DETERMINATION ASSISTEE PAR ORDINATEUR : BASE DE DONNEES, CLE INFORMATISEE OU SYSTEME EXPERT ?

(L. TITO DE MORAIS)

RESUME - Depuis octobre 1988 un groupe informel de travail (1) sur les applications de l'intelligence artificielle à l'aide à la détermination des poissons d'eau douce, s'est réuni régulièrement. L'objectif à court terme est la réalisation d'un système expert d'aide à la détermination des Characidés d'Afrique de l'Ouest. Le travail se déroule en cinq étapes :

- analyse de la démarche de l'expert lors d'une diagnose ;
- "*systématisation*" de la démarche et réalisation de la base de connaissances ;
- mise en place du système expert (écriture des blocs de contrôle et des règles de production, tests, etc.) ;
- mise en place des routines externes de calcul (traitement numérique des données métriques et méristiques) ;
- amélioration de l'interface utilisateur et tests finaux.

Le travail est actuellement dans la troisième phase. Il a permis de tirer certaines conclusions relatives :

- à la démarche de l'expert. Celui-ci procède par "*analyse d'images*", i.e. la mémorisation de certaines espèces et par analogie pour les espèces voisines. En cas d'insuccès, ou pour confirmation, il y a utilisation d'une clé de détermination (mémorisée ou consultée). Les données méristiques (nombre de branchiospines, rayons, dents, etc.) ont alors une grande importance ;
- à la réalisation de la base de connaissances et à l'importance des données écologiques et biogéographiques pour l'orientation initiale ou la confirmation de la diagnose ;
- à la mise en place du système expert. Si la réalisation d'une "*clé informatisée*" permet déjà la mise en évidence d'ambiguïtés, voire d'erreurs, dans les ouvrages de détermination, il importe de dépasser largement ce stade en faisant appel à des routines externes de calcul et à des données bioécologiques régulièrement actualisées (figure 2).

Les mérites comparés, pour ce type d'application, des systèmes à base de règles de production de type ESE d'IBM et "*centrés-objet*" sont discutés.

En conclusion, et au delà du simple cas d'espèce qui fait l'objet du travail actuel, l'intérêt et les contraintes de la réalisation de telles bases de connaissances et de systèmes experts "*ad hoc*" sont discutés.

(1) Ce groupe de travail était constitué par J.F. MURAIL du Muséum National d'Histoire Naturelle, MNHN, et de J. CRUETTE, D. PAUGY et L. TITO DE MORAIS de l'ORSTOM.

Dans certains cas précis, avec la baisse des coûts du matériel et leur "*convivialité*" accrue, la réalisation d'un système expert peut s'avérer intéressante pour fournir à la contrepartie locale un outil de formation et de travail, dans le cadre de programmes censés se poursuivre après le retrait de l'ORSTOM. (Programmes de contrôle biologique à long terme en particulier).

Au niveau des chercheurs, la réalisation de bases de données individuelles est la règle. Une politique de relative standardisation et de mise en place de bases de connaissances régionales, gérées par des systèmes experts, augmenterait les contacts et favoriserait l'échange d'informations au sein des unités de recherches et entre elles. Elle fournirait en outre :

- des outils de formation des jeunes chercheurs et techniciens ;
- des outils de travail pour les chercheurs de spécialités voisines, ou pour des ichtyologues qui changent d'aire géographique d'étude ;
- les maillons d'un réseau de bases de connaissances internationales.

La mise en place de bases de connaissances impliquant plusieurs chercheurs (pour la réalisation comme pour l'alimentation et la mise à jour périodique) demande un engagement des commissions scientifiques autant dans le cadre de leurs attributions relatives à l'animation scientifique et au développement des échanges, que pour la valorisation du travail fourni par les chercheurs pour la réalisation des systèmes.

INTRODUCTION

Dans le cadre d'un groupe de travail ORSTOM-MNHN, nous réalisons une application de l'intelligence artificielle à l'aide à la détermination de quelques poissons d'eau douce : les Characidés de l'Afrique de l'Ouest.

