OPPORTUNITIES FOR AND BARRIERS TO ENGINEERING RESEARCH DATA RE-USE

TOM HOWARD, MANSUR DARLINGTON, ALEX BALL, CHRIS MCMAHON AND STEVE CULLEY

erim3rep100805tjh10.pdf

ISSUE DATE: 1 October 2010

Catalogue Entry

Title	Opportunities for and Barriers to Engineering Research Data Re-use
Creator	Tom Howard, Mansur Darlington (authors)
Contributor	Alex Ball, Chris McMahon, Steve Culley
Subject	Data re-use; confidentiality; intellectual property; software licences; hardware platforms
Description	A survey of researchers in engineering established that descriptive material (i.e. data that reports on and aids understanding of a situation) had more potential for re-use than prescriptive material (i.e. data used to identify shortcomings in a situation and suggest improvements). The latter may still be useful in validating research conclusions and as a starting point for further prescriptive studies. The major barriers to re-use were found to be confidentiality (amid concerns about anonymity, commercial sensitivity and negative publicity), the importance of data ownership to status in the research community, hardware and software licences, and the use of highly specialized technology. Many of these barriers could be reduced for future research provided they are tackled early in the research process.
Publisher	University of Bath
Date	5th August 2010 (creation)
Version	1.0
Type	Text
Format	Portable Document Format (PDF/A-1b:2005)
Resource Identifier	erim3rep100805tjh10
Language	English
Rights	

Citation Guidelines

Tom Howard, Mansur Darlington, Alex Ball, Chris McMahon and Steve Culley. (2010). *Opportunities for and Barriers to Engineering Research Data Re-use* (version 1.0). ERIM Project Document erim3res100805tjh10. Bath, UK: University of Bath.

1. INTRODUCTION

Motivated by the general drive toward accountability and efficiency in public sector activities, there is a developing interest in the ways that the data gathered and generated in publically-funded research might be made more available for use by the research community at large.

In addition to this, it is recognised that researchers themselves can benefit from and wish to have easier access to existing data (Beagrie et al., 2009) yet, because of poor management, social and commercial pressures and legacy sharing practice, such access is often not possible (Birnholtz & Beitz, 2003).

It is quite clear from the scoping work and data case audits discussed in Deliverable 1 (Ball, 2010) and Deliverable 2 (Howard et al., 2010) of the ERIM Project that many research data serve only the purpose associated with the research project for which the data were generated or gathered It is the goal of this research project to enhance the value of these data – that have often been gained through painstaking effort and some considerable expense – by providing means to support their *re-use* or *re-purposin*g (see definitions in Howard et al., 2010).

As an indication of cost, the authors estimate that the cost of the data gathering activity alone for the ERIM Project has been in the order of £14832 (see the Appendix for a breakdown). This figure is arrived at based on the time involved simply in the activities of collecting and generation of what would be classified as 'raw' data, by applying a day rate that is used to approximate industry in-kind contribution for a project engineer. This does not include the time spent by the researchers in preparation for the gathering activity, nor the development and management of the data to make it usable. It also does not include costs associated with the other forms of data generated and collected for the project literature review and the data generated by the researchers' discursive processes. Clearly then, in terms of resources expended and direct costs, this data is *valuable* and providing the means to support its re-use and that of data like it is important.

In order to support the general project goal of making engineering research data more re-usable, the ERIM Project researchers gathered data from the research activities and associated data taken from two research sources. The data gathered were the subject of a scoping survey (see Section 2 for details of sources). Amongst other things, information was elicited from the researchers whose data were being assessed regarding *opportunities for* and *barriers to* the re-use of their data. This report considers the findings of this aspect of the investigation. Firstly, we attempt to assess what, if any, are the potential opportunities for re-use of engineering research data (Section 3). This is perhaps the more fundamental question. If the research were to suggest that the research data had no potential for re-use then the research goal is null. However, if it is shown that the research data has potential for future re-use then we must assess the second issue concerning the barriers that currently or in principle would prevent the data from being re-used. These barriers are identified and discussed in Section 4.

2. SELECTING DATA REPRESENTATIVE OF ENGINEERING RESEARCH

The scoping survey mentioned above was designed to identify engineering research information and data from which a characterization of the broad spectrum of distinctly different engineering research data could be made

2.1 Sources of Data Records

The authors selected as a source of data for their study two repositories. The first of these repositories is constituted of the data assets held by researchers in the Innovative Design & Manufacturing Research Centre (I*d*MRC) located in the University of Bath. The data available is that associated with the I*d*MRC's research projects over a period of nearly a decade, a representative sample of which had already been subject to an audit using the DAF framework (Jones, et al, 2008).

