of experimental studies of database systems? This question incorporates two sub-questions: (i) what is a proper methodology for designing and reporting on experimental studies that facilitates their repeatability? (ii) what is a proper mechanism for evaluating the repeatability of experimental studies presented in scientific research?

Before we present our attempt to answer these questions, we specify what repeatability means exactly. In database research, many experimental studies use measures that are dependent on the experimental environment. For example, performance evaluation measures, such as the execution time of a database system, are hardware dependent. Therefore, when we talk about repeatable experiments we cannot compare directly the performance measurements obtained in similar, but not identical, experimental environments. We define *repeatability* as the property of an experimental study that allows it to be repeated in a similar environment and ensures that the experimental results lead to the same observations and, ultimately, to the same conclusions. This definition implies that, depending on the given experimental study, assessing its repeatability might require human judgement and thus can not always be automated.

In order to achieve repeatability, obviously, the components of the experimental studies, such as the test data and the system under test, should be available or reproducible. There should also be a detailed description of the experimental setup and results. Further, reporting the environmental parameters that influence the experimental results completes the requirements for achieving repeatability. Assembling a standard list of such parameters might help to establish a common practice of reporting their values and impact. With respect to a quality control mechanism, we expect that it can be achieved via a peer-reviewing system, in the same style as the one used to ensure the quality of research papers. As mentioned earlier, requiring human judgment might be unavoidable in an unrestricted domain of experimental studies. The experimental data can be archived in internet-based repositories associated with scientific publication venues.

In this chapter, we report on an attempt at achieving and measuring the repeatability of experimental studies along the lines described above. The key aim of this study is understanding the challenges we face in achieving and promoting the property in database research. Towards this goal, a repeatability review process of the research papers of SIGMOD 2008 [ACM, 2008] was conducted. SIGMOD is an important conference in the database community and many of the submitted articles contain experimental results. The authors were asked to submit, together with the articles, a package containing information about the conducted experimental studies. A committee then assessed the experimental
repeatability. The participation in the repeatability review was optional. Nevertheless, 2/3 of all paper submissions attempted a repeatability review. This was the first attempt to measure the repeatability of experimental studies in the database community. We present the results of this experiment and discuss the lessons it teaches us. We enumerate problems that need to be addressed in order to make this a common practice.

The structure of this chapter is as follows. In Section 4.2, we present the setup of the SIGMOD 2008 repeatability review, including the schema used for describing the experiments and the evaluation protocol. In Section 4.4, we present the results of the reviewing process and the problems the repeatability committee encountered. In Section 4.5, we present an author survey targeted at getting insights about the usefulness of the reviewing process. Finally, in Section 4.6, we discuss the positive and negative sides of this approach and conclude.

4.2 SIGMOD repeatability review setup

For the first time in the history of the conference—SIGMOD has been running since 1975—SIGMOD 2008 introduced an experimental repeatability reviewing process. The goal was to evaluate whether the experiments presented in the submitted articles are repeatable by fellow researchers. A longer term goal was to develop a standard for reporting repeatable experimental studies. The authors of submitted articles were asked to assemble and submit a package containing information about the experimental studies, including the experimental data, a description of the experimental studies and environment. Based on that information and the results presented in the article, a committee formed of volunteer researchers assessed the repeatability.

Participation in the repeatability reviewing process was optional, that is, the submission and the results of the repeatability reviewing process did not have any influence on the paper acceptance to SIGMOD 2008. Nevertheless, if the authors decided not to participate in the repeatability review, they were required to submit a note stating the reasons for not participating. Some papers, especially those submitted to the industrial track of the conference, present research involving Proprietary Data (PD) and/or Intellectual Property (IP), not available for distribution. Therefore, they can not participate in the repeatability review. We consider these reasons as valid reasons for not participating.

The repeatability was evaluated per experiment presented in the paper. If the experiment was successfully executed following a description provided by the authors and if the obtained results led to the same conclusions as presented in the paper, then the experiment was considered repeatable. In Section 4.3, we present the protocol that was used for reporting the experiments and in Section 4.3.1 we present the protocol that was used for assessing the repeatability.

