DEAKIN university

FACULTY OF BUSINESS AND LAW

School of
Information Systems
*Leaders in Business Computing*

## School Working Papers Series 2002

SWP 2002/25

# Automatic Refinement of User Requirements: A Case Study in Software Tool Evaluation

Author:

Jacob L. Cybulski

URL: http://www.deakin.edu.au/infosys/research/working_paper.htm

**Deakin University**

**School of Information Systems**

**Working Papers 2002/25**

# Automatic Refinement of User Requirements: A Case Study in Software Tool Evaluation

Jacob L. Cybulski

School of Information Systems
Deakin University
Email: jlcybuls@deakin.edu.au

**Abstract**

*This paper presents assessment of system effectiveness in automatic requirements refinement by comparing results obtained from experts and novices with those achieved by the system. As the investigated system was a combination of a tightly inter-connected method and a tool, the evaluation framework melded together a number of distinct methodological approaches structured into three empirical studies, which aimed at the construction of a case problem domain, calibrating the system using thus defined domain elements and finally using the calibrated system assessing its effectiveness. In consequence, it was concluded that the evaluated methods and tools were effective in supporting requirements refinement.*

**Keywords**

REQUIREMENTS ANALYSIS, SOFTWARE DEVELOPMENT TOOLS, RESEARCH METHODOLOGIES

## INTRODUCTION

Requirements engineering is an early phase in software development, which aims at complete, consistent and unambiguous specification of user requirements. It is considered difficult, time-consuming, expensive and error-prone - up to 56% of all software defects are due to errors introduced in requirements specifications, taking up to 82% of development time to fix.

One possible approach to improving the process of requirements engineering is to maximise reuse across its activities. This could be done by reusing requirements from the existing software systems or by identifying and utilising reusable designs in the process of refining software requirements for a single project. Both methods have been found to increase analysts' productivity and the quality of requirements specifications and their refinements.

RARE (Reuse-Assisted Requirements Engineering) and IDIOM (Informal Document Interpreter, Organiser and manager) were proposed as a domain-specific method-tool framework with a promise of improving requirements refinement by automatic matching of requirements text with descriptions of reusable design artefacts (Cybulski and Reed 1999). To do so effectively, RARE IDIOM had to deal with terminological difference between requirements and designs, which represent distinct domains of development knowledge, i.e. problem and solution domains, respectively. The main contribution of RARE IDIOM was to propose a novel method of classifying, comparing, searching and retrieving documents sourced from multiple domains. The core of that work was development of a scheme for classifying requirements and designs, and their retrieval using a specially designed domain thesaurus and a method of calculating similarity (affinity) of requirement and design descriptions.

Being a combination of a method and a tool, the research approach adopted in the RARE IDIOM study was based on the system-development research methodology (SDRM) advocated by

Nunamaker, Chen and Purdin (1991). A major difficulty in the undertaken project was empirical assessment of the method-tool operational effectiveness, which in the author's opinion needs to go beyond a simple "test", prevalent in the majority of the reported SDRM projects, and largely inadequate for solid research. The approach taken in the RARE IDIOM study can, thus, help SDRM researchers in structuring their projects to include tool or method empirical evaluation.

The research question of this study was formulated as follows:

- *To what extent is RARE IDIOM effective in refining requirements into reusable designs?*

Effectiveness of requirements refinement tasks was determined by measuring the similarity of RARE IDIOM advice with that obtained from expert and novice analysts. Thus, RARE IDIOM was considered effective when :-

- *Within the scope of its functionality, RARE IDIOM would approach the performance of experts and exceed that of novices.*

## DESIGN OF EMPIRICAL STUDY

The fundamental principles of RARE IDIOM operation relied on the system's ability to search and find reusable designs useful in requirements refinement. This meant that RARE IDIOM had to be evaluated either as a requirements processing tool, as a reuse tool, or as a search engine. Each view of RARE IDIOM could have demanded a specific methodology for the study (Darke, Shanks*, et al.* 1998). For example, studying RARE IDIOM as a requirements engineering process may be considered as grounded in the social context and thus require interpretive research methods capable of analysing information systems processes in their organisational context, as recommended by case researchers, such as Benbasat, Goldstein and Mead (Benbasat, Goldstein*, et al.* 1987). On the other hand, the information retrieval view of RARE IDIOM could indicate a strongly positivist's approach to this study, e.g. relying on inferential statistics, which is advocated by Salton (Salton and McGill 1983) and Keen (Keen 1992). However, as observed by Tague-Sutcliffe (Tague-Sutcliffe 1992), in information retrieval field studies, many elements of the studied phenomena, such as searched documents and the repository architecture, are commonly outside the experimenter's control, in which case a combination of interpretive and positivist approaches could prove more beneficial. In software reuse (as in many other areas of software engineering), a mix of research approaches seems to be a norm even for a single author, ranging from inferential statistics (Frakes and Isoda 1994) to case studies(Frakes 2000) .

