Publ. 197
1993c
Anonyme
"Les systèmes d'information environnementale". Ed. ICALPE,
Le Bourget du Lac, 69-93.
EXEMPLAIRE RÉSERVÉ

ÉCORDRE

Base de données relationnelles pour toutes données écologiques, y compris les activités humaines

P. Soto, M.-B. Bouché*

Résumé

Toute donnée écologique objective, c'est-à-dire issue d'une portion quelconque de l'écosphère peut être repérée dans le champ écologique selon trois référendaires : l'espace euclidien, le temps, la composition (la caractéristique plus ou moins analytique ou globale à laquelle se rapporte la donnée). Ainsi, pour toute donnée, il existe un référentiel parfaitement objectif (lié à l'objet) au plan écologique. Ce constat élémentaire constitue le schéma conceptuel fondamental de la base de données ÉCORDRE. Les systèmes de gestion de bases de données relationnels ouvrent de nouvelles voies à l'écologie ; le traitement et l'accès aux données peuvent maintenant être maîtrisés par l'écologue. L'utilisateur non-informaticien peut aisément récupérer ses données et celles qu'il désire puis les restructurer pour des applications avec des logiciels statistiques, mathématiques, graphiques, etc. Coordonner les efforts, accumuler des données non-redondantes, faire circuler l'information écologique, voilà quelques possibilités offertes par la constitution de la base de données relationnelle ÉCORDRE.

* Laboratoire de zooécologie du sol – INRA, CEPE/CNRS, route de Mende, BP 5051, F-34033 Montpellier cedex

Abstract

ÉCORDRE: a Relational Data Base to all ecological data, including the human actions

NOTA: Only the French text is grounded on a semantico-conceptual study of both word and concept meanings.

Ecology

The recent possibility of using powerful software for data management grounded on relations between these data is a fundamental need of ecology (in the original and precise meaning of Ernst Haeckel, 1866) to describe ecosystem (in the original and precise meaning of Tansley, 1935): a global conservation of analytical data on ecosphere (or "globalyse").

Today, all ecological systems are under human influences, i.e. they have to take in account human actions. Agronomic data, hydraulic management, chemical movements including pollutants, conservation problems... are parts of the description of ecological systems.

Concepts

Nevertheless and consequently, the improvement of uniform and computerized ecological data management requires to a full re-

vision of concepts and the choice of a unique language with a univocal (sharply unambiguous) sense of each word. Every micro-science relating to ecology has specific languages related to its objects of study (chemicals of chemists, soil horizons of pedologists, water flow measures by hydrologists...). But conversely we need a restricted common language about ecological systems and the various steps of data gathering, management and interpretation.

Less the 150 common, univocal words are needed (less than the number in Mendeleev's table) in ecology. Unfortunately, in the absence of this basic rigorous vocabulary, shared between ecologist (i.e. scientist or technician working on true ecology) the fuzzy concepts forbid any shared management of ecological data.

We cannot present here the semantic and conceptual background of such a needed and unavoidable vocabulary. This is given elsewhere (Bouché, in press); only an introduction and some figures illustrate how all ecospheric (biospheric + environmental + agronomical + non-biotic data of the "biosphere") data could be described and then managed, irrespectively of their origins.

Limits

To be clear, ÉCORDRE is designed only for ecological data (but all of them!). For instance, those data describing parameters about factors influencing the decision to introduce a pesticide on the market, or describing psycho-socio-economic "reasons" for the choice of such a pesticide (or any other techniques) by farmers are not ÉCORDRE's data. Only the date, "site", dose, introduced in the ecological system (field) by the farmer, are ecological data.

The same can be said about data from laboratory experiments – only those from models really (not verbally, but factually) closely designed to be validated in one part of the ecosphere (i.e. ecologically falsifiable) could be managed by ÉCORDRE.

Conversely by giving the possibility to gather and manage all kinds of ecological data and pushing us to a univocal ecological language, such a Data Base allows and improves trans-disciplinary transfers. This base, grounded by data from the three exact sciences (biology, physics and chemistry), contributes to ecosystem descriptions and to the promotion of a precise ecological minimum knowledge.

All data are tied to geodetic references, so ÉCORDRE is also a Geographical Information System.

Direct advantages

ÉCORDRE's conceptual models allow us to conserve all "objective" data about the ecosphere.

The use of such a "Relational" Data Base gives in addition, other advantages:

– to oblige us to precisely describe ties between each item of data to every other item of data in the "ecological field",

– to define precisely (and describe) each technical step of the data gathering (techniques, people responsible for such data...),

– to allow us (and in fact oblige us) to eliminate every redundancy. The "normalization" of tables (not data sets or files) eliminates such data repetitions, obligatory in classical computer files,

– because all data is tied of their sources (the "prélevat") it is not necessary to write a program to get them in the system. Just a question (a query) is addressed. The Relational Data Base Management System, (R)DBMS, does the job ! This system optimizes the searching of data you need, and gives answer to your query without any program writing.

All data are initially collected for a research aim. Indeed, ÉCORDRE allows the use of these data for this original aim but, in addition, any re-use for a new purpose could be easily made without limitations. Each item data is an atom of knowledge on our ecosphere which could be, thanks to its linkage, re-used in any new "molecular" combination (making of sets). If a question is given in a totally new form, it is possible to make a query which extracts from ÉCORDRE all data (= objective knowledge) previously accumulated. This query has to be made prior to any filled data collection and consequently can avoid, or limit the need to collect new data in the field.

Falsifiability and ÉCORDRE

The collecting of new data from the "ecological field" needs choices in research aim, research sites, "points" of sampling in the site(s), analytical methods. These choices of hypothesis (options) lead to take the risk to eliminate the best.

In turn, data interpretation needs another kind of hypothesis about the choice of variables, limits of sets, stochastic errors tolerated, design of models, and so on.

Such hypotheses consequently introduce problems of objective knowledges, and the improvment of ecology is obligatorily made by successive approximations.

The management of data in ÉCORDRE, i.e. in (R)DBMS, does not require a hypothesis. Conversely, this type of management allows iterative use of data, at each progress step (and validation of hypotheses for both options and propositions). For example, any item of data previously used for a paper, a thesis, or leading to no results, can be re-used as a piece of information, without connection to the original aim.

Paying only the price for an uniform and logical description of data (= the "data referential"), all ecological data can be managed in ÉCORDRE.

ÉCORDRE's state of the art

First we made an initial training set based on our specific laboratory data sets, devoted to earthworms and their environmental or mesic parameters in relation to soil, climates, vegetation types, field uses. Then, we have tested the generality of our conceptual analysis of data on various data sets, completely alien to our usual types of work, e.g., nesting of birds, prey-predator relations, farming machinery, rice field managements. Data could be quantitative, qualitative or fuzzy and tied to state or "exchange" variables between compartments, mobile objects, etc.

Software and uses

At present, the ÉCORDRE data base uses IBM SQL/DS software running through VM. As an interface can be made between this (R)DMBS and one Expert System (EMICAT) running also with VM, it will be possible to provide, in parallel with good data

management in ÉCORDRE, to possibilities to interpret these data in the Knowledge Base of an Expert System.

We currently use this base to restore data, based on queries, but also through programs (where queries are included) called "macroprocedures", which in routine not only restore data but give them in the needed "form". It is also possible to use the "welcome-program" (written in FORTAN, PL/1...) to interface complex computations with its proper language (SQL). Various graphic and statistic tools are then used.

In addition, the SQL/DS allows a work management tackling various personal positions thanks to limite access, data conservation at a given stage of knowledge, confidential restriction... A list of people working on, and in ÉCORDRE have to be made.

In short, ÉCORDRE allows:

- *the conservation and iterative accumulation of data,*
- *the breaking of exchange limitations between file data,*
- *the "non-fossilization" of data,*
- *non-redundancy,*
- *ecology i.e. transdisciplinarity,*
- *rigour in data management,*
- *limitations of access, according to users,*
- *the interconnection with executive program,*
- *the interface between ÉCORDRE and Expert-Systems (ties with interpretations).*

Introduction : les limites du domaine couvert

Les systèmes d'information en environnement et ÉCORDRE

Les systèmes d'information en environnement sont des moyens techniques permettant la gestion de l'information dont l'homme dispose sur son environnement. Il s'agit essentiellement de moyens informatiques de saisie de données, de gestion-conservation, de traitement et d'édition de l'information.

Selon les objectifs et les moyens disponibles ces systèmes peuvent être très divers quant à leur objet (nature des données prises en compte), à leurs objectifs (répondre à une demande finalisée précise ou générale), à leurs moyens financiers et à la puissance et convivialité des machines et logiciels mis en œuvre.

La base de données relationnelle ÉCORDRE vise, comme son acronyme l'indique, après saisie des données écologiques à assurer la gestion ordonnée de celles-ci. Il s'agit de leur prise en charge exhaustive, leur conservation et leur restitution, cette dernière étant optimisée pour chaque demande spécifique quelle que soit la nature de la demande et l'origine des données.

Cette base a été développée initialement pour répondre aux besoins spécifiques de notre laboratoire qui a accumulé depuis 25 ans une masse de données sur support informatique à partir de préoccupations zoologiques, mésologiques, environnementales, agronomiques, cynégétiques... Cet ensemble, acquis « autour » de l'étude de la première masse animale de France (les lombriciens), nous a obligé à une rationalisation explicite transdisciplinaire (par exemple : zoologie, pédologie, botanique, chimie des sols, contaminants et pesticides, physique des machines agricoles et des propriétés du milieu, itinéraires agricoles, gestion de l'espace...). Cette rationalisation a abouti à une schéma conceptuel unique qui s'est avéré non seulement utile pour nous-même mais applicable à toute donnée venant du champ écologique, autrement dit de l'espace régi par les systèmes écologiques. Toute donnée ayant une origine dans l'espace-temps est géocodée, il s'agit donc aussi d'un SIG (système d'information géographique).

ÉCORDRE nécessite:

1. D'être développé sur des gros systèmes (présentement SQL/DS, prochainement DB2 travaillant aussi avec le langage SQL) car actuellement seuls ceux-ci offrent toutes les fonctionnalités nécessaires à une gestion transdisciplinaire et ouverte (notamment télématique) des données,
2. Un schéma conceptuel transdisciplinaire permettant de décrire, conserver et restituer les données sans connotations spécifiques des diverses disciplines concourant à l'étude objective de notre espace. Ce schéma exclut le flou sémantique mais n'exige qu'un vocabulaire spécifique très réduit.

Quelques limites et concepts

Les études environnementales peuvent porter sur des aspects très divers : législation réglementaire de divers pays, pollution d'une rivière, efficacité de pots d'échappement catalytiques, effets de pratiques agricoles sur l'érosion, optimisation d'un tracé d'autoroute ou de piste de ski dans un paysage, perception

psychosociologique d'une modification faite ou projetée de notre espace, impact de retombées radioactives, influence culturelle de démonstrations dans l'enseignement, augmentation du CO_2 atmosphérique, modifications observées par images satellitaires, etc.