Participation was encouraged by allowing the articles with successfully eval-
uated experiments to include a note that acknowledged the repeatability of their experiments. If only a subset of the experiments could be verified, then the acknowledgement included references to those experiments. If the verified code and data was available for free distribution, the repeatability acknowledgment note contained an URI pointing to where they were stored. More importantly, all the papers that submitted their experiments for repeatability review, got detailed feedback on any problems that arose during the review. The official call for participation in the review can be found at [http://www.sigmod08.org/sigmod_research.shtml](http://www.sigmod08.org/sigmod_research.shtml).

The repeatability program committee consisted of 8 volunteers from the database research community, chaired by Ioana Manolescu and including the author of this thesis. In order to assure independence of the regular review from the repeatability review, the committee did not participate in the regular reviewing process of SIGMOD 2008. The reviewing process was double blind, i.e., the committee members did not know the identity of the paper authors and the authors did not know the identity of the reviewers evaluating the repeatability of their experiments.

### 4.3 Describing experimental studies

In this section, we present the protocol for describing experimental studies used in the SIGMOD 2008 repeatability reviewing process.

The authors were asked to provide a package containing the following information: 

(i) the PDF file containing the anonymized article, 
(ii) the software and data needed to run each experiment subject to the repeatability review, and 
(iii) experiment descriptions that contain all the details necessary for repeating the experiments, e.g., required hardware, software, instructions to install the software, to run experiments.

To achieve the repeatability of experimental studies, the authors need to describe how to set up and execute the experiments. It is also important to determine and list all the environment parameters that influence the results. Though each experiment has different requirements for repeatability, there are generic steps and parameters that the authors need to describe. These include:

- the machines used for the experiments, their hardware and software specifications;
- various third-party software (other than the software being tested) required for running the experiments;
- a detailed description on how to set up the experiments;
- a detailed description on how to execute the experiments; and
Figure 4.1: The DTD containing a schema for describing the experiments presented in a paper for the purpose of repeatability. This schema was used for the repeatability review of SIGMOD 2008.
any comment regarding the repeatability of the experiments, including a list of all the parameters that might influence the results.

The repeatability committee designed a schema for describing experimental studies. The schema is meant as a guideline for presenting experiments and the submissions are meant to be read by humans and are not for automatic processing or execution of the experiments. It is a first step towards standardizing the description of experimental studies for the purpose of repeatability.

The authors were required to submit their descriptions in XML format conforming to the DTD presented in Figure 4.1. Below, we describe a subset of the elements defined by this DTD. We omit the self-explanatory elements.

- The paper element contains an identifier of the paper being tested within the submitted package.
- The machine element describes an individual computer used in the experiments. It consists of hardware and operating system specifications.
- The hardware element describes the machine’s hardware. It is important to describe the resources actually used for the experiments. For instance, the experimental machine can have 2GB of RAM but if the software under test used only 1GB, then 1GB should be indicated here.
- The os element describes the operating system and references to any auxiliary software (e.g., compilers, libraries, database systems). If the experiment under test uses a virtual machine, it describes the characteristics of that virtual machine.
- The windows, linux, and mac elements describe the most common operating systems. The otheros element can be used to describe other operating systems.
- The software element describes any software (other than the code under test) needed in order to run or compile the code under test, for example, compilers used to compile the code, libraries, simulators, software for analyzing the data and producing graphs. If the experiment under test was run using a virtualization platform, then one software element can describe the virtual image player, and another software element can describe the virtual image used.
- The function element describes the functionality of the software in the experimental setting. This allows for the replacement of the software by another one with similar functionalities, in case needed.
- The downloadURL element specifies an URL where a free or trial version of the software can be found.
4.3. Describing experimental studies

- The elements network and machineset can be used for experiments to be run on several machines. The machineset element specifies a set of identical machines: @mkey is a foreign key pointing to an existing machine element; @howmany specifies how many machines of this kind are part of the machine set.

- The network specifies a set of machines used in an experiment. The network may include homogeneous subsets of identical machines, each such subset being described by a machineset element. If all machines are different, 1-sized machine sets can be used.

- The experiment describes an experiment that can be repeated: the @id1 attribute contains the unique id of the experiment (typically “Figure X” or “Table Y”); the optional attribute @id2 is a secondary id (typically “left” or “top”), if any; and, @hardwkey is a reference to either a machine, or a network, depending on the hardware used for this experiment.