Multi-methodological study of development methods and tools is actually the preferred approach in the SDRM methodology (1991), which draws from both positivist and interpretive approaches and which combines them to form and evaluate information systems concepts. In RARE IDIOM evaluation, several approaches recommended by Nunamaker, Chen and Purdin (1991) were therefore considered, i.e. case studies, surveys and field studies, as well as, simulations, field and laboratory experiments. Galliers and Land (1987) further clarify the appropriate use of methods depending on the nature of IS objects under the study, which may include a methodology, technology or an environment. These three categories happen to coincide with the RARE IDIOM project's main deliverables, i.e. the RARE method, the IDIOM tool and domain model used in the system evaluation. After careful analysis of the reported methods, the research framework found most suitable for the task included a single case study involving simulation and system testing, thus, providing rich empirical data that is a direct and explicit reflection of the system behaviour.

One of the most difficult tasks in case study research is the selection of a case that could appropriately demonstrate the issues and the context of interest to the researcher, and which can be effectively investigated, analysed and possibly generalised. The case selection was conducted as follows.

*Selection of case study participants*

As the first step towards a definition of a demonstration case, it was necessary to identify the case study participants. RARE IDIOM's success criteria identify two distinct groups of participants that need to be involved in the evaluation process, i.e.

- *experts* - professional systems analysts with more than 10 years of development experience; and

- *novices* - students with less than 3 years development experience (academic), with some, though limited, knowledge of systems analysis.

Previous studies showed the existence of a significant difference in the performance and quality of results generated by novice and expert analysts (Schenk, Vitalari*, et al.* 1998), esp. on tasks of relevance to this study, such as software reuse (Curtis 1989). Because of this well-known disparity in the outcomes produced by expert and novice developers, it was decided that to judge the quality of RARE IDIOM results, the answers produced by experts could be used to define a range of high quality results, whereas the answers given by novices could set a range of sub-optimal results.

*Selection of a case study domain*

There are few real-life projects with a clear focus on domain analysis and reuse, both of prime importance to RARE IDIOM. Domain analysis, which aims at the construction of a domain model to be used later as the basis for many systems development projects, is a very costly and complex process. Domains similar to those investigated by U.S. Department of Defense, e.g. army movement, command and control, army procurement or quick reaction combat capability (DoD 1995), may require hundreds of domain experts and several years of development to complete. Deployment, maintenance and use of the resulting domain model can also be very expensive and require considerable effort. Gaining access to such real-life domain models is virtually impossible!

The scale and the cost of developing real-life domain models are clearly prohibitive for the study presented in the RARE IDIOM project. Instead, a *simulated application domain* had to be developed and then used in all empirical work. Strict limitations were imposed on the scope and size of the selected domain and on the methodology used in its development. The selected domain, in both problem and solution dimension, had to define a limited vocabulary, consist of a small number of objects and documents, and be developed by few domain experts in a matter of weeks rather than years.

Selection of suitable problem and solution domains was also dictated by the participant's ability to understand domain objects and their relationships, possible computer metaphors for the object behaviour, and their representation in a computer system, as the system properties and functions.

A number of possible candidate *problem domains* were considered, e.g. a domain of commodity prices, a domain of personal banking transactions, and a domain of computer games. It was decided to establish a demonstration case study as part of the simplest domain model, i.e. computer games, which would be readily understood by student participants (novices). The games to be analysed involved metaphors of real-life objects, such as cards, coins, dice, boards and scores, which could all be manipulated and used by players in a competitive environment.

The *solution domain* was also defined in a way so that it could be easily comprehended by novice developers. It consisted of simple data structures, functions and modules that are commonly taught in many undergraduate courses, and which are also available in software development environments used by students in their projects, e.g. to implement a small interactive software system with graphical user interfaces and a data repository.

The size of both the problem and solution domains was chosen to allow students to perform simple analytic tasks in a session that would be no longer than 1 hour. The author's experience with undergraduate teaching of systems analysis and design, indicated that students can be challenged by a single page of user requirements (about 20-50 simple requirements statements) and a couple of pages of design descriptors (about 50-100 design artefacts). The conducted pilot study with students confirmed this intuition of novice analytic abilities and the case study was scoped to refect the required domain contents, size and complexity.

*Selection of case study tasks*

RARE IDIOM performance was evaluated on the basis of analysis tasks performed on a single requirements document for a simple game system. The tasks were undertaken by the system, by expert analysts and by novices, who for every requirement ranked the available design artefacts in the order of their perceived usefulness for the requirement refinement.

The case study was structured into three empirical sub-studies.

- *Test Environment Development.* The first sub-study was exploratory aiming at constructing an operating environment and test data for the investigated method/tool. In RARE IDIOM project, this involved development of a domain model with its domain dictionary, classification scheme, and a domain-specific thesaurus. This study also resulted in a collection of pre-classified design artefacts descriptions, which together with the domain model were used in all the subsequent empirical work.

- *Test Environment Calibration.* The second sub-study aimed at defining a measure that could be used to evaluate results produced by the tested method/tool. The study subjected expert and novice analysts to a small case study that involved analysis and refinement of a small user requirements document of 50 statements. The expert refinement choices were subsequently considered as of a higher quality than those given by novices.
- *Method/Tool Evaluation.* The last sub-study aimed at determining the quality of the method/tool results. This was achieved by comparing RARE IDIOM refinement choices with those obtained from experts and novices. The results of this study were then analysed to assess the success of this research project.

Considering the qualitative/quantitative nature of the RARE IDIOM data, Miles and Huberman approach to data analysis was adopted (1994, pp 10-12). In each empirical study, it therefore included data collection and data reduction, presentation of data, and finally analysis and evaluation, which may result in development of assertions, generalisations and possible verification of results.

