AFRL-IF-RS-TR-2004-339 Vol 1 (of 2)
Final Technical Report
December 2004

FACILITATING SUBJECT MATTER EXPERT (SME)-BUILT KNOWLEDGE BASES (KBS)

Information Extraction & Transport, Incorporated

Sponsored by
Defense Advanced Research Projects Agency
DARPA Order No. K180

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE
ROME RESEARCH SITE
ROME, NEW YORK

STINFO FINAL REPORT

This report has been reviewed by the Air Force Research Laboratory, Information Directorate, Public Affairs Office (IFOIPA) and is releasable to the National Technical Information Service (NTIS). At NTIS it will be releasable to the general public, including foreign nations.

AFRL-IF-RS-TR-2004-339 Vol 1 (of 2) has been reviewed and is approved for publication.

APPROVED: /s/

CRAIG S. ANKEN
Project Engineer

FOR THE DIRECTOR: /s/

JOSEPH CAMERA, Chief
Information & Intelligence Exploitation Division
Information Directorate

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	DECEMBER 2004	Final Jul 00– May 04

4. TITLE AND SUBTITLE
FACILITATING SUBJECT MATTER EXPERT (SME)-BUILT KNOWLEDGE BASES (KBS)

5. FUNDING NUMBERS
C - F30602-00-C-0173
PE - 62301E
PR - RKFM
TA - 00
WU - 02

6. AUTHOR(S)
Edward J. Wright

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Information Extraction & Transport, Incorporated
1911 North Fort Myer Drive, Suite 600
Arlington Virginia 22209

8. PERFORMING ORGANIZATION REPORT NUMBER
N/A

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Defense Advanced Research Projects Agency
3701 North Fairfax Drive
Arlington Virginia 22203-1714

AFRL/IFED
525 Brooks Road
Rome New York 13441-4505

10. SPONSORING / MONITORING AGENCY REPORT NUMBER
AFRL-IF-RS-TR-2004-339 Vol 1 (of 2)

11. SUPPLEMENTARY NOTES
AFRL Project Engineer: Craig S. Anken/IFED/(315) 330-2074/ Craig.Anken@rl.af.mil

12a. DISTRIBUTION / AVAILABILITY STATEMENT
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

12b. DISTRIBUTION CODE

13. ABSTRACT ***(Maximum 200 Words)***

This report describes the efforts surrounding the Challenge Problem Development and Evaluation Management for DARPA's Rapid Knowledge Formation (RKF) program. The goal of RKF was the development of technology by which subject matter experts (SMEs) can develop knowledge bases (KBs) directly, without the intervention of professional knowledge engineers (KEs). IET's primary responsibility under RKF during the first two years was qualitative and quantitative evaluation of RKF-developed technology with respect to challenge problems (CPs) that they developed and administered. RKF's annual evaluation cycle began with new CP discussions in the Fall PI meeting, with subsequent CP specification, and a major evaluation during the summer. This document refers to RKF's first evaluation year (ending all 2001) as RKF Y1 and to its second evaluation year (ending Fall 2002) as RKF Y2.

14. SUBJECT TERMS
Knowledge Base, Knowledge Representation, Capture, Challenge Problem Development

15. NUMBER OF PAGES
43

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

Table of Contents

List of Figures

List of Tables

1. Introduction

Volume I of this report describes progress made by Information Extraction & Transport (IET[1]), Inc. on U.S. Government contract F30602-00-C-0173 during the period July 10, 2001 – July 9, 2002. IET's effort concerns Challenge Problem Development and Evaluation Management for DARPA's Rapid Knowledge Formation (RKF) program. IET's technology development efforts during year three of RKF are covered in Volume II of the RKF final report. RKF's task was the development of technology by which subject matter experts (SMEs) can develop knowledge bases (KBs) directly, without the intervention of professional knowledge engineers (KEs). IET's primary responsibility under RKF during the first two years was qualitative and quantitative evaluation of RKF-developed technology with respect to challenge problems (CPs) that we developed and administered.

RKF's annual evaluation cycle begins with new CP discussions at the Fall PI meeting, subsequent CP specification, and a major evaluation during the summer. We refer to RKF's first evaluation year (ending Fall 2001) as RKF Y1 and to its second evaluation year (ending Fall 2002) as RKF Y2. Note that RKF evaluation years were staggered by several months with respect to IET's contract years.

Note further that RKF Y2 began with a four-month continuation of the RKF Y1 evaluation (*i.e.,* another evaluation using an extension of the Textbook Knowledge CP specification IET developed under RKF Y1)—we refer to this as the RFK Y1.5 evaluation.

The principal investigator (PI) at IET for the evaluations was Dr. Robert Schrag. IET was supported on this contract by subcontractor Veridian Systems Division (VSD) and by consultant Professor Paul Cohen. During RKF Y1, IET also was supported by subcontractor George Mason University (GMU).

Core RKF associate contractors, besides IET, included the following.

- Integrator Cycorp; Dr. Doug Lenat, PI
- Integrator SRI International; Dr. Vinay Chaudhri, PI
- Component Developer Northwestern University (NWU) Institute for the Learning Sciences (ILS); Professor Ken Forbus, PI

Background material regarding IET's work on this effort may be found at http://www.iet.com/Projects/RKF/. Mailing lists are archived at http://www.iet.com/Projects/RKF/RKF-Only, protected by the user id "rkf," password "sme2logic."

1.1 Results Summary & Report-sequel Organization

During RKF's second contract year, IET accomplished the following major tasks, developing corresponding products, selected ones of which we elaborate on in the sequel.

[1] A table of acronyms and expansions appears in Section A.

1. We analyzed and presented results of the RKF Y1 Textbook and Expert [molecular biology] Knowledge CPs.[2]

These results, summarized in Section 2, exhibit our overall style of structured quantitative and qualitative evaluation for the KB authoring enterprise. We apply this evaluation style in the RKF Y1.5 evaluation and in Y2 CPs as well.

2. We specified the RKF Y1.5 CP as an extension of the RKF Y1 Textbook Knowledge CP. We administered the Y1.5 CP and analyzed and presented results.

We refer readers to the Web-posted presentation for details.

3. We developed and presented preliminary RKF Y2 CP concepts at the Fall 2001 PI meeting. At the Government's direction, these included knowledge development challenges supporting modeling of adversarial biological weapons (BW) development and detection of other asymmetric threat activity in intelligence data.

The BW CP also is documented in Volume 2 of IET's first annual summary report under this contract.

4. We developed a new RKF Y2 CP based on Intelligence Preparation of the Battlefield (IPB) with respect to enemy theater missile activity.[3]

The BS CP's rescinded specification is abstracted in Section 3. This description concentrates on the KB-authoring task specification. The detailed metrics and evaluation procedures are substantially similar to those described in Section 2.

5. We developed yet another RKF Y2 CP based on authoring and critiquing military courses of action (COAs).[4]

The COA CP is the basis of the ongoing RKF Y2 evaluation. Its specification is abstracted in Section 4. Again, elided metrics and evaluation procedures are substantially similar to those described in Section 2.

2. RKF Y1 Evaluation

In this Section, we describe the RKF Y1 evaluation—a large-scale experiment in which non-artificial intelligence SMEs—with neither artificial intelligence background nor extensive training in the task—author KBs following a challenge problem specification with a strong question-answering component. As a reference for comparison, professional KEs author KBs following the same specification. This section concentrates on the design of the experiment and its results—the evaluation of SME- and KE-authored KBs and SME-oriented authoring tools.

Evaluation is in terms of quantitative subjective (functional performance) metrics and objective (knowledge reuse) metrics that we define and apply, as well as in terms of subjective qualitative

---

[2] Section A lists uniform resource locators (URLs) for results summarized in Section 1.1. For hardcopy, contact an IET representative listed on the cover page.

[3] The change from the asymmetric threat CP topic noted in Item 3 was directed by the Government—in response to application realignment of the RKF program resulting from a DARPA office restructuring.

[4] The change from the IPB CP topic noted in Item 4 was initiated by the RKF integration teams—who found the IPB CP topic to be a poor match to their existing technology—and approved by the Government.

assessment using several sources. While all evaluation styles are useful individually and exhibit collective power, we find that subjective qualitative evaluation affords us insights of greatest leverage for future system/process design. One practical conclusion is that large-scale KB development may best be supported by "mixed-skills" teams of SMEs and KEs collaborating synergistically, rather than by SMEs forced to work alone.

In the remainder of this section, we first outline our approach to evaluating KBs. Then we describe the "textbook knowledge" challenge problem (TKCP) presented to SMEs and KEs for KB authoring, teams' tools, experimental procedures, and results from each style of evaluation. We close this section with discussion/conclusions.

2.1 KB Evaluation Approach

We consider KB evaluation along three dimensions: functional performance (with subjective metrics), economics (with objective metrics), and intrinsic quality (subjective and non-metric). Here we elaborate on these dimensions and our methodology. In a later section we describe results.

To evaluate functional performance, we follow Cohen et al. [3] in posing test questions (TQs) to authored KBs and scoring their answers against defined criteria. Our criteria fall into three major categories

- **Representation criteria:**
    - Query Formulation
    - Term Quality
    - Compositionality
- **Answer criterion:**
    - Correctness
- **Explanation criteria:**
    - Content Adequacy
    - Content Relevance
    - Intelligibility
    - Organization

While the Answer category obviously addresses a KB's functional performance, we argue that high-quality question representations and explanations also confer valuable (input- and output-oriented) functionality to KBs.

To evaluate economics, we follow Cohen et al. [2] in addressing reuse—the extent to which knowledge created earlier is exploited in the creation of subsequent knowledge. We require that authored knowledge (including constants and axioms) bear labels of authorship and creation time. Other things being equal, greater reuse is considered more economical.

Others—[5],[6]—have suggested (without employing, to our knowledge, in large-scale comparative evaluation) qualitative criteria for assessing intrinsic properties of KBs and

ontologies. Inspired by these, we formed a KB Quality Review Panel from among RKF technology providers and evaluators to assess the following properties.

- Clarity or style
- Maintainability or reusability
- Correctness or accuracy
- Appropriate generality
- Appropriate organization
- Logical propriety

While we discussed making this evaluation quantitative (by adapting our Functional Performance scoring methodology described below), the panel ultimately agreed that free-form commenting along these dimensions would be the most fruitful initial step.

We drew on two other sources, besides the panel, in our subjective qualitative evaluation: post-evaluation SME survey responses and evaluator observations. Findings regarding RKF tools' strengths and weaknesses were consistent across all three sources. RKF tool developers have taken these results seriously and have begun appropriate modifications to their tools.

2.1.1 **Additional Related Work**

The series of Knowledge Acquisition Workshops (KAWs)[5] has emphasized the evaluation of generic problem-solving methods (PSMs) and performance on the associated problem-solving tasks more than that of knowledge for its own sake. This appears to reflect a difference in emphasis or philosophy: whereas the KAW community has focused on the PSM as the primary reusable artifact, the RKF community has focused on KBs themselves as reusable artifacts that should, in principle, be applicable in any task-specific problem-solving setting.

2.2 Textbook Knowledge Challenge Problem

The TKCP's KB authoring task is to:

1. Capture knowledge about deoxyribonucleic acid (DNA) transcription and translation from about ten pages of an introductory undergraduate molecular biology textbook for non-majors [1];
2. Ensure that the authored KBs are capable of correctly answering test questions about the subject material, (extending or revising KBs as necessary).

We chose a textbook source because it serves as a circumscribed reference that offers an intuitively justified basis for required KB content scope. We chose molecular biology because it is a largely descriptive science and because it is of interest to the sponsor. We chose [1] because it largely eschews descriptions of laboratory procedures or scientific history in favor of material phenomena.

[5] See http://ksi.cpsc.ucalgary.ca/KAW/.

The TQs were consistent in difficulty with TQs typically found on Web-available quizzes on molecular biology. Questions appearing in the textbook itself typically required representation of (e.g., hypothetical/counter-factual) situations that were entirely novel compared to the basic material presented in the text. These were judged by the RKF community to be unsuitable (too difficult) for use in evaluation of current SME-oriented KB authoring technology. The TQs were similar in style and difficulty to IET-created sample questions (SQs) covering material in earlier chapters of the textbook. SQs were provided to teams before the evaluation. TQs were not so disclosed.

Besides the primary KB authoring tools described in the following section, RKF teams were required to include facilities for SMEs to pose TQs and to package their answers for evaluation. They also were required to prepare various instrumentation capabilities in support of metrics computations.

