CONCLUSION

8.1 Résultats

Dans cette thèse, nous avons mis au point une méthode opérationnelle d'acquisition des connaissances pour les domaines biologiques. Cette méthode est constituée d'une chaîne en trois parties : acquisition de l'observable avec élaboration d'un modèle descriptif, acquisition de l'observé (les cas) à l'aide d'un questionnaire interactif, puis traitement de ces connaissances (observable et observé) à des fins de classification et/ou de détermination. Les outils permettant de créer le modèle et le questionnaire ont été conçus pendant cette thèse (HyperQuest), ainsi que le module de raisonnement par cas (CaseWork) pour l'objectif de détermination. Pour la classification, nous sommes partis de travaux sur le logiciel KATE [Manago, 1991].

Afin de mettre au point notre méthode, nous nous sommes appuyés sur une application concrète au Muséum National d'Histoire Naturelle de Paris et sur la disponibilité d'un expert du domaine des éponges marines.

Au départ de ce travail, notre objectif était d'obtenir des résultats de consultation robustes face à un utilisateur donnant des réponses «inconnu» aux questions posées par le système expert pour déterminer un nouvel individu. Une méthode de raisonnement par cas, expliquée au chapitre 7, permet de pallier ce type de “bruit” dans la phase de détermination.

Mais nous savions aussi par d’autres expériences menées à l’INRA en pathologie végétale que la robustesse de la consultation dépendait de la qualité des descriptions, c’est-à-dire de la capacité de l’utilisateur à “savoir décrire” à l’aide d’un questionnaire. De même, cette exigence de qualité des descriptions est primordiale pour pouvoir construire des classifications artificielles à partir des exemples.

Or, avant de “savoir décrire”, il faut “savoir observer” : le questionnaire devait donc avoir le rôle de guide d’observation afin d’obtenir des descriptions robustes. La conception d’un guide demande la formalisation d’un bon modèle de description sur lequel on peut ensuite bâtir un questionnaire.

Nous avons alors plutôt accentué notre effort sur la partie “modélisation” des connaissances implicites de l’expert, c’est-à-dire l’observable en amont de la phase de traitement : il s'agit non pas de modéliser le raisonnement de l'expert, mais plutôt son “savoir observer”.

Nous avons donc conçu HyperQuest pour donner la possibilité à l'expert d'expliciter son propre modèle d’observation. Les connaissances de bon sens lui apparaissent alors sous forme graphique et structurée et donnent une vision réelle des trois dimensions des descripteurs : objets, attributs et valeurs.

Avant de constituer un modèle descriptif, l'expert n'est pas toujours conscient de sa propre manière d'observer. Concrétiser un modèle d'observation sur un écran d’ordinateur lui renvoie l’image présente de ses connaissances sur son domaine.

Cette matérialisation prend deux formes :

1) La première, liée à l’observable, lui montre les relations qu’entretiennent les objets entre eux dans des arbres de composition et de spécialisation : c’est une vue globale de son propre modèle de description qu'il ne faut pas confondre avec l'arbre de décision issu de la classification. Nous avons pu dégager ainsi un certain nombre de mécanismes d'observation que l'on retrouve dans la littérature en systématique (chapitre 4) et qui constituent la trame d'un véritable guide de description.

   L'outil permettant de créer et de modifier interactivement ce modèle descriptif ainsi que de le visualiser graphiquement a été développé à partir de la découverte de ces mécanismes.

2) L’autre, liée à l’observé, fait plonger l’expert au niveau des descriptions individuelles grâce au questionnaire instanciant son modèle d’observation. Nous avons montré l'importance de reproduire des descriptions naturelles, c'est-à-dire fondées sur des spécimens et non pas sur des concepts. L'objectif au Muséum est de multiplier le nombre de descriptions par classe pour exprimer sa diversité plutôt que de favoriser les regroupements de descriptions au sein d'une seule définition de concept (ce qui débouche sur des choses non observables, trop larges par rapport à la réalité). Cela permet de plus de valoriser les collections, en déléguant le travail de généralisation des descriptions à un outil d'induction, puis de comparer les résultats avec ceux d'une classification naturelle établie par l'expert.

   Pour acquérir l'observé, nous avons construit un générateur de questionnaire interactif multimédia dont l'intérêt est d’automatiser la fabrication de questionnaires à partir d’un modèle de l'observable tout en tenant compte des capacités d'observation des utilisateurs.

Le questionnaire généré est personnalisable par l'expert et adopte le dialogue structuré de son modèle descriptif (l'ordre des objets). Pour l'objectif de détermination, il est utile de faire participer d'autres utilisateurs au remplissage de la base de cas à apprendre, le classement étant toutefois du rôle de l'expert. En effet, la variabilité des manières d'observer et de comprendre le vocabulaire spécialisé est un obstacle supplémentaire à de bonnes déterminations. Les descriptions restent comparables entre elles puisqu'elles suivent le même schéma, et il est préférable de les intégrer dans la même base de cas même si elles proviennent d'utilisateurs hétérogènes.