Traditionnellement, ces aspects sont abordés après l'énoncé d'une problématique (souvent une préoccupation) et par la définition d'objectifs impliquant une ou quelques disciplines des sciences exactes et humaines. Ensuite, des informations ayant été rassemblées, une interprétation conduit à un ou n scénarios proposés aux décideurs. Cette approche traditionnelle a l'avantage d'être facilement compréhensible, car habituelle et concrète ; elle a l'inconvénient d'être séquentielle (limitée à des objectifs finis) et de ne pas être organisée pour mettre à disposition l'information acquise pour des ré-utilisation. Les informations sont « marquées » par ce qui a motivé leur acquisition et sont fossilisées sur des supports (cahier de notes, fichier informatisé, publications et rapports) dont l'organisation et la disponibilité résultent du but particulier poursuivi et non de la transdisciplinarité.

Cette approche traditionnelle était de toute façon la seule possible car nous ne pouvions jusqu'à présent, au mieux, que travailler qu'avec des fichiers informatiques accumulant l'information. Celle-ci était difficilement restituable, au prix de manipulations de fichiers, longues et très peu conviviales, même avec l'aide de gestionnaires de fichiers sur micro (tel DBase 3).

L'apparition de Systèmes de gestion de bases de données relationnels (SGBDR) a changé radicalement cette situation en permettant « théoriquement » une circulation conviviale des données et une ré-utilisation de celles-ci indépendamment de toute considération « disciplinaire » ou mono-objectif.

Ceci est vrai théoriquement mais ne devient pratiquement possible qu'à condition d'élaborer un schéma conceptuel unique permettant la description des données et leur restitution. Autres conditions : bien percevoir la puissance et les limites d'un tel outil.

Pour que cet outil devienne transdisciplinaire et conserve *ad libitum* les données environnementales, il faut en corollaire ne pas le lier à une étude spécifique particulière c'est-à-dire ne pas entraver son fonctionnement par les pratiques, très variables, d'acquisition des données et celles, non moins diverses, d'interprétation des données.

La réciproque n'est pas vraie : il ne s'agit pas d'acquérir n'importe quoi n'importe où, mais de poursuivre des objectifs précis pour une acquisition et une interprétation optimales des données vers cet objectif fini. Mais une fois les données acquises, il suffit de les décrire objectivement en suivant la rigueur d'un référentiel des données faisant appel à un nombre limité (quelques dizaines) de concepts pour rendre disponible l'information objective.

Cette description implique la connaissance de quelques concepts généraux, si transdisciplinaires qu'ils en paraissent abstraits.

Nous distinguons les Données initiales contrôlées (DICs) qui sont directement issues d'un objet de l'espace écologique étudié. Cet objet est le prélevat. La donnée est objective *stricto sensu* car liée à cet objet. Ce prélevat est situé dans l'espace et le temps, c'est-à-dire géocodé et daté. La DIC est donc liée *via* son prélevat à l'espace-temps : ÉCORDRE constitue donc une base de données écologiques environnementales mais aussi géostatistiques, ou si l'on préfère un SIG.

La DIC est simultanément liée à sa caractéristique (par exemple : poids, couleur, espèce, âge...) et à son mode d'expression. Par exemple, dans les cas ci-dessus, la première caractéristique est exprimée dans le système métrique international par données quantitatives (= 33) liée à une unité (g) ou par appréciation qualitative (rouge), par nomenclature (= *Plantago lanceolata*) liée à des codes de nomenclature ou par des appréciations floues (= vieux).

La DIC est également appréciée *versus* son prélevat par un degré de précision descriptible auquel elle est liée et se rapporte aux dimensions (= cotes) de ce prélevat. Ces cotes, n'étant enregistrées que si nécessaire, se rapportent à l'espace-temps limitant l'objet ayant fourni la donnée. Ce prélevat se repère enfin lui-même directement, ou *via* des prélevats gigognes auxquels il est lié, aux repères géodésiques et à la chronologie universelle.

Les DICs sont également liées aux techniques d'analyses (décrites dans un descriptif technique : DT) par exemple une concentration de cadmium d'un prélevat de sol (ou de plante) est dépendante des modalités de minéralisation mais aussi de la méthode d'analyse chimique de la solution obtenue.

Les DICs sont enfin liées à la motivation d'acquisition qui se traduit par une administration des diverses réalisations et essais de recherche mais aussi à une identification de l'au-

teur responsable de la donnée et aux éventuelles limitations d'accès dans ÉCORDRE.

Concluons sur les DICs avec ces quelques remarques :
– la DIC telle que liée à son contexte comme décrit ci-dessus est totalement indépendante des raisons de son acquisition. Si les options d'acquisition ont obligé les chercheurs à écarter de nombreuses autres données, la donnée n'est pas individuellement dépendante de ces options ;
– la DIC, étant liée à son contexte spatio-temporel et systémique, peut être réinterprétée indépendamment des objectifs initiaux : exemple une donnée relative à un animal observé pour un recensement quantitatif dans une forêt... peut contribuer à la cartographie européenne de l'espèce observée ;
– la DIC est l'atome de connaissance objective le plus fin que nous ayons vis-à-vis de l'espace-temps considéré : elle est nécessairement l'élément sur lequel les interprétations se fondent ;
– la notion de lien, ou relation, joue un rôle-clé dans le fonctionnement d'un système de gestion de bases de données relationnel.

Nous distinguons également les Données secondaires élaborées (DSEs) ; il s'agit de résultats d'interprétation fondés sur des DICs et des hypothèses : tels la moyenne, un paramètre obtenu par l'ajustement d'une fonction, les contours d'une aire d'une espèce. Dans les trois cas, on considère que les données objectives (DICs) regroupées forment un ensemble « suffisant » pour que moyenne, paramètre et contour aient une signification. Les DSEs en condensant l'information sont subjectives car liées aux modalités d'interprétation.

Si l'échantillon de DICs disponibles peut être insuffisant pour constituer un ensemble pertinent, chaque DIC individuellement est objective car liée à son objet. Chaque DSE est liée aux DICs plus les hypothèses liées à la synthèse.

Pour cette raison ÉCORDRE privilégie la gestion des DICs, intrinsèquement objectives, et laisse au stade interprétation (= traitement des données = synthèse) le soin du choix des hypothèses sous-jacentes à l'élaboration de DSEs à partir des DICs.

Toutefois, certaines interprétations standardisées se font en routine (exemple : expression de la biomasse des lombriciens au m^2 à partir des DICs, biomasses individuelles d'animaux liée à des prélevats cotés) et des programmes peuvent associer les requêtes des données (dans notre cas les biomasses individuelles, les liens au prélevat, les cotes des prélevats) et les calculs (en FORTRAN, PL/1, ...) pour obtenir la DSE désirée (biomasse/m^2). De même on peut faire appel aux divers logiciels d'interprétation et édition graphique, après avoir organisé des ensembles de données, à l'aide de requêtes adressées à ÉCORDRE.

En liant pas la base ÉCORDRE ni aux options d'acquisition, ni aux propositions d'interprétation, le gestion des données devient effectivement transdisciplinaire, multi-objectifs... car elle est en fait a-disciplinaire et a-objectif en laissant à chaque problématique le choix de ses objectifs, de ses acquisitions et de ses modalités d'interprétation. Elle permet la conservation et la restitution des DICs avec, si les chercheurs coopèrent, l'interconnexion pour échange de l'information objective acquise entre équipes. Toutes les DICs acquises, par les agronomes, les écologues, les environiciens... sur une portion d'espace-temps quelconque, deviennent regroupables, échangeables et administrées conjointement.

Données, faits et règles

La connaissance scientifique, donc écologique et environnementale, n'a aucune prétention à la vérité mais s'efforce, à partir des faits perçus, d'en donner une interprétation cohérente et rationnelle en liant les faits, aux moyens de règles édictées, à d'autres faits (dans l'espace, le temps et la complexité = composition). Il s'agit d'une approche par approximations successives.

L'apparition récente de moyens informatiques d'aide à l'interprétation, dit systèmes-experts, oblige quelque peu à rationaliser notre approche environnementale. Nous ne l'aborderons ici que succinctement, ÉCORDRE n'ayant aucune prétention à l'interprétation mais étant interfacée à un système-expert spécifique pour celle-ci.

Il s'agit ici plutôt de bien distinguer les faits gérables dans ÉCORDRE de ceux exposés dans un système-expert (ou en son absence dans la phase classique d'interprétation-discussion d'objectifs de recherche).

Toute personne ayant quelques connaissances, quelques pratiques agissant dans l'écosphère est ici un « expert ». Cet expert exprime son opinion en décrivant les faits qu'il juge pertinents et les déductions (= règles) liant ces faits pour décrire, par exemple au point de vue du forestier le devenir de la forêt, sous l'angle de l'agriculteur l'intérêt d'une culture, au sentiment de l'éleveur les qualités des herbages, à l'opinion d'un skieur les propriétés souhaitables d'une piste, etc. L'étude

des sols, des végétaux, des plantes, des cours d'eau est ainsi, outre les données initiales (DICs), accompagnée de faits globaux ou généralisés acceptés par 1 à n experts. Parmi les faits, il y a les DICs, données écologiquement objectives (d'ou l'importance d'interfacer ÉCORDRE à un système-expert), il y a des acquis acceptés usuellement (tel l'effet de la pesanteur), il y a des commodités (comme d'admettre usuellement cette pesanteur $g = 9{,}81\ m/s^2$ alors qu'elle varie en chaque lieu), il y a des croyances ou coutumes empiriques, permettant pragmatiquement de conclure et d'agir.

Au nombre des règles et faits ainsi acceptés, existent les « informations délocalisées », évoquées par F. Spitz. Il s'agit en fait :

1. d'information acquise et depuis longtemps acceptée comme bonne approximation à partir de données localisées antérieures, mais « délocalisées » à la fois par perte des sources (DICs) et par généralisation,
2. d'information acquise dans un contexte non écologique (= non localisé), tel un artifice de laboratoire, supposé applicable en conditions normales sans validation (tel un taux respiratoire établi en conditions standards de respirométrie).

Cartographie, expression des résultats

Nous avons vu pourquoi ÉCORDRE peut gérer toute DIC, c'est-à-dire toute donnée liée à toute caractéristique repérable dans l'espace-temps. L'interprétation de ces données peut faire appel à de nombreux outils mathématiques et informatiques de synthèse, au nombre desquels les statistiques descriptives (AFC, ACC, ACP... régression, classification) importantes à côté des outils d'aide à la décision (statistiques stochastiques) et de synthèse (modélisations diverses). Certains de ces résultats, bruts ou élaborés, relatifs à l'espace peuvent être présentés avec l'aide d'outils graphiques divers et variés permettant la représentation en 4 dimensions (3 dans l'espace + 1 variable Z2 représentée par une variation de couleur). Ces éditions doivent être regardées comme des présentations de résultats visant à répondre à la demande spécifique exploitant l'acquis des données objectives (gérées par un SGBDR) que nous possédions antérieurement sur le milieu considéré. Les DICs complémentaires, éventuellement nécessaires pour cette demande spécifiques, sont acquises pour compléter cet acquis initial et les diverses hypothèses mises en œuvre pour interpréter, y compris les pas d'interpolations et modalités d'extrapolation spatiales, temporelles ou compositionnelles choisies pour ces éditions graphiques. A ce niveau les modalités antérieures d'interprétation (décrites éventuellement par un système-expert) peuvent être mises en œuvre et améliorées.