Teams' tools included substantial TKCP-relevant knowledge before they were handed off to SMEs. Given the premise that a large, general/reusable KB facilitates the construction of more specific KBs, teams were allowed to "prime the pump" of knowledge development by seeding KBs with prerequisite (e.g., pertaining to earlier—largely review—textbook chapters) and background (including high-level/abstract) knowledge or reasoning abilities deemed appropriate (according to defined ground rules) to support the authoring of the textbook's target knowledge.

2.3 Tools under Evaluation

Cycorp's "KRAKEN" tools are supported by a substantial KB based on a higher-order formal predicate logic. The key strategies of SME-oriented KB interaction are natural language (NL) presentation and a knowledge-driven acquisition dialog with limited NL understanding. The KB includes thousands of predicates and understands thousands of English verbs. Cycorp's approach might be described as maximalistic, domain-pluralistic, and conceptually precise. The KRAKEN tools aim to exploit (as leverage) a substantial KB to bring SMEs past an otherwise-steep learning curve by productive collaboration in this sophisticated knowledge representation milieu.

SRI's "SHAKEN" tools are supported by a relatively sparse KB based on the frame formalism. The key strategy of SME-oriented interaction is graphical assembly of components. The KB includes a few hundred predicates serving as conceptual primitives (the components). SRI's approach might be described as minimalistic, domain-universal, and conceptually coarse. The SHAKEN tools may be seen as skirting traditional knowledge representation complexity by presenting an entirely new metaphor with great intuitive appeal.

2.4 Experimental Procedures

IET collaborated with its subcontractor GMU to establish a SME KB authoring laboratory at GMU's Prince William County, Virginia campus. Eight (mostly graduate) biology students participated in the TKCP evaluation, four working with Cycorp's KRAKEN tools, four with SRI's SHAKEN tools. All worked full-time from mid-May until mid-July, 2001. The first week of this period was devoted to classroom-style training of SMEs by teams. The next two weeks were taken up with an evaluation dry run that included shake-down of tools in the installed context and limited additional, informal training. The evaluation-proper was held during the TKCP's final four weeks. It covered about seven pages of the textbook and included 70 TQs (about 3 pages and 10 TQs having been covered in the dry run). The actual test material covered

five subsections of the textbook's target material. SMEs were allowed to author this material in any order they liked, but IET would not release one subsection's TQs to a SME until s/he had completed work on TQs for earlier subsections.

Subsequent to training, SMEs had no direct contact with the teams' KEs. Instead, to deal with tool understanding issues that might arise, IET staffed the SME lab full-time with a "gatekeeper" KE (GKE) who mediated contacts with the teams (including bug reports and fixes). The gatekeeper KE also provided a subjective window on SME activity. Teams were allowed to augment KBs during the evaluation in accordance with the TKCP's pump priming ground rules.

Besides these SMEs, two KEs from each team also participated (off-site from the SME lab) by addressing the same KB authoring tasks using tools of their choice. SRI KEs used the same SHAKEN tools available to the SRI-assigned SMEs. Cycorp KEs usually authored knowledge in CycL (a KE-oriented knowledge representation language) using a text editor, rather than with the SME-oriented tools in KRAKEN. Cycorp KEs did not author all target textbook knowledge during the evaluation. Instead, they relied on a base of target knowledge that Cycorp had first developed in support of its internal pump priming requirements identification, then excised before tool delivery to SMEs. (This was due to unavoidable personnel overlap between Cycorp's pump-priming and TKCP-participating KEs.) SRI KEs were given the same option but elected to author the textbook knowledge during the evaluation. All KEs authored TQ representations and developed answers independently.

SMEs and KEs participants were required to answer at least 75% of the TQs presented for each subsection. In the results below, we include for each subsection the 75% of each participant's answered questions with the highest overall scores, padding with 0s as necessary. One of these subsections ("Signals in DNA Tell RNA Polymerase Where to Start and Finish") was particularly troublesome for the Cycorp SMEs. After they had spent well over a week working on it and were all less than halfway to reaching their answered-TQ quota, IET asked them to proceed to the next subsection to ensure that they had the chance to address most of the target material. (The SRI SMEs had completed their work and performed reasonably well on this 25-TQ subsection.) Because of this gatekeeper KE intervention, the authors have by consensus excluded this subsection from results analyses below.

Our functional performance scoring is both manual and subjective. We employ multiple scorers with expertise both in knowledge representation and in biology. We have historically achieved highly consistent results by articulating specific, value-by-value scoring guidelines for all criteria against the following, relatively coarse, generic framework:

- 0—no serious effort evident/completely off-base;
- 1—mostly unsatisfactory;
- 2—mostly satisfactory;
- 3—(for practical purposes) perfectly adequate.

To arrive at an overall score for functional performance on a given TQ, we set threshold scores for the last two ancillary criteria so that they do not exceed the highest score for (one of) the earlier, primary criteria; average scores for each criterion within a category; then average the category scores.

2.5 Experimental Results

2.5.1 Functional Performance Results

The major functional performance results are reflected in Table 1.

Team	*User type*	Representation	Answer	Explanation	Overall
Cycorp	*SME*	1.66	2.46	2.30	2.14
Cycorp	*KE*	2.54	2.58	2.56	2.56
SRI	*SME*	1.84	2.12	2.08	2.01
SRI	*KE*	2.09	2.48	2.40	2.32

Table 1: Means of teams' KEs'/SMEs' means of TQ scores

KEs' performance (the "gold standard" from RKF's perspective) was better than SMEs' with high statistical significance, but SMEs performed within 90% of the level achieved by their teams' KEs. We take the latter to reflect the relative effectiveness of teams' SME lab-fielded technology. There was no statistically significant difference across teams between the averaged scores of respective SMEs or KEs—either overall or at the criterion category level.

In a more detailed (unpublished/available upon request) treatment, we note statistically significant interactions among scores along the dimensions of individual SMEs, subsections, and question types in a categorization. All of these interactions washed out in the overall scores. We also note a "ceiling" effect, in that answer scoring with respect to several individual criteria exhibits large proportions of (highest-score) 3s. Elements likely contributing to this ceiling include our consistently accessible (i.e., low) quiz-level TQ difficulties and SMEs' consistent efforts to develop (supporting knowledge and) high-quality answers before moving on to additional TQs.

2.5.2 Economic / Reuse Results

Cohen et al. [2] profiled reuse in DARPA's High-performance Knowledge Bases (HPKB) program as the fraction of knowledge items previously existing in a given context. We again have two main reuse contexts to explore: that of constants in axioms and that of axioms in the explanations/proofs of answers to TQs. (To economize, we include only the latter analysis.)

Axiom reuse results appear in Table 2. Results are given for each participant (designated by monikers). Each participants' (KB's) mean overall score and mean number of axiom occurrences used to answer a TQ are included here for reference. "UA" stands for "User-authored Axioms" and "PDA" for "Pre-defined Axioms." "Used" indicates that the noted number of axioms actually appears in the explanation to one of the participant's answered TQs. "Unused" pertains to user-authored axioms that are not so used in a TQ (e.g., because the participant used them to author a subsection's target material before receiving its TQs). "Reused" pertains to user-authored axioms used to answer more than one TQ.

Team	Type	Moniker	Mean Overall Score	Functional Performance TQ count	Reuse TQ count	PDA used	Used: PDA / (PDA+UA)	UA used	UA unused	UA: used / (used+unused)	UA reused	UA: reused / used
Cycorp	SME	Tweety	2.17	39	31	103	49.26%	100	1881	5.05%	34	34.00%
Cycorp	KE	cycMW	2.85	43	35	150	42.53%	111	102	52.11%	13	11.71%
SRI	SME	Amoeba	2.29	30	30	364	62.28%	601	1292	31.75%	95	15.81%
SRI	SME	Celula	1.68	25	25	554	27.77%	213	420	33.65%	157	73.71%
SRI	SME	Iflu	2.26	34	34	473	35.91%	265	2046	11.47%	111	41.89%
SRI	KE	sriPN	2.58	35	35	341	50.29%	345	293	54.08%	313	90.72%
SRI	KE	sriAS	2.50	34	34	377	55.12%	463	919	33.50%	242	52.27%
SRI	SME	Vaccinia	2.59	35	35	402	44.01%	316	1464	17.75%	227	71.84%

Table 2: Reuse data by SME/KE

Table 2 includes only one each KE and SME entry for Cycorp because of difficulties at evaluation time with KB instrumentation and later with information extraction. These reuse results are still incomplete, as may be noted by comparing the numbers of TQs answered/scored for Functional Performance and numbers of TQs scored for reuse. A further issue of note is that the Cycorp KE, cycMW, (legitimately) authored much general knowledge directly into Cyc, as pump priming, where it is counted as pre-defined rather than user-authored.

We present (in Table 2's last columns) three varieties of reuse percentages: of user-authored axioms that appear in more than one TQ; of user-authored axioms that appear (at all) in TQs; and of appearing pre-defined out of all appearing axioms. From an economic standpoint, we comment merely that the latter reuse rate seems (uniformly) sufficiently high to justify the claim that relevant prior content has significant benefit for KB development.

We had an additional motivation (beyond economics) to examine reuse of user-authored axioms across TQs. RKF's functional performance evaluation criteria, being TQ-based, could not address the generality of knowledge across different TQs. Evaluators were interested in quantitative metrics of cross-TQ axiom reuse as a hedge against unprincipled, one-shot axiom "hacks" without lasting value.

Table 3 reports numbers of TQ occurrences for each reused user-authored axiom.

Team	Type	Moniker	TQs > 32	TQs in [17 32]	TQs in [9 16]	TQs in [5 8]	TQs in [3 4]	TQs = 2
Cycorp	SME	Tweety	0	0	0	0	2	20
Cycorp	KE	cycMW	0	0	0	0	5	8
SRI	SME	Amoeba	0	0	2	2	5	86
SRI	SME	Celula	0	0	1	69	28	59
SRI	SME	Iflu	0	0	0	7	53	51
SRI	KE	sriPN	0	0	3	186	67	57
SRI	KE	sriAS	0	1	6	97	57	81
SRI	SME	Vaccinia	0	1	1	20	99	106

Table 3: Incidences of axiom occurrence counts across TQs

Superficially, high axiom TQ-incidences occurred much more frequently for users of SRI's SHAKEN tools than for Cycorp's KRAKEN tools. (The axiom TQ-incidence patterns for pre-defined axioms are qualitatively similar.) However, these data do not appear to indicate cross-team differences in knowledge generality. Cycorp SME Tweety's axiom TQ-incidence profile is quite similar to that of Cycorp KE cycMW whose work—with highly respected representations—received the highest mean overall functional performance score. Axiom TQ-incidence profiles are also similar across SRI's KEs and SMEs. We tentatively attribute the cross-team profile differences to: compactness of (arbitrary-arity) CycL relations compared to binary relations resulting from translating SHAKEN's frames for axiom-counting purposes (suggesting a scaling factor for axioms counted in a given TQ); and conceptual coarseness,

compared to pre-defined predicates in Cyc, of SHAKEN's built-in relations (leading to greater applicability across TQs). Thus, we find no overall quantitative pattern indicating deficiency of appropriate knowledge generality for either team.

2.5.3 Subjective Qualitative Results

Deficiencies Identification: The KB Quality Review Panel concluded that SMEs, working alone, performed quite well at selected KB authoring tasks, but were less effective at others. SMEs with both teams were generally adept at placing and choosing concepts from the pre-existing ontology (i.e., they created and used knowledge at correct levels of specificity) and at general process description (i.e., they implemented Event-Actor vocabulary with accuracy and ease). The Panel highlighted as shortcomings in SME KBs the following major types: incompleteness, redundancy, and non-reusability. After describing these deficiency types below, we take up the question of their sources in the tools and in the KB authoring task.

Both teams' SMEs' KBs exhibit incompleteness, of three different kinds: content incompleteness (failure to describe a process fully, even where the textbook had); hierarchical incompleteness (failure to include natural siblings of a created concept); and interconnectedness incompleteness (failure to articulate obvious relationships between concepts).

SHAKEN SMEs' KBs exhibit significant redundancy attributable to limitations of the evaluated tools' inability to reason about authored concepts from the several distinct perspectives called for by different TQs. KRAKEN KBs also exhibit some redundancy. This usually is not of SME-authored knowledge, owing rather to re-creation by SMEs of pre-defined concepts.

Both teams' SMEs' KBs included concepts of suspect reusability. Mainly these were predicates, attributes, or concepts that combined concepts unnaturally—in a fashion that seemed difficult to reuse.

Suspected Deficiencies Sources: The panel's and SMEs' combined attributions of the above-noted deficiencies to major tool and TKCP task sources include the following (in order of increasing challenge such sources seem likely to pose RKF tool providers): TQ- and textbook-focused SME orientation, absence or inaccessibility of pre-defined knowledge, limited logical expressibility, and inherently difficult representation problems. We consider these in turn below.