Pour l'expert, ce travail répétitif de description peut sembler routinier et peu valorisant comparé à la tâche de classification. Néanmoins, décrire fait partie du travail quotidien du systématicien ; cela est nécessaire pour classifier s'il veut accentuer sa familiarité avec ses objets d'étude, ce qui l'amène un jour à découvrir certains caractères de différenciation des spécimens et émettre des hypothèses sur les classes : ainsi, l'observation et la description peuvent conduire à la découverte en révélant certaines régularités qu'il faudra par la suite mettre à l'épreuve de nouveaux faits. C'est ainsi qu'il applique la méthode scientifique : **conjecturer et tester** [Pólya, 1967] que nous pouvons interpréter en biologie par le schéma suivant (figure 8.1) :

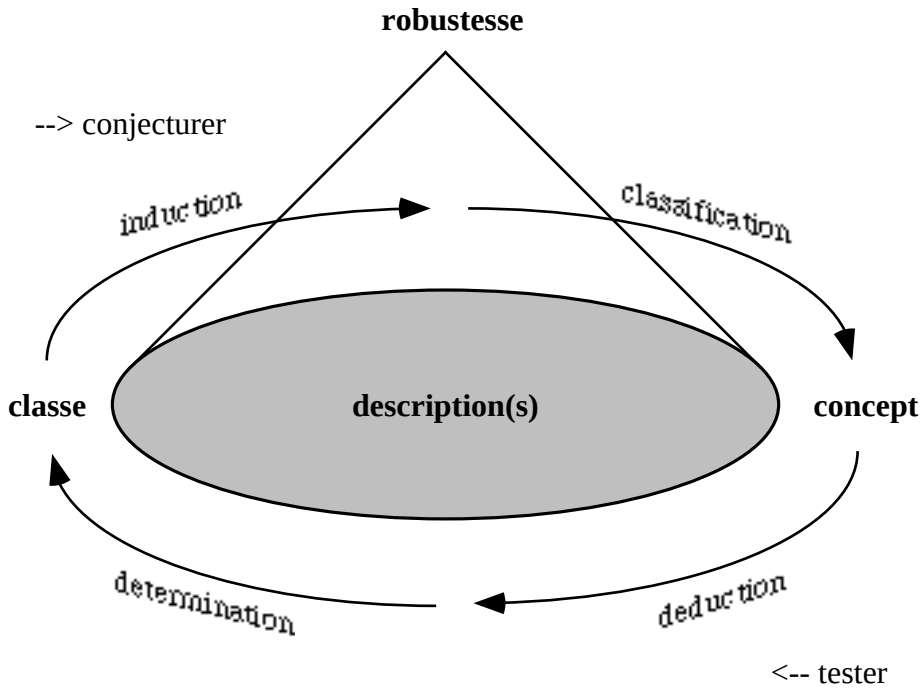

Fig. 8.1 : Conjecturer et Tester

Les tests peuvent revêtir deux formes :

1) la détermination de nouveaux faits par l'utilisation de l'arbre de classification, en utilisant la définition en intension associée au concept.
2) l'identification de nouvelles observations par comparaison avec des descriptions de spécimens représentant l'extension associée à la classe.

C'est par un aller et retour entre l'approche inductive et déductive que l'expert sera capable de valider les connaissances apprises dans le but d'affiner les règles caractérisant ses descriptions.

Le biologiste adopte naturellement la démarche inductive alors que le mathématicien habitué aux démonstrations raisonne plutôt à l'inverse à l'aide de la déduction. Le rôle de l'informaticien est de faire travailler ce système en procurant les outils de mise en œuvre de la méthode scientifique en biologie. L'amélioration de la robustesse tient alors à la capacité de l'informaticien de faire communiquer deux démarches : la première expérimentale (celle des biologistes) et la seconde fondée sur un raisonnement académique (mathématiciens).

Néanmoins, cette méthode de validation après le traitement est nécessaire mais pas suffisante : elle permet principalement de valider l'observé (les descriptions) plus que la validation de l'observable (voir figure 2.4).

Nous nous sommes en effet aperçu qu'une partie non négligeable de l'élaboration du modèle descriptif a lieu en amont de la phase d'induction au moment de l’acquisition des exemples. Par un processus de validation croisée du modèle par des descriptions, au fur et à mesure qu'il décrit, l'expert va penser à des descriptions plus proches de la réalité observée (les exceptions qui traduisent les extrèmes de la variabilité intra-spécifique).