Inversement, dans certaines circonstances des données issues de cartes – donc des données ayant été secondairement élaborées (DSEs) à partir d'une extrapolation spatiale d'observation – peuvent être facilement saisies et gérées dans une base de données. Elles offrent la commodité de la facilité de la saisie automatique (voir Spandau, 1989) et l'inconvénient de ne pas se fonder sur la partie objective ayant initialement permis la carte (= les observations effectives). Ainsi les DICs d'analyse des sondes et carottages pédologiques fondant l'ancienne décision d'une découpe géographique d'une unité pédologique sont généralisées spatio-temporellement ; seule perdure l'édition interprétée de la carte pédologique et souvent les données objectives (DICs) ont été perdues.

Il est possible d'administrer de telles DSEs, à côté des DICs objectives dans ÉCORDRE.

ÉCORDRE est d'abord un moyen d'accumulation et de restitution de données objectives, liées aux objets ayant trait à des variables physiques, chimiques ou biologiques voire économiques (un prix à un moment/lieu donné par exemple). Si l'analyse des mécanismes conduisant à un acte sur le milieu (traitement pesticide, retombée nucléaire de Tchernobyl, récolte de foin, vidange d'un réservoir dans une rivière...) n'est pas toujours évidente pour l'écologue, ces actes humains sont enregistrés comme tels dans ÉCORDRE, selon leur nomenclature spécifique, leur date d'action et leur localisation spatiale. Ainsi les variables socio-économiques agissant sur, ou depuis, le système écologique et environnemental sont-elles enregistrées.

ÉCORDRE et pratiques environnementales

En tant qu'écologues, nous sommes confrontés à une difficulté aussi ancienne que celle de la définition de l'écologie « gesamte Wissenschaft » ou science globale. Nous sommes seulement des spécialistes d'une petite fraction d'un ensemble de l'écosphère et

nous ne pouvons exercer nos démarches, à chaque instant et concrètement, que par une approche sectorielle dont les analyses doivent permettre cependant d'acquérir progressivement une vue générale. En tant que scientifiques, nous fondons nos travaux sur des données issues de notre objet d'étude : c'est la partie objective de notre travail, puis nous fondons nos interprétations sur des « analyses » de données et divers moyens de synthèse.

Ces données sont évidemment acquises, ordonnées, interprétées en fonction de modèles conceptuels, matériels ou/et mathématiques.

Dans toutes ces démarches, les données issues du champ de l'écosphère jouent un rôle pivot car elles sont la seule perception objective que nous en ayons, d'où la nécessaire acquisition de données en relation avec les problèmes que l'on se propose de traiter.

Ces acquisitions participent au progrès des connaissances ; elles doivent être enregistrées de la façon la plus objective possible (= objet d'étude) ; elles doivent rester accessibles, même longtemps après le travail d'acquisition initiale des données, car elles sont une indication temporelle (= historique, évolutive...). L'une des propriétés essentielles des bases de données relationnelles est de permettre cet accès efficace et durable aux sources de l'information écologique.

L'accès aux données

Traditionnellement, pour essayer de répondre à un objectif, on acquiert des données par échantillonnage, c'est-à-dire que l'on effectue un choix plus ou moins conscient de quelques caractéristiques reliées à quelques portions de l'espace-temps du champ écologique. Cette pratique qui permet de fonder un travail, pris isolément, est largement dominante puisqu'elle satisfait les objectifs concrets du spécialiste : elle restera de toute façon indispensable (Frontier, 1982 ; Scherrer, 1984 et sous presse).

Mais ces approches séquentielles (= travail par travail) ont le très grave défaut d'être morcelées, parcellisées et, par leur structure même, de s'opposer à l'intégration transdisciplinaire à propos de laquelle de nombreux spécialistes interviendront non plus tour à tour, en juxtaposant leurs données, mais simultanément.

L'usage intensif de l'ordinateur permet par ailleurs de manipuler de grands tableaux de données. Ainsi, les données de divers spécialistes peuvent être fusionnées, réorganisées, ou manipulées en sous-ensembles, et être présentées dans un même tableau.

Pour trouver les solutions de problèmes complexes, tels par exemple ceux relatifs aux mécanismes conduisant à l'accroissement du CO_2 (ou à la disparition partielle de l'ozone) dans l'atmosphère, ou tels ceux mettant en œuvre les modalités réelles d'action d'une macromolécule après son introduction intentionnelle (pesticides) ou accidentelle dans les systèmes écologiques réels (donc complexes), il est nécessaire de faire appel à de nombreux spécialistes et de manipuler un grand nombre de données. Dès lors, l'accès aux données pré-existantes ou accumulées collégialement est essentiel.

Prenons un exemple concret, en relation avec la préoccupation énoncée ci-dessus, relatif à l'enrichissement en CO_2 de l'atmosphère. *A priori*, on impute celui-ci à l'utilisation accrue de carbone fossile par l'homme, mais en fait les mécanismes en cause font intervenir de très nombreux processus (assimilation végétale, mécanismes de transfert et stockage atmosphère-océan, etc. : Trabalka *et al.*, 1983).

Ne retenons parmi ceux-ci qu'un seul : le changement du stock de la nécromasse (matière organique morte) du sol superficiel. Il se trouve que pour des motivations très diverses, plusieurs opérateurs font des analyses de sol parmi lesquelles le carbone organique est mesuré. Cela peut être l'agriculteur qui essaie d'évaluer l'état organique de son sol face au constat d'une perte de stabilité structurale (= battance) ; cela peut être le chercheur qui souhaite comparer des évolutions de divers types d'humus par le rapport C/N ; cela peut être le pédologue qui cartographie les sols, etc. Des données existent donc et sont acquises pour des motivations très diverses mais totalement indépendantes de la préoccupation d'interprétation de la croissance du CO_2 atmosphérique. Et pourtant, un changement de 1 % du carbone organique endogé représente une couche de CO_2 pur d'environ 6 à 7 m au-dessus du sol ! Cela veut dire que non seulement les sociétés industrielles utilisent intensivement du carbone fossile (pétrole, charbon, etc.), mais de plus qu'elles peuvent libérer, par exemple sous l'effet de l'accroissement des actions mécaniques aratoires oxydantes, des volumes très importants de CO_2 pour chaque pour cent de carbone tellurique (et les pertes de 3 à 10 % sont courantes !).

Cette évolution de la nécromasse dépend encore de l'utilisation des herbicides réduisant la fixation organique du CO_2, ou de l'augmentation des rendements agricoles à l'hectare, ou

de la réduction des surfaces cultivées dans la CEE... ce qui peut entraîner un restockage organique dans les sols mis en friche...

On pourrait ainsi décrire de nombreux exemples de ré-utilisations possibles de données dans le cadre d'interprétations multiples et indépendantes des objectifs d'acquisition initiaux (dans l'exemple ici, du fermier, du chercheur fondamentaliste, du cartographe). Il serait donc logique, après un premier usage ayant motivé l'acquisition des données, de mettre à disposition de la communauté humaine les connaissances objectives que sont les DICs écologiques. Encore faut-il matériellement pouvoir le faire.

En l'état actuel, en général, la donnée est souvent directement perdue... soit qu'elle ne convienne pas pour l'interprétation initiale, soit qu'elle ne soit pas interprétée, ou encore qu'elle soit « diluée » dans un ensemble dont on ne publiera qu'une moyenne (avec parfois l'écart-type), laquelle sera fossilisée dans une bibliothèque dans la majorité des cas. Seules quelques données restent conservées dans des fichiers informatisés dont l'existence est inconnue pour des utilisations totalement étrangères aux objectifs sectoriels initiaux (*cf.* l'exemple ci-dessus). La conservation des données objectives et leur mise à disposition pour d'autres objectifs et d'autres partenaires, en vue de résoudre d'autres problèmes, sont donc loin d'être assurés.

On peut depuis peu, grâce aux bases de données relationnelles, qui constituent une contribution à l'intelligence artificielle de notre environnement, lever cette difficulté. Ceci suppose deux conditions : l'une intellectuelle, dépendant de l'élaboration d'un schéma conceptuel unique et cohérent, mettant en œuvre un langage univoque, et l'autre matérielle, relative à l'accès aux logiciels des systèmes de gestion de bases de données relationnel (SGBDR). Notre propos ici n'est pas d'exposer en détail la résolution de la première condition mais d'illustrer l'applicabilité de ces nouveaux moyens aux données écologiques.

Interdépendance des données : la relation

Une donnée quantitative : 4,5, ou qualitative : vert (ou flore « riche »), n'est pas indépendante d'autres données. Par exemple, la première donnée (4,5) peut être décrite comme le pH d'un sol situé au point géodésique de latitude nord 52,90 et longitude est 3,42 (en grades), par rapport au méridien de Paris, le pH étant mesuré dans une solution aqueuse de rapport volumique sol/eau de 1/1, etc. On peut constater que cette phrase décrit des relations (ou liens) qui, dans l'ordre du texte, sont « 4,5 – unité – pH – sol – latitude nord – 52,90 – longitude est – 3,42 – grade – méridien de Paris – solution aqueuse – rapport sol/eau – 1/1 ». Cet exemple montre que la donnée n'a de sens qu'explicitée par un système de relations ; celles-ci sont présentées ici sous forme linéaire, au fil de la plume, mais elles sont en fait des relations multidimensionnelles dans un hyperespace car, dans cet exemple, le pH n'a de sens que si l'on précise la technique (solution aqueuse, rapport sol/eau 1/1). On aurait pu de même préciser « l'horizon du sol » et rattacher à celui-ci les autres caractéristiques observées (et leurs techniques) et relier ce sol à la « végétation », au « climat », aux « animaux présents », à la « contamination » (dont la biocirculation dépend du pH), etc. Enfin ces données sont liées à des données administratives soit de localité – département – nation... soit de problématique de recherche appel d'offre – contrat – objectif – chercheur, etc.

En conclusion, la donnée « 4,5 » considérée isolément n'a aucun sens. La richesse de l'information dépend des liens (ou relations) qui situent cette donnée dans le champ écologique, (et vis-à-vis de diverses nécessités techniques et institutionnelles).

Idéalement donc, avant d'entreprendre toute recherche, il serait logique de pouvoir disposer de toutes les informations pré-existantes relatives à l'objectif de l'étude spécialisée poursuivie, quelles que soient les raisons initiales de leur acquisition et d'exprimer la signification informative maximale de ces données, grâce aux relations qu'elles ont avec l'objectif de recherche entrepris. Ce n'est qu'après interprétation et si nécessaire ré-interprétation de ces données objectives préalablement acquises, que, si besoin est, un plan d'échantillonnage, éventuellement optimisé d'ailleurs grâce aux données préalables, pourra être établi pour l'acquisition de données écologiques complémentaires.

Intérêt des bases de données relationnelles en écologie

« L'écologie est la science globale des relations des organismes avec leur monde extérieur environnant dans lequel nous incluons,

au sens large, toutes les conditions d'existence » (Haeckel, 1866 ; trad. Bouché, 1980).