Some KB incompleteness (especially of the content variety) are attributable to SME's attempts to tailor authoring in anticipation of unreleased TQs or in consideration of released TQs. (*I.e.,* sometimes authoring favored TQ effectiveness over general applicability or reuse.)

That SMEs were explicitly directed to focus on authoring textbook content may explain some hierarchical and interconnectedness incompleteness.

Some of the above-noted deficiencies resulted from incompleteness in the pump-primed KBs that they received. SHAKEN did not allow SMEs to facet general collections into collections of different kinds of collection subtypes. A SME noted that this would have facilitated clearer hierarchical placement.

While KRAKEN SMEs had access to a substantial background KB and sophisticated representation language, this potential came at a price: access tended to be at times insufficient, during other times overwhelming, thereby limiting and even hampering SME productivity and expressive possibilities. Gatekeeper KE reports and SME surveys mentioned the labor-intensiveness of what turned out to serve as Cycorp SMEs' major axiom entry mode—browsing

through existing axioms to discover one (with an appropriate predicate) to use as a template for editing and assertion.

Both teams' SMEs were—by design—somewhat limited in the logical forms they could use to express knowledge. SHAKEN SMEs were unable to make many assertions that deviated from the form ($\forall x$ (A x $\supset$ ($\exists y$) (B x y)). KRAKEN users had access to more logical forms via a richer vocabulary of rule macro predicates, though interface issues again caused more general rule construction to be prohibitively difficult here.

A major indication frequently occurring in both team's KBs of inherently difficult representation problems is predicates lacking specificity, argument types, or supporting axiomatization. Another indication is impoverished versions of assertions whose formal representation would require complex logical expressions.

Feasibility Assessment: While it is clear that plausible near-term improvements to these tools (and their captured background knowledge) could address some of the above-noted shortcomings, it also seems (to IET) that KB authoring generally does include inherently difficult representation problems whose solution demands well-developed logical skills and the balancing of different engineering principles. The ambition reflected in the present experiment to create tools that can empower a SME to full KB authoring independence—in arbitrary contexts—appears yet too grand.

While we have clear evidence that SMEs can author some high-quality knowledge in a sophisticated domain, we lack evidence that they can author high-quality predicates, analyze and refine background knowledge, develop rule paths to make sophisticated inferences work, or develop complex logical expressions required for some assertions. Also, it is not obvious how the existing tools could be refined to address such requirements.

Recommendation: We suggest that the KB development community's focus ought not be on tools that support KB authoring by "lone" SMEs (except where authoring tasks are relatively precisely defined and tools are fielded to support SMEs in a relatively mature authoring process). On the contrary, it should be on empowering SMEs to perform those KB authoring tasks they can be empowered to perform well. We believe the nascent RKF tools demonstrate a significant advance in such SME empowerment, and we recommend that in future experimental and developmental settings the relative strengths that SMEs and KEs bring to KB authoring should be exploited in a true "mixed-skills" team—a synergistic partnership.

We have some evidence that lightly trained SMEs are capable of significantly enhancing KE efforts to provide background knowledge that will be relevant to a KB authoring task. As a sequel to the TKCP evaluation, IET conducted a separate three-week evaluation intended to allow SMEs to explore teams' tools in a less structured setting. Eight (now tool-savvy) SMEs participated in an "expert knowledge" challenge problem (EKCP), pursuing KB authoring topics related to the life cycle of the Vaccinia virus—for which teams had authored no pump priming knowledge. An IET KE who had prepared some EKCP-supporting background knowledge development (in CycL) found that a Cycorp SME (who had not effectively authored Cyc predicates working alone) was readily able to contribute an informal specification that greatly facilitated the KE's work in extending the background knowledge to support the SME's needs.

We envision such interactions occurring throughout the KB authoring process, with SMEs and KEs contributing dynamically. The KE's role is always to perform sophisticated KB authoring

tasks currently beyond SMEs' reach. We believe that the SME-feasible task set should expand naturally (in a "bootstrapping" fashion) over time, as the talents of SMEs are mined and new tools are developed to meet opportunities presented by existing tools and authoring processes.

2.6 Discussion / Conclusion

All the styles of evaluation are useful in different contexts. Quantitative metrics are genuinely valuable for some purposes—e.g., inspiring a friendly competition among groups working in a common research initiative or demonstrating progress to an uninitiated, numbers-oriented supervisor. By far the long pole in the evaluation tent, however—from a system/process engineering, diagnostic point of view—remains subjective qualitative assessment. This is borne out by the comparative substance of our offered conclusions based on this activity and by the incorporation of insights and adoption of suggestions by technology providers working to develop the next generation of SME-empowering KB authoring tools.

We have seen that the three evaluation styles used here complement one another. The different quantitative metrics assist in each other's mutual interpretation (as, for example, when we appeal to Functional Performance in understanding Reuse), acting together as a synergistic set of reinforcements and consistency checks. We expect our effectiveness in the overall KB authoring enterprise to grow as the collective body of such techniques for understanding quality issues in KB artifacts, tools, and process continues to mature in a science of knowledge development.

3. The Rescinded RKF Y2 BS CP

This battlespace CP (BS CP) focuses[6] on IPB-supporting, knowledge-based models of force structure, behavior, and the battlespace environment and on situational reasoning using these models. The BS CP emphasizes tactical operations, not strategic planning. It emphasizes both fundamental/doctrinal principles (about unit compositions, activities, and generic constraints) and such principles' functional realizations (about real-world, physical constraints over equipment, terrain, environment/weather, existing force disposition, mission, or exceptions to doctrinal norms).

The initial domain focus is on time-critical targets, starting with mobile, theater-range surface-to-surface missiles (SSMs, or theater missiles—TMs).

3.1 IPB Overview

IPB is important to all branches of the military. It is now possible to collect much more threat and environmental information than can efficiently be processed in real-time, real-world battle situations. IPB methods and tools should assimilate and apply this information for their products to be useful and available for planning and execution of operations. The objective of IPB is to provide actionable information for battlefield commanders. The RKF BS CP exploits existing analytic IPB methods for creating IPB products and enhances them with the strengths of KB technology.

---

[6] This section employs both present and future tense reflective of the original specification. The BS CP is not current; the COA CP is (see Section 4).

3.1.1 Situational Reasoning

Military considerations related to IPB follow the "OCOKA" and "METT-T" principles.

OCOKA principles:

- **Observation & Fields of Fire**
- **Cover & Concealment**
- **Obstacles**
- **Key Terrain**
- **Avenues of Approach**

METT-T principles:

- **Mission** (the objective)
- **Enemy** (the threat)
- **Terrain** (land classes, attributes, elevation/slope)
- **Troops** (units and equipment available to the blue force commander)
- **Time Available** (any time constraint imposed)

Both red and blue commanders operate under these principles. Blue force intelligence (S2/G2) is tasked with determining the composition of the red force (Enemy) and projecting the red force Mission (*i.e.*, COAs), given the Terrain and Time considerations and the red forces' perception of the blue forces' capabilities and intentions (Troops). Blue force planning (S3/G3) exploits IPB products generated with attention to these considerations. Notice how all these factors indicate limitations on possibilities.

Situational reasoning goes beyond doctrinal guidelines and data-based threat characteristics/information to apply this knowledge to specific situations. The outcome of situational reasoning is possibilities or probabilities of military action. (The BS CP highlights the possibilistic viewpoint and downplays the probabilistic one.)

3.1.2 SSM / SAM Domain-specific Situational Reasoning

Based on examination of the (SME-produced) CY 2002 BS CP scenario posted at http://www.iet.com/Projects/RKF/BS-CP--TM-IPBScenario.doc, it appears that the OCOKA and METT-T principles are not uniformly applicable in the TM domain under discussion. We make the following notes.

- **OCOKA** notes:
    - We take **Observation (& Fields of Fire)** to apply to proximal terrain. As TM operations typically all occur deep within own-force territory—and generally distal with respect to any invading opposition force—**Observation** is most applicable in the context of penetration by opposition special forces. We downplay this element (at least in initial versions of the BS CP).
    - **Concealment** has obvious implications in TM "hide" activities. (To a lesser extent, this is true of **Cover**.)

- **Obstacles** (commended as barriers between you and your enemy) are of limited relevance based on the same proximity considerations used to rule out **Observation**.
- While **Key Terrain** might be construed to apply to TM-activity suitable terrain, we instead invoke the notion of **Suitability** explicitly (noting again the proximal, force-on-force connotation usual to **Key Terrain**).
- Again based on proximity considerations, we downplay **Avenues of Approach.** It turns out that road networks are the primary way TM vehicles get around. We talk about those, as well as about **Accessibility** of off-road buffer zones.

- **METT-T notes:**
  - Using similar proximity reasoning, within the **Mission** of TM—and given that the BS CP's area of TM operations will have one or more designated terrain bases, we take target locations to be largely irrelevant. (To the extent target types likewise lack impact on operations, we similarly downplay those.)
  - We take the **Enemy** (threat) to be characterized by his counter-measures assets (*e.g.,* combat air patrols, surveillance capability, TM precision counter-strike).
  - With most areas of **Terrain** already mentioned under the OCOKA items, we make note to mention **Weather** here.
  - Within **Troops**, TM operations are concerned mainly with **Equipment** (vehicles).
  - **Time Available** is a uniformly applicable consideration across TM operations.

3.1.3 Doctrinal vs. Functional Knowledge

OCOKA doctrine drives situation-specific reasoning for a given unit, mission, and terrain area. To be captured in part by the BS CP is how the analyst reasons about the application of OCOKA for a given threat unit in a given environment. To provide a foundation for discussing this, we first define/distinguish "doctrinal" knowledge—that which typically appears in field manuals—and "functional" knowledge—that which is required for expert application of the doctrine in particular situations—situational reasoning (as discussed in Section 3.1.1). For convenience, we refer to the combination of doctrinal and functional knowledge supporting reasoning about tactical operations as "tactical behavior knowledge."

3.1.3.1 Doctrinal Knowledge

Doctrinal knowledge captures theories of military actions in general terms—without reference to explicit situations except in idealized terms. It emphasizes general rules without dealing heavily with exceptions. In doctrinal publications, terrain, weather, troop fatigue, equipment loss, and similar realities of battlespace and engagement are treated conceptually, placing the burden of situational analysis on soldiers in the field. Some examples of doctrinal knowledge are listed below.

- Templates indicating typical unit compositions
- Tactics, techniques, and procedures (TTP) for units with given missions

- Generic constraints on TTP
    - Maximum speeds for vehicles of given types over given terrain types

Some of such information (though usually not all of what may be useful) may be available in standard equipment specs (*e.g.*, factory-specified performance for weapons, vehicles)

    - Temporal constraints among activities
    - Spatial constraints among activities or between activities and terrain
    - Combinations of the above

The above can include both hard constraints (those that must be satisfied—perhaps because of physical laws) and soft constraints (those whose satisfaction is generally desirable but maybe not absolutely necessary in every circumstance).

A major goal of the BS CP is to capture such knowledge about unit compositions, TTP, and associated constraints.

3.1.3.2 Functional Knowledge

Functional knowledge provides the practical "how" to doctrinal knowledge's theoretical "what." It supports the application of doctrinal knowledge in a given battlespace characterized including specific mission, terrain, weather, and own/opposition-force deployment. Functional knowledge draws on doctrinal knowledge as the theoretical underpinning from which to proceed. A relevant analogy exists in the field of economics. Most economics textbooks articulate the desirability of maintaining low inflation, *ceteris paribus*. However, policy makers must apply fairly extensive knowledge of the particulars of the economic situation in a region in order to instantiate the principle. We might say that functional knowledge is what the economic policymakers have and rely on to realize the principles agreed upon in economics textbooks.

Note that—while the above says some things functional knowledge is *good for*—it says little (directly) about what functional knowledge *is*. What is involved (in the way of knowledge representation and reasoning) in an expert's specialization of doctrinal knowledge to apply in a given situation?

Suppose doctrine tells us that a missile-launching unit must hide (to avoid likely air attack) after launching (and becoming conspicuous). Functional knowledge should support the enumeration and assessment of different candidate hiding places given a launch location and surrounding terrain description. As such, it needs to include the following.

- Elemental knowledge necessary to generate candidates

This could include characteristics of good hiding places—those that provide concealment for the missile launching unit's equipment (*e.g.,* a forest or a barn).

- Constraints necessary to prune generation of unsuitable candidates

Fundamentally, this includes constraints articulating the relationship between equipment and terrain: Where can missile-launching units vehicles go, based on their height, width, weight, center of gravity, buoyancy, *etc.*?