- The dataset element specifies the dataset(s) used in the experiment under test. It can contain: a path leading to the data file(s) within the submission package, a download URL, or the information necessary to produce it using a data generator.

- The install element contains the installation instructions for the software needed to repeat the experiment, mainly the software under test.

- The howto element contains the sequence of steps to be performed in order to run the experiment. It may include: any necessary pre-processing steps; the commands for performing the experiment; the number of runs, if any; the operations needed to gather and analyse the results; the production of graphs corresponding to those of the paper. The ideal description of a given step is the command line needed to run it.

- The nonrepexperiments describes an experiment that cannot be repeated due to IP rights or other issues. It contains a primary and possibly secondary key as for repeatable experiments.

All the experiments submitted for the repeatability review at SIGMOD 2008 were described using this schema.

4.3.1 Assessing repeatability

For assessing repeatability the committee members adhered to the following steps:

1. Checking whether the repeatability submission package conforms to the requirements given in the previous section;
2. Matching the hardware and software requirements with available machines;

3. Installing the necessary software, setting up the experiments, estimating the total time needed to execute them;

4. Running the experiments. If the execution takes considerably longer than the estimated time, it is considered failed; and

5. Comparing the obtained results with the results presented in the paper. For this step the following procedure is used:

   • If the performance measure deployed in the experiment is deterministic (e.g., measuring the size of data storage in a DBMS) then the committee member compares the absolute values of the obtained results and the results presented in the article.

   • If the measure deployed in the experiment is non-deterministic (e.g., measuring the query execution times) and the obtained results cannot be compared directly with the results presented in the article, then the committee member judges, case by case, whether the results support the observations and the conclusions made in the article. For example, a statement such as “Engine 1 performs twice as fast as Engine 2” can be confirmed or refuted even if the absolute measurements are different.

An experiment is considered repeatable if all these steps were successfully completed. Otherwise, a report is submitted to the authors indicating at what step the assessment failed and how it failed.

Note that the assessment made during the second part of Step 5 might be subjective. To improve the evaluation accuracy, it would have been preferable if two or more reviewers would discuss and agree upon the final assessment. Due to lack of time and (non) availability of suitable experimental environments, it was not possible for each paper to be reviewed by multiple reviewers. Instead, each paper was assessed only by one reviewer, who discussed the assessment with the rest of the committee only when in doubt about its correctness.

4.4 Results

In this section, we present the results of the SIGMOD repeatability review. First, we describe the timeline of the repeatability reviewing process and we give the submission statistics. Then we present the repeatability assessments.

Participation

Nov 16, 2007: SIGMOD’s paper submission deadline. There were 436 paper submissions.
4.4. Results

Table 4.1: Participation in the SIGMOD 2008 repeatability reviewing process. The rows contain the statistics for the accepted papers, for the rejected papers, and for all submitted papers. The parentheses contain the absolute numbers.

<table>
<thead>
<tr>
<th>Category</th>
<th>Submitted experimental descriptions</th>
<th>Submitted valid excuses</th>
<th>No submission or invalid excuses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accepted papers (78)</td>
<td>68% (53)</td>
<td>31% (24)</td>
<td>1% (1)</td>
</tr>
<tr>
<td>Rejected papers (358)</td>
<td>66% (236)</td>
<td>23% (82)</td>
<td>11% (40)</td>
</tr>
<tr>
<td>All papers (436)</td>
<td>66% (289)</td>
<td>24% (106)</td>
<td>9% (41)</td>
</tr>
</tbody>
</table>

Dec 16, 2007: The repeatability review submission deadline. For 289 papers an experimental description was submitted; for 106 papers valid reasons (deploying PD or code under IP rights) for not participating were submitted; the authors of the remaining 41 papers submitted reasons that the committee considers invalid (e.g., losing the experimental code and data) or did not submit any reason at all.

Jan 2, 2008: The regular program committee disclosed to the repeatability program committee a list of 82 papers with high chance of acceptance (2–3 positive reviews). Due to time constraints, the repeatability committee did not verify all of the papers that were submitted for repeatability review. The priority was given to papers with high chance of publication.