## TEST ENVIRONMENT DEVELOPMENT

For the RARE IDIOM system to operate, it required access to the problem and solution domain vocabulary, a classification scheme and a thesaurus that could aid the retrieval tasks. To define the necessary domain information, the study necessitated a separate planning stage with a focus on domain analysis and design. While problem domain terms can be easily extracted from the samples of requirements documents. Solution domain terms, however, are normally hidden in the semantics of design artefacts used in the process of requirements refinement. It was, therefore, important to identify, collect, classify and describe such design artefacts.

### Aim of the Study

***The aim of this study was to develop a domain vocabulary, a classification scheme, a domain thesaurus, and to identify domain artefacts, which could subsequently be used by RARE IDIOM in further empirical studies.***

### Method

The process of building a domain model for this project was inspired by DARE (Frakes, Prieto-Diaz*, et al.* 1998), a well-known domain-analysis approach, which actively utilises information retrieval methods. DARE recommends that domain vocabulary should be identified from domain texts, and then analysed for conceptual proximity to form term clusters, which are subsequently used to form a classification scheme and a thesaurus.

### Data Collection and Reduction

*Identification of solution-domain vocabulary and artefacts*

RARE IDIOM's solution-domain vocabulary was constructed using Luhn's method of term frequency counts (1957). Word occurrence frequencies were counted in a large number of technical and non-technical documents (over 70,000 words). The most commonly found words were subsequently considered as part of the word stop list (or noise words). The most frequently occurring technical terms were used as candidate terms for the classification of objects in a solution domain (1000 terms) and some as candidates for the possible design artefacts in the same domain (126 artefacts).

*Forming a classification scheme*

The next study involved forming a classification scheme using conceptual distance analysis of the solution-domain terms (Besterfield, Besterfield-Michna*, et al.* 1999, Ch 11, Ch 19). Frakes et al. (1998) report that such analysis can be automated, using average-link and seed-based clustering (Rasmussen 1992), which are both supported by DARE tools. RARE IDIOM, however, used a different method of vocabulary structuring, which is based on the QFD (Quality Function

Deployment) team-based affinity analysis, as recommended by the Andersen Consulting Foundation method to elicit, identify and structure system requirements (Andersen Consulting 1994). Thus, three "domain analysts", all experts in object-oriented modelling, were invited to participate in a focus-group-like study to organise the previously collected solution domain terms into eight hierarchically structured clusters, or "facets", of 131 carefully selected classification terms.

*Classifying the solution-domain artefacts*

An experienced commercial systems analyst (over 15 years of industry experience) was then invited to act as a "domain designer" responsible for populating a domain repository with reusable design artefacts (126 design artefacts). Design artefacts were selected from the previously generated list of 1000 solution-domain terms, which referred to data types, functions and modules.

*Construction of a domain-mapping thesaurus*

The study then proceeded with the construction of a small domain-specific thesaurus using a "deductive" approach (ANSI/NISO 1993, p 27). Five documents describing different software systems from different domains (three were from the domain of computer games) were analysed, resulting in 490 problem-domain terms which were then classified to form a domain thesaurus. The thesaurus was later converted into a more convenient form of a list consisting of word pairs, representing associations between terms from the problem and solution domain. As the activity relied on exactly the same tasks as those involved in the previous activity, i.e. classification of concepts, the same domain designer was asked to participate.

### Analysis and Evaluation

As the ultimate aim of the artefact classification was to determine their conceptual distance and similarity (for the retrieval purposes), the evaluation method aimed at assessing the degree of consistency between the conceptual distances between artefacts as calculated using RARE IDIOM and as given by the domain designer, who previously classified these artefacts. To do so, the domain designer proposed 100 artefact triplets of which pairs could be checked for the relative difference in their conceptual distance.

The consistency/disparity data was subsequently collected, analysed and the proportion of the correctly calculated distances established. It was determined that 94% of artefact (relative) conceptual distances were computed by RARE IDIOM in agreement with the domain designer's expectations. What this meant was that for each requirement to be refined, RARE IDIOM would match and order its design artefacts, so that 94% of best matches would be placed high on the list of its recommendations. Assuming an even distribution of the remaining 6% of badly ordered artefacts, the risk that a skilled analyst would select an inappropriate design element or miss a design element suitable for requirement refinement was minimal.

## TEST ENVIRONMENT CAIBRATION

RARE IDIOM will be considered effective when its requirements refinement choices are more like those of experts than those of novices. Thus, it is necessary to identify expert and novice preferences in the refinement tasks and to determine whether or not their refinement choices are significantly different. Should such a distinction be observed, the expert refinement choices can then be regarded as more suitable for RARE IDIOM to return than those obtained from novices.

### Aim of the Study

***This study was conducted to determine which of the previously collected design artefacts would be most suitable for refining sample requirements statements.***

### Method

This study involved two distinct groups of systems analysts, i.e. novices (experience ≤ 3 years, only academic work) and experts (practical work experience in IT ≥ 10 years).

A short requirements document and a pool of design artefacts were also selected for the study.

Both groups of subjects, i.e. experts and novices, were asked to briefly analyse each requirement statement and to assign up to three design artefacts, which in their opinion could be useful in the refinement of that requirement. This task was unaided by any software tool. The answers were recorded and later analysed for their similarity in terms of the classification scheme produced in the previous study.