Nous avons délaissé pour l'instant les Characidés nains (15 genres et plus de 60 espèces), dont la systématique pose actuellement un certain nombre de problèmes non résolus au plan fondamental. La famille se décompose donc en 4 genres (Tableau 1) comprenant 42 espèces.

CHARACIDES D'AFRIQUE				
Genres :				
Alestes	Brycinus	Hydrocynus	Lepidarchus	Nains
5 espèces	31 espèces	5 espèces	1 espèce	env. 60 esp.

Tableau 1 : Genres et nombre d'espèces composant les poissons Characidés d'Afrique.

Nous avons utilisé le générateur de systèmes experts ESE d'IBM.

Le travail se déroule en cinq étapes :

(1) Analyse de la démarche de l'expert lors d'une diagnose. (2) "Systématisation" de la démarche et réalisation de la base de connaissances. (3) Mise en place du système expert : écriture des blocs de contrôle et des règles de production, tests, etc. (4) Mise en place des routines externes de calcul (traitement numérique des données métriques et méristiques). (5) Amélioration de l'interface utilisateur et tests finaux. Le travail est actuellement dans la phase (3).

L'expert biologiste procède très souvent par "*analyse d'images*", i.e. par mémorisation de certaines espèces et par analogie pour les espèces voisines. En cas d'insuccès, ou pour confirmation, il y a utilisation d'une clé de détermination (mémorisée ou consultée). Les données méristiques (nombre de branchiospines, rayons, dents, etc.) ont alors une grande importance. Au cours de la diagnose, l'expert s'éloigne souvent de la stricte phylogénie. Il fait appel à des caractères qui n'ont que peu ou pas de valeur phylogénétique, mais qui à un niveau donné de la détermination peuvent être d'un grand intérêt pratique. Par exemple, des poissons de plusieurs espèces très différentes peuvent avoir le même nombre de branchiospines, cependant, à un niveau bas de la hiérarchie, ce même caractère peut permettre de séparer deux espèces très voisines.

Il y a en effet deux aspects dans la détermination (avec des possibilités d'imbrication multiples) :

- une approche phylogénétique regroupant les espèces suivant des critères systématiques (en genres, familles, ordres, etc.). Elle fait très souvent appel à des caractères difficilement observables (internes, microscopiques, voire ontogénétiques observables seulement à certains stades du développement) ;
- une approche pratique, faisant appel à des critères plus facilement observables (mais pas toujours !). Les regroupements effectués ne correspondent alors pas nécessairement aux groupes zoologiques.

Woolley et Stone ont publié en 1987 un article comportant une brève discussion sur les divers systèmes d'identification taxonomique : tableaux à entrées multiples, clés dichotomiques "*papier*", clés et tableaux informatisés. Ils présentaient également un système expert fondé sur des règles de production (fonctionnant en chaînage arrière) adapté à la détermination des insectes d'une famille d'Hyménoptères. Depuis cette date, d'autres approches ont été développées. Et notamment en France à l'Université Claude Bernard de Lyon et à l'INRIA à Grenoble où sont étudiés des systèmes "*centrés-objet*" (Shirka).

Les tableaux à entrées multiples sont d'une gestion extrêmement complexe sur le papier. Dans leur version informatisée, chaque individu est constitué par un ou plusieurs enregistrements comportant les variables descriptives (figure 1). Leur exploitation par un système de gestion de bases de données se fait de manière non optimale et, si le nombre d'individus est élevé, il faut un très grand nombre d'étapes pour parvenir à une diagnose. Le système est très sensible aux erreurs et ne peut pas fournir des résultats pondérés par un coefficient d'incertitude. Il est en revanche facile à mettre à jour et à modifier.