Feb 22, 2008: The SIGMOD program committee announced the list of 78 accepted papers. Then, the repeatability committee focused on these papers.

Mar 20, 2008: Camera ready submission deadline. By this date, 64 papers covering the accepted papers that participated in the repeatability review, were reviewed and the results were communicated to the authors.

Table 4.1 shows the distribution of participation in the repeatability reviewing process for all submitted papers, for the accepted papers, and for the rejected papers. Note that in the accepted papers category only 1% of the papers did not participate at all in the repeatability reviewing process, while in the rejected papers category 10%. This is easily explainable by the usual percentage of unfinished or invalid papers submitted to a conference. The rest of the percentages are quite similar for the accepted and rejected categories. Considering that this repeatability review experiment was conducted for the first time in the database community and on a voluntary basis, we find 66% participation rate a strong indication of the importance and usefulness of testing repeatability.

The most common invalid reasons the authors presented for not participating in the repeatability reviewing process, are the loss of the experimental data and third party software used in the experiments. Below we include a few quotes from the reasons provided.
Chapter 4. Repeatability of Experimental Studies

Table 4.2: Repeatability assessment results. The first two rows contain the statistics for the assessed papers split between the papers that were accepted for the conference and those that were rejected. The last row contains the aggregate statistics for all assessed papers. The parentheses contain the absolute numbers.

<table>
<thead>
<tr>
<th></th>
<th>All repeated</th>
<th>Some repeated</th>
<th>None repeated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accepted papers (53)</td>
<td>55% (29)</td>
<td>26% (14)</td>
<td>19% (10)</td>
</tr>
<tr>
<td>Rejected papers (11)</td>
<td>36% (4)</td>
<td>55% (6)</td>
<td>9% (1)</td>
</tr>
<tr>
<td>All assessed papers (64)</td>
<td>52% (33)</td>
<td>31% (20)</td>
<td>17% (11)</td>
</tr>
</tbody>
</table>

“We cannot distribute code and data because the authors have moved, making the retrieval of code and data infeasible at this point.”

“We lost some old code. Due to the short notice, we could not reproduce our lost code for these parts.”

“The subsets were chosen randomly from a large dataset, and unfortunately no trace about the identity of the used documents has been kept. The experiments were performed long months ago, and it wasn’t expected to send results to SIGMOD, that’s why we didn’t pay attention about keeping a trace.”

These quotes indicate that the experiments were not conducted with repeatability in mind, which we consider bad practice. Some comments hinted at some misunderstandings of the purpose of the repeatability assessment:

“My experimentation is fully deterministic: if it is wrong, running again my own program would not detect it.”

These quotes underscore how useful assessing repeatability could be for the integrity of our field.

Assessment results  Due to lack of time, only some articles were assessed by two committee members, namely those that required a second opinion for the final assessment.

Table 4.2 contains the results of the repeatability reviewing process. The first two lines present the statistics for the papers that were accepted for publication and for those that were rejected. The last line presents the total. Out of the 64 repeatability assessed papers, 53 papers were accepted to the conference and 11 papers were rejected. Out of the total of assessed papers, for 33 papers (52%) all experiments were repeated, for 20 papers (31%) some of the experiments were repeated, and for 11 papers (17%) none of the experiments were repeated. The percentage of papers of which all experiments were repeatable is slightly higher for the accepted papers than for the rejected papers.

Among the 11 papers with no repeated experiments, 3 required hardware unavailable to the repeatability committee, 2 required unavailable software, the
installation of the necessary software failed for 1 paper, and 5 papers had various runtime failures that prevented the completion of the experiment.

**Assessments costs** One committee member, the author of this thesis, recorded how much time it took her to verify the repeatability of 8 articles. The total amount of time needed for the whole assessment procedure, i.e., steps 1–5 presented in Section 4.3.1, varied from 7 hours to 180 hours. This time includes the running time of the experiments (step 4), which is experiment dependent. Without counting the running time of the experiments, the average amount of time that the reviewer spent on the evaluation of an article is 5 hours. Note that these are not continuous work hours, but rather aggregations of different work sessions. Usually, half of this time was spent on setting up the experiments and the other half on understanding and interpreting the results.