Par exemple, le questionnaire n’oublie jamais de demander une confirmation sur la pertinence de certains caractères non décrits, mais qui devraient l'être pour se conformer au modèle descriptif. Cela oblige l’expert à fournir des descriptions cohérentes et exhaustives, sinon il est amené à modifier son opinion sur son propre modèle de description. Il va alors chercher à l’affiner et à répercuter ses observations dans le modèle descriptif, puis dans les exemples.

8.2 Limites actuelles

Notre méthode demande que le modèle descriptif soit complet par rapport à un domaine bien délimité. L'exhaustivité de l'observable est une exigence théorique très difficile pour l'expert : elle est néanmoins recherchée pour ne pas devoir changer en profondeur la structure du modèle descriptif, ce qui aura pour conséquence de devoir modifier les anciens cas “à la main”.

En effet, nous n'avons pas encore conçu les outils de maintien de la cohérence de l'ancienne base de cas par rapport aux changements effectués dans un nouveau modèle descriptif (élimination d'objets, d'attributs ou de valeurs possibles, rajout d'objets, changement dans la structure de description, etc.). Cette phase de mise à jour des données par rapport à un modèle de l'observable est une des perspectives à prendre en compte dans une prochaine étape pour la robustesse du système global : il n'est pas possible de tout prévoir dès le départ dans le modèle.

Pour ce même modèle, nous n'avons pas non plus conçu l'éditeur permettant de renseigner les règles contextuelles entre les objets et les attributs observables : par exemple, l'expert ne peut pas indiquer le fait que, lorsque le nombre des orifices de la face exhalante est unique, alors ce n'est pas la peine de répondre aux attributs “répartition” et “localisation” des orifices.

Au niveau du traitement des descriptions, nous n'avons pas encore pu mesurer sur notre application l'intérêt d'intégrer les approches inductive et analogique pour “savoir raisonner” à des fins de classification et de détermination en biologie. Cette intégration est l'objet du projet INRECA en cours dont l'annexe 5 donne un aperçu. Plus spécifiquement, nous souhaiterions associer une sémantique au niveau du critère de séparation des classes pour ne pas tenir compte uniquement de son efficacité de discrimination inter-classe : ceci se comprend bien pour la détermination où il faut arriver rapidement à une conclusion mais pas forcément pour la classification : le critère mono dimensionnel du gain d'information est pauvre et peu significatif surtout lorqu'il reste peu d'exemples à comparer. De plus, un choix arbitraire est effectué lorsque deux critères ont un pouvoir de discrimination identique. Il serait bon de faire intervenir d'autres paramètres d’un niveau plus sémantique que la seule entropie dans la mesure (méta-connaissance sur les objets prioritaires par rapport aux autres, facteurs de tolérance aux bruits, etc.).

De même, notre outil d'induction comporte certains biais dans sa manière d'élaborer une classification artificielle. Certains attributs ont un pouvoir de discrimination intrinsèque plus important du fait du nombre de valeurs possibles qu'ils possèdent : la forme du corps de l'éponge contient 17 valeurs lorsqu'elle est traitée sans considérer son type classifié, alors qu'elle ne devrait en compter que 5 en tenant compte de la taxonomie introduite par l'expert

(c'est-à-dire les cinq nœuds intermédiaires). KATE transforme aussi des disjonctions d'imprécision dans les exemples en conjonction de variation au moment de la détermination d'une nouvelle observation. On considère ici la variation comme une forme d'imprécision, ce qui justifie un traitement identique des exemples. De plus, le traitement des intervalles pour les attributs numériques n'est pas optimal quant au choix des seuils : le lecteur peut se référer aux travaux de [Fayyad & Irani, 1993]. Il serait donc utile d'étudier d'autres possibilités de discrétisation que celle de la binarisation de l'attribut dans KATE.

KATE et CaseWork ont été mis à l'épreuve sur d'autres applications non biologiques (attribution de crédits bancaires, aide à la photo-interprétation, diagnostic de pannes, etc.). Dans celles-ci, les connaissances pouvaient se réduire à un tableau de données classique. Dans notre application, KATE doit être capable de traiter les objets multi-instanciés correspondant aux objets *horde* formalisés par [Diday, 1987] et repris par [Conruyt *et al.*, 1992] sous l'appellation *horde composite*. Cela signifie de savoir gérer des appariements multiples entre descriptions pour respecter l'homologie des objets et non pas seulement une unification directe entre deux objets de même nom appartenant à des descriptions différentes : les travaux de [Perinet-Marquet, 1993] sur les structures itératives sont un début de recherche dans ce sens.

