

Record and Re-use in eScience

Chris Greenhalgh, 2004-01-29

[Input paper to EQUATOR record and re-use workshop]

Introduction

This note points out some typical examples of record and re-use-type activity in eScience, with examples from current projects where appropriate. This is intended to complement the other applications (e.g. gaming and entertainment) that have been demonstrated in the MRL's previous work on record and replay.

What is e-Science?

e-Science has been defined by John Taylor (Director General of the UK Research Councils) as,

global collaboration in key areas of science and the next generation of infrastructure that will enable it.

...

e-Science envisages that large scale science will be increasingly carried out in distributed global collaborations enabled by the Internet. A feature of these collaborations is that they will require efficient access to very large data collections and very large scale computing resources and will use distributed visualisation to support a high-level of user access.

Notable examples of e-Science applications are the Large Hadron Collider at CERN and the exploration of the Human Genome and related Life Science data resources.

[<http://www.lesc.ic.ac.uk/admin/escience.html>]

To a first approximation we can consider any use of information and communication technologies to support the conduct of scientific activities.

The (e)Scientific Method

1. The scientific method has four steps

1. Observation and description of a phenomenon or group of phenomena.

2. Formulation of an hypothesis to explain the phenomena. In physics, the hypothesis often takes the form of a causal mechanism or a mathematical relation.

3. Use of the hypothesis to predict the existence of other phenomena, or to predict quantitatively the results of new observations.

4. Performance of experimental tests of the predictions by several independent experimenters and properly performed experiments.

If the experiments bear out the hypothesis it may come to be regarded as a theory or law of nature (more on the concepts of hypothesis, model, theory and law below). If the experiments do not bear out the hypothesis, it must be rejected or modified.

[http://teacher.nslr.rochester.edu/phy_labs/AppendixE/AppendixE.html]

The major potential uses of record and re-use/re-play revolve around the experiments referred to in step 4, above.

1. Supporting Repeatability

The scientific method presumes that the experiments performed in assessing a hypothesis are repeatable, both the initial experimenter and by others. Recording of these experiments can help those experiments to be reproduced in various ways. For example, the various parameters and devices used in setting up the experiment and their organisation and coordination may be recorded, to assist in reconstructing the experimental system and conditions.

2. Supporting the Hypothesis

Acceptance of the hypothesis relies on acceptance of the validating experiments and their results in so far as they bear upon the hypothesis in question. Being able to record, publish and replay individual experiments can in itself be used as a vehicle to establishing trust in the results and hypothesis. However, the detail and accuracy of the recording will be important issues, as will the truthfulness of the recording (and whether this can be verified).

3. Weighing the Evidence

Related to this, subsequent analysis of the recordings may allow the same or another scientist to test for evidence of good practice, or signs of particular possible sources of error or other confounding factors.

4. Supporting Alternative Hypotheses and Hypothesis Formation

Recordings, i.e. “results” in an extended sense, from an experiment may be used to form and test other hypotheses after the event; they effectively form a data corpus which can be mined by researchers. For example, the results from CERN’s colliders and the forthcoming LHC are made widely available as a common resource. Other projects such as various national and international “virtual observatories” [<http://www.ivoa.net/>] are also making relatively “raw” results available to broader scientific communities.

5. Exploring Data Provenance

Recordings may be used to answer more general questions of the form “where did this data come from?” “how was it determined?” “who worked on it?” which may in turn be used to make individual assessments of “should I trust this data?”. Traditionally this is answered by e.g. a textual description of the experimental process which is first recorded in the scientists log book (if they keep one) and later summarised in a part of a scientific publication. In many cases, especially where significant elements of computational analysis are involved (as in bioinformatics), such summaries may not

contain enough detail to be directly repeatable. Recording and exploring provenance is a major theme of the myGrid e-Science project [www.mygrid.org.uk].

This – and many of the prior points – are also related to the notion of “publication at source”, as championed by e.g. the CombiChem eScience pilot project [e.g. <http://www.smarttea.org/>]. This considers the lab instrument data and logs to be a natural – and perhaps extension – to the traditional publication process: each paper has a virtual appendix which allows the readers to explore and test the “raw” experimental data from which the paper’s results were derived.

The Conduct of Science

Record and re-use is also relevant to the individual and social processes by which science is done as an everyday activity. It can potentially provide enhanced linkages in both time and space.

6. What did I do? (and why?)

In some settings (e.g. lab biology) it is normal – usually mandatory – to keep a detailed lab note-book recording the experiments performed, including both methods and results. In other settings the keeping of records is at best ad-hoc. Consequently it can be very hard for a scientist to return to a previous activity – which they can no longer remember clearly with their unassisted memory. For example, they may wish to write it up, check for some possible source of error, perform a minor variant experiment, or simply resume where they left off. In this context automated record and re-play of their everyday activities might allow to review and hence assess and/or resume those activities some time later.

7. What did they do? (and why?)

The same kind of issues apply even more strongly for subsequent (or remote) collaborators. The subsequent collaboration may be anticipated, or not.

The CoAKTing project reflects one aspect of this, specifically the record and replay/reuse of meetings within a broader collaborative process.

Issues

Many of the usual issues of sensing/recording:

- the “cost” of recording, in resources, effort, etc. versus the perceived (at the time) benefit to the one facilitating that recording
- privacy versus reusability/openness
- publishing (with selection, tailoring, polishing, etc.) versus “truthfulness”
- level of detail versus requirements (size, capture, etc.)
- issues of re-use – view, play, transform, partial, test, ...?

The dimension of “science” – “truth”, “reproducibility”, “fraud”, ... - and the potentially diverse questions that may be asked of the data – arbitrary data mining – make this a little different in flavour from e.g. Collaborative VEs for entertainment.