QUEEN
QA
76.76
.E95
D3514
1988

Gouvernement du Canada
Ministère des Communications

Government of Canada
Department of Communications

Le Centre canadien de recherche sur l'informatisation du travail
Canadian Workplace Automation Research Centre

APPROCHE SYSTÉMIQUE DE L'ÉVALUATION DES SYSTEMES EXPERTS

par

Kim Dalkir

Canada

Queen
QA
76.76
E95
D35
1988

Industry Canada
Library Queen
AVR
APR 24 1998
Industrie Canada
Bibliothèque Queen

COMMUNICATIONS CANADA
JAN 6 1993
LIBRARY — BIBLIOTHÈQUE

APPROCHE SYSTÉMIQUE DE L'ÉVALUATION DES SYSTEMES EXPERTS

par

Kim Dalkir

Ce rapport est le résultat de recherches effectuées au Centre canadien de recherche sur l'informatisation du travail (CCRIT) Laval

Communications Canada

Janvier 1988

DD 10236109
DL 12447111

QA
46.76
.95
D35M
1988

No. cat. 28-1/68-1990F
ISBN 0-662-96425-X

TABLE DES MATIERES

1. RÉSUMÉ 1

2. INTRODUCTION 2
   - 2.1 Systèmes experts au gouvernement 2
   - 2.2 Systèmes experts pour l'évaluation de projets de R & D 3
   - 2.3 Le systèmes Satin en France 4
   - 2.4 Le système QUARTZ 5

3. OBJECTIFS 5

4. MÉTHODOLOGIE DES ESSAIS 6
   - 4.1 Théorie de l'évaluation 6
     - 4.1.1 Évaluation formative et sommative 6
     - 4.1.2 Approche systémique de l'évaluation 7
     - 4.1.3 Intégration de l'évalution dans la conception des systèmes 8
   - 4.2 Techniques d'évaluation 9
   - 4.3 Évaluation des systèmes experts 11
     - 4.3.1 Métrologie des systèmes experts (données objectives) 14
     - 4.3.2 Critères qualitatifs (données subjectives) 15
   - 4.4 Conclusions 19

5 ANALYSE ET TRAVAUX FUTURS 20
   - 5.1 Avantages démontrés 20
   - 5.2 Projets de travaux 21

6. RÉFÉRENCES 22

7. APPENDICES I - HIÉRARCHIE DES CRITERES D'ÉVALUATION DE LA THÉORIE MULTI-ATTRIBUT DE L'UTILITÉ (MAUT) 24
   - 7.1 APPENDICES II - MÉTROLOGIE DE LA BASE DE CONNAISSANCES QUARTZ 26

1. RÉSUMÉ

QUARTZ est un système multi-expert qui renferme le savoir de plusieurs experts dans les domaines de l'analyse et de l'évaluation des projets. Il contient les objectifs et critères du programme EDER* du Québec pour le développement économique régional de firmes du Québec. Les utilisateurs par le biais de l'interrogation interactive dans le cadre des propositions soumises, amènent le système expert à cumuler les renseignements qui permettent d'en arriver à une recommandation. Les résultats de la consultation s'impriment avec les renseignements fournis par l'utilisateur, les déductions du système à partir des faits, les règles appliquées et bien sur la recommandation.

Dans le présent rapport, nous décrivons les travaux de développement du projet QUARTZ qui se sont déroulés au cours du dernier trimestre de 1987, en mettant surtout l'accent sur la méthodologie de génie cognitif utilisée. L'acquisition et la représentation des connaissances dans le système QUARTZ sont terminées; elles sont décrites dans un article précédent (Dalkir et Muzard, 1987). Les deux grandes étapes suivantes sont la validation de la base de connaissances et l'évaluation globale du système QUARTZ lui-même. Nous les abordons ici par une étude de la théorie et des techniques d'évaluation et de leur application aux systèmes experts. Nous décrivons une approche systémique de l'évaluation des systèmes experts que nous illustrons avec l'exemple du système QUARTZ. En conclusion, nous formulons les leçons attirées ainsi que les limites du système et proposons des travaux pour la poursuite du projet.

---

* Entente de développement économique et régional

2. INTRODUCTION

QUARTZ est un système expert d'aide au profit des analystes du gouvernement qui doivent évaluer les demandes de subvention dans le cadre du programme EDER. QUARTZ servira d'outil à l'étape de l'analyse du traitement des propositions. Le prototype initial du système sera appliqué à l'évaluation de projets EDER d'un domaine particulier, soit celui du développement de logiciels innovateurs en langue française. L'expertise englobe le savoir combiné des analystes des ministères des Communications de gouvernements fédéral et provincial, qui conjointement sont responsables de l'administration de ce programme.

2.1 Systèmes experts au gouvernement

Le gouvernement est largement vu comme le deuxième plus important client éventuel de la technologie de l'intelligence artificielle, après le secteur privé. Des systèmes experts commerciaux sont déjà exploités avec succès dans l'industrie. Le gouvernement utilise déjà de très grandes bases de données, formées généralement de règles et dont l'exploitation exige de la formation et des connaissances. Toute machine gouvernementale a besoin, pour tourner, d'une infinité de règles, règlements, procédures, lois, etc. Par ailleurs, il y a une tendance à une grande mobilité parmi les fonctionnaires qu'attirent les promesses de situations plus lucratives dans le secteur privé. Les systèmes experts semblent être la solution logique à ce problème de perte d'expertise et un choix possible pour éviter la formation longue et coûteuse.

L'usage généralisé de systèmes experts dans les gouvernements peut présenter un certain nombre d'avantages : offrir au public une information plus uniforme; réduire les exigences de formation professionnelle augmenter la productivité professionnelle, réduire les coûts de traitement de l'information et faciliter au public l'observance des lois et des règlements L'usage généralisé des systèmes experts faciliterait aussi les relations avec le public. Imaginez comme ce serait facile pour vous de présenter votre déclaration d'impôt directement au ministère du Revenu en dialoguant avec un expert-conseil fiscal. Un autre rôle intéressant pour le gouvernement peut consister à établir les standards des systèmes experts. Par exemple, en développant une base de connaissances avant que ne le fasse le secteur privé, les gouvernements peuvent

établir des standards suivant lesquelles seraient mesurées et évaluées toutes ces bases de connaissances.

Malheureusement, la technologie de l'intelligence artificielle reste coûteuse et exige le développement de compétences internes. Aussi, certaines applications gouvernementales peuvent être spécifiques, donc non transférables au domaine public ou universitaire. Pour éviter de tels écueils, la recherche sur les systèmes experts au gouvernement sera vraisemblablement plus appliquée que fondamentale. Selon Brixius [Brixius, 1984], une bonne stratégie consisterait à créer un prototype clé de système expert qui aurait des applications immédiates et qui serait un produit de pointe. Les applications choisies devraient être évolutives afin de pouvoir devenir plus tard des systèmes plus complexes.

