Int. J. Human-Computer Studies (1999) **51,** 729–755
Article No. ijhc.1999.0327
Available online at http://www.idealibrary.com on IDEAL®

The experimental evaluation of knowledge acquisition techniques and methods: history, problems and new directions

Nigel Shadbolt, Kieron O'Hara and Louise Crow
Artificial Intelligence Group, Department of Psychology, University of Nottingham, University Park, Nottingham NG7 2RD, UK.
email: lrc@psychology.nottingham.ac.uk

The special problems of experimentally evaluating knowledge acquisition and knowledge engineering tools, techniques and methods are outlined, and illustrated in detail with reference to two series of studies. The first is a series of experiments undertaken at Nottingham University under the aegis of the UK Alvey initiative and the ESPRIT project ACKnowledge. The second is the series of Sisyphus benchmark studies. A suggested programme of experimental evaluation is outlined which is informed by the problems with using Sisyphus for evaluation.

1. Introduction

> "I wish I could be sanguine—that attention to performance analysis will follow automatically with maturity. But inattention to this matter has been a failing of research with AI programs from the start."
>
> Allan Newell—A Tutorial on Speech Understanding Systems

The field of knowledge acquisition (KA) is attaining a degree of maturity. There is a more or less stable consensus about the value of models of expertise and problem-solving for structuring the acquisition of knowledge, and similar agreement over the general topography of the life cycles of development projects (Wielinga, Schreiber & Breuker, 1992; Shadbolt & O'Hara, 1997). On the consensual view, it is fairly standard to have a phase of requirements analysis, followed by a phase of conceptual modelling when the expertise and its context is modelled, followed by system design and finally implementation. KA is typically performed as part of the conceptual modelling phase, but is also important for requirements analysis. These four phases are not generally linearly arranged, and the knowledge engineer can flip back and forth between phases, sometimes opportunistically, and sometimes in a more structured way dictated by particular methodologies.

So the frameworks directing KA are relatively firmly fixed. Similarly, so are the activities within KA. Most elicitation of knowledge from experts has been performed on a small set of techniques, often embodied in software, that has remained fairly constant since the first sets of KA tools were assembled in the late 1980s: protocol analysis,

1071-5819/99/100729 + 27 $30.00/0

laddering, repertory grids. On top of which, the major structuring implements for KA are skeletal generic models of problem-solving, which encode what is seen as the domain- or task-independent portion of problem-solving of the appropriate type, leaving the knowledge engineer to "fill in" the domain-dependent bit (the better models also tell the knowledge engineer where to look for the domain-dependent knowledge).

Of course, there is a significant degree of variation in the characterization and application of KA techniques. Not only that, but there are arguments taking place both within and outside the general consensus. Within the consensus, there are many competing modelling methodologies, each with its own perspective on the arrangement of life-cycle phases, the types of models to be built, the knowledge contained in the models, the degree of formalization and so on. CommonKADS is obviously the market leader, but the present writers might wave a flag for GDMs, and point also to MIKE, generic tasks, PROTEGE-II and so on (Breuker & van de Velde, 1994; O'Hara, Shadbolt & van Heijst, 1998; Angele, Fensel, Landes, Neubert & Stüder, 1993; Chandrasekaran & Johnson, 1993; Rothenfluh, Gennari, Ericsson, Puerta, Tu & Musen, 1996). There are also prototypes of new KA techniques and tools being demonstrated pretty well constantly on the KAW-EKAW-JKAW circuit, any or all of which might usefully be included within the KA phase of an application, and new (libraries of) problem-solving types promise increased efficiency.

Outside the consensus, there are plenty of interesting arguments being made against the standard view, often rejecting the value of the generic approach. For example, the ripple down rules methodology explicitly rejects the need for modelling, and instead focuses on the evaluation of prototypes developed on the basis of increasing numbers of test cases (Compton & Jansen, 1990).

The field of KA is an active research area. However, there is surprisingly little testing of ideas, or evaluation of the various approaches, models, tools or techniques to inform the debate. In many ways, this is not a problem that is unique to KA: KA is not noticeably more lax in evaluation than software engineering in general. Nonetheless, it is undesirable that the products of KA research receive such little scrutiny. In this paper, we want to explore the reasons for this. We believe the oversight is partly practical, partly political. We hope, in this paper, to sketch the sources of the evaluation problem, and to suggest possible ways forward.

We begin with the difficulties of evaluating KA techniques, tools and methodologies. In the following section, we discuss some of the ways in which small-scale experiments have attempted to evaluate particular tools and techniques, and assess the success of these experiments. As this is an overview, we will focus on the methodology and design issues these experiments raise rather than the results produced. Then we widen our scope to look at the evaluation of what we call KA frameworks, looking at the meagre experimental evidence, and the series of Sisyphus programmes in which particular test problems are undertaken by representatives of various knowledge engineering methodologies. In the discussion, we try to suggest ways forward, both in terms of the construction and design of experiments, and also with respect to the question of funding.

2. Evaluating knowledge acquisition techniques, tools and methodologies

2.1. EVALUATING KA TOOLS AND TECHNIQUES

The main problem in evaluating KA techniques and tools is that they are designed to elicit quality knowledge from human experts. By this we mean that the knowledge elicited is all and only the knowledge required to perform the job for which the expert system (or whatever) is being designed. Therefore, when a tool is evaluated, the knowledge that the expert has delivered must be tested.

Difficulties present themselves immediately. A typical psychological experiment should be performed on a sample large enough to give the results statistical significance. However, any sample of experts large enough (say > 20) will be difficult to assemble and expensive to test. Also, the necessary assumption that all the experts are homogeneous in their expertise will be a distortion of the relative levels of the experts and their different psychological profiles. A further complication is that elicitation sessions may be conducted with groups of experts rather than individuals. Attempted solutions to the problems of expert and domain selection are discussed in Section 3.1.

Secondly, we must consider the nature of the knowledge acquired in the experiment. For example, the obvious way of testing a body of knowledge for the completeness of its coverage of the problem-solving space, a very important parameter, is to measure that body of knowledge against another body of knowledge certified to be complete or adequate. But the expertise that users of KA tools wish to acquire is by and large scarce. Hence, in many cases, there will simply be no source of "gold standard" knowledge available. Other problems in this group might include measuring the inferential power of the knowledge base acquired, or in making sure that the knowledge acquired is on the same level of generality, or grain size. Section 3.2 reviews approaches taken to evaluating the acquired knowledge.

A third group of problems concerns the type of knowledge acquired. At least some KA tools are intended for a wide variety of contexts. For example, a card sort tool should in theory be of value in any domain in which there are objects or concepts or even processes which can be named and shuffled about and sorted. But there may still be variations in its value across domains; a tool that produces valuable knowledge bases for classification tasks, for example, might not necessarily be as valuable in design tasks. Hence, ideally, evaluation experiments should be not only designed within domains, but across them too. The extent to which this has been accomplished is described in Section 3.3.

A fourth group of problems concerns the separation out of questions about the underlying theory and questions concerning the particular embodiment of that theory in the tool. For example, is the poor performance of a tool in an experiment a reflection of the weakness of the tool or the weakness of the user interface? Of course, much software has to contend with evaluation problems of this sort; however, the particular significance of this issue in KA is that with high experimental costs accruing from addressing the first three groups of problems, it becomes even more expensive to have to control for other things like interfaces, implementation platforms, etc. There are related problems controlling for a KA tool's context, too; for instance, most KA tools are intended to be used as part of a suite of tools, rather than standalone. In such a case, it is relevant for an

evaluation to know whether there are any "interference" effects between tools. For example, is a laddering tool more or less effective used after/before/in the absence of a protocol analysis tool? Section 3.4 reviews attempts to answer this kind of question.

Finally, there are problems connected with the quantification of the dimensions over which KA software should be evaluated. The obvious two dimensions are the quantity of knowledge acquired (see above for some remarks about quality), and the effort involved in using the tool. Measuring quantity of knowledge is obviously difficult. It would be nice to measure it by the percentage of coverage of the desired domain, but as we discussed above this is not often practical. Failing that, a simple metric would be the number of items in the acquired KB, but is a crude measure at best, while also being highly variable with the representational format (for example, a well-designed compact frame representation could be quite "small", while a poorly designed representation with many redundant and unconnected production rules would be relatively "large"). Other, more complex measures take into account the concept of "entrenchment"—how central or peripheral a knowledge items is to the whole corpus of knowledge elicited. A highly connected piece of knowledge may have more potential for reuse.

One final point is worth noting explicitly in relation to this. The measures of the knowledge acquired need to be calibrated against the understanding of the uses to be made of the knowledge bases. For example, if we use the number of items in a knowledge base as a measure of gain, then it follows that no attempt is being made to discover whether the knowledge acquired amounts, or can amount, to a cognitive model of the expert. The measure is purely and simply a measure of the knowledge that has been acquired and that is adequate to do the job in hand. If the success of the tool or technique in producing a psychologically interesting model of the expert is to be measured, a more complex metric would need to be devised. Section 3.5 discusses the approaches that were taken to quantification.

The problems we have set out in this section are summarized in Table 1.

2.2. EVALUATING KA FRAMEWORKS

Let us define a *KA framework* as being some item that directs a knowledge engineer in the use of a KA tool or technique. A KA framework might then be a model template, a component library, an ontology or even an entire knowledge engineering methodology.