Les clés dichotomiques (papier et "*automatiques*") sont très complexes à écrire. Toute modification ultérieure conduit très souvent à une refonte globale de

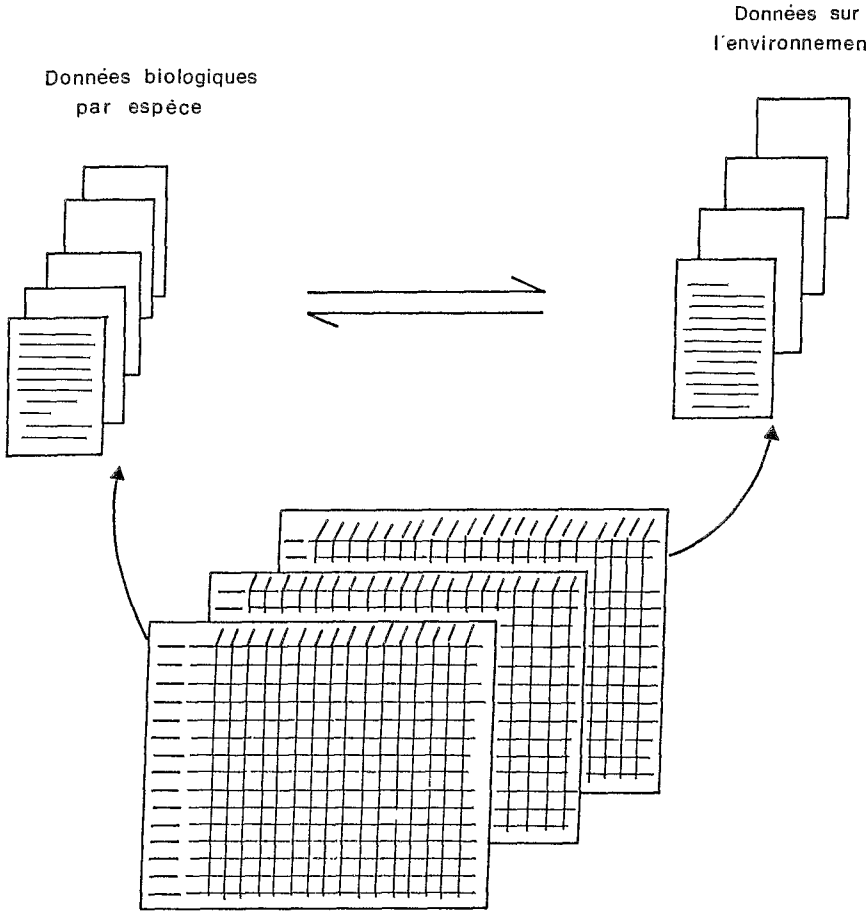

Fig.1: Tableaux à entrées multiples et bases de données (Voir texte).

la clé. Une version informatisée est d'utilisation plus aisée. Mais dans tous les cas, elles sont extrêmement sensibles aux erreurs de réponse, elles ne permettent pas les retours en arrière (sauf reprise depuis le début) et n'acceptent pas les coefficients d'incertitude ni l'absence de réponse. Leur avantage étant qu'il en existe un grand nombre sur papier, qui ont été éprouvées et améliorées au fil des années.

Les systèmes "*centrés-objet*" sont ceux dont la "philosophie" et la structure se rapprochent le plus des critères de la systématique, mais pas nécessairement de la démarche de l'expert réalisant une identification. Les objets et leurs regroupements peuvent être reliés aux notions d'espèce, de genre, de famille, etc.. Leur classification hiérarchique est voisine des structures arborescentes de la classification. La création de la base demande en général une réorganisation de la clé de détermination utilisée, mais les modifications et les mises à jour sont aisées.

Les systèmes à base de règles de production sont, dans leurs versions initiales, des clés informatisées optimisées. Déjà dans ces premiers stades, ils apportent cependant d'importantes améliorations par rapport aux "*simples*" clés (détection d'erreurs, coefficient d'incertitude, meilleure gestion des réponses multiples et des réponses manquantes, historique et justification de la diagnose en fin de parcours, etc.).