Le Centre canadien de recherche en informatisation du travail représente une tentative de la part du gouvernement du Canada de poursuivre ce type alternatif de modèle recherche. Le Groupe des systèmes experts, qui oeuvre dans ce contexte, vise ainsi à produire des systèmes qui sont évolués technologiquement et qui, en même temps, répondent à une demande définie. Il faut, toujours selon Brixius, une masse critique de 9 à 10 individus pour obtenir l'interaction créative suffisante exigée pour le développement d'un système expert. Notre groupe comprend approximativement 15 chercheurs, de formation différente, qui travaillent sur un grand nombre d'applications de systèmes experts, différentes. QUARTZ est une de ces applications, un système expert pour l'évaluation des projets.

2.2 Systèmes experts pour l'évaluation des projets de R & D

Un domaine tout indiqué pour les systèmes experts est celui des projets de recherche et développement. Conçu pour suivre l'avancement des projets de R & D commerciaux, MACIE, est un exemple de système expert appliqué à ce domaine. [Gallant et Balachandra, 1986]. Les projets doivent être soigneusement choisis pour leur potentiel technologique et commercial avant qu'on engage des ressources substantielles dans leur développement. Malgré ce filtrage préalable, de nombreux projets n'aboutissent jamais. Si l'on pouvait, dès le départ, identifier les entreprises à haut risque d'échec, on pourrait utiliser les fonds à des projets moins risqués.

Il y a une grande ressemblance entre les domaines de MACIE et QUARTZ. Prenons, par exemple, les critères de disponibilité de l'expertise, de disposition favorable des groupes visés, de nombre possible d'applications pour le produit; ces critères sont importants pour l'évaluation et le contrôle des projets. Une distinction importante entre les deux cependant vient de ce que QUARTZ a été mis au point au moyen d'interviews classiques d'experts, tandis que le système MACIE a utilisé l'induction de règles à partir de plus de 100 études de cas de projets de R & D différentes pour en arriver aux variables de décision qui ont déterminé la conclusion des projets.

2.3 Le système Satin en France

Un système expert visant les mêmes objectifs est en cours de développement en France; il est l'oeuvre du groupe NITAS [NITAS, 1986]. Le système SATIN réunit de l'expertise en économie, en juridique et en technologie pour décider du financement de projets innovateurs. Son objectif est la promotion du transfert technologique vers les petites et moyennes entreprises, qui, sans cette aide, ne seraient pas en mesure de poursuivre de telles activités de recherche et développement.

Le système SATIN contient deux types de connaissances : une connaissance professionnelle des aspects financier, technique, juridique et commercial et une connaissance interdisciplinaire sur la manière dont tous ces aspects doivent être regroupés. La technologie de l'intelligence artificielle a été choisie pour résoudre le problème en raison de la nature changeante de l'innovation. Les modèles algorithmiques ne suffiraient pas et le système est très difficile à simuler.

Comme MACIE, le système SATIN présente un certain nombre de similitudes avec le système expert QUARTZ. Mais, ici, la différence n'est plus dans l'acquisition des connaissances mais dans leur représentation. Le groupe NITAS a décidé de développer une série de modules de systèmes experts que viendrait coordonner un module superviseur de manière cohérente. La modularité dans le système QUARTZ est interne; en effet, c'est sa base de connaissances qui est divisée en quatre domaines : comme les aspects techniques et innovateurs du produit à développer, l'analyse de l'environnement économique du produit, l'évaluation de l'entreprise qui soumet le projet et l'évaluation du projet en termes de modalités, de faisabilité technique et économique et de ressources de gestion.

Ainsi, les systèmes experts dans le domaine de l'analyse des projets peuvent aider les agents de transfert technologique à visualiser le processus d'innovation à un niveau plus global et à produire des recommandations plus complètes, exhaustives, ainsi que plus faciles à comprendre. L'aptitude universelle des systèmes experts à dévoiler le raisonnement sous-jacent à leurs conclusions, sous la forme d'une liste des règles utilisées dans les deux systèmes, aide l'utilisateur à comprendre le résultat et à décider s'il acceptera ou non le conseil reçu.

2.4 Le système QUARTZ

Le système QUARTZ est actuellement au stade d'un premier prototype opérationnel d'environ 300 règles. Les étapes d'acquisition et de représentation des connaissances sont terminées et actuellement la validation et l'évaluation du système suivant leur cours. Le prototype a été livré aux utilisateurs finals avec un guide d'exploitation et une formule d'évaluation qui servira à recueillir leurs remarques et critiques. Afin de réaliser une évaluation plus systématique, nous avons entrepris une étude exhaustive des théories et méthodologies d'évaluation pour déterminer la meilleure approche auprès des systèmes experts en général et du système QUARTZ en particulier.

3. OBJECTIFS

Le but est non seulement de valider la base de connaissances de QUARTZ (est-ce que les faits et règles sont corrects, précis, etc.) mais également, d'évaluer le système en entier, du point de vue des réalisateurs (des ingénieurs de la connaissance) comme de celui des utilisateurs. Ensuite, il y aura raffinement du premier prototype et développement d'un deuxième prototype opérationnel qui, vraisemblablement, aura deux fois plus de règles et des objectifs. L'objectif ultime est, bien sûr, l'accord de l'utilisateur. Ainsi, la boucle conception-évaluation-raffinement est une boucle continue qui ne s'arrête que lorsque les utilisateurs finals sont satisfaits du système. La faisabilité et l'utilité de l'approche de modélisation systémique ou cybernétique pour l'évaluation des systèmes experts seront également examinées au cours de l'essai du système.

4. MÉTHODOLOGIE DES ESSAIS

Nous avons fait une étude des théories et méthodologies pratiques de l'évaluation pour savoir où en est l'évaluation des systèmes experts. Il y a de nombreux principes généraux de l'évaluation, la plupart dérivés de la recherche en éducation, qui peuvent s'appliquer aux systèmes experts. Ceux-ci ont cependant un certain nombre de particularités dont il faut tenir compte au moment de leur essai. Ainsi la métrologie des systèmes experts : que doit-on mesurer pour évaluer les effets d'un système expert? L'évaluation d'un système expert en est encore à ses premiers balbutiements mais elle prendra une importance de plus en plus grande à mesure que les systèmes de recherche arriveront au stade de l'implantation.

4.1 Théorie de l'évaluation

La plupart des théories d'évaluation utilisées pour évaluer les systèmes d'aide à la décision sont empruntées aux ouvrages d'évaluation et de mesure en éducation. Nous nous inspirerons de ces théories, ainsi que de la théorie générale des systèmes, pour bâtir une théorie d'évaluation des systèmes experts, intégrée à la conception et au développement de ces systèmes.