### Data Collection

The study was preceded by a pilot study involving five (5) students acting as novice analysts. Through the pilot study it was determined that the time sufficient for novices to complete all the tasks was 1 hour. The contents and size of requirements documents, as well as the size of reuse repositories were adequate and remained unchanged.

| | Novices | Experts |
|---|---|---|
| Mean | 89% | 97% |
| Median | 94% | 96% |
| SD | 17% | 3% |

**Table 1: Task completion**

Subsequently, five (5) experts (one later withdrew) and nine (9) novices were asked to analyse and refine two (2) requirements documents. The first requirements document was used to train the participants in the refinement and reuse tasks. The results obtained from the training exercise were subsequently disregarded. The second requirements document was used as a vehicle for the empirical work. The data collected from experts and novices was labelled as E1-E4 and H1-H9 respectively.

Both groups of subject recorded a high level of task completion (see Table 1), though novices displayed a greater variation in the completion rate.

### Data Reduction

To determine distinction between expert and novice refinement choices, it was necessary to assess the semantic distances between pairs of design artefacts that were suggested by both experts and novices. Such distance calculation can rely on the artefact similarity/affinity metric, which was part of the classification scheme developed and validated in the previous study. Direct comparison of the design artefacts selected by the study participants was not deemed appropriate, as it would only reveal the physical overlap between the artefact sets rather than assess their similarity.

The affinity of the refinement sets nominated by the participants was therefore calculated. The affinity of two such sets was determined using a geometric metaphor of shape proximity, whereby the closest pair of member artefacts determines the distance and affinity of the entire two sets. Using this method, it was possible to calculate affinity between any two alternative requirement refinements.

### Analysis and Evaluation

The refinement comparisons were conducted over three groups representing participant pairs, i.e. expert-expert (E-E), novice-expert (H-E) and novice-novice (H-H). The grouping resulted in samples of 250, 1342 and 1206 data items respectively, and the statistics for each group were collected and reported in Table 2, density curves of the three samples were produced and displayed in Figure 2, fitted with lowess curves (Chambers, Cleveland*, et al.* 1983).

*Distribution.* After a brief inspection of density curves for refinement affinity (see Figure 2), it is apparent that the collected samples have non-Gaussian distribution, though they all share a similar distribution shape. The unusual distribution of the three data sets indicates that their means cannot be regarded as good representatives of these sets. This also explains large standard deviations in the data (see Table 2). Instead of the means analysis, the study focussed on the samples' quartiles, which better characterise the collected data (see a boxplot in Figure 1).

The median value of affinities between refinements produced by novices and experts reveals little overall similarity between these two groups, only 64% (H-E). Affinity of results obtained from

| | E-E | H-E | H-H |
|---|---|---|---|
| Number of values | 250 | 1342 | 1206 |
| Mean | 78.70 | 70.05 | 65.49 |
| Std. Deviation | 21.13 | 21.90 | 22.50 |
| Std. Error | 1.337 | 0.5978 | 0.6479 |

**Table 2: Inter-document similarity statistics**

experts, when compared one against the other, however, was 80% (E-E), a much higher level, whereas novice-only affinities lead to an even larger disparity of results, with the median of 60% (H-H). This means that refinement results obtained from experts are similar and highly cohesive, whereas results produced by novices are distinct from those generated by experts, and dissimilar within their own group of results, which indicates randomness and scatter of refinement choices.

It is quite surprising (refer to Figure 1) that when compared with experts, both experts and novices (E-E and H-E) display very high consistency of their refinement choices at the 75% percentile of the collected data - 100%. By looking at the right-hand edge of the density curves (see Figure 2), we can see the explanation of this finding, as nearly 40% of all expert-expert and 25% of all expert-novice refinements are in 100% affinity (which means they share at least one of the three possible refinement artefacts). There is however lack of such high consistency between novices themselves, which presents itself as 83% affinity on the 75% percentile. Such disparity can be accounted by the scatter of novice data, especially in the rank of refinements with the lower similarity. In the 25% percentile range, expert refinements are again much more cohesive (59% affinity) than the refinements generated by novices (48-54% affinity).

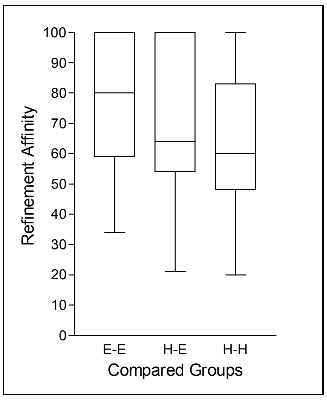


**Figure 1: Refinement affinity between pairs of participants**

*Significance.* Comparison of medians and the major quartiles indicated that there is a clear difference in the refinements produced by experts and novices. It was, hence, decided to statistically test that the observed distinction in the data sets was significant. The Kruskal-Wallis test was selected as the most appropriate for the data, which included three separate sample groups of unpaired data, had non-Gaussian distribution, though all distributions had a similar shape. The test (see Table 3) showed the differences in the samples' medians to be very significant ($P < 0.0001$).