Ce type de système est celui dont la conception et l'utilisation se rapproche le plus de la démarche "*réelle*" d'un expert effectuant une diagnose. Comme lui, il est essentiel que le système puisse faire appel à des connaissances autres que celles d'une clé dichotomique. C'est à dire à des bases de données bio-écologiques, phylogénétiques et à des routines externes de calcul (figure 2). Ces dernières permettent notamment de replacer statistiquement les valeurs métriques et méristiques de l'individu étudié, au sein des intervalles ou des clines de valeurs observées pour les différentes espèces dans différentes régions. *In fine*, le résultat de la consultation doit fournir outre le nom de l'espèce, l'ensemble des informations y afférentes et contenues dans les différentes bases. La modification et la mise à jour des bases annexes sont aisées, celles de la base centrale et des règles de production, proche de la clé dichotomique, reste difficile, car certains changements peuvent altérer la structure même de la base et exiger sa réécriture partielle.

1. DETAIL DU SYSTEME EXPERT DEJA REALISE

Le système est organisé en blocs de contrôle de trois niveaux hiérarchiques (Figure 3).

Le premier bloc est prévu pour tester les intentions de l'utilisateur. Non véritablement opérationnel pour l'instant, il pourra, entre autres, permettre de choisir entre le chaînage avant ou arrière.

Le bloc "*clé-groupe*" permet de déterminer la "*variable-but*" intermédiaire GROUPE-POISSON, c'est à dire de choisir entre un des quatre genres ou le

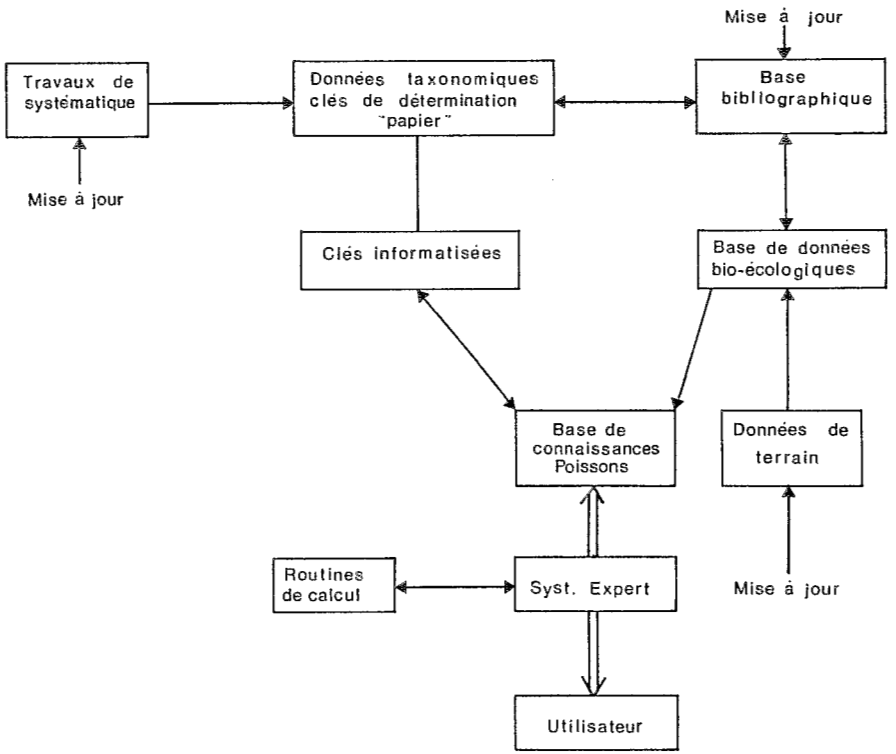

Fig. 2: Schéma d'une base de connaissances sur les poissons d'eau douce couplée à un S.E.