Le mot science étant pris au plein sens moderne du terme. Il y a eu matériellement impossibilité de satisfaire simultanément les exigences d'une telle pratique scientifique portant sur des données ayant trait aux organismes et à leurs milieux dans leur globalité appréhendés par un nombre infini de caractéristiques. Ou bien l'on étudiait une infime proportion de l'écosphère : quelques relations de quelques organismes avec leur environnement, ou bien l'on effectuait des approximations supraspécifiques globales hypothétiques ne se prêtant pas à la réfutation scientifique à partir de données objectives.

Cette impasse opérationnelle est aujourd'hui en voie de disparition grâce au développement de l'informatique, qui après la période classique des années 1960-1980, se développe depuis peu en permettant une optimisation de l'usage des données grâce aux bases de données et une confrontation des interprétations aux hypothèses dans le cadre de bases de connaissances de système-expert. Ceci permet(tra) l'élaboration d'une véritable intelligence artificielle écologique, fondée sur ces données objectives et le traitement intensif de celles-ci par ordinateur (Escoufier, 1987).

Cette possibilité nouvelle permet à l'écologue d'intervenir en respectant les exigences de sa discipline, mais elle modifie profondément les pratiques antérieures d'acquisition, de gestion et de traitement des données. Elle modifie aussi la position personnelle du chercheur ou gestionnaire, nécessairement à la fois celle d'un spécialiste d'un domaine restreint (car il doit être à la fois acquéreur de connaissances en tant que spécialiste) et celle d'un partenaire qui doit inscrire ses connaissances dans un patrimoine commun. Or, précisément, les bases de données relationnelles permettent l'intégration des données dans un ensemble global tout en autorisant leur sélection rapide sur simple requête et une régulation des accès des utilisateurs de données.

Cette nouvelle approche de la pratique scientifique en écologie implique un bouleversement conceptuel et opérationnel qui ne saurait être développé ici. Un livre y est consacré (Bouché, en préparation). Il s'efforce de tirer les conséquences de ce bouleversement. Notamment, à partir du moment où toute donnée écologique (botanique, physique, chimique, climatique, pédologique, zoologique...) peut circuler. Une sémantique aux dénominations conceptuelles univoques s'impose.

Nous nous contenterons simplement de signaler que toute donnée objective (issue de l'objet d'étude de l'écologie, c'est-à-dire une portion quelconque de l'écosphère) peut être repérée dans le champ écologique selon trois référendaires : 1) l'espace euclidien, 2) le temps, 3) la composition, c'est-à-dire la caractéristique plus ou moins analytique ou globale à laquelle se rapporte la donnée.

Nous constaterons ainsi que les données écologiques décrivent les systèmes écologiques quels que soient les actes humains, qui sont des données, et que ces données peuvent être regardées comme agronomiques, environnementales... selon les objectifs.

Si une donnée est ainsi liée à un objet concret, elle est également liée à la situation du prélevat.

La situation est décrite pour un point géodésique dans l'écosphère (figure 1, page suivante) ; ce point ne contenant rien (un point par définition n'a pas de dimensions), une donnée est nécessairement issue d'une portion d'espace-temps matériel dont elle est une des caractéristiques possibles. Cette portion concrète a reçu le nom de prélevat (Bouché, 1975) ; par définition, celui-ci ne peut fournir qu'une donnée d'une caractéristique. Éventuellement, cette portion d'espace-temps peut engendrer plusieurs données dérivées chacune de caractéristiques différentes ; de telles données sont alors liées entre elles par leur prélevat.

Ainsi, pour toute donnée, un référentiel parfaitement objectif (lié à l'objet) au plan écologique existe. Ce constat élémentaire constitue le schéma fondamental du niveau conceptuel de la base ÉCORDRE. A celui-ci se superposent par ailleurs des aménagements pratiques facilitant les nécessaires adaptations entre les pratiques externes et ce niveau conceptuel.

En conclusion, on peut dire que dans son fondement, effectivement, l'écologie n'est pas seulement la science des déterminations et influences physiques issues du biotope (le milieu géophysique) ; elle n'est pas seulement la science des interactions entre les divers et innombrables êtres vivants constituant la biocénose. Elle est la science des interactions (combinatoires/organisatrices) entre tous les constituants (physiques et vivants) des écosystèmes, y compris les actes humains. Une base de données, telle ÉCORDRE, constitue l'un des outils permettant de pratiquer pleinement cette science.

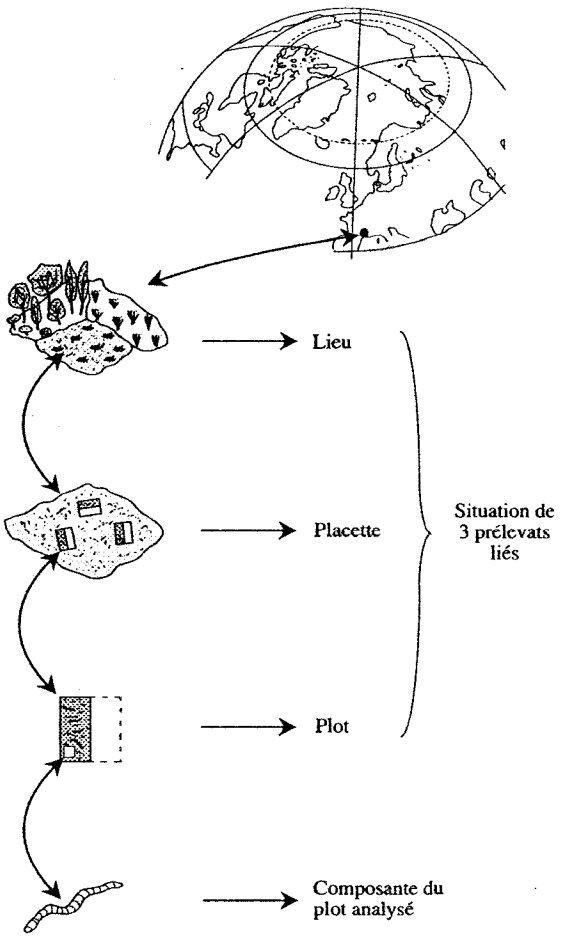

Fig. 1. Références des données écologiques : description de la situation de la donnée pour un point géodésique dans l'écosphère.
Si une donnée est liée à un objet concret, elle est aussi liée à la situation du "prélevat", portion d'espace-temps qui peut engendrer plusieurs données.

ÉCORDRE et manipulation de données

Fichiers et bases de données relationnelles

L'informatique est un moyen puissant pour traiter un grand nombre de données et en extraire l'information. Mais elle a aujourd'hui classiquement en écologie, un inconvénient majeur : chaque nouvelle application engendre ses propres fichiers et programmes et l'utilisateur n'a pas directement accès aux données sur lesquelles il travaille, il est contraint de passer par un programme capable d'extraire les catégories d'informations souhaitées.

De plus, la plupart du temps les fichiers possèdent entre eux des éléments répétitifs ; il y a donc des redondances. Par ailleurs, des associations entre fichiers sont peu exploitées du fait que ces fichiers sont utilisés isolément et indépendamment les uns des autres et que leur recombinaison est une tâche fastidieuse.

Cette redondance de l'information due à l'organisation en fichiers est coûteuse en moyens de stockage ; elle pose des problèmes de cohérence, les mises à jour d'une même information, localisée dans plusieurs fichiers utilisés pour des applications différentes ne pouvant pas toujours être simultanées.

Il est par ailleurs indispensable pour chercher une information de savoir où elle se trouve ! Sans parler des données tombées dans l'oubli, on est fréquemment amené à rechercher l'information souhaitée dispersée dans de nombreux fichiers.

L'organisation des données en fichiers est rigide et coûteuse en maintenance dès que le système informatique devient important, ce qui est courant en écologie, et peut conduire à une situation relativement anarchique.

Aujourd'hui, cette situation est dépassée grâce aux moyens techniques offerts par les bases de données relationnelles. En effet, il suffit que les données écologiques soient stockées dans une telle base pour qu'un utilisateur puisse formuler lui-même une requête, au moyen d'un langage très simple et offrant de très grandes possibilités, pour extraire l'information souhaitée. Cette information est accessible dans un tout, sans les barrières opérationnelles existant entre des fichiers différents, car le SGBDR assure lui-même automatiquement l'accès aux données demandées.

De plus, la façon dont les informations sont organisées dans la base de données est fort simple, la plus familière qui soit pour un chercheur ou technicien : il s'agit de tableaux au sens courant (et non au sens informatique) du terme.

Il faut bien comprendre la différence fondamentale entre la notion de fichier et la notion de base de données. En première approximation, on peut toujours dire qu'il y a des millions de caractères stockés à la suite sur des

disques et que l'accès se fait en séquence ou directement. Il paraît bien difficile d'inventer autre chose.

Pourtant, en regardant de plus près, on s'aperçoit qu'une base de données relationnelle n'est pas un ensemble de fichiers, mais un ensemble d'informations. La grande différence vient du fait que l'information est gérée à partir d'une approche logique de sa structure, sans tenir compte du support physique de cette information, et que cette logique reste indépendante des contraintes liées aux supports.

Grâce à la structure de bases de données, l'ordinateur est en mesure de fournir les informations dont nous avons besoin, et non celles qu'il serait susceptible de nous imposer (par exemple les valeurs inutiles créées pour un autre usage, dans un fichier) !

D'utilisation beaucoup plus simple (au prix d'un logiciel plus élaboré) que les bases de données classiques (hiérarchiques et en réseaux – voir P. Soto, 1985), les bases de données relationnelles offrent aussi aux informaticiens un outil puissant pour l'accès aux données dans les programmes d'applications, avec une prise en compte élaborée de la sécurité (contrôle des accès concurrents et points de reprise) et de la confidentialité (autorisations d'accès). Pour l'écologue, les bases de données relationnelles n'obligent pas à choisir *a priori* une structure générale des données, à la différence des bases hiérarchiques ou en réseaux, ce qui permet de bien dissocier la gestion des données de leur interprétation.

Cette propriété essentielle est acquise grâce à un logiciel spécifique, le système de gestion de bases de données relationnelles. Par inflation commerciale, certains logiciels sur des micro-ordinateurs sont présentés comme des gestionnaires de bases de données relationnelles alors qu'il ne sont que des gestionnaires de fichiers évolués et ne possèdent pas l'essentiel, le SGBDR, ni les fonctionnalités indispensables à une base de données relationnelles.

Malheureusement cet outil d'utilisation facile est usuellement présenté sous une forme très théorique, difficilement abordable, selon un formalisme mathématique particulièrement rebutant car ses développements, très récents, s'adressent aux informaticiens. Dans cet article, nous souhaitons faire une présentation pratique à l'usage des écologues (y compris agronomes, environiciens, etc.).

Nous allons d'abord situer la nature des liens à établir entre les données avant de présenter les bases de données relationnelles. Ensuite, plus spécifiquement, nous présenterons la base ÉCORDRE ce qui, chemin faisant, permettra d'introduire et d'illustrer les concepts et les moyens mis en œuvre pour une organisation générale de données écologiques.