Many such constraints are easily inferred from the combination of equipment specifications (*e.g.*, vehicle width and turning radius) and terrain features (*e.g.*, road width, surface type, and degree of rain saturation). "Could the vehicle ford this stream at this location?"

Other constraints further require some geographical reasoning. If a potential hiding place is further away from the launch location than the vehicle(s) can reasonably travel in the available time, it can be pruned.

Other constraints additionally require consideration of typical unit behaviors to infer. SMEs might represent that a missile-launching unit will need a suitably firm piece of land—ideally hidden from easy observation but with a clear opening to the sky—from which the missile arsenal can be easily reached for reloading.

Other constraints can be inferred from the above by additionally considering common-sense knowledge. "A missile launcher must operate over solid ground, not water or marsh." "A deciduous forest will not provide canopy concealment in the middle of winter."

Other constraints require—beyond garden-variety common sense (as above)—a more specialized kind of common sense that we might imagine to be shared among IPB analysts. "The juxtaposition of trees and open area create the phenomenon of a *tree line*. Backing up—just into the tree line—provides concealment, on the one hand (trees as a background), and observation and fields of fire on the other (open field in front)."

- Assessment knowledge allowing the (absolute/ordinal) rating of nominally feasible candidates (those not failing constraints outright) with respect to general principles (*e.g.*, degrees of concealment or observation/fields of fire)

We may take the above knowledge to comprise a level of "common-sense" battlefield knowledge that applies and extends the knowledge contained in doctrinal field manuals and fact sheets. This application and extension could be characterized as taking place at the intersection of IPB doctrinal knowledge and general, IPB-relevant knowledge nominally touching on fields like geography, botany, and sociology. Our goal is to make such "IPB common sense" available for exploitation by enabling IPB analyst SMEs to capture it in RKF KBs.

Functional knowledge also allows one to address situation-specific exceptions to doctrinal rules, some of which we list below.

- Non-standard unit compositions
- Degraded unit conditions (*e.g.*, fatigue, demoralization, undersupply)
- Non-standard/non-working equipment
- Non-standard/novel applications of units or equipment.

3.2 Reasoning & Problem Solving Requirements

Taken together, authoring of the kinds of functional and doctrinal discussed in Section 3.1.3 should be "constrained activity networks"—where "activities" come from unit/subunit TTPs and "constraints" come from both doctrinal and functional knowledge. When authored, such tactical behavior knowledge can support a variety of complex problem solving tasks, including the following.

- Planning/COA generation
    - Information fusion/battlespace awareness
    - Intent recognition/future prediction
- Simulation
    - Wargaming

The quality of authored tactical behavior knowledge is arguably best evaluated with respect to the performance of (a using system for) such a task. (One vision for a follow-on program would select a few such tasks and develop knowledge/problem-solving application program interfaces—APIs—to prototypical performing systems aimed at specific technology transition opportunities.)

In the following, we express the RKF CY 2002 BS CP's reasoning requirements using parameterized questions (PQs), with the rationale that these can in principle drive either style of evaluation (*i.e.,* they could be construed as elements of a sought API). We also acknowledge that such PQs could be construed as effective elicitation guidelines. *I.e.,* get SMEs to author knowledge answering all possible meaningful PQs.

3.2.1 PQ Spec[7]

We posit a situation description (pseudoclass <SD>) including the following elements.[8]

- Mission elements:
    - {Current, next} transload location
    - Time constraint on any activity type
    - Equipment types in use
    - Mission activities
- Blue COMINT assets:
    - Platform type (combat air patrols, AWACS, satellite)
    - Sensor type (visual, infrared, MTI, foliage-penetrating)
    - Radio signal interception
- History:
    - Red {{pre-launch, post-launch} transload, launch, hide} location, time
    - Blue compromise (of Red) location, time, type (*e.g.,* fly-over, sighting, air-to-surface missile hit)
- Weather:
    - Date (infer seasonal effects)

[7] PQ notation is explained in Appendix B.

[8] To describe <SD>, we currently invoke some PQ notation informally.

    - Time (infer dark/light)
    - {Recent, current, forecast levels for} precipitation, temperature (infer road slickness, soil saturation)
    - Cloud ceiling, density (infer counter-measures susceptibility)
- Terrain:
    - Pixel-based, with attributes indicating elevation, slope, soil type, land class, vegetation type

Along each SD dimension, there are choices that make things easier or harder for Red. We envision successions of SQs/TQs that progressively tighten Red's constraints—thereby invoking more complex considerations regarding Suitability. An easy situation (*e.g.*) for Red would include relaxed time constraints, minimal/non-existent Blue COMINT assets, no history of Red compromise by Blue, summer (with heavy foliation), dark (low visibility), dry, temperate/hot (low relative launch IR signature), and dense clouds with low ceiling.

<StationaryActivity> = {launch, {pre-launch, post-launch} {hide, transload}}

<ActivitySpec> = {<StationaryActivity>, travel from <LocationSpec1> to <LocationSpec2> [within <TimeDurationSpec>]}

<ActivitySpec> = <ActivitySpec> at <DateTimeSpec>

<LocationSpec> = {<CoordinatePair>, location of <StationaryActivity>} %When the latter form is used in <Activity>, it will be constrained w.r.t. TM doctrine (i.e., the TEL activity cycle).

<Criterion> = {Suitability, Concealment, Cover, Observation} % Suitability is always used in the context of a specified activity.

Suitability is our main criterion of interest. Among OCOKA elements, we conceive of Concealment as of primary importance to TM IPB, followed by Cover, with other elements being of marginal utility/relevance. We currently address Accessibility and Trafficability in PQs (below) rather than among these criteria. (We often actually don't need to refer explicitly to any of these things. Consider, *e.g.,* that "hiding" as an activity requires concealment, "launch" requires some open space, and "traveling" requires trafficability.)

1. Given <SD>, rate <LocationSpec> w.r.t. <Criterion> for <ActivitySpec>. % Note: timing constraints initially can be very relaxed.

2. Given <SD>, identify (e.g., highlight on map) locations suitable for <ActivitySpec>.

3. Given <SD> and <ActivitySpec1>, identify (e.g., highlight on map) locations suitable for <ActivitySpec2> (where <ActivitySpec1> immediately precedes <ActivitySpec2> [within <TimeDurationSpec>] in TM doctrine's TEL cycle).

4. Given <SD> and launch at <DateTimeSpec>, <LocationSpec>, identify pairs of locations suitable for pre-launch hide [within <TimeDurationSpec1> of launch] and post-launch hide [within <TimeDurationSpec2> of launch].

5. Given <SD> identify triples of locations suitable for pre-launch hide [within <TimeDurationSpec1> of launch], launch, and post-launch hide [within <TimeDurationSpec2> of launch].

6. Given <SD>, identify (e.g., highlight on map) locations by their trafficability for <VehicleSpec>.

7. Identify maximum speed for <VehicleSpec>, given trafficability rating for <VehicleSpec>.

8. Given that <VehicleSpec> starts at <LocationSpec>, identify locations accessible by <VehicleSpec> within <TimeDurationSpec>.

9. Given SD, origin <LocationSpec1>, and destination <LocationSpec2>, identify paths <VehicleSpec> can traverse within <TimeDurationSpec>.

10. Given <SD>, identify the road buffer zone accessible within <TimeDurationSpec> of off-road travel.

4. The RKF Y2 COA CP

The RKF Y2 evaluation focused on the Course of Action (COA) Challenge Problem (CP), which addresses authoring/representation and critiquing of military COAs. RKF's technology integration teams (Cycorp and SRI International) created knowledge authoring tools (KRAKEN and SHAKEN, respectively) enabling military subject matter experts (SMEs) to author/represent COAs and COA critiques formally. In addition, teams created at least the functional core of a knowledge-based system (KBS) with problem-solving methods to critique formally represented COAs automatically. The team's tools enabled SMEs to extend formal knowledge representations supporting COA authoring/representation and automated COA critiquing. The COA CP spec is available on IET's RKF Web site. The evaluation took place during June – October, 2002.

COA CP evaluation is structured around the notion of a COA problem, whose statement consists of the following.

- Situation sketch (on a map), indicating:
    - Terrain features such as roadways, rivers, lakes, hills, forests
    - Current Blue and Red unit placement
- Scenario narrative, including any details not easily captured on the map (*e.g.,* relevant recent history, current dynamics, expected future evolution, unit status descriptions)

- Mission specification, indicating specific forces under command, required objectives, and constraints (*e.g.*, "Capture Objective JAYHAWK by 1400 hours tomorrow with the following restrictions in place…")
- Situation estimate, used to analyze the terrain.

Faced with such a problem statement, a commander is expected to produce an order to subordinate units that will put into effect the best possible course of action for accomplishing the mission. According to doctrine, several alternative ways of accomplishing the mission are captured in COAs that are sketched out in some detail and analyzed comparatively to facilitate a decision. For the COA CP, a COA solution statement is comprised of the following.

- COA sketch—an overlay on the problem statements situation sketch
- COA narrative—structured description stating the following elements
    - Commander's intent
    - Mission
    - Desired end state
    - Main attack
    - Supporting attack
    - Fire support
    - Reserve

  Each element above must address what units perform what actions for what purposes.

A commander's staff will compare the candidate COAs in a subjective critiquing process, usually resulting in a matrix comparing the viable COAs, and present the results to the commander for a decision. The manual process uses a small, arbitrary set of critiquing criteria, chosen because they seem to be important in a specific scenario. For the COA CP, we developed detailed/extensive criteria suitable for deliberate knowledge engineering. We suggested some of these as the nominal COA CP focus, but SMEs and integration teams were free to focus where they wished. See Figure 1.

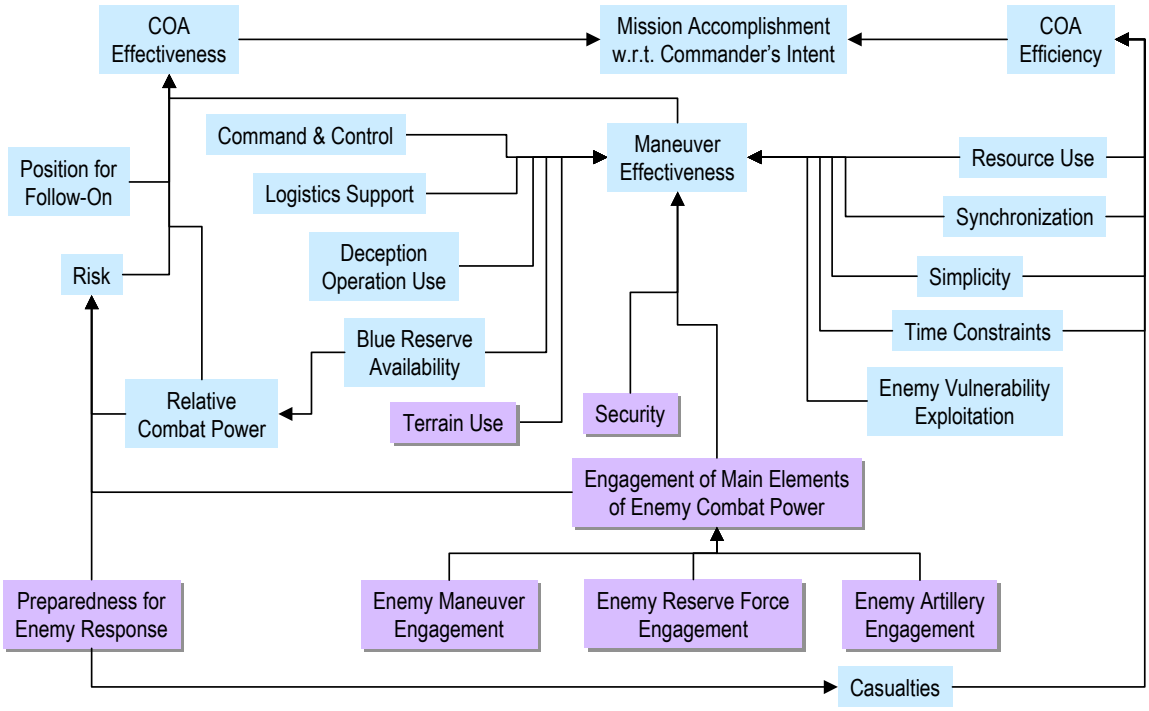

Figure 1: COA critiquing criteria & dependencies

In Figure 1, arrows point from dependees to dependents, so that (*e.g.*) Mission Accomplishment with respect to Commander's Intent (the top-level critiquing criterion) is dependent upon both COA Effectiveness and COA Efficiency. IET's nominal focus criteria are highlighted.