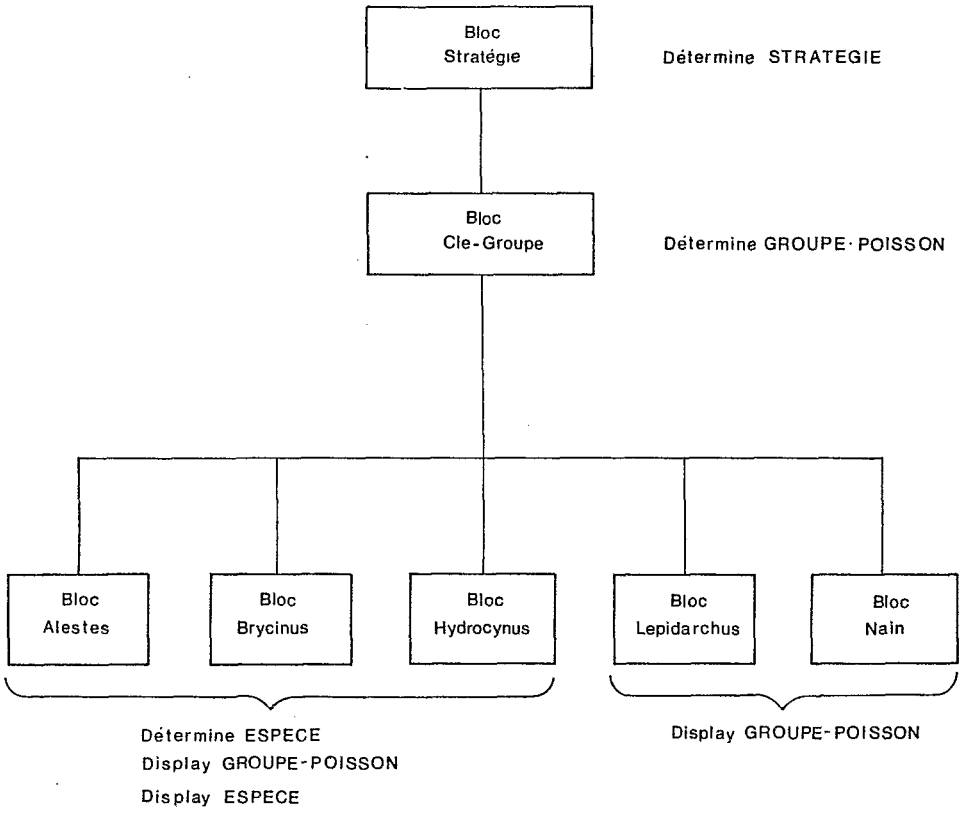

Fig.3 Hiérarchie des Blocs de contrôle et textes de contrôle.
(Voir texte)

groupe des Characidés nains. Ce dernier n'étant pas pris en compte et le genre *Lepidarchus* ne comportant qu'une seule espèce, leurs blocs de contrôle respectifs ne comportent qu'un ordre d'affichage mais pas d'autre ordre "*détermine*".

A titre d'exemple, je présente le bloc le plus simple : le bloc *Hydrocynus* qui comprend cinq espèces. Le tableau 2 indique pour ce bloc : (1) les paramètres qui y apparaissent pour la première fois, (2) le nom des règles de production (une par espèce), le texte de contrôle du bloc (qui indique le but et les actions à effectuer). S'il existait un bloc "*descendant*", le dernier ordre dirait de passer la main à un tel bloc ("*establish*"..), (3) le nombre de fois où il faut parcourir le bloc, (4) le bloc "*parent*", (5) un interrupteur pour la remise en ordre permanente des règles (s'utilise en chaînage avant), (6) un interrupteur permettant de fixer un temps d'exécution, (7) la liste des objets auxquels le bloc fait référence, (8) la liste des objets qui font référence au bloc.

La structure d'une règle est la suivante (Tableau 3) : (1) le texte de la règle avec les prémisses "IF...AND..." et une ou des conclusions "THEN..." ("*clause d'action*") ; (2) le bloc auquel la règle est rattachée ; (3) le type de règle ; (4) la liste des objets auxquels la règle fait référence et (5) la liste des objets qui font référence à la règle.