When we move on from tools to frameworks, even greater challenges arise. Each framework is designed to guide KA practice over a range of problems, if not the complete problem-solving space, then at least a relatively wide subset thereof. Hence, the sheer scale of the sampling problem that applies to the experimental evaluation of KA tools will be multiplied by orders of magnitude. Only a whole series of experiments across a number of different tasks and a number of different domains could control for all the factors that would be essential to take into account.

Even then, it is possible that only a sample of the scope of a framework might be possible. For instance, with today's decompositional modelling methodologies such as CommonKADS and GDMs, the number of models offered is orders of magnitude more than the original modelling methodologies from the 1980s, such as generic tasks, which

TABLE 1
Problems associated with the experimental evaluation of KA tools

Problem type	Basic problem characterization	Typical questions
How many experts?	The difficulty of getting a large enough sample of experts together to give statistical significance.	How many experts do we need to use? Can we avoid using experts altogether? If so, what level of expertise do we need to demand from experimental subjects? What can we tell from domains where expertise is very common?
The nature of the acquired knowledge	The difficulty of comparing, understanding or validating the knowledge acquired during an experiment.	Is there a "gold standard" of knowledge available against which to compare knowledge acquired? If not, how can we measure the coverage of the knowledge? How can we measure the value of the knowledge in terms of its inferential power?
Knowledge acquired in different domains and tasks	The need to discover the range of tasks and domains for which tools/techniques are useful.	How many domains do we need to run the evaluation over? Which domains? How many tasks do we need to run the evaluation over? Which tasks?
Forms of tools, context of use	The difficulty of isolating the value-added of a technique.	Can we complete evaluation before the tool/technique is outdated? How do we separate out intrinsic problems with the KA technique from problems caused by inadequate implementation, interface, platform? Do we measure a tool in isolation or in a KA context with other tools? If the latter, how do we separate out the individual contributions of each tool?
Quantification	The selection of a reliable and informative metric.	How do we quantify knowledge? How do we compare the quantity of knowledge represented in different ways? Can we expect translation between representation to yield equivalent quantities of knowledge? If not, how can we select one representation over the others? How can we quantify knowledge engineering effort? How can we quantify effort spent pre- or post-KA session? How many elicitation sessions are optimal?

gave us six models, and KADS-I, which offered around 20. It is, of course, not feasible to test all the models in the more recent methodologies—but that does raise the question of how best to sample.

Furthermore, the selection of metrics is problematic: measurements have to distinguish between the contribution of a framework and the contribution of the particular tools that the framework is organizing. In other words, the problem is to measure the value-added of the framework.

We don't want to claim that all or even some of these problems are insoluble. For example, Uschold, Clark, Healy, Williamson and Woods (1998) produced interesting results from their investigation of the process involved, and the costs incurred, in the reuse of an existing ontology in a new domain. Their results are positive, but their conclusions necessarily tentative.

Nevertheless, any experimental approach to the evaluation of KA techniques, tools and methodologies will have to deal with all of these issues, and probably more. Opinions will differ in the field as to which solutions are better, and which evaluations more reliable. We can expect much intense discussion of the evaluation issue before there is agreement on how best to evaluate our tools and framework. If our field is approaching maturity, our set of evaluation tools is in its infancy. This is not a healthy state of affairs.

3. Attempts to evaluate tools: successes and drawbacks

In the late 1980s and early 1990s, a series of experiments was carried out by Nottingham University and partners. Some of this work was done under the remit of an Alvey project (Burton, Shadbolt, Hedgecock & Rugg, 1987; Schweickert, Burton, Taylor, Corlett, Shadbolt & Hedgecock, 1987; Burton, Shadbolt, Rugg & Hedgecock, 1988; Shadbolt & Burton, 1989; Burton, Shadbolt, Rugg & Hedgecock, 1990*b*), and some under the ESPRIT project ACKnowledge (Burton *et al*., 1990*a*; Rugg & Shadbolt, 1991; Rugg, Corbridge, Major, Burton & Shadbolt, 1992; Corbridge, Rugg, Burton & Shadbolt, 1993; Corbridge, Rugg, Major, Shadbolt & Burton, 1994). See also Shadbolt and Burton (1990).

This series of experiments was generally intended to address two questions. Firstly, there was the question of whether KA techniques and tools were differentially effective. This is not just the crude question of whether one technique is better or worse than another, but includes also such considerations as the complementarity of pairs of techniques, i.e. questions about whether particular pairs of techniques will tend to acquire knowledge of a similar sort, and therefore run the risk of redundancy in the acquired KB, or alternatively whether the type of knowledge typically missed by one technique might be picked up by another.

Secondly, there was the question of the possible differential effects of using techniques in different domains. Attention was paid to getting a range of domains each of which might set different problems for KA techniques. Some domains were real-world/commercial, such as the domain of lighting for industrial inspection (Schweickert *et al*., 1987), or the identification of corrosion of metals (Corbridge, Rugg, Burton & Shadbolt, 1993). Other domains were real-world/academic/scientific, such as the identification of flint artifacts from Stone Age tool production (Burton *et al*., 1990*b*) or the diagnosis of acute

abdominal medical conditions (Corbridge *et al*., 1994). Others were "toy" domains, such as the identification of fruit (Rugg *et al*., 1992), or methods of transport (Burton *et al*., 1990*a*).

The purpose of this paper is not to review the results of these experiments. There were a number of interesting features, such as the relatively poor performance of protocol analysis (Burton *et al*., 1988, 1990*b*), the good performance of so-called contrived techniques against natural techniques [Shadbolt & Burton, 1989; see Shadbolt & Burton (1990) for a definition of the distinction], and the failure to show that software implementations of techniques outperformed pen-and-paper versions (Rugg *et al*., 1992; Corbridge *et al*., 1994), but these features are discussed more fully in the review paper (Shadbolt & Burton, 1989), and the discussion sections of the various experimental reports. Our intention here is to try to evaluate the evaluations, to see how well the experimental designs coped with the inherent problems of software evaluation. In Section 2.1, we set out a series of five groups of problems that an experimental approach to the evaluation of KA software in particular would meet, and we shall now examine the experiments in the light of those problems.

3.1. HOW MANY EXPERTS?

The first group concerned the problems of amassing sufficient quantities of experts for a genuine comparison to be made. A number of different approaches were tried. One approach was to make the attempt to use genuine experts, or near-experts, at postgraduate level or above. For instance, expert metallurgists were used for experiment 1 of Corbridge *et al.* (1994), and expert archaeologists for experiments 1 and 2 of Burton *et al.* (1990*b*). The problem, as expected, is bringing together and co-ordinating large enough numbers of such experts to provide statistical significance: Corbridge *et al.* used eight experts in corrosion, while in the Burton *et al*. experiment 1, although 16 archaeologists were used, they were divided across two domains, with eight experts on flints and eight experts on pottery. In the Burton *et al.* experiment 2, five flint experts were used. Greater numbers (24) were used by Corbridge *et al.* (1993). It should be noted that all these experiments focused on individual rather than group elicitation sessions.

The results were valuable, and in particular, where relevant, they seemed by and large to corroborate other studies. For example, Burton *et al*. experiment 1 seemed to confirm the relative lack of efficiency of protocol analysis in terms of resource consumption, despite the large amount of knowledge eventually captured (Burton *et al*., 1990*b*, p. 171; see also Burton *et al*., 1987, 1988). Furthermore, the experimental designs were created to facilitate detection of effects among small groups of subjects (Corbridge *et al*., 1994, pp. 324–325). However, the smallness of the samples entails the desirability of reproduction of the experiments to increase confidence (Corbridge *et al*., 1994, p. 322). For instance, experiment 1 of Burton *et al*. reported some different patterns between the flint domain and the pottery domain (Burton *et al*., 1990*b*, p. 171). Furthermore—perhaps more worryingly—there was an important discrepancy between the results on feedback effects between Burton *et al.* (1990*b*) and Corbridge *et al.* (1993), the latter having a sample three times the size of the former. Corbridge *et al*. found that giving the subjects feedback in the form of "pseudo-production rules" had no effect

on the gain of the elicitation sessions, where Burton *et al.* had found feedback effects on gain.

A variation of this approach was to use a single genuine expert (Schweickert *et al.*, 1987; Rugg & Shadbolt, 1991). This allowed a deeper interrogation to take place, but of course made generalization problematic. A further problem, common to these first two approaches, is that it is necessarily difficult to assess the quality of the knowledge acquired because of the absence of a gold standard (Burton *et al.*, 1990*b*, p. 171); there are no "greater" experts to critique the knowledge acquired. However, an attempt was made to get around this problem in Rugg and Shadbolt (1991) by running two experiments on a single expert, asking him in the second experiment to critique his own performance in the first. The authors, rather than drawing too many conclusions from the statistics, used the experiments as a case study to point up a number of problems and practices to be avoided.