On average, each committee member assessed 8 articles. Supposing that every committee member needs the same amount of time for assessing a paper, then the total amount of time spent on the assessment is 40 hours per person and 320 man-hours in total.

Since the repeatability review has been organized for the first time, a large amount of time was spent on the setup. In the future, this time can be reduced by building an automated system for collecting and distributing the articles among the reviewers.

### 4.5 Authors survey

After the repeatability evaluation was completed, the authors of the papers were asked to participate in a short survey. The purpose was to get insights into whether the SIGMOD repeatability reviewing process was useful, whether it should be continued, and in what format. In this section, we present the results of this survey.

The survey participation request is given in Figure 4.2. The survey contained four questions. We discuss only the first three, the fourth question is outside the scope of this chapter. The first question recorded the authors’ participation to the assessment (did not participate; participated and all experiments repeated; participated and some experiments repeated; or, participated and none of the experiments repeated). The second question asked the authors if they found the process useful, and elicited suggestion on how to improve it. The third question asked the authors whether they would participate in a repeatability reviewing process in the future SIGMOD conferences, assuming it would remain optional. The results of the survey were anonymized to encourage the authors to speak their mind.

Most answers were clear Yes/No answers. Less than 20% of the answers were ambiguous, in the style of “Yes and no; on one hand...but on the other hand...”
Chapter 4. Repeatability of Experimental Studies

This is meant to be a sub-5 minute survey about experimental repeatability. In the case of multi-author papers, only one of you needs to answer (though we are happy to receive comments from more than one). We will strip your email headers from your responses programmatically, so please speak your mind.

1. Did your paper succeed on all/some/none of the repeatability tests? Or did you not submit for intellectual property reason?

2. If you submitted, was the repeatability experience helpful? If so, how? If not, how could it be improved?

3. Would you attempt repeatability in the future if it remained voluntary (i.e. had no effect on acceptance decision but you would be allowed to mention success in your paper) and you had no intellectual property constraints?

4. Do you think it would be useful to have a Wiki page for each paper so the community could comment on it, you could post code etc.?

Warm Regards, Ioana (repeatability chair) and Dennis (program committee chair)

Figure 4.2: The request for participation in the authors’ survey.

For such answers, half a point was counted.

Table 4.3 presents the number of positive answers to the second and third questions about the usefulness of the repeatability review. The results are grouped according to the answer to the first question about the authors’ participation to the repeatability assessment and their results. Most of the participants to the review (80%) found the repeatability process useful, though the participants with all experiments repeated were more positive (85%) than the ones with none of the experiments repeated (56%). Even more participants (84%) reported that they would participate in the repeatability reviewing process at future SIGMOD conferences. This indicates that the authors believe that such a repeatability review is useful.

Below we present a list of representative quotes extracted from the authors’ answers about the usefulness of the process.

“Yes, it was helpful to organize the source code properly for future use.” (All experiments repeated.)

“It was helpful. It forced me to write documentation which I would otherwise have postponed indefinitely.” (Some experiments repeated.)

“It was helpful. It required us to further clean up my code and scripts and prepare documentation.” (Some experiments repeated.)

“Helpful? Greatly yes. Some scripts written for this test could be used to append additional experimental results immediately. To package experiments in a script form, at first, seemed bothersome, but we
4.6 Lessons learned and conclusions

In this chapter, we pursued Question 4.1 “How to ensure the repeatability of experimental studies of database systems?” The SIGMOD 2008 repeatability reviewing process provides a solution and shows that repeatability is possible to achieve and measure.

Referring to Question 4.1 (i), the methodology used for describing and reporting experiments that the repeatability committee developed was enough to cover the 289 papers that were submitted for the repeatability review. Out of 64 papers that were assessed by the repeatability committee, 33 (52%) achieved the repeatability of all presented experiments and 20 (31%) achieved repeatability of some of the experiments. Considering that we strive for all experiments to be