Since the main aspect of this evaluation was to determine the differences and scatter of results obtained by experts and novices, Dunn's multiple comparison post-test was also conducted to specifically determine the differences in E-E vs. H-E and H-E vs. H-H data sets. The test showed that differences in the affinity of requirements refinements, as observed in the results produced by experts and novices, was also very significant (with $P < 0.0001$ overall, and $P < 0.001$ for all data pairs, in particular E-E vs. E-H and E-H vs. H-H).

In conclusion...

*Based on the Kruskal-Wallis analysis and Dunn's post-test, it is evident that the difference between refinement results produced by experts and novices is significant. As experts produce much higher quality results than novices, thus, results produced by experts and novices could also be used as a measure of refinement quality, e.g. "good" and "suboptimal", respectively.*

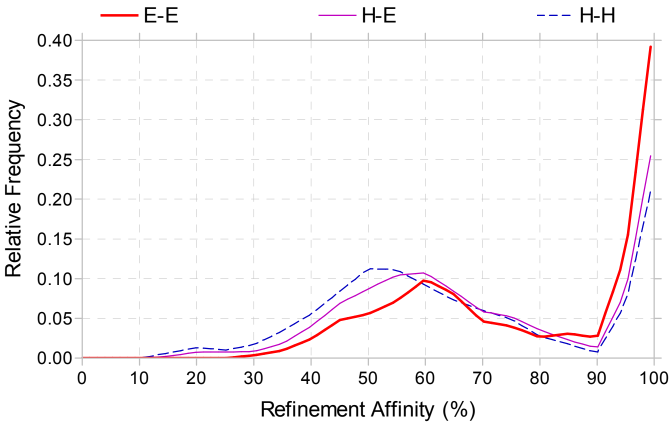


**Figure 2: Comparison of refinement affinity between participants**

| | |
|---|---|
| Number of groups | 3 |
| P value (in Gaussian approximation) | *** P<0.0001 |
| Kruskal-Wallis statistic (H) | 78.07 |
| Do the medians vary signif. (P < 0.05) | Yes |

**Table 3: Kruskal-Wallis test results (E-E, H-E and H-H)**

## METHOD/TOOL EVALUATION

The previous study determined that experts and novices make significantly different choices in requirements refinement and it is known from other sources that refinements made by experts are of a higher quality than those of novices. What needs to be shown next is that RARE IDIOM refinement choices are more alike those made by experts than those by novices, which would mean that the quality of the method/tool recommendations can also be regarded as high. Such recommendation would be demonstrated by the RARE IDIOM ranking of its design artefacts during requirement refinement, in such a way that the top ranking artefacts would include a greater number of those preferred by experts and the low ranking artefacts would include more artefacts favoured by novices.

### Aim of the Study

***To determine if in response to the task of requirement refinement, RARE IDIOM could consistently rank its reusable design artefacts, so that the artefacts which had been selected by experts would be ranked significantly higher than those which had been chosen by novices.***

### Method

The study was initiated by entering the text of a sample requirement statement into the RARE IDIOM system and using it in ranking all of the available reusable design artefacts as possible candidates in refining a subset of all requirement statements. The ranking positions of recommendations made by the experts and novices were noted and consequently analysed.

### Data Collection and Reduction

The ranking tables were produced for 20 randomly selected requirements from the case study. Each ranking table represented RARE IDIOM analysis of a single requirement. It listed all design artefacts that were available for refinement of the requirement, and sorted them by their affinity to the requirement (as calculated by RARE IDIOM). The table also indicated individual requirement refinement choices as given by the experts (labelled "E1"-"E4") and novices (labelled "H1"-"H9"). Each refinement choice was subsequently converted into its ranking number between 1 and 127 (the number of the available design artefacts) and then turned into a relative position (as a percent) in a ranking table, starting with 0% (position 127 - worst) up to 100% (position 1 - best). The relative ranking position was regarded as RARE IDIOM measure of refinement quality.

### Analysis and Evaluation

All of the refinement quality data was subsequently classified into two groups, i.e. 148 of expert generated data items (labelled "E1"-"E4") and 260 of data items generated by novices (labelled "H1"-"H9"). The basic statistics for each group have been collected in Table 4 and their density curves given in Figure 4 (smoothed with cubic splines).

*Distribution.* The density curves (Figure 4) of expert and novice refinement quality showed that 53% of all expert advice was located in the 80-100% quality range (as established by RARE IDIOM). This is in contrast to the novice analysts, who produced only 37% of recommendations in the same range. At the other end of the scale are the artefacts considered by RARE IDIOM to be of lesser value. Although RARE IDIOM classified

| | Experts | Novices |
|---|---|---|
| **Number of values** | 148 | 260 |
| **Mean** | 64.28 | 54.96 |
| **Std. Deviation** | 32.28 | 33.71 |
| **Std. Error** | 2.736 | 2.090 |

**Table 4: Refinement quality descriptive statistics**

**Table 5: Mann-Whitney test for significance of median difference**

| | |
|---|---|
| P value (in Gaussian approximation) | ** 0.0016 |
| One- or two-tailed P value? | One-tailed |
| Sum of ranks in two samples | 33651 , 49785 |
| Mann-Whitney U | 15860 |
| Are medians signif. different? (P < 0.05) | Yes |

18% of all expert recommendations as of the lowest quality (in the 0-20% quality range), it also rejected many more of the novice recommendations - 27% (in the same range).