A third approach was to use a larger group of people who might approximate to experts in technical domains: typically such people would be undergraduates who have reached a particular standard. So, for example, Burton *et al.* (1988) took 32 final year undergraduate geographers, and experiment 2 of Corbridge *et al.* (1994) took 32 final year medical students. The advantages of this sort of experimental method, of course, are the greater numbers can be mustered, and they can be assessed against a gold standard provided by their teachers. The disadvantage is that there is evidence that students may not behave like experts—Chi, Glaser and Rees (1982) found significant differences in the problem-solving approach and performance of undergraduate students and advanced Ph.D. students in the domain of physics problem solving. Burton *et al.* (1988) respond that the students "had been equipped with the necessary skills for [the experimental] task very early on in their university careers. Furthermore, the necessary factors and dimensions had been built upon many times, and so we claim that these students function as models of real-world experts" (Burton *et al.*, 1988, p. 88). Clearly, the usefulness of the students for such studies is very closely linked to the experimental design.

The fourth approach was to use trivial, though rich, domains in which everybody is an expert, such as the identification of fruit (Rugg *et al.*, 1992), or means of transport (Burton *et al.*, 1990*a*). The study reported in Rugg *et al.* (1992) assembled 75 subjects, experiment 21 in Burton *et al.* (1990*a*) managed 40. This is a very good way to get large numbers of experimental subjects, but there must be reservations about the use of such domains where expertise is usually not valuable nor experts established and "centrified" (Rugg *et al.*, 1992, p. 282). Ideally, in a carefully designed programme of experiments, such large-scale experiments in toy domains might be used in tandem with smaller-scale experiments with genuine experts to measure the extent to which such types of experiment confirm each other's results. This is an example where empirical evaluation work on evaluation techniques would be valuable.

There is certainly a potential for using less-trivial domains in which expertise is commonplace. Wagner and Stanovich (1996) discuss the conceptualization of reading as an expert skill, pointing out that any definition of expertise which relies on excellence in comparison with the rest of the population will necessarily produce few experts. However, as we learn more about the characteristics of expertise across domains, we can begin to characterize expertise in terms of these agreed characteristics. For example, Ericsson,

Krampe and Tesch-Römer (1993) assert that since expert performance across domains requires a decade of intense preparation, it can be considered to be application of the same acquisition mechanisms responsible for more ordinary levels of performance intensively over a long period of time. In this case, skills such as reading, which are practiced intensively by a large proportion of the population might be more suitable domains for such experiments, producing comparatively large numbers of expert subjects.

3.2. THE NATURE OF THE ACQUIRED KNOWLEDGE

The possibility of getting a gold standard, as discussed in Section 3.1, depends largely on the use of the third or fourth approaches to getting experimental subjects—i.e. it is only possible to use gold standards to assess the work of hypothesized models of experts, not established experts themselves [although Rugg & Shadbolt (1991) used a single expert to critique creatively his own results].

As a consequence, there is a question mark over results using gold standards. For example, Shadbolt and Burton (1989) make the following assessment of the completeness results of Burton *et al.* (1987).

> … although the union of subjects [32 geology students] gave a good coverage of the domain, individual subjects a relatively poor coverage (around 30%). This is not because our subjects were behaving 'inexpertly'—for example in the protocol analysis there were very few mis-identifications. The problem is that individual elicitation sessions do not reveal explicit information about much of the necessary decision making underlying identification. This may be a general problem in knowledge elicitation. [Completeness information] is not usually available, as knowledge engineers obviously do not normally have access to a gold standard. It is only through this kind of experimentation that such issues may be examined. (Shadbolt and Burton 1989, p. 17)

This illustrates how difficult it is correctly to interpret the experimental results when using models of experts. Burton *et al.* (1988) confirmed the coverage results of Burton *et al.* (1987) to some extent, using 32 geography students in the same experimental design.

Another issue we might briefly mention is that of the inferential power of acquired knowledge bases. There are various ways of measuring this. As part of the evaluation work performed in the ACKnowledge project, a set of criteria was drawn up to guide the evaluation (Berger, Burton, Christiansen, Corbridge, Reichgelt & Shadbolt, 1989). Inferential power is relative to the formalism chosen for representations, and Berger *et al.* discuss different measures of inferential power depending on the KRL used. There were three suggestions: firstly a pseudo-production rule formalism was suggested, in which case the generative power of the rule set needed to be measured, and formal grammar theory would provide appropriate metrics; the second suggestion was an intuitive frame formalism, in which cases the inferential power would be connected to the amount and type of inheritance; the third idea was to use restricted semantic nets, which would require a fairly complex measure of local connectivity.

There is no obvious reason why these three metrics, with their different properties, will always give us similar results for the same base of acquired knowledge. This is an important issue for the evaluation of this property, therefore, as the representational formalism chosen may determine how useful knowledge is measured to be.

3.3. KNOWLEDGE ACQUIRED IN DIFFERENT DOMAINS AND TASKS

The problems in covering the space of domains and tasks are chiefly logistical. Conceptually, the ideal is to conduct as many experiments as possible over as wide a range of domains and tasks as possible. The experiments reported here managed a fair range of domains: fruit, mineralogy, metal corrosion, abdominal diagnosis, driving hazards, lighting for industrial inspection, flints, glacial features, sport, transport, pottery and igneous rock. However, few of the experiments were reproduced exactly in new domains [one exception was the connected series of experiments recorded by Burton *et al.* (1987, 1988, 1990*b*) which carried out the same experimental design in the domains of igneous rocks, glaciers, flints and pottery].

On the other hand, most of the tasks involved in these experiments were analytic tasks, usually some type of identification, classification or diagnosis. Ideally, the task context of a KA experiment should also be varied. This would obviously lead to a scaling up of any experimental programme (even if we understood task space well enough to allow us to take a principled and informative sample of tasks). Such a massive overhead on experimental evaluation is clearly one of the major drawbacks of the approach.

3.4. FORMS OF TOOLS, CONTEXT OF USE

This group of problems is tricky to address; experimental designs have to be very carefully managed to separate out background issues. For example, a series of experiments we are reporting investigated the technique of laddering. Sometimes pen-and-paper versions of laddering were used, and sometimes a software tool. The tool used was an early version of ALTO (Major & Reichgelt, 1990), the laddering tool which was ultimately incorporated into the ACKnowledge workbenches ProtoKEW and KEW. Corbridge *et al.* (1994, p. 335) describe how ALTO was improved after the experiment was performed, and now if the same experiments were performed at Nottingham—or if the experimental design was reproduced with one or two variables changed—the obvious software to choose would be the laddering tool in Epistemics Ltd's portable knowledge engineering workbench PC-PACK (O'Hara *et al.*, 1998).

In particular, given the pace of innovation in software development, there is a trade-off between thoroughness in the design of a programme of experiments, and early results. It would be pointless to continue to evaluate ALTO in 1998, now that its development and maintenance team has dispersed and that the tool itself is more or less out of use. However, series of experiments covering a wide variety of domains and task types will take time to conduct, and there is a real danger that the appropriate tool in the early stage of such a series will have been superseded before the series has been completed.

Such technological change can also affect results that are theoretically independent of actual implementation. The results of Corbridge *et al.* (1994) comparing two pen-and-paper laddering techniques against software, and those Rugg *et al.* (1992) comparing sorting techniques, were necessarily tentative, precisely because of the anticipated progress of KA technology (Corbridge *et al.*, 1994, p. 338; Rugg *et al.*, 1992, p. 288).

Other issues like the quality of the interface, which will clearly have a large effect on the quantity of knowledge acquired, will be hard to disentangle in the context of KA software, where so many tools are prototypes. In such circumstances, it is near-impossible to control for side-effects of the presentation of a technique, and a sound technique

may receive a poor evaluation because it has been poorly implemented (or alternatively a flawed technique may get a great evaluation because of a clever implementation).

Other contextual problems are easier to address, though at the cost of driving up the scale of the experiments. For example, it is unusual for a KA tool to be used in the standalone fashion with no other tool support. Therefore, a straight-forward experimental evaluation of a tool or technique might produce distorted results. Moreover, what may seem to be a poor gain in terms of knowledge acquired might in fact be very valuable if the knowledge is very difficult to acquire by other means.

In many of the experiments reported in this section, the effects of tools used in tandem were evaluated, as noted above, by using different combinations of tools together (Burton *et al*., 1987, 1988, 1990*b*; Corbridge *et al*., 1994). Subjects were divided into four groups, each one of which would have to use one of the four combinations of either protocol analysis or interview followed by either laddering and card sort. This design ensured that each subject got to use one natural technique, and one contrived technique.

Although the design seems to obscure the effects of any single tool, in practice the results were quite clear, perhaps due to the small numbers of tools and combinations available. For instance, the good showing of the two contrived techniques, and the way that they tended to elicit knowledge that complemented the knowledge acquired by the natural techniques, came out very conclusively, as did the resource intensive nature across contexts of the protocol analysis technique (which did acquire knowledge difficult to acquire by other means) (Burton *et al*., 1990*b*, p. 177). Nevertheless, there would be difficulties of scale-up if a larger group of tools were simultaneously to be evaluated in a series of KA contexts.

3.5. QUANTIFICATION

Finally, there are difficulties involved with the quantification of results. Knowledge is not something, strictly, that you can have a quantity of, so different metrics were devised within the ACKnowledge project (Berger *et al*., 1989) for different knowledge representation formalisms. Three types of formalisms were proposed, as outlined in Section 3.2, pseudo-production rules, frames and semantic nets. The measures of gain would be different, of course, in each case. In the end, it was decided that production rules would be used for the experiments, mainly because of the greater interpretative effort required for the other two formalisms. In particular, in many of the experiments, knowledge acquired by experts was fed back to them, and it was felt that the relative simplicity and intuitiveness of production rules would make things easier.