The second repository is 'virtual' in that it consists of an inventory of data assets associated with a large-scale research project distributed geographically over eleven universities. The project in question (the KIM Project) was of a highly inter-disciplinary nature covering a range of engineering research topics looking at such diverse things as product modelling, document management, information archiving, information modelling and engineering standards, the nature of learning organisations and HR policy for product-service systems. The diversity of the subject matter is reflected in the data assets.

The assets were subjected to a selection process by which means twelve cases (encompassing about 50 different individual data record types) were selected which the authors believed display a good mix of research topics, methods, data size and data format, from research activities which themselves were representative of the data gathering and data generation processes to be found across engineering research.

3. OPPORTUNITIES FOR RE-USE

There are two forms of engineering research (Blessing & Chakrabarti, 2009). One is essentially prescriptive, that is the research provides an understanding of current shortcomings and then the means that will potentially improve an engineering situation and provide accompanying information so that it can be applied. The second is descriptive, in that it provides information that is descriptive of the engineering situation such that it can be analysed, characterized and generally better understood. Thus the outputs of both of these forms of research by definition have opportunities for re-use (if not usable then there is no point in the research). However, in both forms there are large amounts of intermediary information that is collected, generated and developed in order to reach the final, often published state (for evidence see Howard et al., 2010).

This section will discuss the research associated with identifying the opportunities for the re-use of this intermediary research data and information, as well as of the raw data that is to be found in all research activities. In particular an attempt is made to understand the characteristics of the data assets and the Research Activity Information Development (see Howard et al., 2010, for details of a generic approach to modelling this development process) that make the particular data assets more or less re-usable.

The following sub-section will first discuss the methodology used to investigate the opportunities to re-use. This will be followed by a discussion of interview results based on the four key questions and the observations about the data types and the Research Activity Information Development (RAID) diagrams for each case.

3.1 Methodology

The methodology used to uncover the opportunities for re-use used a combination of methods taking the case study analysis described in Howard et al. (2010) and a semi-structured interview conducted with the case study 'owner' (this being the principle researcher on the associated project).

Each of the case study participants was asked the following questions regarding the data assets described within their case study:

1. Can you think of any other purposes to which these data can be put?

2. What would the benefits be to using these data for that purpose? (prompts: time? cost? availability? originality?)

3. What are the shortcomings of these data for the purposes for which you intended them? (prompt: think at level of data, record and case)

4. What are the shortcomings of these data for the purposes for which others may intend? (may be answered at the data, record or case level)

The interviews were recorded and notes were taken using MS OneNote. The discussions and answers to the questions were compiled and summarised in the following sub-sections.

3.2 Results discussion

In this sub-section each of the questions posed in section 3.1 are discussed along with a summarised list of the answers given by the interview participants.

3.2.1 Can you think of any other purposes to which these data can be put?

Put simply the answer was yes, and yet for no case was this idea of re-using the data met with a great deal of interest by the interviewee. It was clear that the future use of their research data was something that they had spent little time considering and attributed little importance to, thinking it would be of little value to anyone else. After further discussion all interviewees identified possible scenarios in which these data could be re-used. However, these scenarios were couched in phrasing such as 'if someone was researching into X then these data could be used for Y' rather than being able to identify a ready-made market and recipient.

Of the potential uses of the interviewees' research data, the following were identified:

- The data could be used as a case to draw new insights or conclusions from another perspective

"People could take it, use it and analyse it," "I'm not sure what particular research question, but could use for anything in the area of knowledge transfer and organisational learning".

- Could be used for its original purpose.

  In many instances, mainly the prescriptive studies, the data assets comprise a tool or intervention that are to be used for a preconceived purpose but in different contexts. In some cases the data assets are only useful for the same purpose and context, though if the data are unique then this is still valuable. *"The machine parameter data is useful as it doesn't exist elsewhere."*
- This is quite a strange case, where descriptive material is used in order to trial and refine prescriptive material. If there are enough trial cases then the descriptive trials can be used for comparison.
- Could use if they were to carry on and build on the same research
- Could use some of the trials to compare cases
- Data used as a manufacturing output

  This is where descriptive material is used as expected describing the same context for different purposes.
- Could be interesting for someone starting research in this area
- Could use the methodology

3.2.2 What would the benefits be to using these data for that purpose?

There were a number of benefits of the data that could be imagined. The following list is of those potential 'benefits' that were voiced during the interviews:

- Taster sample of data could help describing a case in the topic area.