A tout moment il est possible d'interrompre le déroulement du programme pour faire appel à une base extérieure ou à une routine de calcul.

2. INTERET ET COUT HUMAIN DE LA REALISATION DE BASES DE CONNAISSANCES ET DE SYSTEMES EXPERTS EN ECOLOGIE

Par sa durée, la réalisation d'un système expert d'aide à la détermination "*per se*" ne se conçoit que dans des cas précis :

- existence d'un plan important de formation de techniciens et de chercheurs locaux. Le rapport intérêt/temps investi peut alors devenir favorable ;
- existence d'un programme à long terme au sein duquel il est prévu que se succèdent plusieurs biologistes dont la connaissance de la faune locale n'est pas acquise. Un système expert peut alors grandement faciliter la tâche des nouveaux arrivants ;
- existence d'une volonté de suivi à long terme, censé se poursuivre après le retrait de l'ORSTOM. La réalisation d'un système expert peut s'avérer intéressante pour fournir à la contrepartie locale un outil de formation et de travail (programmes de contrôle biologique à long terme en particulier, suivi des lacs de barrage, etc.). Dans ces cas, si le produit est bien fini, il est envisageable de vendre le système expert à l'organisme finançant l'opération ou à la société d'exploitation.

En revanche, il serait souhaitable que tout biologiste venant à opérer une révision systématique d'un groupe ou de la faune ou flore d'une région débouchant sur une nouvelle clé de détermination le fasse sous forme de système expert. Avec un bon générateur, ce travail n'implique pratiquement aucune surcharge de travail (hormis l'apprentissage du fonctionnement du générateur) et peut même le faciliter et constituer un excellent test de la cohérence de la clé.

La réalisation de "*simples*" bases de connaissances bio-écologiques est d'un tout autre ordre. Liées ou non à un système expert, ces bases permettraient de regrouper les informations dispersées sur un grand nombre de supports parfois introuvables, voire non publiées, et destinées à disparaître avec la cessation d'activité du chercheur.

FOCUS CONTROL BLOCKS

CONTROL BLOCK:F_HYDROYNUS

Parameters PARAMETER:P_H_ESPECE
PARAMETER:P_H_DENTS

Rules RULE:R_HTANZANIAE
RULE:R_HBREVIS
RULE:R_HVITTATUS
RULE:R_HFORSKALII
RULE:R_HGOLIATH

Control text display p_groupe_poisson;determine p_h_espece; display p_h_espece;

Max instances 1

Parent F_CLE_GROUP

Dyn Rule Order FALSE

DisposeWhenDone FALSE

If ref it list PARAMETER:P_H_ESPECE-DisplayToParam
PARAMETER:P_H_ESPECE-DetermineToParam
PARAMETER:P_GROUPE_POISSON-DisplayToParam
RULE:R_HGOLIATH-FCBToRules
RULE:R_HFORSKALII-FCBToRules
RULE:R_HVITTATUS-FCBToRules
RULE:R_HBREVIS-FCBToRules
RULE:R_HTANZANIAE-FCBToRules
PARAMETER:P_H_DENTS-FCBToParams
PARAMETER:P_H_ESPECE-FCBToParams
FCB:F_CLE_GROUP-DesctoParent
It ref me list FCB:F_CLE_GROUP-ParenttoDesc

Tableau 2 : Exemple de structure d'un bloc de contrôle dans ESE.

RULES

RULE:R_HBREVIS

Rule text IF p_h_dents = '12' and p-h-ecailles = '3'
THEN p_h_espece = 'H.brevis'

Owning FCBs :FCB:F_STRATEGE
:FCB:F_CLE_GROUP
x :FCB:F_HYDROCYNUS
:FCB:F_NAIN
:FCB:F_ALESTES
:FCB:F_BRYCINUS
:FCB:F_LEPIDARCHUS

Rule type Inference

I ref it list PARAMETER:P_H_ESPECE-LHSPaInThen
PARAMETER:P_H_ECAILLES-TestPaInPrem
PARAMETER:P_H_DENTS-TestPaInPrem

It ref me list FCB:F_HYDROCYNUS-FCBToRules

Tableau 3 : Exemple de structure d'une règle dans ESE.