The representation chosen was not a computer-readable language, but a version of English understandable by a novice and providing a level of structure above that of the raw data. All the knowledge was coded into rules of the form "IF condition (AND condition AND …) THEN action." Various metrics could be applied to the knowledge base code in this way to quantify the knowledge; in particular, the measures used were the number of IF-clauses, the number of AND-clauses and the combined total of IF- and AND-clauses (Burton *et al*., 1990*a*, p. 21).

There is no best or right way of measuring the knowledge acquired, but it is worth noting that if a frame-based language had been used instead of the rule-based one, with with the measure of gain being the number of frames acquired together with the degree of

inheritance (Berger *et al.*, 1989, p. 14), the same experimental knowledge acquisition session might have produced very different results. But a decision had to be taken, and then adhered to, in order that good comparative results could be achieved. If the experiments were repeated now, it would be more sensible to use one of the emerging standards for knowledge representation, such as Ontolingua (Gruber, 1993), though even then there would be problems with over-representation of some aspects of knowledge and under-representation of others, such as task and procedural knowledge.

In measuring gain-for-effort, the parameter used was gain per minute (Burton *et al.*, 1987, 1988, 1990*b*). The time was calculated in two parts. First the time taken with each elicitation session was recorded. The results of each session were then transcribed and formulated in the pseudo-production rule language [see Shadbolt & Burton (1990) for a description of the method of encoding], and the time taken to do this was recorded as well. Then the measure of effort is the number of conditional statements in the knowledge base divided by the sum of the time taken to elicit the knowledge and the time taken to transcribe it. One effect of this idea is that techniques that produce relatively structured knowledge (e.g. laddering, sorting) will, all things being equal, produce greater gain per unit of effort, because of the greater time taken translating relatively unstructured knowledge acquired using other techniques such as interviewing or protocol analysis (Burton *et al.*, 1988, pp. 86, 87).

Other relevant issues were generally ignored in these experiments, although in practice it would not be difficult to incorporate measures of them into the gain per minute measure. Examples, in which extra effort would be spent include: the acquisition of sufficient knowledge to run the KA session by the knowledge engineer in advance of the session; preparation of materials (e.g. prior to a self-report); selection of a suitable expert [Berger *et al.* (1989, p. 12) give a discussion of these issues].

Indeed, some of the experimental results using psychometric testing procedures strongly implied that it was worth spending time evaluating experts before going into a KA session. For instance, Burton *et al.* (1987) found positive correlations between both the total amount of effort and the effort required to code transcripts of laddering sessions, and subjects' embedded figures test (EFT) scores. The EFT test involves finding a particular shape amongst others, and measures the ability to manipulate spatial constructs. The term "field-dependent" is used with reference to the test to indicate subjects who tend to be overwhelmed by context. Burton *et al.* deduced from their result that field-dependent individuals with high EFT scores would have difficulty with a spatial technique such as laddering. Corbridge *et al.* (1994) found that introverted subjects (on the Eysenck personality inventory test) took longer to interview but produced more rules and clauses than extroverts.

Another issue relevant to gain-for-effort is that of the optimal number of elicitation sessions with the same expert. A number of experiments investigated this, including Corbridge *et al.* (1993) and experiment 1 of Corbridge *et al.* (1994). The results suggest that neither feedback nor training have particularly marked effects on the gain per effort in the techniques investigated. However, further investigation may be required to establish when effort has to increase to extract similar quantities of gain.

It can be seen that several different approaches have been taken to counter each problem discussed here. The responses that experimenters made to the problems we have outlined are summarized in Table 2.

TABLE 2
Responses to the problems of experimental evaluation of KA tools

Problem type	Response
How many experts?	Use small numbers of domain experts Use a single expert and interrogate him or her in depth Use undergraduates to model expert behaviour Use trivial, rich domains with large numbers of experts
The nature of the acquired knowledge	Use gold standards to measure coverage and quality where possible (i.e. when large numbers of non-experts are being experimented on) Measure inferential power of knowledge represented as production rules using metrics from formal grammar theory Measure inferential power of knowledge represented as frames using inheritance Measure inferential power of knowledge represented as semantic nets using metrics of local connectivity
Knowledge acquired in different domains and tasks	Experiment in a wide sample of domains Experiment in the context of as many tasks types as possible, maybe using small task hierarchies (e.g. KADS-I) to guide the sampling of the task space
Forms of tools, context of use	Test different implementations of the same technique against each other Test software implementations against pen-and-paper versions of the same technique Test groups of tools in complementary pairings Test different orderings of the same set of tools Test the value of a single session against multiple sessions Test for feedback effects in multiple sessions Exploit techniques from the evaluation of standard software to control for effects from the interface, implementation platform, etc.
Quantification	Count numbers of rules in a rule-based representation Use a standard representation, such as Ontolingua Time the session, and add on preparation time before the session and any coding time afterward, to give a measure for effort of rules/minute Calibrate results against psychometric tests of the experimental subjects

4. Attempts to evaluate frameworks

The design of a experimental programme to evaluate individual tools or small workbenches is not trivial, enmeshing the designer in a whole web of questions, some logical, some logistical, for which there is probably no one best solution. When a framework is to be measured, the problems scale up dramatically from even this high point. Breuker and Boer (1998) report that "what initially appeared as a straight-forward implementation and testing exercise [validating the problem-solving methods collected and indexed in

the CommonKADS library for expertise modelling] easily grows to unmanageable complexity when all issues that come up are taken into account" (p. 1).

One very significant problem is the switch from the specific to the general. Specific systems, and even specific KA tools, have specific purposes, and their performance can be measured, maybe with difficulty, against these purposes. Frameworks have a wider specification, and as Breuker & Boer (1998) point out, we have to establish the scope for reuse.

In addition, a methodology, method or model set is always incomplete. There will be conditions that have to apply for them to be used, and then some domain knowledge will need to be acquired. Each framework will have something different to say about which knowledge or how much knowledge to acquire, and strict comparisons will be hard. Even reusing an ontology cannot be strictly automated, and work has to be done to customize an ontology for a particular task (Uschold *et al*., 1998).

This massive scale-up of the degrees of freedom means that it is even harder to apply the experimental approach to the evaluation of KA frameworks. Nevertheless, there have been attempts to manage it. Most famously, there have been the Sisyphus experiments in which benchmark problems are set and the various solutions collected. We discuss these in Sections 4.2 and 4.3. But we begin with a rare attempt to continue using the methods of experimental psychology along much the same lines as the experiments reported in Section 3.

4.1. AN EXPERIMENT ON THE VALUE OF MODELS

As part of the ACKnowledge project's investigation into the efficacy of KA techniques, an experiment was performed to evaluate the use of models in knowledge engineering (Corbridge, Major & Shadbolt, 1995). The experimenters, while recognizing the major theoretical problems associated with evaluating models and methodologies, tried to keep to a simple design; a between-subjects design (to keep experimental conditions entirely separate), where the subjects would perform a KA task with either a detailed model, a less-detailed model or no model at all.

The detailed model was a model systematic diagnosis that had been developed as part of the ESPRIT project VITAL which had originally been intended as descriptive of the unsystematic, heuristic elements of systematic diagnosis. It had been developed from actual protocols of medical diagnosis, and captured the revisions of rankings of hypotheses that continued throughout the problem-solving , together with the element of routine that accompanied the diagnosis, where hypotheses which were barely plausible were still tested for because of either their seriousness or the simplicity and minimal cost of the testing procedure.

The less-detailed model was an epistemological model that merely divided knowledge into three layers. Subjects were told that knowledge could be divided into expertise concerning the ordering and choosing of tasks to perform (strategic knowledge), expertise about how to perform the required tasks (task knowledge) and the concepts in a field of interest and their structures (domain knowledge). This last could be divided into knowledge about objects and their interrelationships (static knowledge) and knowledge about actions and events in the domain (dynamic knowledge).

The domain chosen was the diagnosis of respiratory diseases, the domain from which the detailed model had been drawn. This is a real-world domain in which significant expertise is required. Because of the requirement for subjects to be familiar with typical AI knowledge representation techniques, it was essential that they be experienced knowledge engineers; the nature of the experiment effectively ruled out the use of students or members of the general public. Fifty five knowledge engineers agreed to be subjects (12 from the ACKnowledge project, 18 participants in the Banff KAW workshop, six participants in EKAW, 14 university-level academies and five from industry), of whom 29 actually returned usable responses.

The task to be performed had to be simple enough to be performable in a small time, while also representative of actual KA practice. It was decided to ask the subjects to edit a transcript of a diagnostic interview between a doctor and a patient, including comments added afterwards by a doctor, into a knowledge base. Two transcripts were given. Each subject who completed the experiment had 3 hours to study the protocols and to build a knowledge base. Nine subjects were also given details of the epstimological model, nine were given details of the model of systematic diagnosis, while the other 11 received no guidance about modelling at all.

Four metrics were used in the analysis. The first was the number of disorders correctly identified, out of the 20 possible. The second was the number of knowledge fragments extracted, based on a gold standard scoring sheet. The third was the organization of the knowledge fragments; a list of the most commonly used terms by subjects to organize the knowledge extracted from the protocols was tabulated for each condition. The fourth was based on a post-experimental questionnaire which assessed aspects of the subjects' responses to the experimental task.