4.1.1 Évaluation formative et sommative

Nous pouvons décrire les théories d'évaluation selon la dichotomie générale d'évaluation formative ou sommative. L'évaluation formative représente une tentative de relier la structure interne d'un système à son comportement externe en vue de guider le processus de développement. La question à poser est la suivante : "Que dois-je faire pour améliorer mon produit?" L'évaluation formative est ainsi une approche prescriptive qui se concentre sur le procédé ou moyen utilisé pour développer le système. Le système évalué est, ici, une boîte de verre qui, dans le cas de l'évaluation de logiciels, contiendrait le code du programme.

L'évaluation sommative représente une approche conceptuelle très différente et donne un résultat très différent de l'évaluation formative. L'évaluation sommative ne

mesure la performance du système que lorsqu'il est complet. Elle donne des mesures externes du succès, sous forme de diagnostiques. La question posée en évaluation sommative est : "Le système fonctionne-t-il bien?". L'accent est sur le produit, le résultat. Les analyses coût-bénéfices, par exemple, sont des évaluations sommatives. Le système évalué est ainsi vu comme une boîte noire et seuls les comportements ou résultats observables sont utilisés comme données d'évaluation.

Les deux approches offrent des perspectives différentes du fonctionnement du système étudié et devraient donc être considérées comme complémentaires et non comme mutuellement exclusives. D'un point de vue global, il est essentiel de les appliquer toutes deux afin d'évaluer adéquatement le système expert.

4.1.2 Approche systémique de l'évaluation

L'approche systémique de l'évaluation préconise l'analyse de la structure du système et la synthèse des résultats pour étudier le fonctionnement du système ("l'anasynthèse") à différents niveaux hiérarchiques, soit :

1. le niveau du problème - limites conceptuelles du problème étudié
2. niveau de décomposition ou d'analyse - subdivision du problème
3. niveau du plan - description et explication de la solution du problème
4. composition ou synthèse - interactivité des parties du système.

Un système expert doit être vu comme un sous-système d'un système plus grand, tel le groupe de travail ou l'organisation intéressée (voir figure 1). Il est très important de définir et d'analyser le contexte dans lequel le système doit fonctionner afin d'évaluer efficacement son rendement. Il est aussi important d'analyser le grand système selon les même principes, afin d'être en mesure non seulement d'évaluer le système expert mais aussi le rendement et l'utilisation possible du système entier.

Ainsi, l'évaluation doit mesurer l'efficacité du système (jusqu'à quel point permet-il d'accomplir les fonctions prévues?), son efficience (de quelle manière exploite t-il les ressources pour fonctionner?) et sa viabilité (peut-il continuer à accomplir ces fonctions?). Les conclusions auxquelles est arrivée l'auteure de cette étude indiquent que la "théorie du système viable" [Beer, 1979] peut constituer une bonne base

théorique pour l'évaluation des systèmes experts. Les critères de fiabilité, de conservation et de viabilité du système font des systèmes experts des candidats idéals pour cette approche conceptuelle, non seulement pour ce qui est de l'évaluation, mais aussi pour la conception initiale.

4.1.3 Intégration de l'évaluation dans la conception des systèmes

Si la conception des systèmes vise des objectifs stratégiques, les évaluations, elles, tendent généralement à être presque exclusivement opérationnelles de nature. Il ne suffit pas de se demander si le système fonctionne convenablement : il faut également définir les processus sous-jacents et les évaluer. D'où la nécessité d'intégrer entièrement l'évaluation des systèmes experts dans la méthodologie de conception et de développement, et donc de s'entendre sur les critères de rendement et sur les jeux d'essai dès le stade de la conception.

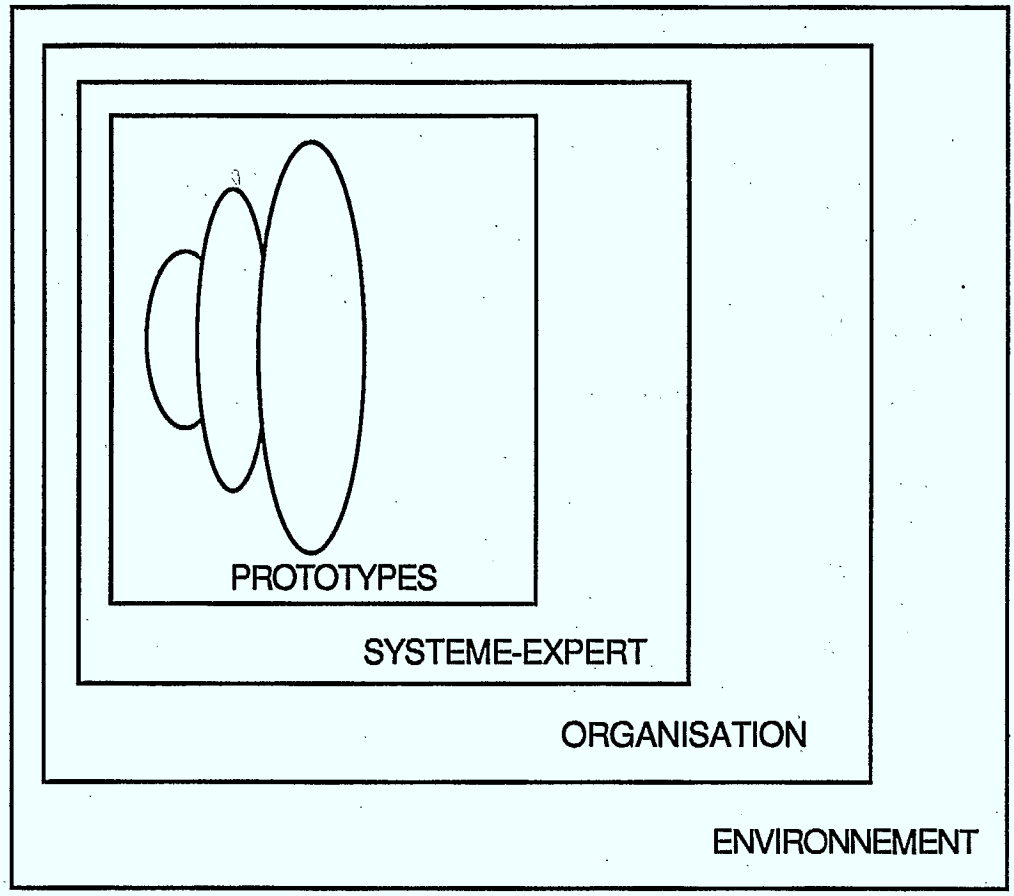
Un plan d'évaluation doit être élaboré au début du projet. Trois niveaux d'objectifs doivent être traités dans ce plan :

1. le système fonctionne-t-il convenablement?
2. les utilisateurs finals visés comptent-ils l'utiliser?
3. l'approche ou la technologie utilisée est-elle appropriée?