Au niveau de ceux-ci, la réalisation de bases de données individuelles pour usage personnel est courante. Une politique de relative standardisation et de mise en place de bases de connaissances régionales, gérées ou non par des systèmes experts, augmenterait les contacts transversaux et favoriserait l'échange d'informations au sein des UR et entre celles-ci. Elle fournirait en outre :

- des outils de formation des jeunes chercheurs et techniciens ;
- des outils de travail pour les chercheurs de spécialités voisines ou changeant d'aire géographique d'étude ;
- les maillons d'un réseau de bases de connaissances internationales.

La mise en place de bases de connaissances régionales trans-disciplinaires passe par l'existence d'une politique informatique souple qui favorise une relative standardisation sans imposer trop de contraintes rebutantes. C'est là un des "*serpents de mer*" du développement informatique des grandes institutions ou sociétés. L'expérience a montré que les services informatiques centraux ont une tendance naturelle à développer selon une finalité et une problématique propres, en réduisant les liens avec les utilisateurs extérieurs. Une autre tendance étant le repli sur l'informatique "*administrative*", par nature plus codifiée et moins "*fantasque*" que celle des chercheurs.

Par ailleurs, le développement et les prix de la micro-informatique ont habitué les utilisateurs à une très grande variété de configurations matérielles et surtout logicielles. Toute tentative de standardisation sera nécessairement ressentie comme une brimade et tout choix contesté.

Cependant, deux voies qui, à ma connaissance, ne sont pas utilisées systématiquement pourraient être explorées :

- la standardisation matérielle est en fait une réalité (deux options possibles). Néanmoins, lors de l'achat de quantités importantes de matériel informatique (équipement de services au siège, d'un laboratoire, etc.) des contrats sont négociés (ou devraient l'être) pour un achat groupé à prix réduit. Une enquête préalable pourrait permettre à ceux à qui la configuration convient, de s'équiper d'un matériel donné, au meilleur prix sans que cela apparaisse comme une contrainte ;
- la question des logiciels est plus complexe. Cependant, il devrait être possible, après enquête, de négocier auprès des éditeurs de progiciels des licences d'utilisation multisites pour l'ensemble du personnel de l'Institut. Trois conditions essentielles devraient présider à tout choix : (1) être un logiciel répandu internationalement, (2) être d'un prix abordable, (3) pouvoir transmettre facilement les données sous forme utilisable sur un mini-ordinateur.

Au sein de l'Institut, l'augmentation des échanges transversaux et la réalisation d'outils de formation sont des points d'intérêt peu discutables. Mais, et c'est l'un des points de réflexion du présent séminaire, quel est l'intérêt de bases de connaissances au plan international ? Le passage d'un "*outil de travail et de formation*", à usage interne de quelques dizaines de personnes, vers une base connectée à un réseau international est un saut qualitatif important. Il ne se justifie

à mon avis qu'en liaison avec les efforts d'autres organismes, notamment le MNHN, dans le cadre du programme existant d'informatisation et d'interconnexion entre les Museums européens. Celui-ci devrait aller outre les aspects "*conservation*" et "*catalogue*" de spécimens pour devenir une vraie base de données bio-écologiques.

En tout état de cause, de tels projets impliquant plusieurs chercheurs (pour la réalisation comme pour l'alimentation et la mise à jour périodique des bases), demande un engagement des commissions scientifiques, autant dans le cadre de leurs attributions relatives à l'animation scientifique et au développement des échanges, que pour la valorisation du travail fourni par les chercheurs pour la réalisation des systèmes.