KB-authoring SMEs performed two basic tasks.

1. **Given text/graphical COA problem statements, formally represent selected elements of these in the KB.**
2. **Given a formally represented COA solution statement, author (conceive of and formally represent) knowledge supporting an effective COA critique by the evaluated team's KBS.**

The experimental procedures for these tasks are outlined below.

Before the evaluation, IET's evaluator-collaborating SMEs prepared text/graphical versions of COA problem and solution statements. Teams authored foundational (AKA "pump-priming") knowledge elements of the problem statements. IET provided the terrain portion of the situation sketch graphically, as output from Northwestern University's (NWU's) nuSketch Battlespace component.

During the evaluation, for each COA attempted by a KB-authoring SME, Task 1 included the following sequence of events.

1. The SME completes the situation sketch (if not already completed for a previous COA addressing the same situation), extends this to complete the COA sketch, and develops formally in the KB the COA narrative. See Figure 2.[11]

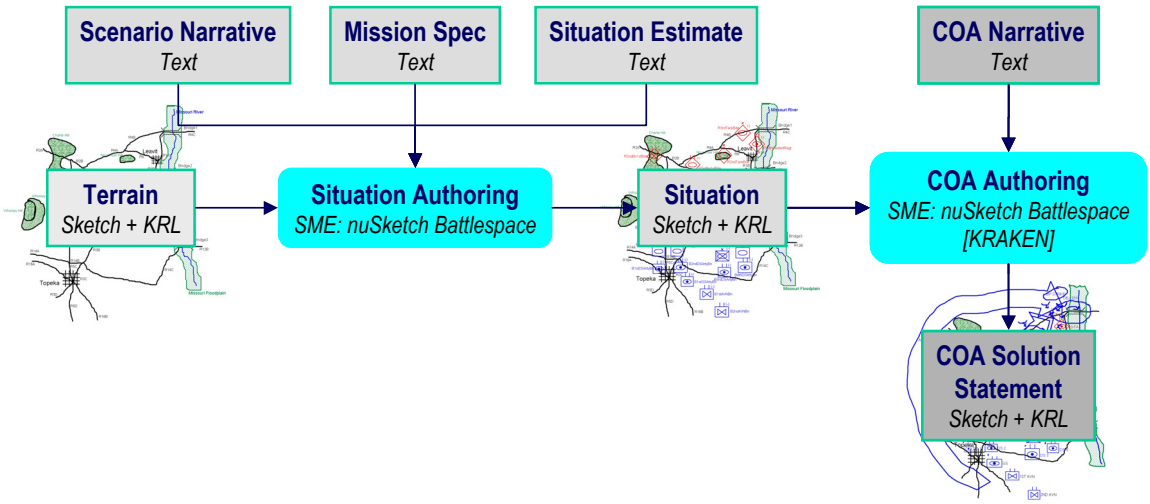

Figure 2: Task 1—COA representation

2. The SME administers COA diagnostic questions (DQs). The SME assesses the KB's performance on the DQs. See Figure 3.

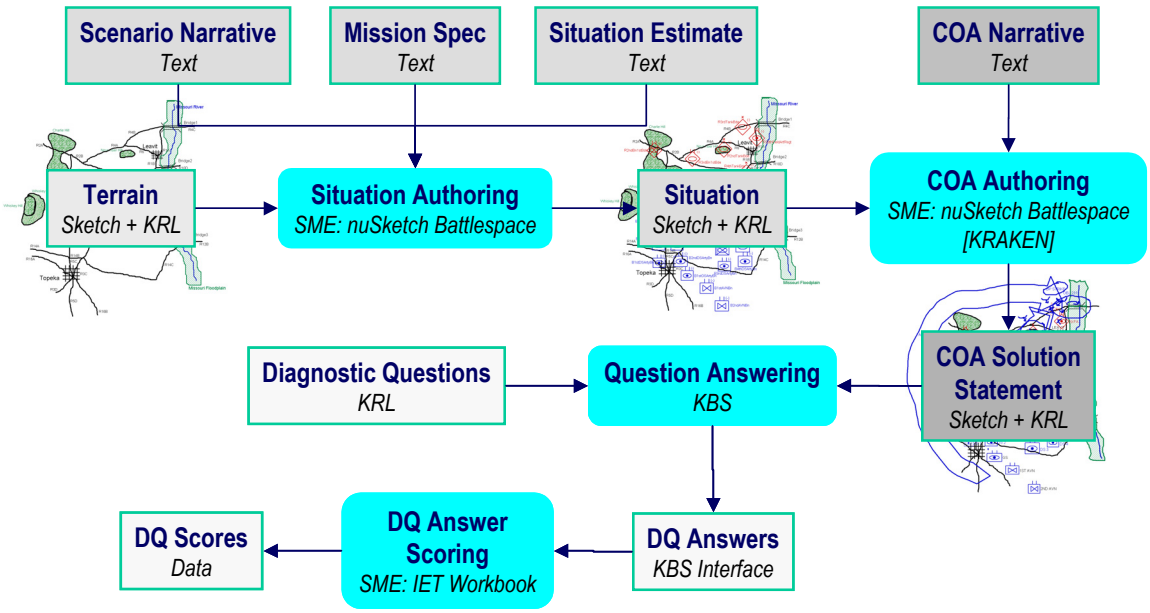

Figure 3: Task 1.2—COA representation evaluation

KB-authoring SMEs executed Task 2 for each of the four COAs whose representation they completed during the final evaluation week. For each COA attempted by a KB-authoring SME,

[11] Square-cornered boxes represent artifacts (inputs/outputs), with subtext indicating form. Round-cornered boxes represent processes, with subtext indicating performer(s) (tool/human).

this task included the following sequence of events. Steps 1–5 are depicted in Figure 4, Steps 6 and 7 in Figure 5.

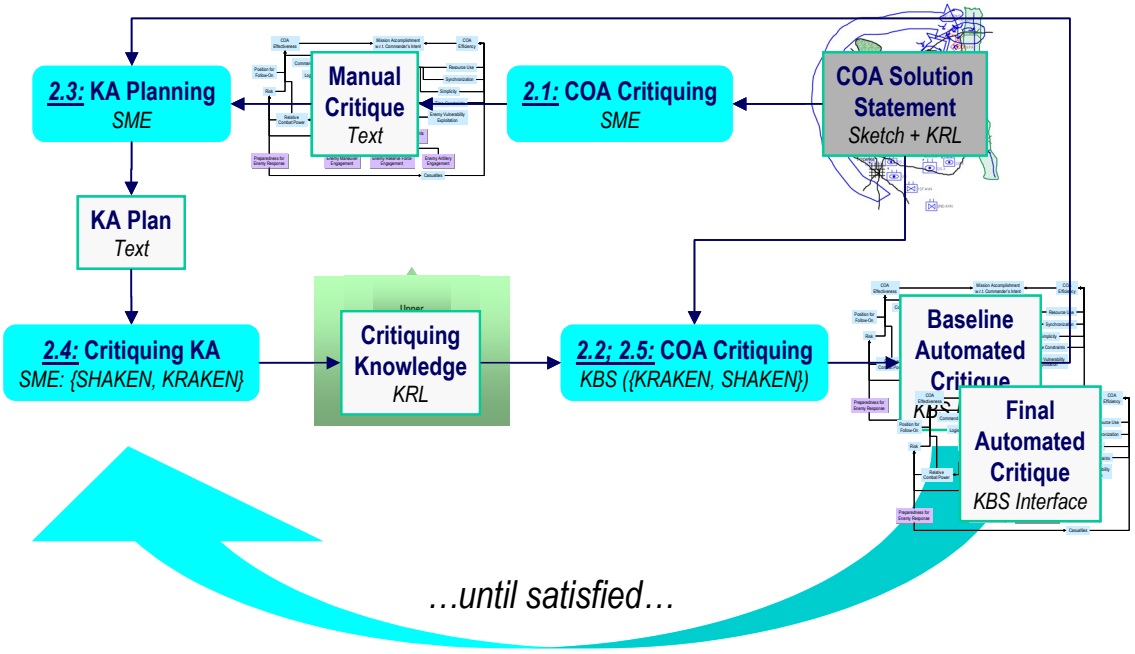

Figure 4: Task 2—COA critiquing

1. The SME creates a baseline manual critique of his represented COA. This critique addresses the same criteria as the corresponding team's KBS does, applying the same rating scale. The SME records minimalistic rationale supporting his individual ratings for each criterion.
2. The team's KBS produces a baseline automated critique of the SME's represented COA—before the SME has authored any knowledge to support its critiquing particularly.
3. The SME, reviewing the baseline automated critique (including both criteria ratings and supporting rationale), prepares a minimalistic knowledge development plan. This plan should cover discrepancies between his own baseline manual critique and the KBS's baseline automated critique that he believes (after reviewing the latter) are important for the KBS to address.
4. The SME authors knowledge using the team's knowledge authoring tools until he is satisfied that he has to the best of his ability improved the KBS's performance in critiquing the COA.
5. The team's KBS produces the ultimate automated critique.

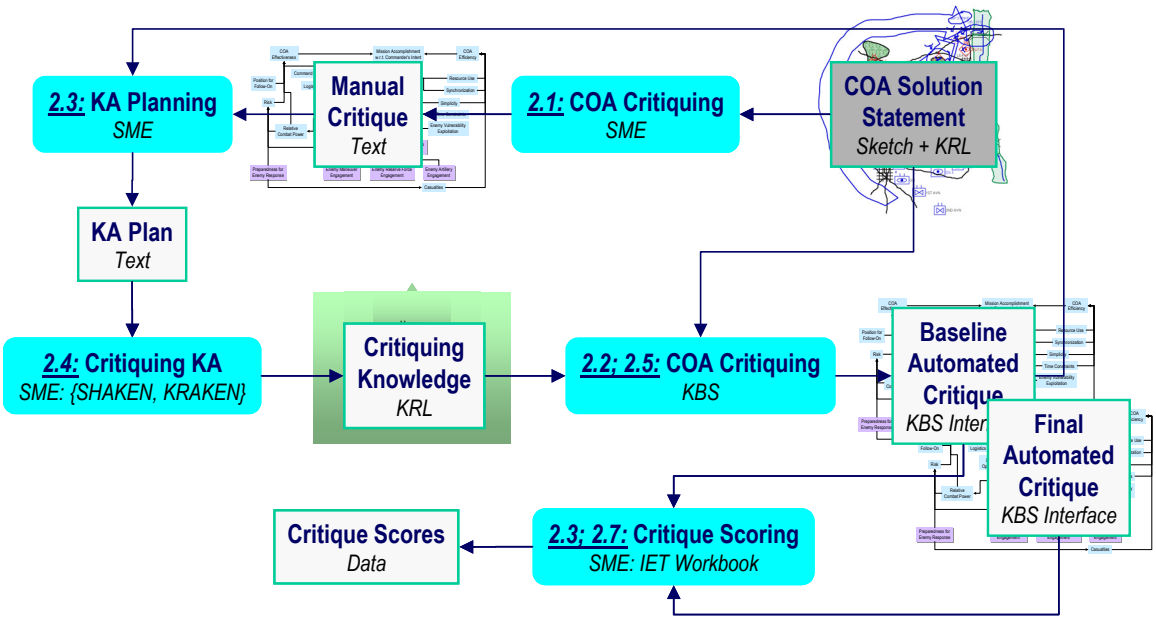

Figure 5: Task 2—COA critiquing evaluation

6. The SME assesses his impressions of the disposition of the original version of his knowledge development plan with respect to its ultimate version.
7. The SME, using a format provided by the evaluator, assesses the KBS's performance on the baseline and ultimate critiques.

The two different integration teams and the various SMEs associated with each team approached the dry runs and evaluations differently—within the guidelines specified by IET—as depicted in Figure 6.