Par exemple, le premier niveau consisterait à faire vérifier les résultats par le système; le deuxième, à vérifier si les utilisateurs se sont engagés à employer le système et le troisième, s'il était indiqué, au départ, de recourir aux technologies de l'intelligence artificielle?

Cette approche intégrée de l'évaluation donnerait une meilleure compréhension de la technologie utilisée pour développer le système, un meilleur système en général (la réaction de l'utilisateur servirait à affiner le système) et ferait ressortir les difficultés (par exemple, le système est-il trop difficile à utiliser?).

Figure 1. Vue systémique des systèmes experts



4.2 Techniques d'évaluation

On retrouve dans toutes les théories d'évaluation trois principes de base. Il s'agit des principes suivants :

1. Aucun processus ou système complexe ne peuvent être évalués à l'aide d'un seul critère ou d'un seul chiffre ("une seule note")
2. Plus nombreux sont les critères, et plus les mesures à prendre le sont aussi, meilleure sera la qualité de l'évaluation
3. L'unanimité n'est pas rencontrée quant à la signification relative des critères parce que les personnes n'ont pas la même formation ou expérience ni les mêmes intérêts.

La technique d'évaluation utilisée doit être basée sur un nombre de critères quantitatifs et qualitatifs suffisant par rapport à la complexité de la performance du système. Tout comme le développement d'un système expert est un processus itératif dynamique, l'évaluation doit être continue et adaptative. Finalement, en choisissant un bon échantillon d'utilisateurs potentiels du système, avec une expérience et des intérêts variés, on réussirait à mieux tester le système. Par exemple, il semble y avoir une tendance parmi les utilisateurs potentiels à se concentrer sur la manière dont le système peut s'adapter à leurs méthodes (critères principalement subjectifs), tandis que les chercheurs se penchent surtout sur les questions d'interface, comme le temps de réponse (critères principalement objectifs). Les deux aspects doivent être évalués.

Les deux principales techniques d'évaluation qui peuvent servir à évaluer à la fois les données subjectives et objectives sont les sondages des utilisateurs et les revues post-implantation. Quant aux sondages auprès des utilisateurs, les questionnaires ou les interviews peuvent servir à mesurer la quantité d'information et l'importance de l'assistance que, selon les utilisateurs, le système leur procure (critères subjectifs). Notons ici que cette appréciation comprendrait aussi l'évaluation de toute documentation ou de toute formation fournie avec le système. Les évaluations post-implantation sont essentiellement sommatives et révèlent le succès avec lequel le système exécute les fonctions prévues (critères objectifs).

La plupart des techniques d'évaluation sont empruntées à l'évaluation des logiciels, et, en particulier, à celle des systèmes d'aide la décision. Les systèmes experts tombent sous l'étiquette générale des aides à la décision; le problème qui se pose est le même : comment mesurer l'amélioration dans la prise de décision? Les techniques d'évaluation répondent à des questions qui ressemblent à ceci : "Dans quelle mesure le système facilite-t-il ou complique-t-il le travail de l'utilisateur? Dans quelle mesure facilite-t-il les aspects cognitifs de la prise de décision? Le système aide-t-il à améliorer la qualité globale des décisions prises? Quels critères doit-on évaluer? Comment doit-on les évaluer?

Certaines méthodologies d'évaluation sont basées sur des techniques d'analyse de décision, comme celles des analyses de sensibilité et de valeur attendue. D'autres utilisent l'approche de l'homme rationnel, avec l'hypothèse que l'information parfaite est toujours disponible et que la solution optimale peut être trouvée au moyen de méthodes quantitatives. Les approches incrémentielles sont plus indiquées pour les

systèmes experts. Une de ces approches est PAM (Project Appraisal Methodology) [Leibowitz, 1985]. Celle-ci combine la notation à des méthodes d'évaluation économique où chaque critère sert à évaluer un aspect distinct du système. Si tous les critères sont mutuellement exclusifs, il n'est toutefois pas nécessaire de les évaluer tous pour étudier le système.

Une autre technique d'évaluation des systèmes couramment utilisée est MAUT ou théorie multi-attribut de l'utilité [Adelman, 1986]. Voir l'exemple à l'appendice 1. Les critères sont, selon cette méthode, évalués à l'aide d'un questionnaire de 121 questions, notées de 0 à 10 (0 signifiant "pas du tout d'accord" et 10, "tout à fait d'accord"). Le questionnaire a été validé et donne des résultats fiables. En général, plus la note globale du système est élevée, meilleurs sont les résultats de l'évaluation. Cette approche permet l'évaluation d'un grand nombre de critères et de plusieurs points de vue différents. PAM et MAUT semblent toutes deux être de bonnes techniques pour l'évaluation de systèmes experts.

4.3 Évaluation des systèmes experts

Comment doit-on évaluer les systèmes experts? L'évaluation formative semble être essentielle et a été proposée comme le meilleur moyen d'évaluer les systèmes experts [Hofmeister, 1986]. Toutefois, une brève étude de la manière dont les systèmes experts sont évalués indique que l'évaluation sommative est plus fréquente [Soloway et Littman, 1987]. Comme je l'ai déjà dit, les deux sortes doivent être appliquées si l'on veut une évaluation exhaustive de la performance du système expert.

Les prototypes de systèmes experts en réalité servent d'étude de faisabilité de l'application de la technologie de l'intelligence artificielle à la résolution de problèmes donnés. Un système expert n'est jamais vraiment "arrivé à terme"; il est en évolution continue. Ainsi, l'évaluation sommative, dans le sens strict du terme, est impossible. Toute critique ou suggestion formulée par l'utilisateur n'est pas simplement mise de côté et classée, mais incorporée au système. L'évaluation formative est nécessaire pour guider le développement.

Il est bon de tester le système expert sur le terrain dès que l'on a un prototype opérationnel. L'essai en milieu opérationnel permet en effet la collecte de données

évolutives le plus tôt possible. Ensuite, le système peut être revu par un expert-conseil qui jugera du choix des outils et de la conception générale. Pour vérifier les connaissances, on peut valider le premier ensemble de règles de la base de connaissances en les vérifiant en regard des sources de connaissances factuelles et en confiant à un groupe d'experts la vérification des connaissances heuristiques. Le raisonnement peut être évalué avec un ensemble de jeux d'essai, développé par les utilisateurs en collaboration avec l'ingénieur de connaissances. On peut comparer les résultats obtenus avec le système avec ceux d'experts humains travaillant sur les mêmes jeux d'essai.