  Though the data were not considered enough to fully analyse but could produce research questions. *"You miss a lot without being there."* Having said this, one interviewee was *"using someone else's ethnographic data without being there… though its hard, I'm constantly asking what is that, who is this person, what do they do, etc. So I have to spend a lot of time looking around for context.*
- Sharing data would support communities of practice.
- Time and cost savings.

  "Could save another researcher six months worth of work in terms of temperature profiling"
- Benefits to learner to transfer knowledge from experts.

  "The opinions of the data capturer on what were the key events were vital".

This means the derived data was considered very important and should be captured and offered as well as the raw data.

3.2.3 What are the shortcomings of these data for the purposes which you intended?

For this question we tried to get an idea of how good, reliable and interpretable the data was for the purpose which it was gathered. Some of the answers given were summarised as follows:

- Would always like more quantity.

  However, there is a point in cases "*where there is unlimited access at some point you can reach saturation in terms of the observations you can make*".
- Need more cases.
- The medium of the data – *"some users preferred more visual representation"*
- In one case, *"Not many problems with the data at all"*
- A limit to the computational capabilities (cost) – *"calculating using dynamic methods was too expensive so used FEA instead"*.
- Not verified.

  "*It was difficult to verify dynamically as it was all done statically*".

  Also, verifiability was difficult due to lack of context – *"multiple choices for the answers to questions – no written comments"*.
- Missing data points

  "*Some data plots of the temp profile missing*".

  However, it was suggested that though this was not a problem for the current use it could have been in terms of how re-usable the data is.

3.2.4 What are the shortcomings of these data for the purposes which others intend?

For this question we tried to get an idea of how good, reliable and interpretable the data was for other purposes. Some of the answers given were summarised as follows:

- Missing context – "*you'd need to know about the industry and the domain specific knowledge*".
- Research participants were not profiled.

  However, the participants' "*experience and history was profiled on a demographic questionnaire*". This seems to be a problem for both use and re-use. It was not clear why some context was gathered and some ignored.
- Loss of context

  Of the context that was gathered (the experience history example in the bullet above) the history was converted to numerical rating which was useful to the research but conceivably less useful – because difficult to interpret – to future researchers.
- Lack of accuracy

  "*For initial experimental work then it's all complete but if going on to more complex work then there could be issues*."

3.3 Summary of opportunities for data re-use

It is clear that there are a number of potential opportunities for the re-use of engineering research data. However, although there have been instances of successful re-use of data identified both in literature – for example in the Delft protocol studies (Cross et al., 1996) – and in the case studies in reality the frequency of re-use is low and the value of these data re-use opportunities has not been measured. The real value of data re-use might be found not at the individual level of data use, but when considering multiple re-use of the same data. In that same way that with data, the whole is greater than the sum of the parts (Fisher, 2009) so might be the whole greater when considering as one the output of all research based on using the same data.

It was established that the descriptive material had much more potential for re-use than the prescriptive material. If enough context is supplied, the descriptive material always has the potential for further interpretation. However, it was also established that the descriptive material in prescriptive studies, such a trials, pilots and case studies could also be of use.

It must also be acknowledged that the intermediate data is also vital for evaluating both prescriptions and descriptions and would often be required for verification, validation and the repeatable testing of results. This supports the view that data is least useful when it stands alone, without supporting context as the basis for further interpretation.

It is expected by the authors that opportunities for engineering data re-use will usually be in a broader and more complex (zooming out) situation than their original use. The rationale for this is that data are usually gathered, and certainly generated, with a very specific use in mind. The likelihood of such specificity of data application being found again for a data set seems small.

4. BARRIERS TO RE-USE

The goal of the ERIM project is to maximise the value of engineering research data. Here we are referring to raw and intermediary data, which is often not re-used, rather than the data that would be considered to be the output of a research project (i.e. publications reporting the findings of research). In Section 3 it was established that there are some situations in which data could be re-used. In this section the research will be discussed which was undertaken to indentify some of the barriers to the re-use such research data.