CONCLUSIONS

Je reviendrai pour terminer sur le travail que nous avons réalisé. Les systèmes experts ont des avantages certains sur les autres procédés en tant qu'outils d'aide à la détermination. Le choix entre tel ou tel système expert se fait en fonction du critère que l'on souhaite privilégier, l'objet et sa description (systèmes centrés-objet), ou la démarche de l'expert (systèmes à base de règles de production).

Nous avons choisi cette dernière approche parce que :

- les outils informatiques y sont plus développés actuellement et que notre approche est plus biologique et utilitaire qu'informatique. Les systèmes centrés-objet sont encore entre les mains des chercheurs en informatique. Ce que nous ne sommes pas ;
- ces systèmes permettent très facilement l'accès aux bases et routines extérieures ;
- il nous a paru important dans cette première approche de décortiquer le mécanisme du "*savoir*" de l'expert biologiste, et c'est à cela que nous avons tout d'abord consacré nos efforts.

Les premières conclusions auxquelles nous sommes parvenus, et qui sont brièvement citées au début de ce texte, nous posent cependant un problème fondamental, celui de l'"*analyse d'images*" réalisée par l'expert biologiste pour orienter sa diagnose. La reconnaissance visuelle d'un individu intègre très rapidement un grand nombre de paramètres sans que l'expert en ait même véritablement conscience. Cela se traduit par une phrase du type "*cette individu a un aspect de X, ou ressemble à un X*". Or, dans l'état actuel, aucun système informatique n'est en mesure de réaliser une telle opération. Et le seraient-ils un jour, que de tels systèmes resteront longtemps d'un coût élevé et disproportionné pour l'usage envisagé ici.

Si en poussant plus loin l'analyse de la démarche de l'expert, on le prive du contact visuel avec l'échantillon (en passant par un intermédiaire que l'expert interroge par exemple), la démarche change. L'expert privilégie les aspects bio-écologiques (lieu de récolte en particulier,...) pour réduire l'éventail des possibili-

tés. Puis il cherche à connaître l'aspect général de l'individu à déterminer (est-il trapu, élancé, etc.). Mais ces notions sont subjectives, et si l'on cherche à les quantifier l'on s'aperçoit vite que la variabilité peut être importante au sein d'une même espèce. Chez les poissons par exemple, entre des individus bien ou mal nourris, mâles ou femelles, et pour un même sexe en fonction de l'état de maturité. Passé ces stades, l'expert rentre beaucoup plus vite dans une démarche de type "*consultation de clé*" que lorsqu'il voit l'objet à déterminer. La présentation d'écrans d'images ou de silhouettes schématisant les diverses formes pouvant se rencontrer, pourrait peut-être faciliter la reconnaissance dans les premières étapes de la diagnose. Mais pas dans tous les cas. Dans celui des Characidés par exemple, l'expert discrimine immédiatement des individus dont l'allure générale est très proche.

Devant la difficulté de formaliser le savoir et la démarche "*réelle*" de l'expert, ma conclusion serait que pour toutes ces approches d'application des systèmes experts à la biologie il faut garder présente la très grande complexité de tout ce qui touche au vivant. Pour en tenir compte, le moyen terme n'est pas admis. Il faut soit adopter une démarche ultra-réductrice et se contenter (dans notre cas) d'une clé dichotomique plus ou moins optimisée, soit faire un grand bond qualitatif et profiter des atouts des systèmes experts, pour gérer des bases de connaissances variées qui cherchent à être aussi exhaustives que possible.

Ceux qui se sont intéressés à la modélisation des interactions biologiques, se sont vite retrouvés face à des phénomènes analogues. Des modèles "*simples*" peuvent être extrêmement féconds (Lotka-Volterra, modèles de gestion des stocks, etc.), mais dès que l'on veut aller plus loin, il faut aller très vite très loin.

REFERENCES BIBLIOGRAPHIQUES

WOOLEY, J.B. & STONE, N.D., 1987. *Application of artificial intelligence to systematics : SYSTEX, a prototype expert system for species identification.* Syst. Zool., 36(3) : 248-267.