Of course, such an experimental set-up can only begin to address the big issues underlying model use. There are other KA tasks than extracting knowledge from protocols to be tested, other domains than respiratory medicine, other models and model-types than the ones presented, and other problem-solving types than diagnosis. But in the absence of an experimental design that would cover all the important bases this was a start. Hence, the results cannot be called conclusive, but equally the experimental design itself, being relatively straight-forward, is easily reproducible. This single experiment would be much more valuable as one of a set of similar experiments about modelling and KA.

In many ways, it is unfortunate that similar work has not been performed since the Corbridge *et al.* study. This is particularly the case because in fact the experimental results showed that the subjects using no model at all outperformed the others! For instance, those using no model identified a mean of 75% of the diseases and 41% of the available knowledge fragments, better than those who used the systematic diagnosis model (55 and 34%, respectively) or the epistemological model (50 and 28%).

Given the consensus in KA about modelling, these results should have been worrying. The present writers, identified recently as being among the main figures in the consesus (Menzies, Cohen & Waugh, 1998, p. 1), are particularly sensitive to the potential significance of the results. However, a possible explanation could be the effects of a U-shaped learning function in which the use of unfamiliar or new models initially impairs performance, but over a longer period of time causes improvements. This claim could be associated with a longitudinal study of model use.

In the absence of duplicating experiments, it is clearly impossible to rule out any of the competing explanations of the results. Maybe the experimental design was flawed. There was a skewing of the composition of the three subject groups (the epistemological model group contained fewer people currently active in knowledge acquisition, although the mean duration of KA experience was about the same as for the other two groups) which could have had an effect. It is possible that the particular models chosen were confusing or misleading (although most subjects claimed, in their post-experimental questionnaire, to understand them and feel comfortable with them). A fourth possible explanation is that the experiment did not model the use of models well enough, since an actual KA context a model would have been used to guide the taking of the protocol and structured interviewing. Perhaps subjects should have been able to use models of their own choice to extract knowledge in some way.

But of course there is a fifth explanation and that is that the use of models certainly does not improve and may even hinder the acquisition of knowledge at least in some contexts. We need to know whether this is true, and if it is true, we need to know in which contexts it is true. The KA community owes it to itself and its clients to investigate this issue promptly and effectively.

4.2. SISYPHUS I & II

The main approach to the evaluation of methodologies during the period under review has been the series of Sisyphus studies. Sisyphus was an initiative that began at the 1990 EKAW to try to unify the disparate spread of approaches to structured knowledge engineering. It was thought that, despite the obvious diversity, there should be shared concepts underlying the approaches, partly because there seemed, intuitively, to be a lot of common ground at meetings of the community, and partly since it is reasonable that methodologies that spring up to address the same problems will by and large develop in similar directions.

To test this, and to help consolidate common ground, it was decided to distribute a problem to the various research groups. Once each group had applied their methods to to the problem, the result would be a number of case studies which could be used as a basis for detailed comparisons. The domain chosen was a room allocation problem, a self-consciously toy domain, though not without interest. For years, people across the globe devoted all their working hours to explicating the Delphic utterances of Siggi the Wizard, and wondering what would happen if Monika X took up smoking. The final set of papers were collected in this journal (Linster, 1994).

It was always clear that a single experiment would be of less interest than a series. Hence Sisyphus II was begun. Here, it was thought that the toy domain used in Sisyphus I might possibly have been misleading, and the domain was made somewhat more realistic and knowledge intensive by "borrowing" the application data from an already existing expert system, VT (Marcus, Stout & McDermott, 1988), which was in the domain of elevator design (also providing a synthetic task type in contrast to the analytic task of Sisyphus I). The Sisyphus II papers appeared in this journal, together with an introduction (Schreiber & Birmingham, 1996) which includes a detailed history of the Sisyphus initiatives to date.

The Sisyphus initiative is continuing, and Sisyphus III, which is in progress, will be discussed in Section 4.3. So far, it has had many beneficial effects in the KA/KE community. The strengths and weakness of the various approaches are becoming more visible and vocabularies are moving together instead of drifting apart. Additionally, the early Sisyphus studies have provoked participants into "tightening-up" discriptions of their KA methods and indeed the methods themselves.

However, we must distinguish these beneficial effects from the specific evaluation issues. Sisyphus I and II may have been important prerequisites for experimental evaluations of KA/KE methodologies, but did not in themselves constitute evaluations. There were two sets of reasons for this, one set to do with the structure of the studies, and the other set having to do with the specific details of the problems chosen.

The studies themselves were unlikely ever to bring any formal results that could be used to evaluate the participating methodologies. In the first place, no formal evaluation was performed; the technique of comparison was to bring papers together, firstly in an informal collection as a workshop track, and then, in more polished form, in dedicated special issues of this journal. The result was, in each case, half a dozen or so papers collected together. No "referees" were appointed to decide who did well, or who badly, in which respects, and any conclusions were tentative and provisional.

In the second place, there were no formal evaluation metrics against which performance could be measured, except a couple of test cases. In Sisyphus I, the system needed to reproduce a particular room allocation, and then deal with a minor change in the data. In Sisyphus II, a test case was provided. Other than hitting those performance targets—which all the systems did—there were no metrics for the effort involved in building the system, or the time the system took to perform the task. All the authors for the special issue on Sisyphus II, for example, were asked to evaluate their systems, but the metrics and dimensions were not fixed and the results across the various papers are not commensurable. There was no requirement to assess the amount of knowledge engineering effort that was involved. It is arguable that if there had been more formal assessment metrics, the level of participation might have been considerably lower.

The second set of difficulties with the evaluation followed from the problem content rather than the structure. For example, both Sisyphus I and II were concerned mainly with the modelling of knowledge. The knowledge was provided, in Sisyphus I via a very structured protocol, and in Sisyphus II by a detailed and meticulous description of the process of designing an elevator. Hence, in neither case could a realistic process knowledge acquisition take place. Not only were the participants not asked to evaluate the effort required to produce a model of the domain knowledge, but also the experiments failed to model many of the KA phenomena of real applications, such as the gradual accumulation of knowledge, the calculations of cost-effectiveness that may rule out certain techniques or tools, or the different formalisms in which knowledge acquired by different techniques (e.g. machine learning, laddering, elicitation from textbooks or protocols) may be presented.

None of the above is intended to imply that Sisyphus I and II were poorly designed or otherwise deficient. They had their own purposes, and were very beneficial to the community. But as far as evaluation issues in particular are concerned, they only scratched the surface. Sisyphus III was intended to provide much more evaluation and quantitative comparison of the participating methodologies.

4.3. SISYPHUS III

The third Sisyphus experiment was an attempt to "fill in the gaps" in the kinds of evaluation knowledge collected by Sisyphus I and II and to supplement the relatively small amount of information they had produced about the earlier stage of KBS construction, as mentioned in the Sisyphus II editorial.

> On the negative side, most researchers have concentrated mostly on problem solving, and have not considered knowledge acquisition. Since getting a problem solver to work is a prerequisite to getting a knowledge-acquisition tool working, emphasis on problem solving is natural. We hope that the next round of Sisyphus will encourage more research on knowledge aquisition (Schreiber & Birmingham 1996, p. 280).

Therefore, it was decided that Sisyphus III would focus less on the knowledge modelling and implementation aspects of KBS construction and more on actual knowledge acquisition and the process of transforming acquisition results into knowledge models. This process information could then be related back to the final systems produced by participants, giving previously absent information about the cost–benefit trade-off of the various activities that had been pursued. In this way, the methodologies and activities that contributed to systems could be evaluated, as well as the implementations.

In order to make a quantitative comparison of systems and methodologies, Sisyphus III was more strictly experimental in design than its predecessors had been, representing a move in the series away from the exploratory component of empiricism (Cohen, 1995). According to Cohen's fairly standard characterization of empirical method, preliminary exploratory studies yield causal hypotheses that are tested in observation or manipulation experiments. As discussed above, while the earlier experiments had allowed different knowledge engineering methodologies to be applied to the same problem, they did not provide a formal structure within which to evaluate and compare the approaches objectively. Sisyphus III explicitly stated the criteria on which systems were to be evaluated from the outset, and proposed that complete systems be submitted to a "referee" for evaluation with a standard "test set" of data at the conclusion of the experiment.

This change of emphasis was achieved in several ways. Firstly, Sisyphus III followed the general trend of the series towards more realistic and complex domain information—a wider variety of KA techniques were applied to a panel of experts in the chosen domain (igneous petrology). In addition, this material was not presented in a single coherent document, but in a staged series of releases in order to model genuine knowledge acquisition as a process of discovery occurring over time. To add a simple simulation of pragmatic constraints, an economic model was introduced, in which resources had an associated cost in terms of time and money. Finally, an additional requirement was made of the participants—that they should keep records (or knowledge engineering meta-protocols) concerning the processes that they went through in the KBS construction process (Shadbolt, 1996).

The concept of outcome evaluation metrics was explicitly introduced at the outset for quantitative comparison of the submitted systems. The dimensions chosen were the following.