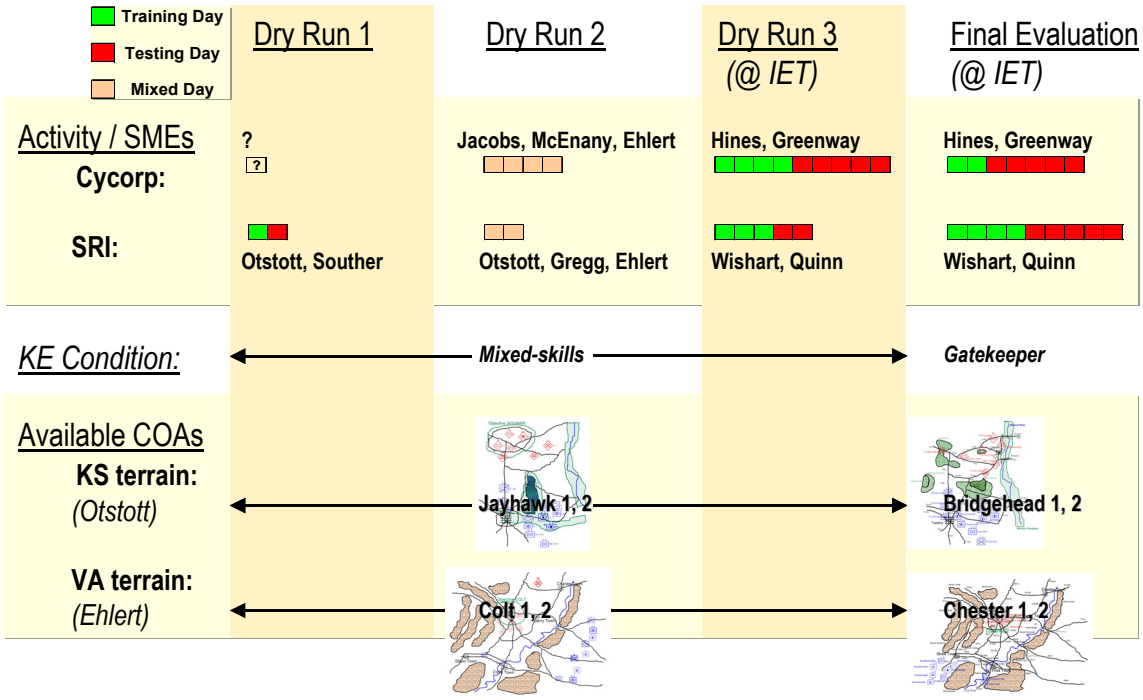

Figure 6: Evaluation structure

As shown in Figure 6, the final evaluation employed different COAs than did the dry runs—albeit over the same terrain. Additionally, IET monitored (with its "gatekeeper" KE) SME-KE interactions during the final evaluation than during the dry runs.

Figure 7 illustrates the structure of the final evaluation in more detail.

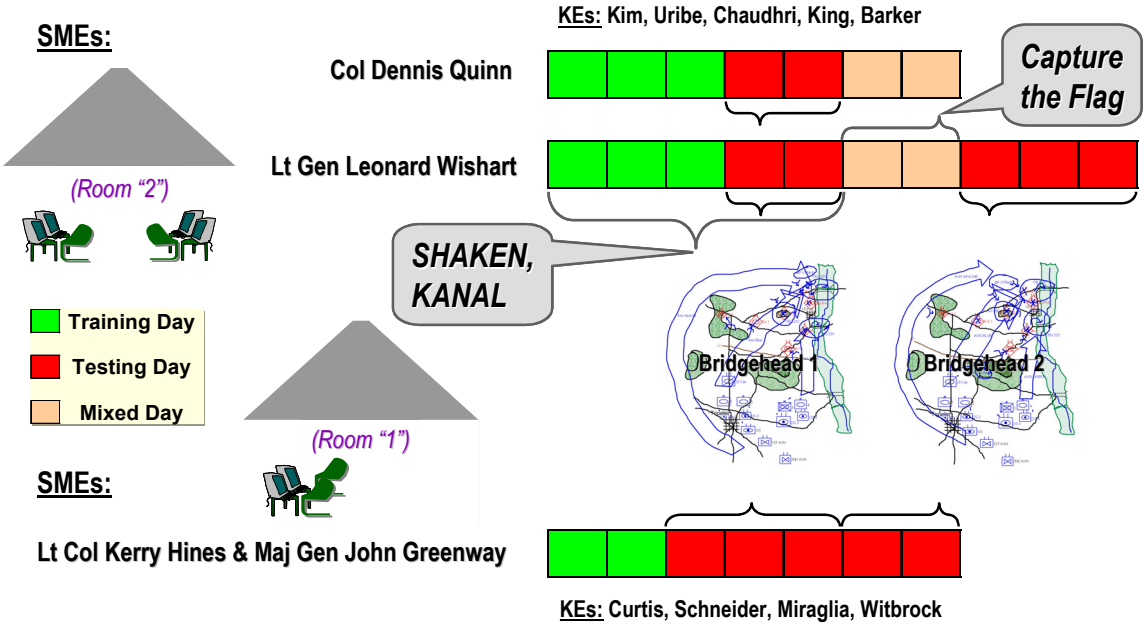

Figure 7: Final evaluation detailed structure

As Figure 7 shows, the SMEs associated with each team had separate working areas. The SME's associated with the Cycorp team collaborated, whereas those with the SRI team worked individually. Several KEs from each team also participated (on-site) with training and with evaluation facilitation. SMEs for both teams worked on COAs for one terrain background, without proceeding (lacking time) to address COAs for the other.

4.1 Evaluation Dimensions

RKF systems and resulting KBSs were evaluated on the following basis.

- Tools' effectiveness in supporting SME authoring/representing of COA solution statements
    - Emphasis was be on COA task specification and temporal dependencies rather than on the display of the COA sketch using NuSketch.
    - We tested the representation using fairly simple questions aimed at verifying the COA representation (rather than testing the systems' inferencing abilities).
- Tools' effectiveness in supporting authoring of COA critiquing knowledge
- Resulting KBSs' effectiveness in critiquing authored COAs

For each basis above, evaluation considered the following.

- Comparison of authored knowledge to intended knowledge
    - Accounting both for observed deficiencies and for observed enhancements beyond a SME's specification
- Comparison of systems' question answers to answer keys'

Question answers were scored as in the CY 2001 RKF evaluation.

4.2 Y2 Results

4.2.1 Functional Performance Results[12]

In this section we summarize the quantitative results from the Y2 challenge problem. As noted in above, SMEs were to formally represent COAs. This involved the representation of the COAs themselves as well as representing the critiquing principles. Below we summarize the results. IET presented a paper discussing some of these results at the 2003 K-CAP conference, see [8]. In considering the below it is important to note that the SMEs using the KRAKEN system worked together. The SMEs using the SHAKEN system worked separately.

4.2.1.1 COA Representation Functional Performance Results

The first evaluation task for the Y3 SMEs was the representation of COAs. SMEs used the elicitation tools and the NuSketch program to render a formal representation of these COAs. COAs were represented and 9 questions were posed, examples of the questions can be found in **Figure 8**. The system responses were then evaluated for answer correctness and answer evaluation. Results were scored by domain experts and by knowledge representation experts

[12] Further discussion of Y2 results can be found at http://www.iet.com/Projects/RKF/#Y2EvalationResults.

(IET). Results were graded on a range from -2 (lowest quality) to +2 (highest quality). A summary of these results is given in **Table 4**. This table represents the average for each COA representation for the 9 diagnostic questions posed.

Table 4: COA Diagnostic Results

COA	Team	IET-Correctness	SME-Correctness	IET-Explanation	SME-Explanation
1	Cycorp	1.86	1.43	1.57	1.57
2	Cycorp	1.57	1.43	1.43	1.71
1	SRI	1.21	1.5	.57	1.21
2	SRI	1.14	N/A	.43	N/A

The correctness of those COAs were evaluated in comparison to an answer key. COA quality in the Cycorp/Kraken representation system was quite high. This may have reflected the fact that in the Kraken system SMEs spent a larger proportion of their evaluation time representing the COA. SMEs represented a COA via NuSketch and then supplemented the representation with further interaction with the tool. In the SRI/Shaken system, SMEs utilized only the NuSketch sketching tool which may have resulted in less complete answers.

	Diagnostic Question
DQ1	*Which avenue of approach is exploited as the main attack axis of advance in the COA?*
DQ2	*What is the objective end state of this COA?*
DQ3	*Does the supporting attack occur before or after the main attack?*
DQ4	*Which enemy units are being attacked in the supporting attack?*
DQ5	*Which friendly units are being deployed in the main attack?*
DQ8	*Which unit(s) provides fire support during the main attack?*
DQ9	*Where will units engaged in the supporting attack be located during the main attack?*

Figure 8: Examples of COA diagnostic questions

4.2.1.2 COA Analysis Results

The second phase of functional evaluation considered the system-generated critiques by comparing them to (textual) manually authored critiques for the same COAs (prepared by the evaluation-participating SMEs in advance of critiquing rules capture). In economic evaluation (following [3]), we measured in-situ knowledge reuse as an indication of knowledge generality.

It is important to note that this evaluation was not designed as a "bakeoff" between the two systems, i.e., program participants and evaluators were not interested, or at least not primarily interested, in the correctness of a hypothesis of the following form, "System A better facilitates the elicitation of SME COA critiquing knowledge than does System B". The amount of data that can be collected in the kind of evaluation described, i.e., one involving complex knowledge representation for a large comprehensive reasoning task, makes it very difficult to generate enough evidence to be able to confidently reject the related null hypothesis. In addition, the tools

were bringing potentially complementary KA methods to the evaluation. For example, the KRAKEN participants devoted far more effort to detailed COA description than did the SHAKEN participants. The evaluation served to subject the systems to a comprehensive KA test to show the system designers how effective their tools were for a nontrivial knowledge representation and reasoning (KR&R) task of interest to tool-using SMEs.

The Evaluation-participating SME themselves evaluated system critiques for correctness and quality, as determined by a manual comparison to a critique generated by the participating SME. Beyond just considering the raw judgments, evaluators also considered explanations—whether each system gave appropriate rationale and considered relevant circumstances. Quality considerations involved correctness and completeness of explanation. Table 5 presents system success with respect to criteria that the pair of SMEs associated with each team's system was able to address. The first three columns indicate the system, SME and COA being critiqued. The third and fourth columns indicate the average score for correctness and explanation quality over all criteria considered for a given COA (on a -2 to +2 scale). The last column indicates the number of criteria considered for the specified COA and for the specified critique.

Table 5. COA Critique Results

System	SME	COA	Correctness	Quality	# of Criteria
SHAKEN	1	1	0.2	0.42	12
SHAKEN	2	1	0.17	0.17	6
KRAKEN	1&2	1	0.8	0	6
KRAKEN	1&2	2	0.5	-0.2	5

The elicited KBs contained sufficient knowledge, in terms of volume and complexity, for each system's reasoning tools to offer high-level critiques (that were plausible, in the judgment of the military SMEs)—from a number of diverse perspectives—of knowledge-rich COAs.

As noted, the scores reported for COA critiquing performance do not lend themselves well to cross-system comparison because the averages are over the number of criteria that the SME addressed in creating critiquing rules. No attempt was made to reward or penalize with respect to the number of criteria addressed. For example, SME 1 of the SHAKEN team attempted to represent critiquing principles for 12 different criteria, far more than any of the other SMEs. Attempting to address more critiquing criteria may have brought down this SME's average score. Similarly, SMEs were given complete latitude in choosing the criteria for which they represented rules.

The number of rules created by SHAKEN's SME users, an average of 30 per complete elicitation session, reflects well on its graphical elicitation process. The rule elicitation interface on the KRAKEN system is robust and functional and includes a number of useful tools that ensured effective quality control and system integration, but SMEs using it were not able to use it to create rules at the same pace, i.e., 8 per complete elicitation session. As the examples below indicate, the elicited rules were comparable in terms of complexity, as measured by the number of distinct terms appearing in the rules, and granularity in the level of representation. The rules

had similar critiquing foci but a noteworthy difference is the fact that KRAKEN SMEs did create different pairs of rules that produced a reasoning chain, i.e., a rule in which the consequence of one rule could be used to help determine whether the antecedent of another SME rule held. Also, the SMEs working on the KRAKEN system spent a larger proportion of their time on another component of the evaluation, i.e., formulating COA descriptions that went beyond the rudimentary descriptions elicited via NuSketch. Hence, while the quality of rules produced in KRAKEN was quite adequate and KRAKEN COAs were very well axiomatized, the number of critiquing rules created was low compared to the SME-created critiquing rules number that SMEs were able to create in SHAKEN.

4.2.2 Economics/Reuse Results

During the Y2 evaluation, IET also implemented metrics for assessing the amount of knowledge accumulated and the utility of that knowledge for purposes of the reasoning tasks designated as part of the challenge problem. IET tallied the amount of knowledge generated by SMEs and by KEs during the evaluation period. The overall production figures are given in **Table 6**. See 6 in the Web-accessible Materials for more discussion.

Table 6: Production Figures for SMEs and KEs

Team/Vendor	Subject	Type (SME or KE)	Constants Produced	Axioms Produced
Cycorp	SME1 and 2	subject matter expert	81	926
Cycorp	KE	knowledge engineer	250	3346
SRI	SME1	subject matter expert	38	421
SRI	SME2	subject matter expert	22	1580
SRI	KE	knowledge engineer	2433	141

These production figures, when compared with the production figures for Y1, (see 2.5.2), demonstrate significant further progress towards the goal of rapid knowledge elicitation. An important contributor to this increase in production was hypothesized to be the addition of the NuSketch tool that allowed users to represent COAs by using a graphical interface. See [4] for a fuller description.