Enfin, il reste aussi la limite suivante : nos outils d'acquisition de l'observable et de l'observé ont été testés à partir d'un modèle de description issu d'une seule application (Hyalonema). Il faudrait étudier d'autres classes zoologiques pour expérimenter les logiciels et montrer ainsi le bien fondé de notre méthode d'acquisition de connaissances descriptives pour aider les systématiciens dans leurs recherches. Si KATE et CaseWork sont déjà commercialisés par AcknoSoft, le logiciel HyperQuest a quant à lui atteint un niveau de prototype avancé avec une documentation associée [Conruyt & Dumont, 1993].

8.3 Perspectives

L'expérience nous montre que la robustesse n'est pas simplement un résultat lié au traitement des données, qui s'arrêtera à la validation des connaissances apprises. C'est pour nous un processus incrémental qui s'inscrit dans la continuité, en appliquant la méthode hypothético-déductive sur un même domaine d'expertise, de manière itérative. Les domaines naturels sont incomplets par nature car ils sont caractérisés par une grande variabilité (multiples exceptions), une évolution des phénomènes à décrire (ex : maladies) et des techniques d'observation de plus en plus précises (cytologie, biochimie, ADN...). Il est alors difficilement concevable de modéliser "tout" l'observable à un moment donné : le modèle descriptif est une photographie qui

reflète le domaine de discours et les connaissances instantannées de l'expert : cela évolue nécessairement.

La validation des connaissances apprises (règles, arbre de décision) n'est pas ainsi seulement un processus post-opératoire sur les données comme nous pouvions le penser avant cette thèse. La qualité d'une classification artificielle est dépendante de la précision et de l'exhaustivité des descriptions fournies. En introduisant des connaissances "de fond" (le modèle descriptif), il s'agit pour l'expert de valider l'expérience acquise mais non toujours explicite (les "savoir observer" et "savoir décrire") avant d'appliquer un raisonnement. Cette caractéristique est à prendre en compte pour les perspectives de développement d'outils d'aide à la validation de ce savoir en phase d'acquisition des exemples. N'oublions pas que le temps consacré à cette phase est de loin le plus important dans la méthode d'apprentissage utilisée.

Dans l'avenir, le rôle de l'informaticien désireux de concrétiser son travail de recherche sur l'acquisition des connaissances ne se bornera pas à fournir des outils de traitement des données ("classez, nous classifierons ensuite !"). Il lui faudra assumer un rôle de cogniticien, prêt à s'investir avec la curiosité nécessaire pour comprendre les difficultés inhérentes au domaine étudié. Il est préférable qu'il parte d'ailleurs de problèmes concrets à résoudre et qui sont posés par l'expert (par exemple, celui de traiter le biais introduit par la quantité d'information des attributs classifiés dont on ne considère que les feuilles de la taxonomie des valeurs possibles). C'est une démarche coopérative et pluridisciplinaire qui doit partir des travaux existants pour améliorer la robustesse des systèmes d'aide à la classification et à la détermination en biologie.

Cette amélioration passe par la revalorisation de la notion de description dont le schéma 8.1 montre le rôle central. Elle doit exprimer toute la richesse du domaine naturel et refléter l'état des connaissances de l'expert à un moment donné. Il ne suffit pas de savoir représenter des connaissances à l'aide d'un langage à objets pour obtenir un système de détermination robuste. Il faut pouvoir expliciter correctement la connaissance de l'expert en facilitant sa structuration (facteurs de compréhension et de précision), apprécier sa diversité (exhaustivité et redondance), et connaître sa sémantique pour les autres utilisateurs de son système (compréhension, ergonomie et tolérance aux bruits).

Le progrès technologique des ordinateurs permet de reconsidérer des pratiques anciennes considérées comme utopiques à l'époque d'Adanson : les descriptions de spécimens. Ces dernières sont compatibles avec les capacités de stockage des machines actuelles, ce qui permet de conserver le maximum d'information par rapport à des "descriptions" de concepts. Posons-nous donc la question de savoir ce que sont les véritables qualités d'une donnée après le

travail énorme réalisé dans le domaine de leur analyse ! L'expert devra disposer d'outils permettant de développer sa familiarité avec les spécimens. La transmission de son savoir par un système expert de détermination passe alors par une valorisation de son expérience. Celle-ci pourra s’acquérir à l'aide d'outils de modélisation de son domaine pour acquérir des descriptions robustes, puis de mise à l'épreuve de ses opinions par la construction de classifications artificielles.

Ayant toujours comme référence le modèle observable et disposant intégralement des exemples issus du modèle, ceux qui auront à utiliser ces classifications profiteront de toute la connaissance explicitée à un moment donné. Cela devrait permettre d'éviter de raisonner à partir de connaissances comprises hors de leur contexte, non maîtrisées ou trop abstraites, puisque la source même de ces connaissances aura été préservée.

L’amélioration de la robustesse des systèmes d’aide à la description, à la classification et à la détermination des objets biologiques est donc le préambule à l’élaboration d’outils de Taxonomie Assistée par Ordinateur plus performants.