Figure 4 also clearly shows that the distribution of the collected data is non-Gaussian. Thus, as in the previous study, the evaluation was conducted based on the analysis of median and the major quartiles, which can be represented in a boxplot (see Figure 3). The boxplot shows the distribution of expert and novice refinement quartiles, from which it is evident that expert refinements choices are placed 15-20% higher on the RARE IDIOM quality scale than the corresponding novice choices.

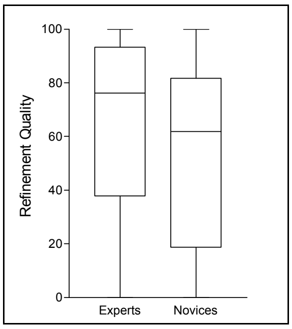


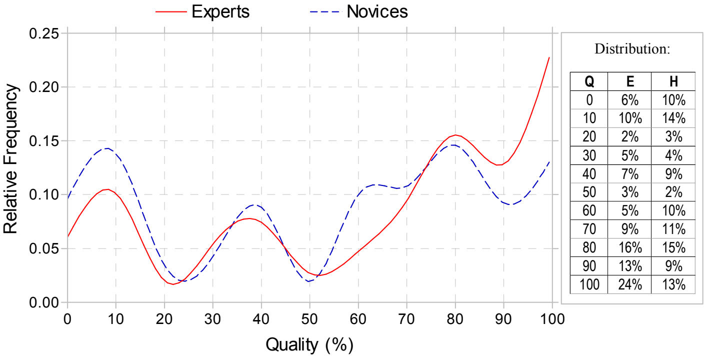
**Figure 3: Boxplot of the refinement quality by experts and novices**

*Significance.* From the distribution of expert and novice refinement data it is clear that the medians of the two samples are different, though it is not known whether the observed difference was statistically significant. As we were dealing with only two samples of unpaired data, of non-Gaussian though similar distribution, a one-tailed Mann-Whitney test was used to determine the significance of the difference between the medians. A one-tailed test was selected because from the outset we were testing the hypothesis that RARE IDIOM performance is closer to that of experts rather than novices, which can be translated into the hypothesis that on the RARE IDIOM quality scale, the median of expert refinement choices is significantly higher than the median of the novice choices. The test (see Table 5) clearly showed the difference of two medians to be significant (P < 0.05).

*Exceptions.* The density curves in Figure 4 show that on occasions RARE IDIOM rejected expert recommendations in favour of its own preferred refinement artefacts, with 18% of expert refinements being judged by RARE IDIOM as in a 0-20% quality range. An in-depth study of such cases revealed that the RARE IDIOM search failed to rank its design artefacts in line with expert recommendation because of faults and limitations of its domain thesaurus. Should such failure be consistent, the domain model ought to be reviewed, in particular its thesaurus and, in the worst-case scenario, the entire classification scheme.

In conclusion...

*A Mann-Whitney test provides evidence that RARE IDIOM ranks expert refinement choices*



Distribution:

| Q | E | H |
|---|---|---|
| 0 | 6% | 10% |
| 10 | 10% | 14% |
| 20 | 2% | 3% |
| 30 | 5% | 4% |
| 40 | 7% | 9% |
| 50 | 3% | 2% |
| 60 | 5% | 10% |
| 70 | 9% | 11% |
| 80 | 16% | 15% |
| 90 | 13% | 9% |
| 100 | 24% | 13% |

**Figure 4: Density curves of expert/novice refinement quality (as ranked by RARE IDIOM)**

*significantly higher than those selected by novices. Since experts make higher quality analysis decisions than novices, it is possible to assert that through its refinement and ranking methods RARE IDIOM can assist analysts in finding high quality reusable design artefacts for the purpose of requirements refinement.*

## DISCUSSION OF RESULTS

Three empirical studies have been conducted and together they have claimed success in developing a demonstration case which provided a sample classification scheme and a domain thesaurus, then classifying designs and requirements, calibrating an artefact space into expert and novice refinement choices, and finally determining RARE IDIOM effectiveness as compared with that of experts. While the list may be impressive, we still need to recall that all these tasks have been conducted as part of a single demonstration case study, of which results are hard to generalise. Thus, before we lay any claims to RARE IDIOM as delivering possible solutions to requirements refinement with reuse, we first need to re-assess the three studies for their generality and scalability. This will be accomplished by identifying and evaluating the limiting factors that have been imposed on the demonstration case study with respect to the size and complexity of its domain model, reuse repository, requirements documents and the skills of the study participants.

*The size and the complexity of the selected domain model is reflected in the size of a RARE IDIOM lexicon (1000 terms), the classification system (eights facets of 131 terms), and the domain thesaurus (of 490 terms). Should the domain size grow, would the selected domain analysis method be scalable?*

The most commonly used domain-engineering method, FODA (Kang, Cohen*, et al.* 1990), has been applied with great success to very large domains, such as Army Movement Control (Cohen, Stanley Jr.*, et al.* 1992). DARE (Frakes, Prieto-Diaz*, et al.* 1998) is quite different from FODA, but it shares many similarities with RARE IDIOM. As reported in the literature, DARE is still in its research phase and it is capable of capturing domains of the size similar to those studied in the RARE IDIOM project, i.e. where domain lexicons, facets and thesauri contain 100s of terms. So, when it comes to the research-quality systems, RARE IDIOM compares well with other systems in its class.