L'organisation relationnelle des données

Toute conception de la recherche en écologie doit commencer par une analyse de la situation existante tant sur le plan des informations disponibles que sur celui des organisations. Les résultats de cette phase initiale donnent naissance à une représentation logique des données depuis un schéma conceptuel unique, qui devra aboutir à une description informatique d'un schéma d'informations écologiques.

Les schémas relationnels permettent de faire progresser de manière significative la conception et la spécification de ces schémas.

A l'aide du modèle relationnel, le schéma logique pourra être représenté par des associations d'ensembles d'objets fondées sur la notion mathématique de relation et visualisées sous forme de tableaux (figure 2).

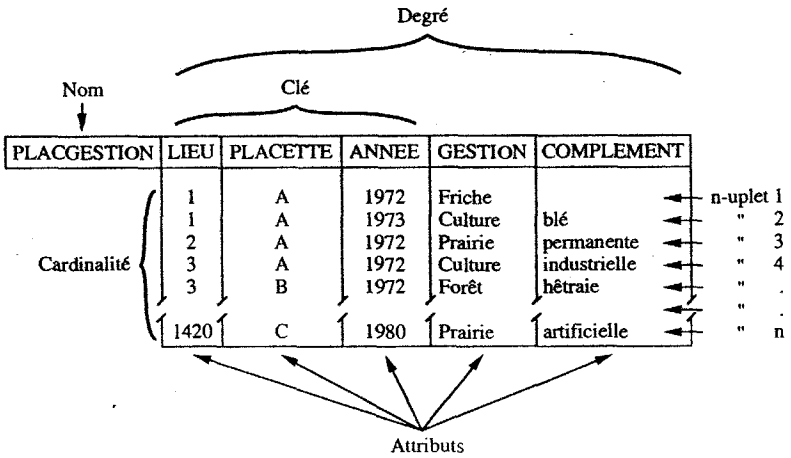

Fig. 2. Schéma logique de l'organisation relationnelle des données : exemple de relations relatives à l'organisation des parcelles. A l'aide du modèle relationnel, le schéma logique pourra être représenté par des association d'ensembles d'objets fondées sur la notion mathématique de relation et visualisées sous forme de tableaux.

Le modèle relationnel des bases de données est beaucoup plus proche du raisonnement humain que les autres modèles d'organisation classiques (hiérarchiques et en réseaux).

Par exemple, lorsque l'on nous pose une question, il va de soi que nous n'allons pas consulter la totalité de notre mémoire, ni même toutes les informations que nous possédons sur un sujet donné, avant de donner une réponse. Lorsqu'on nous demande quelle est notre profession, il est évident que nous n'avons pas besoin de passer mentalement en revue tous les métiers que nous connaissons pour trouver le nôtre. Nous le trouvons immédiatement parce que la question est posée de telle façon que les mots « profession » et « nôtre » et la relation établie réduisent, entre ces deux données, le champ de recherche à son strict minimum. Pour l'homme, c'est le contenu de la question qui va servir pour en déduire la réponse.

Il y a ici un processus déductif, lequel est à l'opposé du processus inductif utilisé dans la recherche à partir d'une base de données hiérarchisée.

Les SGBDR fonctionnent selon ce principe : c'est le contenu de la question qui va déterminer quelles sont les voies d'accès à utiliser et les liens à établir, puis à utiliser. Ces liens, ou pointeurs, ne seront plus fixés une fois pour toutes dans de telles bases ; ils seront créés dynamiquement selon les besoins.

Mais cela n'est pas suffisant, car il faudra que le langage d'interrogation du système soit capable de comprendre correctement une question. Pour cela, une seule possibilité : l'utilisation des opérateurs logiques de la théorie des ensembles, où la notion de lien entre données joue un rôle essentiel, comme c'est le cas, par exemple, pour le concept banal d'intersection de deux ensembles.

Propriétés des bases de données relationnelles

Dans une base de données relationnelle, les informations écologiques sont regroupées dans un ensemble cohérent de tableaux, conjointement avec les moyens de relier entre elles les informations réparties dans des tableaux distincts (au lieu d'être dispersées dans de multiples fichiers indépendants).

On peut donc définir une telle base de données en écologie comme étant un processus d'organisation du stockage d'informations écologiques reliées entre elles. Cette organisation doit être la plus exhaustive possible, dans le but de supprimer les redondances et d'obtenir une certaine indépendance entre la logique des relations entre les données et le support informatique physique.

Il est usuel aujourd'hui de présenter une base de données selon trois niveaux, le niveau interne avec le schéma physique, le niveau conceptuel avec le schéma logique et le niveau externe avec les schémas externes :

a) le niveau interne représente le schéma physique, c'est-à-dire, l'organisation physique des données sur les supports de stockage (disques, bandes...) de l'ordinateur. On peut considérer qu'il est obtenu par projection du schéma logique sur les supports physiques. Ainsi, une réorganisation de ceux-ci, par exemple pour un changement de matériel, ne devra pas affecter le schéma logique et, par conséquent, sera sans incidence sur les applications. Un algorithme d'accès règle le rangement physique des données indépendamment de l'organisation conceptuelle choisie : il y a indépendance entre l'organisation que l'on choisit pour les données écologiques et le « rangement dans la machine ».

b) le niveau conceptuel représente le schéma logique, c'est-à-dire, la description de l'organisation des données en associations d'ensembles d'objets, indépendamment de chacune des applications les utilisant. Il est indépendant du stockage physique, mais il est influencé par le type de SGBD utilisé. Le Schéma conceptuel unique (SCU) d'ÉCORDRE est le calque des relations entre données existant dans les systèmes écologiques et vis-à-vis des pratiques de l'acquisition d'information.

c) Le niveau externe représente les schémas externes ; chacun d'eux constitue une vue partielle du SCU et est spécifique à une application donnée. Ce niveau est susceptible d'une évolution permanente, au gré des applications. Les modifications d'un schéma externe sont sans incidence sur le SCU d'ÉCORDRE, car il n'en sont qu'une application propre à chaque spécialiste. Ainsi sont assurées l'indépendance logique des données et une remarquable souplesse pour tout usage écologiquement cohérent.

Cette organisation se traduit par une représentation de la structure des informations selon un schéma logique (niveau conceptuel) et par une transposition de celui-ci dans le stockage physique des données, formant le schéma physique (niveau interne). Pour ce faire, on utilise un logiciel de description des données et d'accès à celles-ci. C'est le Système

de gestion de bases de données (SGBD) qui permet principalement d'organiser les données sur les supports périphériques conformément au schéma logique et fournit les procédures de recherche et de sélection de ces mêmes données.

Les fonctions principales que doit assurer un SGBD en écologie sont :
a) mettre à la disposition de l'écologue un outil pour décrire l'ensemble des données qui seront stockées dans la base de données ;
b) offrir aux écologues un outil d'interaction avec la base de données sous forme d'un dialogue pour rechercher, sélectionner et modifier des données ;
c) offrir à l'écologue la possibilité de définir des règles qui permettent de maintenir l'intégrité de la base de données. Ces règles sont appelées des contraintes d'intégrité ; elles correspondent à des propriétés qui devront toujours être vérifiées dans la base, quelles que soient les valeurs enregistrées (exemple : le poids d'un individu d'une espèce précisée de ver de terre doit être inférieur à 2 000 mg) ;
d) permettre l'indispensable mise en commun des connaissances sous forme de données, c'est-à-dire permettre à chaque spécialiste-écologue d'échanger pour l'intégration effective des données avec tout autre ensemble, entre écologues, y compris par des échanges télématiques ;
e) offrir des mécanismes permettant de vérifier les droits d'accès des utilisateurs ;
f) offrir des mécanismes qui permettent de détecter les cas où il y aurait conflit d'accès et de les traiter correctement ; exemple : deux utilisateurs accèdent aux mêmes informations de la base de données en même temps ;
g) offrir la sécurité de fonctionnement : en cas d'incident (matériel ou logiciel) le système doit permettre la prise de point de contrôle pour remettre la base de données dans un état satisfaisant.

Les liens (ou relations)

Les liens qui existent entre les informations en écologie, sont finalement plus importants que les données elles-mêmes. Dans l'exemple, déjà décrit différemment, le pH d'un sol est lié à un horizon particulier, lié à son profil, lié à sa position géodésique, mais aussi aux végétaux produisant la litière, aux organismes la transformant et aux minéraux interférant avec ce pH. Donc, dans les schémas relationnels, comme l'expression l'indique, l'importance est donnée aux liens entre données. L'information écologique pourrait être définie comme les DICs et leurs liens.

Le lien offre également un mécanisme d'accès qui, à partir d'un élément d'un ensemble, permet d'accéder aux éléments d'autres ensembles qui lui sont reliés.

Quels sont les types de liens entre les informations ?
a) le lien 1-1 : permet de relier deux types d'informations, une à chaque bout, parfaitement symétriques. On parle de lien binaire ;
b) le lien 1-n : dans ce lien, on trouve à une extrémité une seule information et à l'autre extrémité il y en a « n ». C'est un lien orienté. Ce lien est hiérarchique ;
c) le lien n-1 : nous disons que l'on a un lien n-1 entre un ensemble A et un ensemble B si tout élément de A détermine au plus un élément de B et si à un élément de B correspond un ou plusieurs éléments de A ; on dit que A détermine B et B « est déterminé » par A. On parle de lien déterminant ;
d) le lien n-m : constitue une réalisation de liens où chaque élément d'un ensemble A peut être associé à « m » éléments d'un ensemble B, et chaque élément de l'ensemble B peut être aussi associé à « n » éléments de l'ensemble A ; on parle de lien maillé.

Essayons de voir ces différents liens dans l'association de deux ensembles. Ces associations illustrent le fait qu'entre deux ensembles il peut exister plusieurs associations de significations différentes, avec des liens distincts.

Exemple 1, prenons l'ensemble HOMME et l'ensemble FEMME et créons des liens par type de mariage :
- Monogamie (1-1)
- Polygamie (1-n)
- Polyandrie (n-1)
- Polygynandrie (n-m)

Avec la première association, lien 1-1, nous considérons une société judéo-chrétienne, qui exprime la monogamie comme le « mariage conventionnel ».

Avec la deuxième association, lien 1-n, nous considérons une société islamique dans laquelle la polygamie existe, nous avons le lien « mariage » où un homme est marié simultanément à plusieurs femmes.

Avec la troisième association, lien n-1, nous considérons une société dans laquelle la polyandrie existe ; nous avons le lien « mariage » où une femme a plusieurs maris.

Avec la quatrième association, lien n-m, nous pouvons considérer une communauté hippie dans laquelle le « mariage en groupe » existe.

Exemple 2 : prenons l'ensemble des LIEUX étudiés et l'ensemble des ESPÈCES-VÉGÉTALES, et créons les liens : Espèce-dominante, Description-floristique, Aire-répartition et Associations-végétales.
- Espèce-dominante (1-1)
- Description-floristique (1-n)
- Aire-répartition (n-1)
- Associations-végétales (n-m)

La première association, lien 1-1, exprime le fait que dans un lieu existe une seule espèce dominante et qu'une seule espèce végétale domine dans ce lieu.