4.1 Methodology

The methodology used to uncover the barriers to re-use used a combination of methods taking the case study analysis described in Howard et al. (2010) and a semi-structured interview conducted with each case study owner.

The first question of the semi-structured interview was to find out whether there were any barriers at all to the interviewees' data being re-used. Once it was established that there were barriers to re-use present, several more specific topics were discussed on the following types of barriers: Confidentiality barriers, Ownership barriers, Licensing barriers and Technical barriers. The interview closed with a brief discussion regarding the barriers the interviewees had experienced preventing them from gathering/using data they would have liked.

The interviews were recorded and notes were taken using MS OneNote. The discussions and answers to the questions are summarised in the following sub-sections.

4.2 Results discussion

In this sub-section each of the questions posed are recorded with representative responses together with accompanying discussion. This investigation into the barriers to re-use started by asking the following very broad question:

1. I'd like to take these data away and use for my own purposes, may I?

The answer was often 'no' but for quite different reasons. The following questions investigated in more detail the different types of barriers to re-use.

4.2.1 Confidentiality barriers

Confidentiality was expected to be a frequently encountered barrier to reuse (since it was the author's own experience that his was a common prohibition) yet this was found not to be true for many of the cases.

2. What level of formality is the confidentiality?

The formality ranged from no explicit agreement (rather an expectation of behaviour in the mind of the researcher) to a verbal agreement with the individual representing the collaborating company, to a written constraint on data sharing by the collaborating company, to a legal document signed by all parties. In some cases the confidentiality constraint prevented the data being gathered and taken away for analysis, that is, it had to be inspected on-site. In these cases field notes were allowed but often richer methods of data capture, such as video recording, were not.

3. What is the reason for the confidentiality? (prompt: commercial sensitivity, anonymity)

- To allow freedom to talk – company personnel were afraid the data could give a *"warped version of reality"*. They may also not want others in the organisation to see the data.

- Commercial sensitivity – some of the data are valuable to customers or contain trade secrets that give the company a competitive advantage.

- Anonymity – some companies were happy for the data to be shared provided that the association of their company with was reliably hidden.

4. Could it have been possible to avoid these issues?

In two cases it was thought that the basis for data provision were the only ones that could have been negotiated. In another case it was thought that a non-disclosure agreement would have helped, since it would have clarified the basis for data sharing. Interestingly, in one case the researcher felt that the research case might have been constructed using data that was not commercially sensitive had this been considered an issue at the time.

4.2.2 Ownership barriers

5. Is there value to having sole rights to these data?

- Not relevant as data belongs not to the researcher but to the source company.

- Research competitive advantage – *"This is our data, why should we want anyone else to benefit from our data and relationships?"*

- Happy to give the outputs but reluctant to give away the code (prescriptive study).
- Yes, data may be used to develop a spin-off company.

6. What would the circumstances have to be for there to be more value in sharing these data than not?

- If other sources/users would contribute to grow the data.
- To bring something to future collaborations.
- If a strategic collaboration was found with legal constraints to help protect the intellectual property.

4.2.3 Licensing barriers

7. Are there any licensing or contractual issues to overcome for the purposes of sharing these data?

In two cases it was found that there would be software licensing issues in attempting to re-use data.

8. Could it have been possible to avoid these issues?

No suggestions were forthcoming.

4.2.4 Technical barriers

9. Are there any technical barriers to sharing these data? (prompts: software, hardware)

In one case an HGI machine had been used; likewise in other cases other expensive and specialist hardware. In further cases, specific measurement and machining equipment would be needed if experiments were to be re-run. However, assuming access to the appropriate equipment the technical barriers were limited.

10. How could these technical barriers have been avoided?

In the first case cited, software could have been coded for a standard Windows-based machine, rather than a HGI machine. From the cases audited, it would seem that avoiding the technical barriers will always be very case specific. However, simply considering the need for supporting the re-use of data will help promote behaviour which prevents technical barriers to re-use being erected in the first place.

4.2.5 Barriers to other desirable data

11. Are there any other data you would like?

- There were types of data the researchers of each case were looking for. However, they did not know whether these data exist or not. They had little expectation that such data would exist or could be easily found if indeed they already existed. This is a good justification for the endeavours of the ERIM project. Only by making existing data visible and accessible, and by providing the necessary contextual support, will researchers consider data re-use first, before data gathering or generation is carried out. At the moment useful data may exist but frequently there is no way of finding them, and there is no culture of first seeking what already exists.