Dans ce test préliminaire de la fiabilité du système, on peut vérifier si un même utilisateur obtient des résultats constants à des moments différents et si des personnes différentes travaillant sur le même problème obtiennent les mêmes résultats. Le système expert peut alors être graduellement raffiné et la documentation mise à jour en fonction des changements apportés.

Comme il est impossible d'énumérer et d'évaluer toutes les entrées possibles du système expert, il faut générer un échantillon représentatif de jeux d'essai. Il peut être aussi difficile d'essayer le système avec tous les utilisateurs finals éventuels. On devrait donc choisir également un bon échantillon d'utilisateurs pour évaluer le système.

Dans les deux cas, on doit appliquer les méthodes statistiques appropriées afin d'obtenir des données valides (par exemple, choisir des échantillons de taille convenable et de bonnes techniques d'échantillonnage). Pour vérifier le système, on peut choisir des candidats parmi les utilisateurs finals visés, d'autres parmi les ingénieurs de connaissance, des étudiants universitaires du deuxième cycle, des personnes n'ayant aucune notion d'informatique et des experts d'autre milieu [Krcmar, 1985]. Ces différents types d'utilisateurs devraient fournir un éventail de données de toutes sortes différentes.

Une bonne façon de rassembler des critères d'évaluation consiste à demander à des experts de noter des études de cas ou des décisions prises à partir d'une variété d'attributs. Par exemple, la complexité relative de chaque cas, sa qualité (les risques de situations semblables? a-t-on de la rencontrer de nouveau?), le temps passé par les

experts à noter, etc. Ce type d'analyse de tâches générera un bon éventail de problèmes de types différents pouvant servir à vérifier le système expert.

Quelques systèmes experts sont arrivés au stade de la mise en marché et ont été évalués. Un bref aperçu de certains d'entre eux illustre les méthodes d'évaluation appliquées aujourd'hui [Soloway et Littmann, 1987]. Le premier et le plus connu des systèmes experts, MYCIN, a été évalué avec le jugement d'expert sur l'exactitude des résultats. La performance du système a été comparée en regard de trois groupes : experts, médecins résidents et étudiants en médecine; MYCIN a été trouvé aussi bon ou meilleur que les trois groupes pour résoudre un ensemble de dix problèmes réels complexes. Le critère utilisé a été le pourcentage de diagnostics corrects. Cette évaluation sommative a aussi révélé que les utilisateurs trouvaient le système trop difficile à employer.

Le système INTERNIST a également été évalué avec des techniques sommatives et ses résultats ont été comparés à ceux des médecins. Le système PROSPECTOR a fait l'objet d'une évaluation sommative de son intégrité et de l'organisation de sa base de connaissances, toujours en comparaison avec des experts humains utilisant les mêmes connaissances. Des programmeurs novices ont essayé le système PROUST sur 200 programmes et ont fait une évaluation sommative du pourcentage d'erreurs (de virus) détectées par le tutoriel intelligent. De plus, une évaluation externe des avantages pédagogiques a été réalisée et l'on a trouvé que PROUST donnait aux étudiants un avantage de 16 % aux examens!

R1/XCON est le seul système qui a été évalué avec des techniques d'évaluation formative. Pendant trois ans, des évaluations trimestrielles ont été faites de ce projet de développement d'une durée de 6 ans (avec plus de 6 000 règles dans la base de connaissances). Les critères utilisés ont été la diminution en pourcentage des erreurs factuelles et des règles incorrectes dans la base de connaissances. Ainsi, l'évaluation formative tend à être relative (toujours en termes de degré d'amélioration sur le prototype précédent), tandis que l'évaluation sommative donne un jugement de valeur absolue ou la certification du système expert.

4.3.1 Métrologie des systèmes experts (données objectives)

Peu de travaux ont été faits à ce jour sur le développement d'une métrologie pour l'évaluation formelle des systèmes experts, probablement parce que si peu d'entre eux sont des produits commercialisés. En outre, certaines contributions uniques de systèmes experts méritent d'être évaluées par des experts. Nous pouvons cependant utiliser la métrologie pour spécifier un certain nombre de paramètres objectifs du système expert, qui nous serviraient à définir les objectifs du système, à prévoir les résultats que nous pouvons en attendre et à mesurer les résultats réellement obtenus [Kaisler, 1985]. Des critères quantitatifs ou quantifiables peuvent aussi servir à mesurer l'écart entre les résultats attendus et réels, ainsi qu'à comparer différents systèmes.

Les critères quantitatifs sont particulièrement utiles lorsqu'on doit mesurer le progrès évolutif des systèmes experts dans une évaluation formative. Les deux types généraux de métrologie sont la métrologie des applications, qui est particulière au domaine, et la métrologie technologique, qui traite d'efficience et d'implantation et qui est particulière au noyau ou au langage utilisé pour développer le système. Parmi les catégories métriques qu'il peut être utile de considérer, on compte la complexité de la base de connaissances, l'exécution de la base de connaissances et la métrologie du système.

Avec la complexité de la base de connaissances, on mesure l'ampleur du travail de développement. Les mesures peuvent être déterminées par le nombre d'objets, le nombre de cases par cadre, le nombre de règles, le nombre moyen d'instructions si dans une règle, etc. En nous inspirant encore une fois de l'approche systémique, nous pouvons considérer la base de connaissances comme un sous-système du système expert. La complexité du système est généralement définie comme étant égale à (N-1)/S, N représentant le nombre de noeuds ou de points de décision, et S, le nombre de liaisons entre les noeuds. En la modifiant légèrement, nous pouvons adapter cette formule au schème de représentation des connaissances utilisé, N devenant le nombre d'objets et S, le nombre de schémas par exemple. (L'appendice 2 donne les mesures de complexité des trois bases de connaissances développées pour le système QUARTZ).

La métrologie de l'exécution d'une base de connaissances mesure, entre autres, le temps requis pour le démarrage, la mise à jour et le temps de réponse du système. La

métrologie de l'exécution mesure la vitesse globale de traitement des connaissances, le nombre de règles considérées pour chaque conclusion, la quantité d'informations comprise dans les règles individuelles, le temps moyen de consultation par utilisateur, le nombre de concepts distincts représentés dans la base de connaissances, etc.

Dans l'évaluation quantitative des systèmes experts, nous devons tenir compte d'un certain nombre de points importants, notamment :

- la métrologie n'est pas statique étant donné que le système évolue et se développe
- les outils qui ont servi à élaborer le système doivent aussi être évalués
- le nombre et les types de problèmes que peut résoudre le système doivent être identifiés, évalués quant à la précision et la fiabilité du système
- l'ampleur des changements à apporter à la conception doit être mesurée, avec, par exemple, le nombre de règles par changement
- une analyse des tendances doit être faite à partir d'instantanés incrémentiels du système en évolution (par exemple, un graphe chronologique de la complexité de la base de connaissances).