La deuxième association, lien 1-n, exprime le fait qu'un lieu a plusieurs espèces végétales et qu'une seule description floristique se réalise pour chacun des lieux à un moment donné.

La troisième association, lien n-1, exprime que l'aire d'une espèce végétale est caractérisée par plusieurs lieux.

La quatrième association, lien n-m, exprime le fait que dans un lieu peuvent exister plusieurs associations végétales et qu'une même association végétale peut exister dans plusieurs lieux.

Données et perceptions dans l'espace-temps

Rappelons que les exigences de l'écologie placent l'écologue dans un dilemme *a priori* insoluble : être un bon spécialiste et inscrire son savoir dans un ensemble global. Nous avons indiqué que les bases de données relationnelles lèvent en grande partie cette difficulté consubstantielle à l'écologie. Aujourd'hui, des logiciels puissants de base de données relationnelle nous permettent de surmonter ce handicap. Encore faut-il les mettre en œuvre. Nous avons donc entrepris la réalisation concrète de la base de données ÉCORDRE dans cette perspective.

Nous allons succinctement présenter et illustrer certains aspects de la pratique de cette base pour en faire percevoir les potentialités, sans pouvoir concrétiser exhaustivement cette présentation (chaque application serait spécialisée et nécessiterait au moins quinze pages d'explications spécifiques !).

Nous ne prétendons pas aborder nous-mêmes tous les aspects des interactions combinatoires/organisatrices entre chacun et tous les constituants physiques et vivants des écosystèmes. Concrètement, nous avons abordé la construction de la base à partir de notre situation particulière, c'est-à-dire à partir de l'écologie des lombriciens et de certains travaux préliminaires sur la logique des données écologiques (Bouché, 1975, 1980).

Fort du schéma conceptuel unique adopté dans ÉCORDRE, dans cet objectif, on peut construire une base de données relationnelle à partir d'un domaine quelconque du champ écologique (par exemple, des végétaux, des contaminants, des lombriciens, des oiseaux, des travaux aratoires, des limites administratives, etc.) et à partir de là, on peut commencer à concevoir des relations qui permettront de lier les différents aspects objectifs des écosystèmes. On aura la possibilité donc de déplacer le centre d'intérêt de la base de données, et grâce à cette propriété, chaque spécialiste pourra se greffer sur ÉCORDRE.

Deux types de données permettent la description-interprétation de tout système écologique :
1) les états de « compartiments », qu'on va comparer entre eux,
2) les cinèses, qui correspondent aux échanges entre compartiments.

Les états correspondent aux données qualitatives ou quantitatives, présentes à un instant « t » donné. Par exemple : couleur d'un horizon de sol, hauteur de la végétation, concentration de zinc, etc. Il s'agit de caractéristiques liées à des formes ou structures. La durée du prélevat est négligée : la donnée est instantanée et se rapporte à une date.

Les cinèses correspondent à des débits, ou flux, liés à l'une des caractéristiques exprimées en fonction du temps. Par exemple : quantité de CO_2 absorbée à l'heure, quantité de pluie journalière (mesurée en mm), nombre de vaches par ha/jour, etc. La donnée est liée à une durée.

Nous les avons appliqués, par exemple, aux lombriciens, aux caractéristiques du milieu et aux itinéraires techniques agricoles (= succession des travaux agricoles).

A partir de ces données, on va construire des relations dans la base sans se donner de limites.

Basée sur une conception unitaire et globale de l'écologie (Bouché, 1980), on peut dire que la conception d'ÉCORDRE est tout à fait indépendante des lombriciens. Nous avons commencé, par commodité et besoins spécifiques, avec des applications relatives aux vers de terre mais la base peut accepter toute relation nouvelle ; il suffira qu'il y ait un point commun avec l'ensemble des relations existantes, tel le lien spatiotemporel.

ÉCORDRE peut subir dans le temps une accrétion qui peut-être recouvrira l'ensemble du champ écologique, parce que cette base a été conçue pour se développer de cette façon.

Le champ écologique correspond à l'ensemble qui recouvre les trois référendaires de l'espace euclidien, le temps et la composition (analytique ou global) des objets écologiques repérés dans l'espace-temps.

Dans cet hyper-espace que constitue le champ écologique, il y a des branches qui vont vers le global et d'autres vers l'analyse, c'est-à-dire qu'on va du particulier au général et du général au particulier, autrement dit qu'on compose et décompose selon les objectifs. On peut en outre gérer des DSEs obtenues après une interprétation synthétique se référant au champ écologique.

Pour toute information, on peut se placer à un niveau très global ou bien plus détaillé. Par exemple, on pourra indiquer le poids d'une espèce de plante au niveau d'une région, ou au niveau d'une prairie ; ensuite on pourra s'occuper du carbone dans ce poids, et à l'intérieur du carbone, du ^{14}C et des différents isotopes, etc. En détaillant, on augmente le nombre de paramètres potentiellement mesurables. En pratique, on ne peut pas tout analyser, on est conduit à faire des choix « réductionnistes ».

En résumé, toute donnée peut se rapporter à une position dans l'espace, à une date dans le temps et à la précision de l'analyse effectuée qui peut être très globale ou bien détaillée.

Dans la conception d'ÉCORDRE, on ne choisit pas *a priori* le niveau d'analyse. Cet avantage nous permet de recevoir toutes les données écologiques sans limites dans la gestion de l'information.

Grâce aux bases de données relationnelles, l'écologie a pu trouver pour la première fois un outil informatique adapté au développement *ad libitum* de la gestion des données relatives à cette science.

Il est nécessaire de présenter d'abord la place des données et de leur gestion dans la démarche scientifique et technique , le vocabulaire d'une base relationnelle illustré d'exemples spécifiques, la normalisation la plus courante des tableaux de données et enfin un exemple simple d'utilisation d'ÉCORDRE. .

Élaboration des connaissances et rôle des données

Par souci de clarté et de cohérence plusieurs étapes ont été définies dans le travail sur l'information écologique au niveau de notre laboratoire.

1re phase : l'acquisition des données

La première étape est une étape de réflexion pour savoir quelles sont les données qu'on va prendre dans le champ écologique. Elle dépend de la problématique, des connaissances antérieures, de la sélection d'objectifs précis et des moyens disponibles (voir Frontier, 1982 ; Scherrer, 1984 et sous presse).

La deuxième étape est la collecte des données dans le champ écologique en mettant en œuvre les techniques retenues. On établit des descriptifs techniques qui indiquent à chacun la manière de procéder pour acquérir certaines données et chaque spécialiste (et là on redevient spécialiste) explicite ses choix techniques dans le domaine correspondant (exemple : relevé de plantes, analyse de métaux lourds, prélèvement de lombriciens, etc.).

La troisième étape est l'analyse des prélevats collectés sous forme brute (enregistrement, prédonnées, prise, etc.) pour obtenir à terme des DICs.

La quatrième étape est la saisie des données. Un ensemble de données peut être saisi directement à partir des appareils électroniques (exemple : balance, centrale d'acquisition, etc.) ou à l'aide du clavier d'un ordinateur.

La cinquième étape est le contrôle des données. Dans cette étape on fait des tests de corrections et on crée des fichiers qui correspondent aux images des relations qui vont être chargées dans la base. On entre dans ÉCORDRE ces données initiales contrôlées (DICs).

Une fois que ces DICs sont entrées dans la base, elles sont « neutres » et bien sûr valent ce qu'elles valent, au plan de leur futur usage en interprétation (ce qui ne peut être discuté ici, car ceci varie en fonction de chaque objectif de recherche).

2e phase : la gestion des données

Une sixième étape est la création ou la mise à jour des relations (tableaux) existant dans la base et un dernier test de correction des données, test propre aux contraintes d'intégrité définies pour chaque relation. Cette

étape fait partie de l'analyse conceptuelle de la base ; chaque nouvelle donnée doit être intégrée conceptuellement et à ce niveau, il doit y avoir dialogue avec les spécialistes pour établir les nouvelles relations ou les nouvelles contraintes d'intégrité qui doivent être introduites.

3e phase : l'interprétation

Une septième étape est la synthèse ou l'interprétation des données. L'interprétation (comme le choix des techniques d'acquisition, d'ailleurs) est une étape riche en hypothèses. On choisit les hypothèses (exemple : assemblages, prédicats, suppositions, modèles, etc.) et les méthodes (exemple : analyses statistiques, modèles mathématiques, représentations graphiques, etc.) et, si les résultats ne nous satisfont pas, ou si on s'aperçoit qu'il nous manque des données, on peut revenir en arrière, les ajouter, et refaire l'interprétation avec des données adéquates.

Le rôle de la base permet de réorganiser à tout moment les données. A ce niveau on peut avoir plusieurs possibilités : l'interprétation nécessite de nouvelles données. Soit on obtient les données fournies par d'autres équipes de recherche (par télématique par exemple), soit encore on les réorganise à partir de données pré-existantes, soit enfin on fait de nouvelles acquisitions au terrain.

On peut organiser des sorties normalisées (par exemple, spécifiques aux lombriciens).

On peut faire subir aux données initiales (DIC) qui se trouvent déjà dans la base, différents types de traitements qui permettent de créer à partir de ces données, de nouvelles données. On peut avoir des données élaborées standards issues de tableaux de DICs, par exemple on pourra créer l'ensemble rapport C/N à partir des ensembles carbone organique et azote total, de façon à accéder plus rapidement à l'information.

4e phase : formalisation des connaissances

Les connaissances acquises dans ce processus général de développement de l'information écologique sortent du champ de la base de données relationnelle pour entrer dans celui d'une base de connaissances, impliquant la création d'une base de faits et d'une base de règles (qui composent la base de connaissances d'un système-expert).

La base de règles représente les connaissances spécifiques d'un domaine d'expertise indiquant quelles conséquences tirer, ou quelles actions accomplir lorsque telle situation est établie ou est à établir : il s'agit des règles que l'on se donne.

La base de faits accumule ce que l'on considère comme établi ou à établir ; les données gérées dans ÉCORDRE contribuent à celles-ci.

On peut en effet distinguer différentes catégories de connaissances. Les connaissances heuristiques (le plus souvent représentées en base de règles) produisent des hypothèses à partir de faits établis précédemment dans le but d'en produire de nouveaux par généralisation. Les connaissances de type procédural opèrent aussi des hypothèses à partir de faits, mais ces connaissances ont pour conséquences le choix d'un groupe de règles par rapport à un autre dans le but d'optimiser l'inférence, c'est-à-dire l'exécution des règles selon une logique bien précise. On appelle ces dernières connaissances, des méta-connaissances mises en œuvre par des méta-règles. Celles-ci traduisent de façon explicite les modèles d'interprétation des données.

Cette quatrième phase, en cours de développement dans notre laboratoire, complète le contexte informatique permettant l'utilisation en intelligence artificielle des données écologiques.

Vocabulaire et concepts de la base (figure 2)

RELATION : c'est le lien entre deux ou plusieurs ensembles d'informations. Elle est caractérisée par un nom. Dans SQL une relation s'appelle TABLE (tableau).