The greatest reassurance of the scalability of RARE IDIOM's methods, however, is the existence of some major undertakings in the construction of domain-specific faceted thesauri of 1000s of terms. Spiteri (1999) gives examples of several large projects, where the methodology taken, is very much alike that adopted in RARE, i.e. it is based on the effort of domain experts, working in teams, to sort domain terms into facets. It is worth noting that, similarly to RARE IDIOM, virtually none of such large-scale undertakings are based on automatic thesaurus construction techniques, which could quickly generate large, though inaccurate, thesauri (Srinivasan 1992).

*Our demonstration repository is quite small as it contains only 127 reusable artefacts. Is this repository size representative of reuse systems and can it be scaled up if so required?*

The best-known commercial reuse system, which is based on a faceted classification system, similar to that used in RARE IDIOM, has been deployed at GTE (Prieto-Diaz 1991). In two years of its commercial exploitation, the system led to the savings of US$1.5 million. Yet, the initial number of reuse artefacts was only 190 and over the years it shrunk to 128 - the size comparable to that used in RARE IDIOM evaluation. Furthermore, Prieto-Diaz states that building very large reuse libraries is ineffective, and instead several smaller domain-specific repositories should be built, each with its own specific vocabulary and a faceted scheme.

*Only a small and unstructured requirements document has been processed by RARE IDIOM in its evaluation studies. Is the selected document representative of the commonly processed user requirements? Can the size of requirements documents increase without affecting the system performance significantly?*

In refinement tasks, RARE IDIOM retrieval method is totally insensitive to the size of the processed requirement document. It should be noted, however, that at this point of time, RARE IDIOM assumes the flat, i.e. unstructured, requirements documents. Such structure is consistent with that of informal user requirements documents, as exemplified by some large-scale requirements documents for real-life systems, e.g. "Remote Sensing Based Spatial Information for the Sustainable Management of Forests" (Ministry of Economic Affairs, Ministry of Foreign

Affairs DGIS*, et al.* 1999). Should the need arise to handle other types of documents, the development effort to provide such facility would be minimal.

*The system evaluation relied on the performance of novice and expert analysts who conducted small requirements refinement tasks, the results of which have been subsequently compared against those produced by RARE IDIOM. Should the demonstration case study be scaled up in the size of its domain model, reuse repository and requirements documents, would we observe RARE IDIOM to be equally effective?*

It is anticipated that with the size and complexity of the tasks, the expert performance may slowly degrade, while the performance of novices is likely to rapidly deteriorate. Thus the effectiveness gap between the two groups of participants can be expected to be much more pronounced than that observed in the conducted studies. At the same time, as pointed out above, RARE IDIOM can be adapted to the complexities of its working environment without a major loss of its effectiveness. Although large-scale studies of RARE IDIOM performance are yet to be undertaken to test this assertion, it can be hypothesised that compared with experts and novices, the RARE IDIOM effectiveness will remain stable or will slightly improve.

*As can be seen from the preceding discussion of the limiting factors imposed on RARE IDIOM for the purpose of its evaluation, the results obtained from the empirical studies can be generalised to larger-scale test environments.*

## SUMMARY AND CONCLUSIONS

This paper presented a systematic approach to evaluating outcomes of the SDRM research project (Software Development Research Method - Nunamaker, Chen*, et al.* 1991). The empirical method adopted in this evaluation is grounded in the belief that simple system testing is insufficient in a development-based research. Typically testing focuses on showing the method or a tool to be correct or efficient, however, the investigation of the proposed method or a tool impact on its target audience is frequently either missing or is conducted in an indirect way. As an alternative to the commonly used approaches, this paper explored and fused a number of different empirical methods to define and develop a testing environment for the proposed method/tool, to calibrate test data so that the test results could be assessed and compared, and finally to test and evaluate the method/tool in the well-designed and calibrated environment.

As the RARE IDIOM project focuses on using information retrieval techniques to support requirements refinement into reusable designs, the evaluation process involved definition of suitable problem (for requirements) and solution (for designs) domains, populating this domain with fully described and classified concepts and artefacts, calibrating the quality measures of requirements refinement with respect to the expert and novice design decisions, and finally evaluation of the method/tool by showing its performance to be approaching that of experts and significantly exceeding that of novices.

The paper thus demonstrated the effectiveness of the RARE IDIOM method/tool. Even more importantly, it also illustrated that development-based research, as exemplified by the SDRM method, can be grounded in the sound empirical process to show the ultimate worth of the proposed methods and systems to their users (Cooper 1973).

## REFERENCES

Andersen Consulting (1994): *FOUNDATION Methods Version 2.0*, Manual V 9.0, Andersen Consulting, Arthur Andersen & Co.

ANSI/NISO (1993): *Guidelines for the Construction, Format, and Management of Monolingual Thesauri*, Standard Z39.19-1993 (Revision of Z39.19-1980), American National Standard Developed by the National Information Standards Organization: Bethesda, Maryland.

Benbasat, I., D.K. Goldstein, and M. Mead (1987): "The case research strategy in studies of infomration systems". *MIS Quarterly*. **11**: p. 369-386.

Besterfield, D.H., C. Besterfield-Michna, G.H. Besterfield, and M. Besterfield-Sacre (1999): *Total Quality Management*. Second ed. Upper Saddle River, New Jersey: Prentice Hall.