Of course, axiom and constants are of limited value in measuring the success of knowledge elicitation tools. The value of constants and axioms lies in their utility in performing reasoning tasks. We present data on reuse in some of the figures below. **Figure 9** presents data concerning the extent to which SME-produced constants were reused in the axioms created. We see that a large number of constants were used more than 64 times in axioms and very few were used infrequently. **Figure 10** indicates how SME-created axioms were reused in test questions, i.e.,

the number of times that axioms were used to reason about the answer to a TQ. Only a small proportion of axioms were used in TQs but this is consistent with attempts to build a general KB in a limited testing domain as **Figure 11** shows a comparable fate for the axioms produced by KEs. **Figure 12** shows the extent to which constants produced by SHAKEN SMEs were reused. We can see that one of the SMEs, SME2, had difficulty in creating usable constants and relied entirely on constants created by others in the axioms s/he created. **Figure 13** shows, by comparison, that KE produced constants tended to be used much more frequently in axioms. Data for the implementation of SHAKEN axioms and constants in TQs was unavailable.

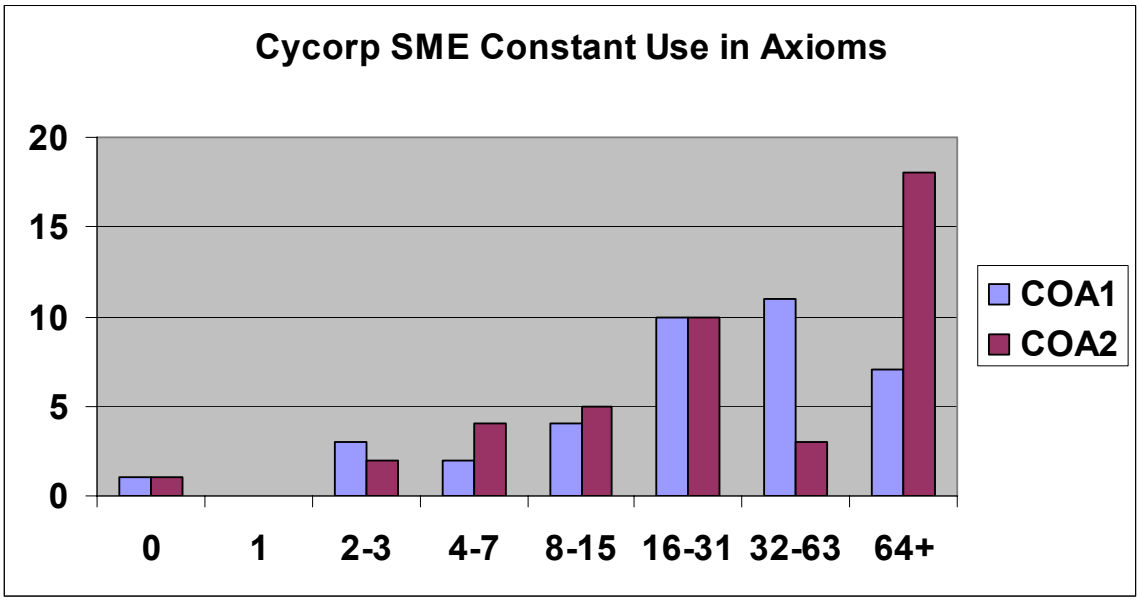

Figure 9: Number of appearances of KRAKEN SME constants in axioms

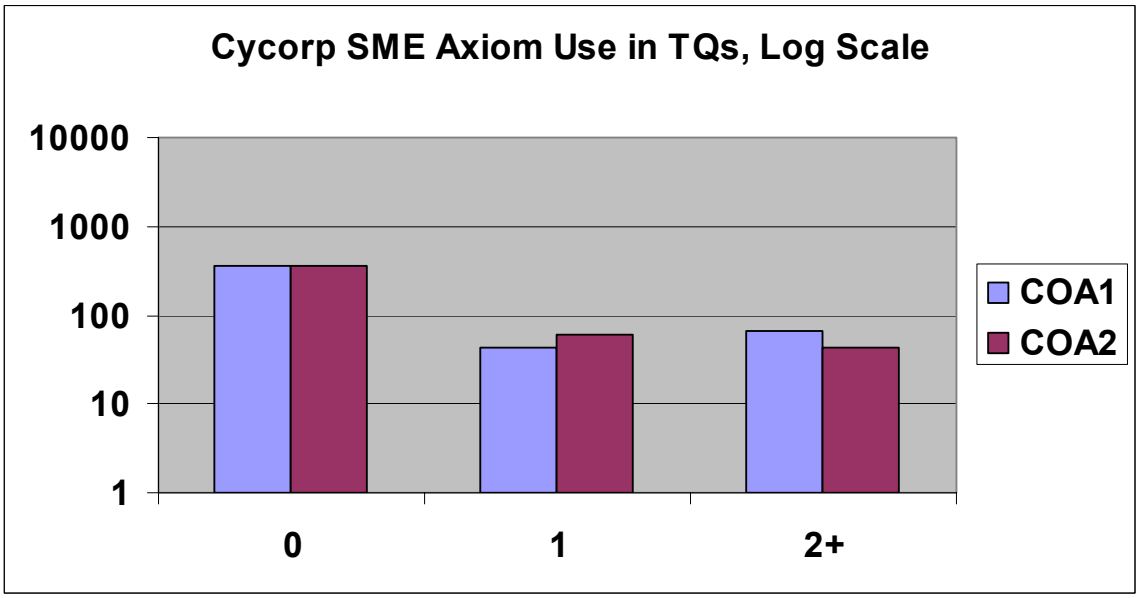

Figure 10: KRAKEN SME axiom usage figures

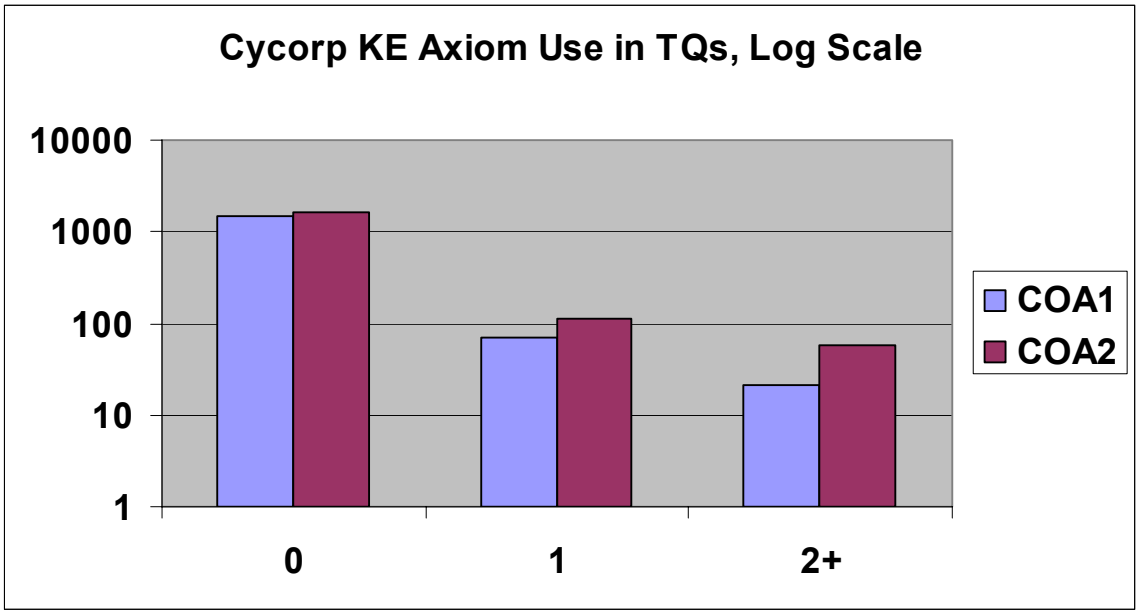

Figure 11: KRAKEN KE axiom usage figures

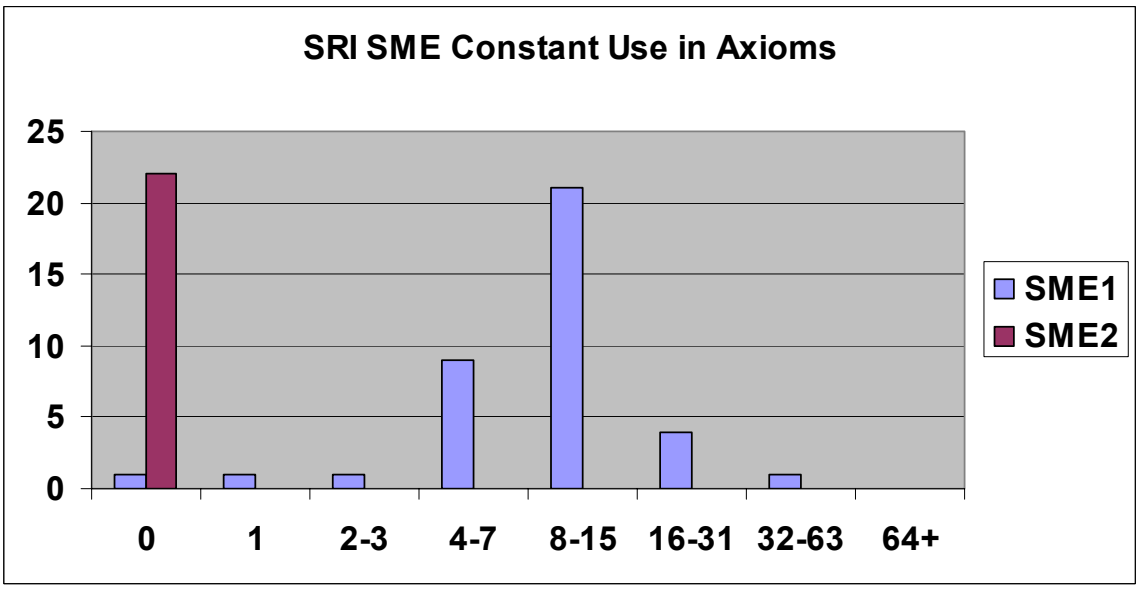

Figure 12: SHAKEN SME Constant Usage

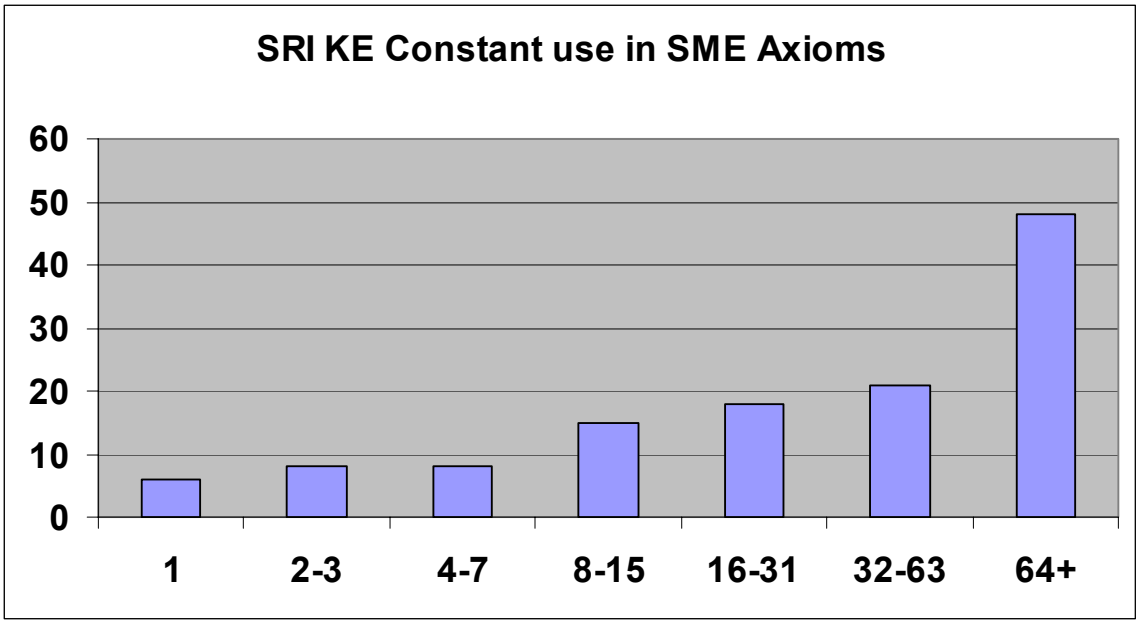

Figure 13: SHAKEN KE Constant Usage

In addition to the number of axioms and constants created it is useful to consider the number of rules that SMEs were able to produce explicitly for purposes of COA assessment. These are important because they are very complex and their elicitation required a significant amount of effort and consideration by the teams. These axioms tended to have complex antecedent and were characterized by a consequent with some conclusions about COA quality with respect to a given COA critiquing criteria. The number of rules created by SHAKEN's SME users, an average of 30 per complete elicitation session, or a total of 60, reflects well on its graphical elicitation process. The rule elicitation interface on the KRAKEN system is robust and functional and includes a number of useful tools that ensured effective quality control and system integration, but SMEs using it were not able to use it to create rules at the same pace, i.e., 8 per complete elicitation session. As the examples below indicate, the elicited rules were comparable in terms of complexity, as measured by the number of distinct terms appearing in the rules, and granularity in the level of representation. The rules had similar critiquing foci but a noteworthy difference is the fact that KRAKEN SMEs did create different pairs of rules that produced a reasoning chain, i.e., a rule in which the consequence of one rule could be used to help determine whether the antecedent of another SME rule held. Also, the SMEs working on the KRAKEN system spent a larger proportion of their time on another component of the evaluation, i.e., formulating COA descriptions that went beyond the rudimentary descriptions elicited via NuSketch. Hence, while the quality of rules produced in KRAKEN was quite adequate and KRAKEN COAs were very well axiomatized, the number of critiquing rules created was low compared to the SME-created critiquing rules number that SMEs were able to create in SHAKEN.