- *Efficiency*—the time and effort spent to produce a system of a certain fidelity.
- *Accuracy*—the discriminatory and exploratory power of the built system.

- *Completeness*—the coverage of the system with respect to certain fixed points in the domain.
- *Adaptability*—the ability to extend the problem-solving functionality of the system.
- *Reusability*—the ability to reuse elements of the system in either extending the domain coverage or the problem solving functionality.
- *Traceability*—the ability to relate elements of the final system behaviour to its provenance in original KA material.

The first stage of Sisyphus III has been completed, and the second set of material released. Although the experiment is not finished, it is possible to describe in general terms the story so far. The first clear trend that can be extracted is that as the Sisyphus experiments have moved further towards realism and complexity (and therefore resource-intensive), the level of participation has dropped. Both the previous Sisyphus studies showed a "critical mass" phenomenon regarding participation—once a significant proportion of the community had put forward a Sisyphus solution, the problem became a community benchmark. It was then in the interest of all researchers to propose their own solution to maintain credibility. This has not yet been the case with Sisyphus III—it has seen the lowest level of participation of all the experiments. Additionally, the groups that did participate chose by and large not to complete any protocols of the activities that had been undertaken, used external sources of domain information, and voiced concerns about the design and execution of the experiment.

Some of these concerns related to the domain material—it was argued that this was in some senses both too realistic and not realistic enough. It was artificial in that as the material was released in stages over time, there was no provision for the knowledge engineers to interact with the domain experts and pursue lines of inquiry, or fill obvious gaps. At the same time, the inconsistencies that arose from using a variety of genuine domain experts whose styles of description and levels of expertise differed also caused dismay and resulted in some participants abandoning this materials altogether.

The low level of participation has meant that Sisyphus III will not be the large-scale structured evaluative comparison of methodologies it was intended to be. Although increased participation in the later stages would no doubt be informative, any groups joining now will have the advantage of observing the results of stage 1, so the evaluation of process may be impeded. However, that is not to say that it has not been instructive. There is certainly some value in speculating about why the experiment has unfolded in this way, and what this implies for future efforts.

Firstly there are pragmatic concerns. One reason that some groups were unable to participate in such a large project as Sisyphus III was that their funding bodies were unwilling to finance an endeavour which did not directly contribute to their current projects. Without a source of funding, it was impractical to participate considering the amount and complexity of the material and the longitudinal design of the experiment. Those researchers who did participate had managed to square participation in Sisyphus III with their own research agendas; for example, Ph.D. students and researchers intending to make a demo KBS already.

Another contributing factor to the low participation rate in Sisyphus III was the suggested use of quantitative metrics for comparative evaluation of the systems built. From the point of view of the individual or research group, if it takes a lot of effort to

participate in the experiment and it is possible that the result may be a detailed and unfavourable comparison with other methodologies, then participation is very high risk, particularly if funding difficulties will mean that the system built will not be a thorough demonstration of the methodology concerned. To be seen to "fail" against a community benchmark could be disastrous.

This obstacle is compounded when we consider that KA is a field which is very open to commercial possibilities, and could be said to have a foot in both the camps of academe and industry. This can mean that some members of the community might serve their own interests better by avoiding complete openness with respect to their system building efforts. In contrast, the academic emphasis on using intellect to overcome challanges could be held responsible for the fact that researchers "took the initiative" when faced with imperfect elicited material and turned to textbooks and other sources. Whilst this was a reasonable "lateral thinking" solution to the problem in terms of outcome, it makes process evaluation very difficult as there is a new confounding factor—the use of different materials. It could be suggested that the use of quantitative metrics for comparative evaluation produces such a great pressure on producing a good final product that it becomes very difficult to evaluate process.

The results of Sisyphus III at this stage also seem to indicate that if future large-scale evaluation attempts of this kind are to take place, it would be advisable to find some way to integrate them into the practical goals of participants as well as the long-term theoretical needs of the community.

5. The way forward

This has been an extensive review of the experimental work on evaluation performed from or connected with the Psychology Department at the University of Nottingham. To reiterate, we do not wish to claim that good work has not been done elsewhere, nor do we want to say that all evaluation should be experimental. However, we feel enough has been done to show that experimental evaluation in knowledge engineering and acquisition is a worthwhile endeavour, and should continue. In this section, we will build on the decade or more of experience we have amassed to suggest ways in which a fully fledged programme of experimental evaluation might continue.

5.1. THE EXPERIMENTAL EVALUATION OF INDIVIDUAL TOOLS

Section 3 above discussed much of the experimental work performed on individual (or small combinations of) tools or techniques for knowledge acquisition. We listed various problems that were inherent in any experimental programme, and showed how these problems had been approached.

In general these experiments appear to be largely correct and for the evaluation of individual tools there seems to be no reason why such evaluations should not continue where funding can be found. The main problem was that of getting sufficient numbers of experts together for an experiment. As we discovered, it was possible to draft in larger numbers of lower-grade experts such as students, or to trivialize domains to make sure that there would be sufficient numbers of "experts" available (e.g. fruit identification). Then, however, there would always be a lurking suspicion that something qualitatively

different was being left out of experiments that would have been present had "real" experts been used.

There is no best or right answer to this dilemma (although the greater the amount of funding the greater the probability of amassing enough experts to sidestep the problem). However, it would make sense to design a series of experiments which not only evaluated KA tools but also the evaluation process itself, by comparing results of experiments with "genuine" experts with those from the "ersatz" experts. There would clearly be many degrees of freedom, but if it could be shown that "ersatz" experts did act as models of experts, the result would be a big boost towards the experimental evaluation effort. Indeed, as we have seen, there was a tendency for experiments with experts to confirm the results of experiments with students, which is encouraging. Equally, if expertise cannot properly be modelled, then this also has a great deal of impact on our understanding of the experimental results discussed in Section 3.

One further possibility is evaluation during development. This is common practice in software design methodologies, such as SSADM, which incorporate evaluative review stages (albeit non-empirical ones) throughout the design process (see, for example, Ashworth & Goodland, 1990). Specific expert system design lifecycles (e.g. Gaschnig, Klahr, Pople, Shortliffe & Terry, 1983) can also include evaluative stages, but as Menzies *et al*. (1998) point out, as the lifecycle proceeds the practice of these evaluations becomes less and less common. If small evaluation experiments were routinely built into the project management cycles of tool development programmes, over time an impressive body of evaluative material could be created. Clearly, such results would have to be reproduced and checked by other members of the community to be properly confirmed, but such a development methodology can only be beneficial to the community in terms of both of its success and reputation.

5.2. THE EXPERIMENTAL EVALUATION OF FRAMEWORKS

In Section 4, we discussed two types of experimental evaluation of frameworks. The first type of experiment consisted of experiments of the type discussed in Section 3, asking subjects to perform set tasks, while controlling variables and then performing statistical analyses on the results. It seems clear that such an experimental methodology could not provide all the evaluation that KA frameworks need; the scale-up problem would be too great, and the number of experiments that would have to be performed to isolate the value added by KA methodologies too large. However, given that the one experiment of this type performed gave the extremely inconvenient, yet statistically significant, result that using models for KA was positively harmful (Corbridge *et al*., 1995), it would seem imperative to try to show where the experiment went wrong (if it did), or to confirm it (if it did not). Hence, it would seem to be incumbent on our community to continue to perform experiments of this type to establish once and for all whether or not (Corbridge *et al*., 1995) was a rogue result.

5.3. THE FUTURE OF SISYPHUS

The main effort, though, for the evaluation of frameworks should be concentrated through Sisyphus or Sisyphus-like programmes. However, given the disappointing

results from Sisyphus III, the first Sisyphus study designed specifically for evaluation, we should review how best to design a plan for Sisyphus that will maximize cooperative participation to allow proper evaluations to be performed.

One problem with the study was that the participants were reluctant to produce protocols of their own problem-solving. There are, no doubt, a number of reasons for this. KA researchers will generally enter a programme such as Sisyphus with priorities different from the organizers, and may be more interested in concentrating on the problem at hand and engineering a decent result. Further, the preparation of protocols can seem time-consuming and tedious (and, when it exposes mistakes or wrong turnings, sometimes embarrassing). On top of this, it is clear that, given the funding structures that currently dominate technology-based science, it is not in any researcher's interest to expose his or her methodology to negative assessment.

Nevertheless, there is great deal of evidence that while protocols may not be the optimum tool for eliciting perceptual expertise (Burton *et al*., 1988), they can increase our understanding of process. Verbal protocols have become a standard method for illuminating process-oriented areas such as accounting (Belkaoui, 1989) and user testing of computer poducts (Denning, Hoiem, Simpson & Sullivan, 1990), and the more we know about the process of applying a KA methodology to an application, the better our methodologies should be. In order to share our knowledge of how to conduct knowledge engineering, we must be able to reflect on our practices, articulate them and document them. Protocols are a useful tool for doing just that. We see the aim of using protocols as one of KA process (not expertise) acquisition. The people who use the technologies and methodologies we create need guidelines and descriptions of the characteristics of the KA process. If our ideas are to be taken up, we must meet these needs. A final point to note on the subject of protocols is that the designers of methodologies may not be the best people to supply information about the process of KA in practice, as they do not necessarily practice KA with experts on regular basis.