Ces mesures peuvent nous indiquer si la base de connaissances est trop compliquée ou trop simple, et, par là même, si l'application de la technologie des systèmes experts au problème était opportune, si les utilisateurs passent trop ou trop peu de temps avec le système (ce temps doit se comparer au temps que l'on passe avec les experts humains, soit 20 minutes environ par consultation). Une autre sorte de métrologie consisterait à mesurer l'efficience du génie cognitif en termes de nombre de règles par jour (ou de nombre de jours par règle!). Toutefois, alors que les mesures quantitatives sont nécessaires, elles ne sont cependant pas suffisantes pour évaluer les aspects uniques du système expert. Nous devons également obtenir des données qualitatives ou subjectives.

4.3.2 Critères qualitatifs (données subjectives)

Un ensemble généralement accepté de critères qualitatifs ou subjectifs reste à développer pour évaluer le système expert. Un grand nombre de critères peuvent être empruntés à l'évaluation classique des logiciels, comme la maniabilité, la capacité

d'évolution, la rentabilité, etc. A cette liste, il faudra ajouter des critères spécifiques à la technologie des systèmes experts. Pour classer le grand nombre de critères applicables, nous utiliserons la taxinomie proposée par Krcmar [Krcmar, 1985].

Cette taxinomie reconnaît trois grandes catégories : les aspects techniques (de matériel et de logiciel utilisés pour développer le système), la base de connaissances (la validité des faits et règles du système) et ce que l'on appelle les facteurs environnementaux ou humains (sensibilisation des cadres, l'acceptation par l'utilisateur, analyses coût-bénéfices-risques, etc.). Dans le tableau qui suit, nous élaborons une liste des critères les plus importants dans chaque catégorie.

Tableau 1. Critères qualitatifs d'évaluation des systèmes experts

Aspects techniques

- souplesse du système expert et adaptation aux changements de l'environnement (par exemple, les objectifs du système ne sont plus les mêmes)
- l'addition de connaissances exige-t-elle des changements majeurs?
- aptitude du système à prendre de l'expansion avec l'apport d'expertise et de rétroaction
- fiabilité de fonctionnement du système (par exemple, est-ce qu'il tombe en panne?)
- efficience de l'utilisation des ressources par le système (par exemple, l'utilisation de la mémoire)
- indépendance par rapport au matériel ou maniabilité (le système peut-il tourner sur des configurations de matériel différentes?)
- intelligibilité ou compréhensibilité du système (le code est-il facilement lisible?)
- testabilité (le fonctionnement du système est-il facilement testable?)
- facilité d'interaction avec le système (qualité technique de l'interface)

Base de connaissances

- applicabilité de la base de connaissances (quels problèmes peuvent être résolus?)
- la base est-elle compréhensible pour les utilisateurs comme pour les réalisateurs (par exemple, ambiguité des locutions)
- précision du contenu (généralement déterminée par des experts du domaine)
- qualité des déductions et des avis du système
- exactitude des techniques de raisonnement utilisées
- les résultas produits sont-ils utiles? (par exemple, présentation des rapports)
- facilité d'interaction (par exemple, fonction d'aide)
- qualité non technique de l'interface (par exemple, fenêtres, mise en évidence)
- intégralité de la base de connaissances (y a-t-il des cas non résolus?)
- cohérence de la base de connaissances (par exemple, terminologie, notation uniformes)
- structuration de la base de connaissances (par exemple, est-elle bien structurée?)
- autodescription (y a-t-il assez d'information dans la base de connaissances pour déterminer les objectifs, les hypothèses, les contraintes, l'état à tout moment?)
- les conflits sont-ils résolus de manière adéquate? (par exemple, tous les conflits entre experts ont-ils été résolus ou des métamodèles d'expertise ont-ils été construits? Dans ce dernier cas, les utilisateurs ont-ils le choix de l'expert à consulter dans le système?)
- pouvez-vous faire la différence entre une bonne base de connaissances et un mauvais moteur d'inférence et une base de connaissances médiocre mais avec un bon moteur d'inférence?
- la représentation des connaissances est-elle trop riche? trop simple? compatible avec les organisations des connaissances des experts?

Facteurs environnementaux ou humains

- les utilisateurs acceptent-ils le système?
- ont-ils l'intention de s'en servir?
- ont-ils confiance dans son fonctionnement et dans ses résultats?

- le système aide-t-il les utilisateurs dans leur travail?
- étant donné que les utilisateurs ne sont prêts à s'adapter aux besoins d'aucun système, le système peut-il, lui, s'adapter à l'utilisateur, au contexte, au niveau d'expertise et aux stratégies préférées?
- le système expert peut-il être utilisé dans d'autres applications? par d'autres groupes? (jusqu'à quel point est-ce un produit générique ou spécialisé?)
- le temps de conception convient-il aux développeurs? aux utilisateurs? à la direction?
- le système fonctionne-t-il bien du point de vue des utilisateurs, des ingénieurs de la connaissance, des consultants externes?,
- les suggestions des utilisateurs ont-elles été intégrées dans le système?
- le système est-il rentable?
- quels sont les avantages du système? (par exemple, sauvegarde de l'expertise de l'entreprise, formation, prise de décision meilleure et plus uniforme, meilleure qualité de vie au travail, protection contre les erreurs, gain de temps, meilleure communication entre les utilisateurs humains, aptitude à servir une clientèle plus large)
- quels sont certains des risques associés à l'utilisation du système? (par exemple responsabilité civile)
- combien a coûté le système? (coût du développement, de la livraison, de l'entretien).

Il est aussi important de connaître les principales raisons pour lesquelles les systèmes experts échouent. Le critère d'applicabilité semble jouer le plus grand rôle dans le succès ou l'échec de ces systèmes. La raison pour laquelle la plupart des systèmes ne sont jamais implantés avec succès est l'inapplicabilité du problème traité. Ou bien celui-ci a été mal compris, ou bien il ne requiert que le sens commun; il n'existe peut-être pas de solution au problème (de l'avis même des experts humains); ou bien nous faisons face à une situation en vase clos, et de ce fait c'est-à-dire, le système ne sait pas ce qu'il ne sait pas (contrairement au médecin, par exemple, qui sait quand vous recommander d'aller voir un autre spécialiste).

L'acceptation de l'utilisateur est le dernier obstacle décisif a franchir en regard des systèmes experts. Afin de maximiser les chances de succès, le système doit être conçu de manière à ce qu'on puisse le soumettre à des tests; les utilisateurs doivent faire confiance au système et ne pas penser que le système les remplacera au travail.