4.2.3 Subjective Qualitative Results

Another aspect of the evaluation involved the implementation of subject matter experts as a Quality Review Panel (QRP), to assess the rules that SMEs had created. The full presentation is available at http://www.iet.com/Projects/RKF/QRP02/#PI_Meeting. This analysis was intended

to provide further evaluation feedback to DARPA and developers than that which is available strictly from a performance based critique. In particular, this analysis allowed for a more comprehensive evaluation including consideration of redundancies, scalability of elicited knowledge (e.g., reusable for scenarios similar to the challenge problem, i.e., other COAs), general characterization of KR approaches and deficiencies, etc.

A detailed discussion of the QRP's observations can be found in [8]. By way of summary, the QRP observed that SMEs occassionally had difficulties distinguishing necessary from sufficient conditions in terms of implementing these for purposes of rule representation. At times SMEs confused necessary conditions with sufficient conditions. In addition, they sometimes wrote rules for the purpose of detecting a large number of necessary conditions for a positive assessment with respect to a criterion while a more efficient representation may have been to simply query for the presence of sufficient conditions for a negative assessment.

The QRP also observed errors in generality. In the Y2 evaluation, SMEs were working on fairly focused concrete scenarios. While this focus on a particular task seemed to usefully motivate the rule representation effort, an unfortunate consequence of the scenario focus is that SMEs occasionally tended to overly restrict a rule by including unnecessary details from a particular scenario when these details were not relevant to the general principle being applied. By the same token, SMEs also, in some instances, wrote rules that were too general, suggesting, for example. that a particular principle applied to any military task, when a large range of exceptions existed.

The implementation of negation also seemed to present difficulties for SMEs. They used negation infrequently and when it was implemented the implementation often failed to represent the presumed intended meaning of the SME.

The QRP also observed occasional inconsistencies across different knowledge representations, suggesting the need to keep the work of different SMEs somewhat isolated, and even within the work of a particular SME. SME's KR efforts occasionally contained gaps in knowledge representation. SMEs created many new classes and relationships and a common error when doing so was incorrect placement of the new concept into the preexisting KB. Also, SMEs did not always state sufficient information about a new concept to differentiate it from other preexisting concepts in the KB. A more helpful interface would better aid the SMEs in managing the status of their work to help make it clear when there are incomplete formalizations that need to be removed or completed.

SMEs occasionally failed to give enough information concerning spatial and temporal constraints. Often whether one event is salient to another is dependend upon whether the two events are spatially or temporally proximate, but SMEs representations often failed to consider this factor

Also noteworthy are other kinds of content that SMEs failed to consider. Several patterns authored by SHAKEN SMEs checked dependencies between actions. For example, one assessed a unit based on the ordering of its assigned tasks, and another checked for the presence of a follow-on mission. Another pattern linked the actions of friendly and opposing forces by checking whether an aviation attack is used to inhibit an opposing counterattack. However, no SMEs authored any knowledge that checked causal or parent/child relationships in the plans.

With respect to improving KA tools, several of the errors made by SMEs stem from the inability of KB systems to adequately explain the meaning of knowledge contained in the system. SMEs

regularly reported that the systems did not tell them clearly enough all they needed to know. In some cases, the knowledge was in the KB, but the interface left it difficult for SMEs to find. In others cases, such as system-generated explanation of predicates, rules, and inference results, their interpretation required KE-level skills.

Another important factor, noted by SMEs, was the inability of the systems to represent uncertainty, distributions over different scores and the means to implement a method for rolling up rules into a single evaluation.

The evaluation presented a significant challenge to the integration teams in terms of designing systems that SMEs could use to articulate reasoning principles requiring fairly complex logical representation. Evaluation results showed that SMEs were able to write KB rules that enabled correct analyses of COAs for some criteria, and in a manner that presented coherent explanations for the critique. Above we have analyzed some of the shortcomings in those representations and possible means of addressing them in terms of the focus of future work on development of KA tools.

Some of the elicited formal rules reflect KA challenges that might motivate interface extensions, including helping to manage work in progress and removing unwanted work, making the knowledge actually captured more transparent, connecting authored knowledge to prior knowledge, facilitating consistency checking within an SME's work and across SME efforts, extending dialogue tools to help ensure completeness and implementing more complete integration of tools for regular cycles of querying and resolution. One of the interesting questions concerns ways in which these systems present complementary KA tools that might be reintegrated into an even more effective system. The QRP noted a number of useful tools that exist within KRAKEN, for example, the Salient Descriptor, query suite, and generalization tool, that could be applied to some of the knowledge elicitation problems arising in SHAKEN. However, also notable is the fact that the SHAKEN rule elicitation tool – that is, the graphical interface within which a rule is mapped to a graph -- facilitated the elicitation of large numbers of patterns or COA critiquing rules.

5. References

[1] Alberts, B.; Bray, D.; Johnson, A.; Lewis, J.; Raff, M.; Roberts, K.; and Walter, P. *Essential Cell Biology: An Introduction to the Molecular Biology of the Cell*, Chapter 7, Pages 211–222, Garland Publishing, Inc., 1998.

[2] Cohen, P.; Chaudhri. V.; Pease, A. and Schrag, R. "Does Prior Knowledge Facilitate the Development of Knowledge-based Systems?" In *Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI-99),* Pages 221–226, 1999.

[3] Cohen, P.; Schrag, R.; Jones, E.; Pease, A.; Lin, A.; Starr, B.; Gunning, D., and Burke, M. "The DARPA High-Performance Knowledge Bases Project." AI Magazine 19(4), 1998.

[4] Forbus, K., Usher, J., and Chapman, V. 2003. "Sketching for Military Courses of Action Diagrams", Proceedings of the Intelligent User Interfaces Conference (IUI-03), January 2003. Miami, Florida.

[5] Gruber, T. "Toward Principles for the Design of Ontologies Used for Knowledge Sharing," Technical Report, KSL-93-04, Department of Computer Science, Stanford University, 1993.

[6] Gruninger, M.; and Fox, M.S. "Methodology for the Design and Evaluation of Ontologies," in Proceedings of the IJCAI-95 Workshop on Basic Ontological Issues in Knowledge Sharing, 1995.

[7] Mahoney, S., and Laskey, K. Constructing Situation Specific Belief Networks. UAI 1998: 370-377.

[8] Pool M., Murray, K., Fitzgerald J., Mehrotra M., Schrag R., Blythe J., Kim J., Hans Chalupsky H., Miraglia P., Russ T., Schneider D. "Evaluating Expert-Authored Rules for Military Reasoning," in Proceeding of 2003 Knowledge Capture Conference.

6. Web-accessible Materials

This section lists uniform resource locators (URLs) for results summarized in Section 1.1. For hardcopy, contact an IET representative listed on the cover page.

1. Results of the RKF Y1 Textbook and Expert Knowledge CPs: http://www.iet.com/Projects/RKF/IET-RKF-Y1-Evaluation.ppt
2. Results of the RKF Y1.5 Textbook Knowledge CP: http://www.iet.com/Projects/RKF/IET-RKF-Y2-SanDiego-MidtermEvaluation.ppt
3. Rescinded/preliminary RKF Y2 CP concepts regarding modeling of adversarial biological weapons (BW) development and detection other asymmetric threat activity in intelligence data: http://www.iet.com/Projects/RKF/IET-Y2-RKF-CP.ppt
4. Rescinded RKF Y2 IPB CP: http://www.iet.com/Projects/RKF/BSCP-spec--v1.0.doc
5. Current RKF Y2 COA CP: http://www.iet.com/Projects/RKF/COA-CP-spec--v1.1.doc
6. Y2 COA Evaluation Charts and Data : http://www.iet.com/Projects/RKF/RKF-COACritiquingEvaluation-SummaryCharts.xls

7. Acronyms & Expansions

Acronym	*Expansion*
AI	Artificial Intelligence
API	Application Program Interface
BS CP	BattleSpace Challenge Problem
COA	Course of Action
CP	Challenge Problem
DARPA	Defense Advanced Research Projects Agency
DNA	DeoxyriboNucleic Acid
EKCP	Expert Knowledge Challenge Problem
E2E	End-to-End
GKE	Gatekeeper Knowledge Engineer
GMU	George Mason University
HPKB	High Performance Knowledge Bases (DARPA program)
IET	Information, Extraction and Transport, Inc.
ILS	[NWU] Institute for the Learning Sciences
IPB	Intelligence Preparation of the Battlefield
KAW	Knowledge Acquisition Workshop
KB	Knowledge Base
KBMC	Knowledge-Based Model Construction
KBS	Knowledge Based System
KE	Knowledge Engineer
KR	Knowledge Representation
KR&R	Knowledge Representation and Reasoning
MDMP	Military Decision Making Process
METT-T	Mission, Enemy, Terrain, Troops, Time available
NWU	Northwestern University
OCOKA	Observation & fields of fire, Cover & concealment, Obstacles, Key terrain, Avenues of approach
PDA	Pre-Defined Axioms
PI	Principal Investigator
PQ	Parameterized Question
PSM	Problem-Solving Method
RKF	Rapid Knowledge Formation (DARPA program)
RKF Y1	RKF's [Evaluation] Year 1
RKF Y1.5	RKF's [Evaluation] Year 1.5 (mid-year evaluation)
RKF Y2	RKF's [Evaluation] Year 2
SD	Situation Description
SME	Subject Matter Expert
SQ	Sample Question
SRI	(no expansion—non-acronym name of SRI International, Inc.)
SSM	Surface-to-Surface Missile
TKCP	Textbook Knowledge Challenge Problem
TM	Theater Missile

TQ	Test Question
TTP	Tactics, Techniques, & Procedures
UA	User-Authored Axioms
URL	Uniform Resource Locator
VSD	Veridian Systems Division

8. PQ Notation

In the following, "expr" stands for an arbitrary PQ grammar expression.

"<HPKB:Country>":	an instance of the collection Country as defined in the Cyc Integrated KB (IKB)
"{UN, NATO, USA, ...}":	one of "UN," "NATO," "USA..."
"<MultilateralAgent> = {UN, NATO, USA, ...}":	A pseudo-class definition.
"[or why not]":	The subexpression "or why not" is optional; it can occur here, in this question or class definition, or just be omitted.
"expr+" (or "+expr"):	one or more patterns allowed by "expr", as in the following example (where –> denotes a possible rewriting). "{<Country>, the world}+" –> "France Algeria the world"
"expr*" (or "*expr"):	zero or more patterns allowed by "expr".
"^expr":	expr {and expr}*
"|expr":	expr {or expr}*
"[[0.0 .. infinity]]"	The closed interval between 0.0 and infinity.
"%comment":	The remainder of this line, following "%," is a comment, not part of a definition.

PQ Variable Scoping

If an expression like "<Agent>" appears twice in a question, below, it means that the user must choose the same binding or parameter choice for both occurrences. If, however, what appears is "<Agent1>" and "<Agent2>", that means the user has two choices to make; the two agents selected might or might not be the same. We also refer to occurrences of "<Agent>" and "<Agent1>" as "class variables."