Une base de données relationnelle est, tout simplement, composée d'un grand nombre de tableaux, chacun étant défini par son nom et par la liste de ses colonnes ; du type de celui présenté ci-dessous.

Exemple 1 : tableau représentant les catégories et les sous-catégories écologiques des lombriciens (nom : CATECOLOG, deux ensembles : Code, Dénomination).

Pour désigner une relation (tableau) parmi toutes celles que contient la base de données, il suffit de donner son nom. Cependant, on la décrit souvent en ajoutant à son nom la liste de ses colonnes.

CATECOLOG	*Code*	*Dénomination*
	0	Inconnu
	10	Épigés
	12	Corticoles
	15	Straminicoles
	18	Pholéophiles
	20	Endogés
	21	Polyhumiques
	22	Épiendogés
	23	Carnivores
	24	Rhéophiles
	25	Mésohumiques
	28	Oligohumiques
	30	Anéciques
	32	Épianéciques
	35	Endoanéciques
	38	Euanéciques

Exemple 2 : voici 4 relations de plusieurs centaines de lignes dans ÉCORDRE qui ont comme nom : LIEU, COMMUNEFRA, LISTEFLORISTIQUE et FLORE_FOURNIER.

LIEU (Lieu, Pays, Département, Commune, Latitude, Longitude, Échelle)

COMMUNEFRA (Département, Commune, Nom)

LISTEFLORISTIQUE (Lieu, Placette, Année, Quantième, CodeEspèce, CodeFlore)

FLORE_FOURNIER (Code, Genre, Espèce)

ATTRIBUT : les attributs d'une relation sont les ensembles d'informations qui participent à cette relation (colonnes d'une relation caractérisées par un nom). Par exemple:
PLACETTE (NoLieu, CodePlacette, NomPlacette, Altitude, Précision).
Le numéro de Lieu, le code Placette, le Nom de la placette, l'Altitude de la placette (en mètres) et le degré de Précision de l'altitude sont les cinq attributs de la relation PLACETTE.
Les colonnes d'un tableau ne peuvent être subdivisées en sous-colonnes. Elles sont donc dites élémentaires.

RÈGLE GÉNÉRALE : à l'intérieur d'une même relation l'ordre des attributs est non significatif, c'est-à-dire que les attributs peuvent être dans n'importe quel ordre ; l'ordre des attributs n'exprime aucune particularité. Mais il faut nécessairement que tous aient des noms différents car tous les attributs doivent être distincts.

DOMAINE : c'est l'ensemble des valeurs possibles que peut prendre un attribut. Dans une opération, à l'intérieur d'une même relation on peut avoir deux attributs prenant une même valeur pourvu qu'elle soit à l'intérieur du domaine. Par exemple :
TAXONS_LOMBRICIEN (Code, Genre, Espèce, SousEspèce, Variété, CatEcolog, ResiSech, CertResi).
Dans ce cas, on admet qu'il y a quatre chaînes de caractères (genre, espèce, sous-espèce et variété) dont les domaines des attributs sont les mêmes (il y a des noms de sous-espèce qui sont des noms d'espèces, *Lumbricus rubellus rubellus*, etc.). Ainsi, CatEcolog, ResiSech, et CertResi qui sont numériques prendront des valeurs du domaine des nombres entiers. Dans le modèle relationnel on distingue bien attribut et domaine et on ne peut pas avoir deux attributs qui portent le même nom, alors que les valeurs dans le domaine peuvent être identiques.

DEGRÉ : le degré d'une relation est le nombre d'attributs de la relation. Par exemple :
ESSAI_AUTEUR (Réalisation, Essai, NumAuteur, NomAuteur) est une relation de degré quatre, ou encore, PAYS (Code Nom) est une relation binaire (degré deux).

N-UPLETS : ce sont les valeurs d'une ligne pour une relation de degré « n ». Pour une relation R(A,B,C), ils sont notés par convention (a,b,c). Par exemples :
FLORE_FOURNIER (Code, Genre, Espèce)... un triplet sera (5, Abies, alba)

PAYS (Code, Nom)... un duplet sera (250, France)

CARDINALITÉ : c'est la taille de la relation exprimée par le nombre de n-uplets présents dans la relation (nombre de lignes d'un tableau). Dans l'exemple suivant la cardinalité de la relation est de cinq (et son degré 2).

STADELOMBRIC	*Stade*	*Dénomination*
	0	Inconnu
	1	Cocon
	2	Juvénile
	3	Subadulte
	4	Adulte

Le temps de traitement dépend de la cardinalité et du degré des relations. Au niveau de la conception des relations, il faut chercher un compromis entre ces deux termes.

CLÉ : la clé correspond à un ou plusieurs attributs de la relation où la validité de l'unicité des n-uplets sera vérifiée. Pour cette raison, la connaissance de la clé permet d'identifier un n-uplet précis dans la relation considérée (= sélectionner une ligne).

Autrement dit, on entend par clé une colonne ou un ensemble de colonnes dont les attributs permettent d'identifier de façon unique chacune des lignes sans exception. Les colonnes constituant la clé sont soulignées dans la présentation littéraire.
L'unicité des n-uplets sera vérifiée seulement sur la clé mais cela entraîne *ipso facto* l'unicité de la totalité des n-uplets (= lignes) puisqu'ils se distinguent au moins sur les attributs de la clé.
Par exemple, dans la relation :
TAXONS_LOMBRICIEN (Code, Genre, Espèce, SousEspèce, Variété, CatEcolog),
la clé est Code et le n-uplet correspondant au code unique 2661 est : (2661, LUMBRICUS, RUBELLUS, FRIENDOIDES, TYPICA, 32). Il exprime le fait qu'il n'existe qu'un taxon lombricien qui a un certain code et qui simultanément correspond à un certain nom et appartient à une certaine catégorie écologique. Donc qu'il y a une combinaison unique de ces six valeurs qui indique que la relation est vraie, parce qu'elle correspond à un seul taxon.
Autre exemple :
COMMUNEFRA (CodeDépartement, CodeCommune, NomCommune).
Dans cette relation qui représente l'ensemble des communes de France, la clé ne peut pas être seulement le code de la Commune, pour la raison que le même code est utilisé dans plusieurs départements ; elle ne peut pas être non plus le code du département tout seul, parce qu'un même code de département est associé à différents communes. Par contre l'ensemble des deux va former la clé et la relation sera vraie car elle correspond à un seul nom de commune dans le département.
Dans ce dernier exemple, les deux colonnes CodeDépartement et CodeCommune sont soulignées, ce qui signifie que leurs valeurs conjointes identifient de façon unique chaque ligne.
Dans une relation il ne peut pas y avoir deux n-uplets identiques. Donc il ne peut y avoir qu'un seul n-uplet correspondant à une combinaison donnée des valeurs des attributs. Si cette opération est indiquée au système et si on essaie de rentrer deux n-uplets avec la même clé (on a oublié l'avoir déjà fait !), même si la suite est différente, le système refoulera l'opération, il ne peut y avoir redondance.
Il faut bien comprendre que la clé d'une relation n'est pas une clé d'accès, mais une clé de contrôle de l'unicité des n-uplets. Par contre, l'accès aux n-uplets d'une relation peut se réaliser par n'importe lequel des attributs de la relation.

Reste à dissiper un dernier malentendu possible. Les relations des bases de données relationnelles sont des tableaux au sens courant du terme, c'est-à-dire un ensemble de lignes et de colonnes (n-uplets), une liste d'éléments dans un ordre quelconque sans numéro d'ordre pour chaque ligne. Il ne s'agit pas d'un tableau au sens informatique du terme, qui est (pour un tableau à une dimension) une liste ordonnée de lignes, l'accès à une ligne déterminée se faisant exclusivement en donnant son numéro d'ordre (indice).

Au contraire, pour accéder, par exemple, aux informations concernant *Lumbricus terrestris* de la relation TAXONS_LOMBRICIEN, on ne se soucie absolument pas du numéro d'ordre de la ligne (qui n'est d'ailleurs pas connu ici). Il suffit d'indiquer que l'on veut obtenir la ligne (ou les lignes) dont les valeurs prises respectivement dans les colonnes Genre et Espèce sont *Lumbricus, terrestris.*

Les formes normales et la normalisation de tableaux

Si l'on veut obtenir une base de données relationnelle bien construite, il faut que les tableaux respectent certaines normes pour éviter au maximum la redondance des informations.

Les formes normales sont des contraintes auxquelles doivent se plier les relations pour permettre un traitement optimal de leurs contenus. En quelque sorte, elles « obligent » à un énoncé explicite des relations et constituent donc un cadre méthodologique.

Première forme normale

Une relation sera en première forme normale si tous les attributs (colonnes élémentaires) de cette relation sont en valeur unique, et non multiple, et si elle admet au moins une clé, donc si le contenu de la relation peut être représenté par un tableau bidimensionnel.

Dans la réalité on énonce parfois initialement des relations qui ne sont pas dans la première forme normale : il faut alors les normaliser pour les mettre en première forme normale.

La normalisation va consister en fait, à décomposer la relation initiale, au moins, en deux relations qui sont de degré inférieur à la relation de départ. Par exemple :
LIEU (NoLieu, CodePays, CodeDépartement, CodeCommune, Latitude, Longitude, Échelle, CodePlacette, SurfacePlacette, SurfacePrécision)

Dans cet exemple, la clé est le numéro de lieu (NoLieu) parce que pour chaque point d'étude repéré par la latitude et longitude dans l'écosphère le « NoLieu » est unique. Dans notre système il ne peut exister deux numéros de lieux différents avec les mêmes valeurs de latitude et longitude (et *vice versa*). Cependant, cette relation présente une anomalie qui résulte du fait que les trois derniers attributs sont à valeurs multiples, car on peut associer à chaque lieu (NoLieu) 1 à N placettes (par exemple, parcelles ayant reçu un traitement agronomique varié).

Pour éviter cette répétition à chaque placette d'un même lieu des valeurs identiques des sept premiers attributs, la normalisation de cette relation s'effectue de la manière suivante :

LIEU (NoLieu, CodePays, CodeDépartement, CodeCommune, Latitude, Longitude, Échelle)
PLACSURFACE (NoLieu, CodePlacette, SurfacePlacette, SurfacePrécision)

La redondance des informations en est réduite ; elle se limite à la répétition des colonnes identifiant de façon unique chaque ligne.

Deuxième forme normale

Une relation est en deuxième forme normale, si elle est déjà en première forme normale et de plus si l'ensemble des attributs ne formant pas la clé dépendent globalement des attributs qui forment la clé et non partiellement. On voit apparaître ici une nouvelle notion : la notion de dépendance. Par exemple :
CODEGÉOGRAPHIQUEFRA (CodeDépartement, CodeCommune, NomDépartement, NomCommune)

Dans cet exemple, correspondant à la codification géographique de la France, l'ensemble CodeDépartement et CodeCommune va former la clé, mais l'attribut NomDépartement a une dépendance partielle vis-à-vis de la clé. NomDépartement dépend seulement du CodeDépartement et non du CodeCommune. La normalisation de cette relation se fait en la coupant en deux :

DÉPARTEMENTFRA (CodeDépartement, NomDépartement)
COMMUNEFRA (CodeDépartement, CodeCommune, NomCommune)

Troisième forme normale

Une relation est en troisième forme normale, si elle est déjà en deuxième forme normale, et si la dépendance des attributs ne formant pas la clé, est une dépendance directe et non transitive de la clé.