D'autres raisons peuvent influencer les utilisateurs à ne pas accepter le système expert, entre autres :

- la méconnaissance de la technologie informatique
- la difficulté à s'adapter à l'interface utilisateur
- la conviction que les humains peuvent faire le travail mieux que le système
- le sentiment qu'ils doivent s'en servir que pour les cas très difficiles
- la tendance à se rappeler les problèmes (même très anodins) qu'ils ont déjà eus avec le système
- le manque de confiance ou trop de confiance dans les résultats
- la diminution de la satisfaction au travail.

Il a de tout temps été prouvé [Lane, 1986] que l'introduction de toute innovation dans un système donné doit tenir compte de la volonté des utilisateurs visés à accepter le changement et à l'intégrer à la structure de leur travail. Il y a encore beaucoup de méfiance à l'endroit des ordinateurs et la question la plus critique entre toutes est peut-être le degré de transparence ou d'opacité du processus de raisonnement du système expert. Tant que les utilisateurs ont l'impression de pouvoir accéder aisément au raisonnement interne du système, ils l'accepteront plus volontiers. D'où l'importance des fonctions d'explication du système (son aptitude à répondre aux "comment" et "pourquoi").

4.4 Conclusion

L'évaluation complète d'un système expert exige des mesures à la fois objectives et subjectives. Les mesures objectives comprennent l'analyse des erreurs, le temps passé sur les tâches et les tendances d'utilisation, tandis que les mesures subjectives ont trait à la fiabilité du système, à son utilité, à son dogmatisme, à la manière adéquate ou non dont il justifie ses conclusions, etc. Comme le système expert a autant besoin de l'acceptation de l'utilisateur que de performance pour réussir, il faut en faire une évaluation à la fois formative et sommative. Actuellement, il reste, semble-t-il, beaucoup de recherche à faire pour développer les mesures quantitatives et qualitatives qui s'appliquent aux systèmes experts. Une méthode systématique de saisir et d'intégrer la rétroaction diagnostique et prescriptive de l'utilisateur doit être mise au point pour l'évaluation des systèmes experts. Une réalisation intéressante à ce

propos est EVA (Expert System Validation Associate), qui est un système de validation automatique des systèmes experts [Stachowitz et Combs, 1987].

5. ANALYSE ET TRAVAUX FUTURS

5.1. Avantages démontrés

QUARTZ a déjà démontré son utilité à sauvegarder l'expertise de l'entreprise, puisque plusieurs des experts de la première heure ne travaillent plus à l'évaluation des projets EDER. Leur expertise particulière ainsi que leur perspective générale sur l'analyse des propositions font aujourd'hui partie de la base de connaissances de QUARTZ. Aussi, QUARTZ a servi avec succès d'étude de faisabilité de l'utilisation des systèmes experts pour surveiller le cycle de vie d'un projet, de même que, dans cette recherche, d'un système expert devant aider les analystes à l'analyse et la sélection des projets admissibles aux programmes de subvention du gouvernement.

Un certain nombre de possibilités semblent s'ouvrir au système QUARTZ. Celui-ci peut être modifié et devenir un système expert pour l'analyse des projets soumis au programme EDER du Manitoba, par exemple, ou pour l'analyse interne à notre propre centre de recherche. Il est aussi concevable que QUARTZ puisse être utilisé directement par les entreprises dans la préparation de leurs soumissions. Cette possibilité s'inscrirait dans le sens des politiques gouvernementales d'accès du public à l'information.

Finalement, la base de connaissances de QUARTZ aide à produire des analyses plus uniformes et plus exhaustives et des recommandations qui sont basées sur une analyse plus systématique de tous les critères qualitatifs et quantitatifs utiles. Cette expertise formalisée peut aussi être mise à la disposition des analystes débutants comme aide ou outil de formation. Le groupe d'experts de Québec a accepté de poursuivre l'étude de cette application possible de QUARTZ. Le principal avantage tant attendu du système QUARTZ est d'améliorer de façon certaine l'efficience et la qualité de la rétroaction dont peuvent bénéficier les entreprises intéressées.

Les avantages évoqués ici seront évalués au cours du prochain trimestre, selon les méthodologies décrites dans ce rapport.

5.2 Projets de travaux

Les recherches effectuées, ce dernier trimestre, ont servi à formuler une base théorique solide pour l'évaluation formative des prototypes QUARTZ. L'approche systémique de l'évaluation - ainsi que le point de vue qu'un système expert doit être viable - semble être une méthode pragmatique justifiée pour le développement et la validation des systèmes experts. L'évaluation du système QUARTZ se fera ainsi selon l'option proposée d'intégrer le processus d'évaluation dans la conception du système et avec les critères quantitatifs et qualitatifs déterminés à l'appendice 1. Dès qu'un prototype aura été accepté par l'utilisateur, une évaluation sommative sera réalisée sur cette version "définitive" de QUARTZ.

Les leçons tirées de cette étude des méthodes d'évaluation indiquent que QUARTZ devrait être testé selon quatre niveaux différents au moins. Ces niveaux sont, dans l'ordre croissant de généralité :

1. base de connaissances de faits et règles; mécanisme d'inférence; interface utilisateur
2. le système expert en entier - ses entrées, sa capacité de traitement, ses sorties
3. sous-système d'analyse du programme EDER - efficience, qualité
4. le processus décisionnel EDER en entier.

Parmi les jalons posés, nous comptons atteindre les réalisations suivantes au prochain trimestre:

- la validation de la base de connaissances à l'aide des premiers experts (en autres, d'un expert de notre propre centre de recherche)
- l'évaluation de la performance du système, du point de vue des utilisateurs (questionnaire et interviews)
- la métrologie du système afin d'estimer le travail de développement requis.

La prochaine étape consiste à dresser un plan d'évaluation pour QUARTZ. Quelques-uns des repères principaux prévus pour les prochains mois seront les essais de validation et de fiabilité du questionnaire de rétroaction, l'incorporation des suggestions des utilisateurs et des corrections et la livraison du deuxième prototype opérationnel.

6. RÉFÉRENCES

Adelman, L. (1986). On the basis of what criteria should decision support systems be evaluated? Proceedings Expert Systems in Government Symposium, Washington, D.C. pp. 348-350.

Bell, M. (1985). Why Expert Systems Fail. Journal of Operational Research Society, 36, 7, 613-619.

Borg, W. and Gall, M. (1983). Educational Research (4th ed.) New York : Longman, Inc.

Beer, S. (1979). The Heart of Entreprise. Devon : Pitman Press.

Brixius, L. (1984). Artificial Intelligence Applications in Civil Government. Proceedings IEEE EASCON'84 conference. pp. 45-49.

Dalkir, K. et Muzard, J. (1987) QUARTZ : système expert d'aide à l'analyse et à l'évaluation des projets. Actes de la conférence IEEE Montech COMPIN'87. Montréal, Québec. pages 208 à 210.