Chambers, J.M., W.S. Cleveland, B. Kleiner, and P.A. Tukey (1983): *Graphical Methods for Data Analysis*. Pacific Grove, CA: Wadsworth and Brooks.

Cohen, S.G., J.L. Stanley Jr., A.S. Peterson, and R.W. Krut Jr. (1992): *Application of Feature-Oriented Domain Analysis to the Army Movement Control Domain*, Technical Report CMU/SEI-91-TR-28, ADA 256590, Software Engineering Institute, Carnegie Mellon University: Pittsburgh, Pa.

Cooper, W.S. (1973): "On selecting a measure of retrieval effectiveness: Part 1". *Journal of the American Society for Information Science*. **24**: p. 87-100.

Curtis, B. (1989): "Cognitive issues in reusing software artifacts", in *Software Reusability: Concepts and Models*, T.J. Biggerstaff and A.J. Perlis (Editors). ACM Addison Wesley Publishing Company: New York, New York. p. 269-287.

Cybulski, J.L. and K. Reed (1999): "Automating Requirements Refinement with Cross-Domain Requirements Classification". *Australian Journal of Information Systems*(Special Issue on Requirements Engineering): p. 131-145.

Darke, P., G. Shanks, and M. Broadbent (1998): "Successfully completing case study research: combining rigour, relevance and pragmatism". *Information Systems Journal*. **8**: p. 273-289.

DoD (1995): *Software Reuse Initiative: Technology Roadmap, V2.2*, Report http://sw-eng.falls-church.va.us/reuseic/policy/Roadmap/Cover.html, Department of Defense.

Frakes, W. and S. Isoda (1994): "Success factors of systematic reuse". *IEEE Software*. **11**(5): p. 15-19.

Frakes, W., R. Prieto-Diaz, and C. Fox (1998): "DARE: domain analysis and reuse environment". *Annals of Software Engineering*. **5**: p. 125-141.

Frakes, W.B. (2000): "A case study of a reusable component collection". in *3rd IEEE Symposium on Application-Specific Systems and Software Engineering Technology (ASSET'00)*. Richardson Texas: IEEE Computer Society, p. http://dlib.computer.org/conferen/asset/0559/pdf/05590079.pdf.

Galliers, R.D. and F.F. Land (1987): "Choosing appropriate information systems research approaches". *Communications of the ACM*. **30**(11): p. 900-902.

Kang, K., S. Cohen, J. Hess, W. Novak, and S. Peterson (1990): *Feature-Oriented Domain Analysis (FODA) Feasibility Study*, Technical Report CMU/SEI-90-TR-21, Software Engineering Institute, Carnegie-Mello University.

Keen, E.M. (1992): "Presenting results of experimental retrieval comparisons". *Information Processing and management*. **28**: p. 491-502.

Luhn, H.P. (1957): "A statistical approach to the mechanised encoding and searching of literary information". *IBM Journal of Research and Development*. **1**(4): p. 309-317.

Miles, M.B. and A.M. Huberman (1994): *Qualitative Data Analysis: An Expanded Sourcebook*. Second ed. Thousands Oaks, California: Sage Publications.

Ministry of Economic Affairs, Ministry of Foreign Affairs DGIS, and Ministry of Agriculture Nature Management and Fisheries (1999): *User Requirements Study for Remote Sensing Based Spatial Information for the Sustainable Management of Forests: User Requirements Versus Existing Capabilities*, Technical Document 7, http://www.itc.nl/forestry/URS/td7.pdf, Kingdom of Netherlands.

Nunamaker, J.F., Jr, M. Chen, and T.D.M. Purdin (1991): "Systems development in information systems research". *Journal of Management Information Systems*. **7**(3): p. 89-106.

Prieto-Diaz, R. (1991): "Implementing faceted classification for software reuse". *Communications of ACM*. **34**(5): p. 88-97.

Rasmussen, E. (1992): "Clustering algorithms", in *Information Retrieval: Data Structures and Algorithms*, W.B. Frakes and R. Baeza-Yates (Editors). Prentice Hall: Englewood Cliffs, New Jersey. p. 419-442.

Salton, G. and M.J. McGill (1983): *The SMART and SIRE Experimental Retrieval Systems*. New York: McGraw-Hill.

Schenk, K.D., N.P. Vitalari, and K. Shannon Davis (1998): "Differences between novice and expert systems analysts: what do we know and what do we do?" *Journal of Management Information Systems*. **15**(1): p. 9-50.

Spiteri, L.F. (1999): "The essential elements of faceted thesauri". *Cataloging & Classification Quarterly*. **28**(4): p. 31-52.

Srinivasan, P. (1992): "Thesaurus Construction", in *Information Retrieval: Data Structures and Algorithms*, W.B. Frakes and R. Baeza-Yates (Editors). Prentice Hall: Englewood Cliffs, New Jersey. p. 161-176.

Tague-Sutcliffe, J. (1992): "The pragmatics of information retrieval experimentation, revisited". *Information Processing and Management*. **28**: p. 467-490.

## ACKNOWLEDGEMENTS

I would like to thank Assoc. Prof. Karl Reed and Assoc. Prof. Graeme Schanks for their critical comments useful in the preparation and conducting of the study reported in this paper.