<table>
<thead>
<tr>
<th></th>
<th>Number of participants</th>
<th>Found it useful</th>
<th>Would do it again</th>
</tr>
</thead>
<tbody>
<tr>
<td>Did not participate</td>
<td>16</td>
<td></td>
<td>12 (75%)</td>
</tr>
<tr>
<td>All experiments repeated</td>
<td>24</td>
<td>20.5 (85%)</td>
<td>21.5 (90%)</td>
</tr>
<tr>
<td>Some experiments repeated</td>
<td>12</td>
<td>10 (83%)</td>
<td>11 (91%)</td>
</tr>
<tr>
<td>None of the experiments repeated</td>
<td>8</td>
<td>4.5 (56%)</td>
<td>6 (75%)</td>
</tr>
<tr>
<td>Total</td>
<td>60</td>
<td>35 (80%)</td>
<td>50.5 (84%)</td>
</tr>
</tbody>
</table>

Table 4.3: Authors survey on the usefulness of the repeatability review at SIGMOD 2008. The survey participants are split in 4 categories depending on their participation and their results in the repeatability review. The last two columns contain the counts of positive answers to the respective questions. A half point (0.5) was given to the answers of the form “Yes, and No.”

found out that it is good for ourselves, and improves our productivity.” (All experiments repeated.)

“It is a great thing for the community that this service is available, and I hope that it will have a very positive effect on both the trustworthiness of SIGMOD results and the quality of publicly-available research tools.” (All experiments repeated.)

“It’s only helpful in the sense that it provides some extra credibility to the paper. It was not helpful to myself in any way.” (Some experiments repeated.)

“We are happy to see that our algorithms show consistent results through machines with different hardware/software configuration.” (All experiments repeated.)

In the next section, we list the lessons learned and conclude.

4.6 Lessons learned and conclusions

In this chapter, we pursued Question 4.1 “How to ensure the repeatability of experimental studies of database systems?” The SIGMOD 2008 repeatability reviewing process provides a solution and shows that repeatability is possible to achieve and measure.

Referring to Question 4.1 (i), the methodology used for describing and reporting experiments that the repeatability committee developed was enough to cover the 289 papers that were submitted for the repeatability review. Out of 64 papers that were assessed by the repeatability committee, 33 (52%) achieved the repeatability of all presented experiments and 20 (31%) achieved repeatability of some of the experiments. Considering that we strive for all experiments to be
repeatable, 52% is a low fraction. Nevertheless, we consider these results to be a good start towards achieving the repeatability of experimental studies in the database research field.

Referring to Question [4.1] (ii), the high percentage of participation in the optional review, 66% of the total submissions to the conference (289 out of 436), hints at the usefulness of a peer reviewing process. The positive feedback from the authors of the papers recorded by the survey also indicates that such a review is useful for the community: 80% of the surveyed authors found the process useful, while 84% would participate in such a process in the future.

Lessons learned  Our experience with the SIGMOD experiment showed that there are a few problems that need to be addressed, if ensuring repeatability of experimental studies is to become a common practice in the database community.

The first problem is the effort the authors need to undertake to ensure the repeatability of their experiments. Currently, due to a lack of common practice, achieving repeatability is tedious and time consuming. A software tool can be very useful for automating (parts of) this task. In Chapter 5 we present a tool for automating benchmark tasks in the context of XQuery, including the automated recording of the testing environment specifications for the purpose of repeatability. More such tools are needed.

The second problem is the tremendous amount of handwork that the reviewing committee had to do in order to set up and verify the repeatability. We estimated that it took 320 (40 × 8) human hours for the assessment process alone. It is not realistic to pursue such a review in the future without a proper framework that would facilitate the task. We also expect that the assessment will become easier as the practice of presenting experiments in a way that facilitates repeatability will increase.

A third problem is reducing the number of papers that claim non-repeatability of their experiments due to legal reasons. At SIGMOD 2008, 24% of the paper submissions could not attempt the repeatability review because they deploy PD and/or use software under IP rights that does not allow free distribution. This number can be reduced by designing and using openly available benchmarks. In Chapter 3 we show that the standard benchmarks are rarely used for experimental studies, while it is possible and advisable to design synthetic data and workloads that emulate the properties of real data [Jain, 1991].

Future work  One aspect of repeatability that we did not discuss in this chapter is what is a proper archiving mechanism for ensuring the accessibility of experimental data and results. Long-term preservation and even experimental results curation is another key factor of scientific proliferation. This question is actively being addressed in other database related fields, such as Information Retrieval [Agosti et al., 2007]. We leave this as future work.