- Transcripts and data that were observed but not recorded.

- Another layer of data capture on top of the case (testing/observing another variable). This is more a need for the technology than data.

- More data and data from external vendors.

12. What are the barriers to you getting these other data?

Similar barriers were suggested as those given above. There was also a mention of financial barrier – *"we'd need extra funding of £50-£100k"*.

13. Can these barriers be removed?

"In the future we could be clearer from the start how the data will be captured and used" – However, this is the fundamental problem with confidentiality agreements and the more exploratory sort of research: it is not always possible to predict at the beginning what will be of interest and how it will be captured and used.

In one case it was suggested that using a different software package could have removed these barriers.

4.3 Summary of barriers to research data re-use

Four major forms of barriers were identified which could be split into further types:

- Confidentiality (when industry is involved)
    - Anonymity
    - Commercial sensitivity
    - Fear of misrepresentation (or true representation)
- Ownership barriers – researchers being afraid of losing their research status by sharing their valuable data.
- Licensing barriers (licence to software and machinery).
- Technical barriers (not always possible to obtain/understand the same machinery and code, especially when non-standard).

5. CONCLUSIONS AND FUTURE WORK

A number of data assets were identified that would potentially be of use to new research projects, and the circumstances explored in which they might be reused. It is thought that the descriptive material in the research projects had the most potential for re-use and thus require more attention in terms of the management of the data. Prescriptive material may be of use for any validation of the prescription or if there is a continuation researcher project where the prescription is further developed.

In terms of the barriers to re-use, several were identified. It is envisaged that these barriers could be tackled at a number of levels. Many of the barriers could have been removed by giving consideration to re-use during use, or prior to the gathering of the data. The authors envisage a progression of tools, each one more specializing than the last which will assist in supporting better data management during the research activity, embryonic versions of which form part of the output of the ERIM Project. The first tool is a set of principles to guide behaviour and actions in relation to research data management.

At one level beneath, and informed by the principles, an *engineering research data management plan requirement specification* will provide a basis for the development of the data management plan which should be implemented for each research project. Implementation could be assisted by provision, on the one hand, of a template or a set of templates each of which is customised to suit a particular class of research activities. On the other hand, it is possible to envisage an application environment that would prompt the user for information suitable to their research activity, the end product of which would be a *bespoke data management plan* for that research activity or project. Finally, it is possible to envisage tools, or perhaps one *integrated research data management tool*, which would assist in the execution of the data management plan during the research activity and thus promote opportunities for data re-use and at the same time seeking to limit the erection of barriers to data re-use. In particular, the authors believe that the recording and preservation of data and data asset *context* is fundamental to making engineering research data more reusable.

6. REFERENCES

Ball, A. (2010). *Review of the State of the Art of the Digital Curation of Research Data* (ERIM Project Document erim1rep091103ab12). Bath, UK: University of Bath. Retrieved September 21, 2010 from http://opus.bath.ac.uk/19022

Blessing, L. & Chakrabarti, A. (2009). *DRM: A design research methodology.* London: Springer. ISBN-13: 978-1-84882-586-4.

Cross, N., Christiaans, H., & Dorst, K. (1996) *Analysing Design Activity*. Chichester, UK: John Wiley & Sons.

Fisher, T. (2009). *The Data Asset: How smart companies govern their data for business success*, Hoboken, NJ: John Wiley & Sons.

Howard, T., Darlington, M., Ball, A., McMahon, C., & Culley, S. (2010). U*nderstanding and Characterizing Engineering Research Data for its Better Management* (ERIM Project Document erim2rep100420mjd10). Bath, UK: University of Bath. Retrieved September 21, 2010 from http://opus.bath.ac.uk/20896

7. APPENDIX

Cost of gathering raw data for the ERIM Project.

SCOPING QUESTIONNAIRE

Participants

Half days	Full days	Cost/day (FEC*)	Total
13	0	£510	£3312

Researchers

Half days	Full days	Cost/day (FEC)	Total
0	13	£510	£6630

Travel + subsistence

Total
£500

CASE STUDIES

Participants

Half days	Full days	Cost/day (FEC)	Total
6	0	£510	£1530

Researchers

Half days	Full days	Cost/day (FEC)	Total
0	6	£510	£3060

Travel + subsistence

Total
£250

Overall total of cost of ERIM data gathering activity:	**£14832**

* FEC = Full Economic Cost