Defense Systems, Inc. The Decision Aid Measurement and Evaluation (DAME) Methodology Assessment of the Integrated Tactical Decision Aid (ITDA). 29 Apr. 1985.

Gallant, S. and Balachandra, R. (1986). Using Automated Techniques to Generate an Expert System for R & D Project Monitoring. IFAC conference on Economics and Artificial Intelligence, Aix-en-Provence, France. pp. 61-66.

Giezl, L. (1987). The Expert System Applicability Question. Proceedings. Artificial Intelligence and Simulation conference, San Diego, CA. 17-20.

Gordon S. (1988). The Human Factor in Expert Systems. AI Expert, January, 1988, pp. 55-59.

Hofmeister, A. (1986). Formative Evaluation in the Development and Validation of Expert Systems in Education. Comput. Intell., 2, 65-67.

Kaisler, S. (1985). Expert System Metrics. Proceedings Expert Systems in Government Symposiun, Washington, D.C. pp. 114-119.

Krcmar, H. (1985). Enterprise-wide Information Management: Expert Systems for Information Management. IBM LA Scientific Centre, Report G320-2767.

Lane, N. (1986). Global Issues in the Evaluation of Expert Systems. Proceedings IEEE Conference on Systems, Man and Cybernetics, Alexandria, VA. pp. 121-125.

Leblanc, L. (1987). An Analysis of Critical Success Factors for a Public Sector Decision Support System. Evaluation Review, 11, 1, 73-83.

Liebowitz, J. (1985). Evaluation of Expert Systems: an approach and case study. Proceedings 2nd Annual Conference on Artificial Intelligence Applications: the Engineering of Knowledge-based Systems. Miami Beach, Florida, December 1985.

NITAS (1986). A Reasoning System for the Guidance of Technological Transfer. IFAC Conference on Economics and Artificial Intelligence, Aix-en-Provence, France. pp. 183-186.

Soloway, E. and Littman, D. (1987). Expert System Evaluation. Tutorial, IJCA187 conference, Milan, Italy.

Stachowitz, R. and Combs, J. (1987). Validation of Expert Systems. Proceedings of the 20th Annual Hawaii International Conference on System Sciences, 1987

7. APPENDICE I. HIÉRARCHIE DES CRITÈRES D'ÉVALUATION DE LA THÉORIE MULTI-ATTRIBUT DE L'UTILITÉ (MAUT)

[Tiré de Adelman, L. (1986). On the Basis of what Criteria Should Decision Support Systems Be Evaluated? Procs. ESIG Conference.]

0 utilité globale (6)

1. interface système/utilisateur

1. rapport avec le personnel
   1. formation et expérience technique (3)
   2. style de travail, charge de travail et intérêt (4)
   3. besoins opérationnels (5)
2. caractéristiques du système
   1. caractéristiques générales
      - facilité d'emploi (4)
      - compréhension des processus du système (3)
      - facilité d'initiation (2)
      - temps de réponse (2)
   2. caractéristiques particulières
      - interface utilisateur (2)
      - intégrité des fichiers de données (3)
      - précision des jugements experts (2)
      - aptitude à modifier les jugements (2)
      - compréhension des algorithmes du système (2)
      - utilité des graphes (2)
      - utilité des imprimés (2)
      - compréhension du texte (2)

2. interface organisation/utilisateur-système

1. facteurs d'efficience
   1. acceptabilité des temps pour :
      - l'accomplissement des tâches (3)
      - la gestion des données (2)
      - la configuration du système (2)
   2. fiabilité perçue dans des conditions réelles (2)

disponibilité des compétences (2)
disponibilité du matériel (2)

2. rapport avec les facteurs organisationnels
   1. effet sur la structure et les procédures de l'organisation (2)
   2 effet sur les postes d'autres personnes dans l'organisation
      acceptabilité politique (2) charge de travail d'autres employés (2)
   3. effet sur le flux d'information (2)
   4. répercussions indirectes
      valeur dans l'exécution d'autres tâches (2)
      valeur pour les organisations affiliées (2)
      valeur pour la formation (2)

3. interface organisation/environnement

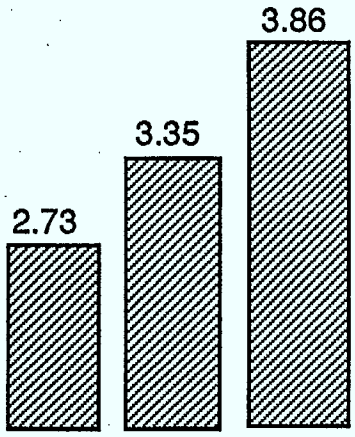
1. précision des décisions (8)
2. rapport entre l'approche technique du système et les aspects du problème(7)
3. qualité du processus de décision (2)
   1. qualité du plan d'incorporation des jugements (2)
   2. nombre de solutions (2)
   3. comparaison des différentes solutions (2)
   4. évaluation des conséquences des différentes solution(3)
   5. réexamen du processus décisionnel (3)
   6. utilisation de l'information (3)
   7. considération d'un plan d'implantation et de mesure d'urgence (2)
   8. effet sur les discussions de groupe (3)
   9. effet sur la confiance du décideur (2)

Note : les chiffres entre parenthèses indiquent le nombre de questions utilisées dans le questionnaire pour évaluer le critère.

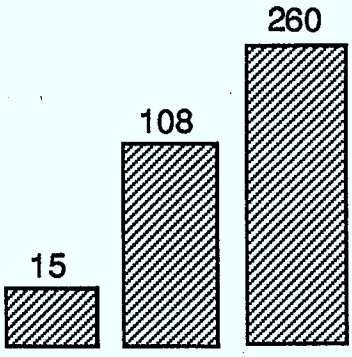
7.1 APPENDICE II METROLOGIE DE LA BASE DE CONNAISSANCE DE QUARTZ

Date	Nb d'objets	Nb de schémas	Complexité	Nb de règles
FÉVRIER 87	83	30	2.73	15
MARS 87	58	17	3.35	108
MAI 87	82	21	3.86	260

COMPLEXITÉ STRUCTURELLE



NOMBRE DE REGLES



QUEEN QA 76.76 .E95 D3514 19
Dalkir, Kim
Approche systémique de l'é

INDUSTRY CANADA/INDUSTRIE CANADA
97015

DATE DUE
DATE DE RETOUR

CARR McLEAN 38-296

Pour plus de détails, veuillez communiquer avec :

Le Centre canadien de recherche sur l'informatisation du travail
1575, boulevard Chomedey
Laval (Québec)
H7V 2X2
(514) 682-3400

For more information, please contact:

Canadian Workplace Automation Research Centre
1575 Chomedey Blvd.
Laval, Quebec
H7V 2X2
(514) 682-3400