The result of protocolophobia is that the only products available to an evaluation programme are completed systems. But completed systems in KA are not always built in accordance with the engineering principles in the development methodology; more often than not they are built by researchers, familiar with their own methodologies, who can reinterpret their work creatively to produce impressive results. In that case, evaluating completed systems is not necessarily a good way to evaluate methodologies.

A third problem we noted was that the complexity of the domain, which is a consequence of its realism, meant that any attempt to solve Sisyphus III would be resource-intensive, and therefore that funding would be difficult to secure for the several man-months required. This reduced participation, whereas of course the more groups that take part in such initiatives, the better.

One solution to these problems would be no run Sisyphus in such a way that the groups who designed methodologies—i.e. those most interested in their getting a good evaluation—do not participate. The participants should be disinterested parties, and one sufficiently numerous group suggests itself immediately: students on courses of knowledge engineering or knowledge-based system development.

This would meet most of the objections to the design of Sisyphus III. Students can be ordered to keep protocols of their problem-solving as part of their course requirements. Poor or inadequate students could be identified by virtue of their results in other

examinations, and the numbers taking part should be large enough to produce statistical significance (particularly if the experiment was performed across several sites). Funding would not be a problem, because the Sisyphus problem could be set as a coursework examination. In that event, students would have an incentive to produce an adequate system if possible, while still not having a strong connection to the methodology they were assigned to. Indeed, the fact that untrained people would have to use methodologies would entail that KA methodologies would have to be made clearer, and their steps more precise, which would force methodology developers to sharpen their ideas, and focus more on the essential step of producing adequate documentation. The results would also be reproducible (different domains and task types could be tried from year to year). The downside of such methodological regimentation is that KA methodologies would have to become more rigid in form, leaving less room for creativity.

Of course, this structure for Sisyphus would not solve all problems. Expertise would still be a problem, although the staged and quantified release of material is still the simplest model of a genuine KA process. The realism of the material (i.e. its contradictory nature) should not be an objection, because that is a pervasive problem in representations of expertise. But for evaluation purposes, it is essential that everyone has, at least in principle, access to the same expertise; seeking our textbooks, etc., is of course good knowledge engineering practice, but is fatal to the experimental structure. Students, we suspect, would be less likely to do this.

One helpful side-effect would be that a massive expansion of Sisyphus would be possible. This would mean, for example, that individual results would be less damaging, since it would be possible to control for different types of domain, and methodologies unsuited to one domain might be allowed to shine in the next. It would also mean that improvements in methodologies over time could be seen objectively.

The evaluation of frameworks, under the current system, is a risk, and if Sisyphus III is evidence, an unacceptable risk in the view of many KA researchers. The Sisyphus experiments we have described should help alleviate that risk while also providing greater benefits to the knowledge engineering community. This need not be the only type of Sisyphus experiment—the community has always used Sisyphus in different ways and that need not change—but it should certainly form an important core of work.

6. Discussion

Evaluation is a as tricky and complex issue. As we have tried to show, this is partly due to intrinsic problems with the field, but also, as has been implicit through the paper, there are extrinsic problems that the knowledge acquisition community may not be capable of dealing with alone. We cannot discuss evaluation without briefly mentioning those extrinsic factors.

First among these extrinsic problems is the political problem of the funding for knowledge acquisition and knowledge engineering work. Primarily, funding goes towards the development of new prototypes. However, for a prototype to be of any use, it has to be evaluated and the problems ironed out, before being brought to market. This virtually never happens.

Second, there is the problem that people in knowledge engineering often do not follow a traditional scientific method, by which results are supposed to be reproducible.

Third, there is the technological problem that the field moves on so quickly that evaluations may take too long to perform; by the time they were complete, the tool on which they were focused could easily be out of date.

Fourth, there is the logistical problem of the assembling of the numbers of experts required to give statistical significance to experiments. With such a small and heterogeneous user base, getting reproducible and generalizable results is problematic.

These extrinsic, pragmatic problems might raise more serious questions about the value of experimentation. Throughout this paper, we have discussed empiricism in KA under the more or less implicit assumption that experimentation is worthwhile. Nonetheless, one might question this assumption, claiming either that empiricism is unlikely to yield interesting results in KA, or that the interesting results that are obtained take too much effort and time. In response to the former claim, it seems reasonable to point to the many results of the experiments discussed in Sections 3 and 4.1, results which have not been uncovered by any other non-empirical method of analysis.

In response to the latter claim (that of overwhelming resource demands), it is important to be realistic about the effort involved in experimentation, and not to raise expectations about the ability of the empirical approach to solve every problem that might arise. Intrinsically, the biggest problem in the empirical evaluation of KA/KE, as we have seen, is with the larger-scale, reusable, elements of knowledge engineering practice, the methodologies, the models, the ontologies. Because they are supposedly reusable, they need to be evaluated across a range of possible domains and task types, Even if sufficient resources are available for this, designing experiments that can pinpoint the value added is not trivial.

It is therefore probably unrealistic to expect KA research to become Popperian disinterested hypothesis testing; there are too many variables and competing interests for that. Only the painstaking development of a culture of evaluation would give us a complete picture of the efficacy of the full range of KA tools and frameworks, and provide the necessary links to funding authorities. But as it stands, experimental results are too sparse to base serious funding decisions on, and as such the results are easily ignorable.

In the face of this lack of evaluative activity, our suggestion is merely that more experimental evaluation should be done, and ideally should be incorporated into the outline of KA tool development. In this way, the practice of KA research should become a little bit more like traditional scientific method. It is too much to expect experimentation to solve all the evaluative problems in KA, but we should certainly expect an advance.

The major obstacle lies with frameworks rather than tools. We suggest a continuation of a coordinated series of benchmark studies on the lines of Sisyphus. Thus far, none of the Sisyphus experiments have yielded much evaluation information (though at the time of writing Sisyphus III is not yet complete). Each Sisyphus design has been informed by the experiences with the previous studies, and our experiences with Sisyphus III led us to propose a cheaper yet more evaluative experimental structure in Section 5.

Where the evaluation of individual tools and techniques is concerned, the empirical approach can yield important results from (short series of) small-scale experiments that are relatively inexpensive and simple to design. Section 5.1 advocated the making routine

of experimental evaluation during tool development; of course such "in-house" experimentaion would have to be supplemented by independent studies to reproduce positive results. But there is no doubt that a greater volume of experimental results would improve our understanding of empirical evaluation (for example, helping to answer the question of how reliable results are that do not use highly ranked domain experts), thereby helping simultaneously to cut the costs of experimentation (e.g. by making experiment design routine) and to increase our confidence in experimentally obtained results.

We recognize that empiricism is not the only evaluation solution, and indeed the positioning of practical KA within larger fields such as software design opens up other evaluation methodologies, such as the systematic validation and verification approach suggested by Preece (1998). In terms of tool evaluation, this approach seems more easily integrated into an existing product development cycle. However, we assert that software evaluation techniques are not entirely adequate for the field of KA. Whilst tool evaluation is a significant and worthwhile endeavour, the evaluation of methodologies and models cannot be conducted independently from their embodiment in software tools using these techniques. Software engineers conducting evaluations may be faced with the problem of separating weaknesses in the functionality of a computer program from problems with the user interface. The speedy maturation of the field of KA has meant that techniques and their software embodiments have grown up almost side by side. For knowledge engineers, their problems may stem the user interface, the functionality of the program or indeed a flaw in the technique that the program has been written to embody.

To conclude, we suggest evaluation in KA uses two parallel strategies. First, there should be routine tool evaluation using software evaluation methodologies. Second, the Sisyphus series should be continued as we suggest above to try and elicit implementation-independent results for frameworks. This would have the important effect of generating valuable process knowledge as a by-product. Having considered the particular external circumstances related to the expert practice of KA as mentioned above, we feel that empirical evaluation in the Sisyphus series will be most fruitful when conducted under more controlled conditions. The power of this kind of experimentation in assessing the effect of changing one variable (be it methodology, model or ontology) is directly proportional to the degree to which other potentially confounding variables (skill of the practitioner, material used, time spent) can be kept constant. This idea is at the root of our vision for the future of Sisyphus.

Currently, a sort of Darwinian evaluation operates, whereby tools or methodologies that are perceived to have underperformed run out of research funding, or die of lack of academic interest, or cause the companies that built them to go bankrupt. But, owing to vagaries in funding policies, variable academic attention spans and an unpredictable KA software market this is surely an inefficient way of weeding out bad knowledge engineering equipment and practice from the good. As we noted earlier, there is at least as much evaluation in KA as there is in software engineering generally, but this should not function as an excuse. If we as a community want to maximize our scientific potential and our influence in the world outside, then we need to take responsibility for deciding objectively what we are doing right, and what we are doing wrong.

References

ANGELE, J., FENSEL, D., LANDES, D., NEUBERT, S. & STÜDER, R. (1993) Model-based and incremental knowledge engineering: the MIKE approach. In J. CUENA, Ed. *Knowledge Oriented Software Design. IFIP Transactions A*-27, pp. 139–168. Amsterdam: Elsevier.

ASHWORTH, C. & GOODLAND, M. (1990) SSADM—*A Practical Approach*. London: McGraw-Hill.

BELKAOUI, A. (1989) *Human Information Processing in Accounting*. New York: Quorum Books.

BERGER, B., BURTON, A. M., CHRISTIANSEN, T., CORBRIDGE, C., REICHGELT, H. & SHADBOLT, N. R. (1989) *Evaluation Criteria for Knowledge Acquisition*, ACKnowledge project deliverable ACK-UoN-T4.1-DL-001B. University of Nottingham, Nottingham.

BREUKER, J. & BOER, A. (1998). So you want to validate your PSMs? In B. R. GAINES & M. MUSEN, Eds. *Proceedings of the 11th Banff Knowledge Acquisition for Knowledge-Based Systems Workshop*, University of Calgary, Calgary: Also at http://ksi.cpsc.ucalgary.ca/KAW/KAW98/breuker/

BREUKER, J. A. & VAN DE VELDE, W. (1994). *The CommonKADS Library for Expertise Modelling*. Amsterdam: IOS Press.

BURTON, A. M., CORBRIDGE, C., DE HOOG, R., VAN HEIJST, G., KERR, I., MAJOR, N., POST, W., REICHGELT, H., RUGG, G. & SHADBOLT, N. R. (1990*a*). *Final report on KA techniques evaluation*. ACKnowledge project deliverable ACK-UoN-T4.2-DL-003-A, University of Nottingham, Nottingham.

BURTON, A. M., SHADBOLT, N. R., HEDGECOCK, A. P. & RUGG, G. (1987). A formal evaluation of knowledge elicitation techniques for expert systems: domain 1. In D. S. MORALEE, Ed. *Research and Development in Expert Systems IV*. Cambridge: Cambridge University Press.

BURTON, A M., SHADBOLT, N. R., RUGG, G. & HEDGECOCK, A. P. (1988). Knowledge elicitation techniques in classification domains. In Y. KODRATOFF, Ed. *Proceedings of 8th European Conference in Artificial Intelligence*, pp. 85–90. London: Pitman.

BURTON, A M., SHADBOLT, N. R., RUGG, G. & HEDGEOCK, A. P. (1990*b*). The efficacy of knowledge elicitation techniques: a comparison across domains and levels of expertise. *Knowledge Acquisition*, **2,** 167–178.

CHANDRASEKARAN, B. & JOHNSON, T. R. (1993). Generic tasks and task structures: history, critique and new directions. In J.-M. DAVID, J.-P. KRIVINE & R. SIMMONS Eds. *Second Generation Expert Systems*. pp. 232–272. Berlin: Springer-Verlag.

CHI, M. T. H., GLASER, R. & REES, E. (1982) Expertise in problem solving. In R. J. STEMBERG, Ed. *Advances in the Psychology of Human Intelligence*. Hillsdale, NJ: Lawrence Erlbaum.

COHEN, P. R. (1995). *Empirical Methods for Artificial Ingelligence*. Cambridge, MA: MIT Press.

COMPTON, P. & JANSEN, R. (1990). A philosophical basis for knowledge acquisition. *Knowledge Acquisition*. **2,** 141–157.

CORBRIDGE, C., MAJOR, N. P. & SAHDBOLT, N. R. (1995). Models exposed: an empirical study. In *Proceedings of the 9th Banff Knowledge Acquisition for Knowledge-Based Systems Workshop*, University of Calgary, Calgary.

CORBRIDGE, C., RUGG, G., BURTON, A. M. & SHADBOLT, N. R. (1993). Feedback and training effects in knowledge elicitation. In M. A. BRAMER & A. L. MACINTOSH, Eds. *Proceedings of Expert Systems 93, The 13th Annual Technical Conference of the British Computer Society Group on Expert Systems*, pp. 183–194. Oxford: BHR group.

CORBRIDGE, C., RUGG, G., MAJOR, N. P. & SHADBOLT, N. R. & BURTON, A. M. (1994). Laddering: technique and tool use in knowledge acquisition. *Knowledge Acquisition*, **6,** 315–341.

DENNING, S., HOIEM, D., SIMPSON, M. & SULLIVAN, K. (1990). The value of thinking-aloud protocols in industry: a case study of Microsoft. *Proceedings of the Human Factors Society—34th Annual Meeting*. Vol. 2, pp. 1285–1289. Santa Monica, CA: The Human Factors Society.

ERICSSON, K. A., KRAMPE, R. T. & TESCH-R-MER, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review*, **100,** 363–406.

GASCHING, J., KLAHR, P., POPLE, H., SHORTLIFFE, E. & TERRY, A. (1983). Evaluation of expert systems: issues and case studies. In HAYES-ROTH, F., WATERMAN, D. & LENAT, D., Eds. *Building Expert Systems*. pp. 241–280. Reading, MA: Addison–Wesley.

GRUBER, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, **5,** 199–220.

LINSTER, M. (1994). Sisyphus '91/92: models of problem solving. *Intenational Journal of Human-Computer Studies*, **40,** 189–192.

MAJOR, N. P. & REICHGELT, H. (1990). ALTO: an automated laddering tool. In B. J. WIELINGA, J. BOOSE, B. GAINES, G. SCHREIBER & M. VAN SOMEREN, Eds. *Current Trends in Knowledge Acquisition*, pp. 222–236. Amsterdam: IOS Press.

MARCUS, S., STOUT, J. & MCDERMOTT, J. (1988). VT: an expert elevator designer that uses knowledge-based backtracking. *AI Magazine*, **9,** 95–111.

MENZIES, T., COHEN, R. F. & WAUGH, S. (1998). Evaluating conceptual modeling languages. In B. R. GAINES, & M. MUSEN, Eds. *Proceedings of the 11th Banff Knowledge Acquisition for Knowledge-Based Systems Workshop*. University of Calgary, Calgary.

NEWELL, A. (1975). A tutorial on speech understanding systems. In D. R. REDDY, Ed. *Speech Recognition: Invited Papers Presented at the* 1974 *IEEE Symposium*. pp. 3–54. New York: Academic Press.

O'HARA, K., SHADBOLT, N. & VAN HEIJST, G. (1998). Generalized directive models: integrating model development and knowledge acquisition. *International Journal of Human-Computer Studies*, **49,** 497–522.

PREECE, A. (1998). Building the right system right. In B. R. GAINES & M. MUSEN, Eds. *Proceedings of the 11th Banff Knowledge Acquisition for Knowledge-Based Systems Workshop*. University of Calgary, Calgary. Also at http://ksi.cpsc.ucalgary.ca/KAW/KAW98/preece/

ROTHENFLUH, T. E., GENNARI, J. H., ERIKSSON, H. PUERTA, A. R., TU, S. W. & MUSEN, M. A. (1990). Reusable ontologies, knowledge-acquisition tools, and performance systems: PROTÉGÉ-II solutions to Sisyphus-2. *International Journal of Human-Computer Studies*, **44,** 303–332.

RUGG, G., CORBRIDGE, C., MAJOR, N. P., BURTON, A. M. & SHADBOLT, N. R. (1992). A comparison of sorting techniques in knowledge acquisition. *Knowledge Acquisition*, **4,** 279–291.

RUGG, G. & SHADBOLT, N. R. (1991). On the limitations of repertory grids in knowledge acquisition. *Proceedings of the 1991 Knowledge Acquisition for Knowledge Based Systems Workshop*. Banff, Canada.

SCHREIBER, A. Th. & BIRMINGHAM, W. P. (1996). Editorial: the Sisyphus-VT initiative. *International Journal of Human-Computer Studies*, **44,** 275–280.

SCHWEICKERT, R., BURTON, A. M., TAYLOR, N. K. CORLETT, E. N., SHADBOLT, N. R. & HEDGECOCK, A. P. (1987). Comparing knowledge elicitation techniques: a case study. *Artificial Intelligence Review*, **1,** 245–253.

SHADBOLT, N. R. (1996). Sisyphus III. Problem statement available at http://psyc.nott.ac.uk/research/ai/sisyphus/

SHADBOLT, N. R. & BURTON, A. M. (1989). The empirical study of knowledge elicitation techniques. *SIGART Newsletter*, **108,** 15–18.

SHADBOLT, N. R. & BURTON, A. M. (1990). Knowledge elicitation. In J. R. WILSON & E. N. CORLETT, Eds. *Evaluation of Human Work: A Practical Ergonomics Methodology*, pp. 321–345. Taylor and Francis.

SHADBOLT, N. R. & O'HARA, K. (1997). Model-based expert systems and explanations of expertise. In P. J. FELTOVICH, K. M. FORD & R. R. HOFFMAN, Eds. *Expertise in Context*, pp. 315–337. Menlo Park, CA: AAAI Press/MIT Press.

USCHOLD, M., CLARK, P., HEALEY, M., WILLIAMSON, K. & WOODS, S. (1998). An experiment in ontology reuse. In B. R. GAINES & M. MUSEN, Eds. *Proceedings of the 11th Banff Knowledge Acquisition for Knowledge-Based Systems Workshop*. University of Calgary, Calgary.

WAGNER, R. K. & STANOVICH, K. E. (1996). Expertise in reading. In K. A. ERICSSON, Ed. *The Road to Excellence: The Acquisition of Expert Performance in the Arts and Sciences, Sport and Games*. Mahwah, NJ: Lawrence Erlbaum.

WIELINGA, B., SCHREIBER, A. Th. & BREUKER, J. (1992). KADS: a modelling approach to knowledge engineering. *Knowledge Acquisition*, **4,** 5–53.