Revenons à l'exemple de la relation LIEU et essayons d'y ajouter à la fin l'attribut NomPays.

LIEU (NoLieu, CodePays, CodeDépartement, CodeCommune, Latitude, Longitude, Échelle, NomPays)

La clé de cette relation est toujours le numéro de lieu. Rappelons que dans ÉCORDRE « NoLieu » est unique, puisque ne peut pas exister deux point d'étude dans l'écosphère avec le même numéro.

Cette relation est en première forme normale parce qu'aucun attribut n'est répétitif ; cette relation est aussi en deuxième forme normale pour la raison que si la clé est formée d'un seul attribut, la dépendance des autres attributs ne formant pas la clé vis-à-vis de la clé ne peut être que global et non partiel.

Dans l'exemple, un lieu (NoLieu) donné est associé à un seul code pays, à un seul code département, à un seul code commune, à une seule latitude, à une seule longitude et à un seul code d'échelle correspondant à la carte utilisée dans la détermination des coordonnées géographiques. Donc là, la dépendance est bien une dépendance directe, c'est-à-dire, que « NoLieu » fixe automatiquement les attributs décrits précédemment.

Mais, la dénomination en clair du nom du pays (NomPays) pour chaque lieu, dépend seulement du code pays (CodePays) : il s'agit d'une dépendance indirecte de la clé. Là encore il y a redondance d'information, parce que si plusieurs lieux sont associés au même pays, ce nom sera répété dans autant de n-uplets que de lieux existants, et on se retrouvera avec les mêmes inconvénients que si la relation n'était pas en deuxième forme normale.

Comment normaliser cette relation ?
LIEU (NoLieu, CodePays, CodeDépartement, CodeCommune, Latitude, Longitude, Échelle)
PAYS (CodePays, NomPays)

En conséquence, pour connaître le nom en clair des pays où se trouvent les différents lieux d'études il faudra accéder à deux relations. La première pour savoir les numéros des lieux et les codes des pays, la deuxième pour savoir les noms en clair des pays correspondants.

Exploitation de la base de données

Connaissant la composition de la base de données relationnelle, nous allons maintenant étudier un exemple simple de consultation de nos données, en illustrant nos exemples pour simplifier par seulement deux relations de la base.

TAXONS_LOMBRICIEN (Code, Genre, Espèce, SousEspèce, Variété, CatEcolog, ResiSech, CertResi)

CATECOLOG (Code1, Dénomination)

Question 1 : sélectionner à partir du tableau TAXONS_LOMBRICIEN les noms de tous les taxons appartenant à la catégorie écologique des épigées, plus précisément on pourra demander de sélectionner les colonnes genre, espèce et sous-espèce lorsque le contenu de la colonne « CatEcolog » est compris entre 10 et 18. Cette requête traduite en langage SQL est la suivante :

```
SELECT GENRE, ESPÈCE, SOUS-ESPÈCE
FROM TAXONS_LOMBRICIEN
WHERE CATECOLOG BETWEEN 10 AND 18
```

Question 2 : sélectionner par ordre alphabétique toutes les colonnes du tableau TAXONS_LOMBRICIEN où les sous-catégories écologiques sont définies.

```
SELECT( *)
FROM TAXONS_LOMBRICIEN
WHERE CATECOLOG > 10
AND CATECOLOG NOT IN(10,20,30)
ORDER BY GENRE, ESPECE, SOUSESPECE, VARIETE
```

Lorsqu'on traite toutes les colonnes d'un tableau, on se contente de mettre un astérisque (*) au lieu de taper tous les noms de colonne.

Question 3 : on veut connaître le nombre de taxons (sous-espèces et variétés) à l'intérieur de chaque espèce lombricien.

```
SELECT GENRE,ESPECE,COUNT(*)
FROM TAXONS_LOMBRICIEN
GROUP BY GENRE,ESPECE
```

Dans cet exemple la clause GROUP BY permet de regrouper les taxons lombriciens ayant le même nom de Genre et Espèce, puis d'effectuer le comptage du nombre d'éléments à l'intérieur de chaque groupe.

Question 4 : on veut connaître la dénomination en clair de la catégorie écologique pour chaque taxon, en sélectionnant aussi son nom en entier.

```
SELECT CODE, GENRE, ESPECE, SOUSESPECE,
VARIETE, DENOMINATION
FROM TAXON_LOMBRICIEN, CATECOLOG
WHERE CATECOLOG=CODE1
```

Dans cet exemple on a combiné les deux tableaux ayant une colonne en commun (CatEcolog avec Code1). Ces colonnes permettent d'établir la relation entre les deux tableaux.

Avec ces exemples on peut constater que l'accès aux données est simple. Il suffit en effet de définir précisément ce que l'on souhaite obtenir, sans décrire la procédure que l'ordinateur utilisera. Le SGBDR se chargera d'assurer une navigation automatique dans la base de données. Les requêtes s'expriment dans un langage informatique simple, facile à comprendre et donc à acquérir (figure 3, ci-contre).

Les caractéristiques de cet outil font qu'il peut être utilisé aussi bien par l'utilisateur final (non-informaticien) que par des informaticiens.

Ainsi, les bases de données relationnelles ont deux caractéristiques originales :

a) Elles sont peu procédurales, en ce sens que la définition de ce que l'on veut obtenir remplace la description détaillée d'une procédure que l'ordinateur devra suivre.

b) Elles sont proches de l'utilisateur et d'emploi facile, donc conviviales.

Tout ceci est possible si l'organisation rationnelle des données est assurée par l'adoption d'un schéma conceptuel unique qui dans le cas d'ÉCORDRE doit être cohérent avec l'écosphère et par le processus de normalisation des relations, dont les règles doivent absolument être respectées.

Les moyens techniques actuels d'ÉCORDRE, et leurs coûts

La base de données écologique ÉCORDRE utilise les moyens techniques mis à disposition par le Centre national universitaire sud de calcul (CNUSC) de Montpellier. Ce centre s'intègre dans le réseau de moyens informatiques mis en place par le ministère de l'Éducation nationale afin de servir les besoins de la recherche et de l'enseignement de toutes disciplines.

L'ordinateur de base

Le CNUSC (1989) dispose d'un ordinateur de grande puissance, un IBM 3090-400 VF avec 64 millions d'octets de mémoire centrale, 32 canaux ; sa puissance de traitement est de 30 millions d'instructions par seconde, et deux

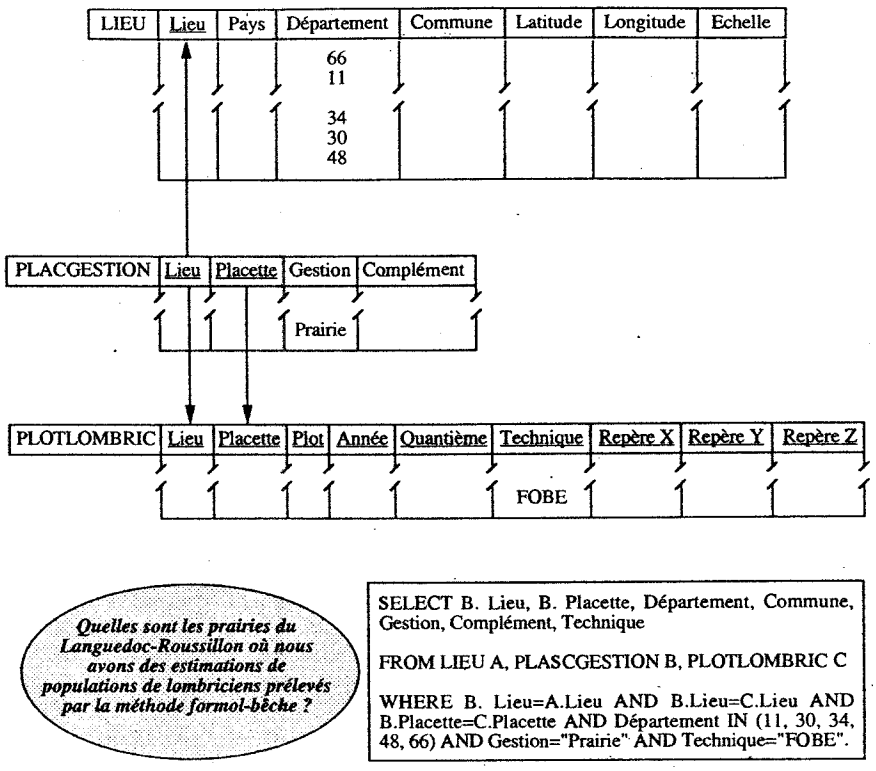

Fig. 3. Exemple de consultation de la base de données.
Les requêtes s'expriment dans un langage informatique simple, facile à acquérir.
Ainsi, à la question "quelles sont les prairies du Languedoc, etc.", les ordres successifs et imbriqués du type SELECT, FROM, WHERE, DEPARTMENT, IN(HERAULT, GARD), etc., permettent la sélection successive des lieux d'étude de la région puis des seules prairies régionales où la technique "formol-bêche" a fourni des données.

unités centrales de calcul vectoriel. Les unités de disques magnétiques permettent l'accès direct à 80 milliards de caractères. Cette configuration de base est complétée par diverses unités périphériques et par les moyens techniques de transmission de données à distance.

Les logiciels de base

SQL/DS (Structured Query Language/Data System) est le système de gestion de bases de données d'ÉCORDRE utilisant le langage SQL. Il fonctionne sous VM/CMS et supporte le modèle relationnel.

VM (Virtual Machine) est surtout un système interactif. Il comprend le système d'exploitation CMS (Conversational Monitor System) totalement conversationnel, c'est-à-dire, qu'il est l'interlocuteur direct de l'utilisateur avec son terminal. Pour le traitement par « lot » (gros calcul) VM/CMS n'est pas actuellement bien adapté à ce mode de travail ; c'est pourquoi actuellement nous faisons de tels traitements « hors SQL/DS » sous MVS (Multiple Virtual Storage) qui est orienté pour ce type de traitements. A MVS sont associés TSO et SPF.

TSO (Time Sharing Option) est le sous-système « temps partagé » standard d'IBM. Il est doté d'un éditeur ligne à ligne et donne accès à toutes les fonctions de l'OS (Operating System). Il est installé, au CNUSC, avec le produit SPF (System Productivity Facility) qui apporte des facilités importantes pour les utilisateurs d'écrans fonctionnant en mode pleine page (« full-screen »).

Par ailleurs, SPF aide surtout à la gestion des fichiers (données, langage source, bibliothèque, etc.). Il travaille directement sur les fichiers OS résidant sur disque magnétique.

En outre, nous disposons du logiciel QMF d'IBM qui permet de formuler des requêtes portant sur des données organisées en mode relationnel et gérées par le système SQL/DS.

A partir de QMF, l'utilisateur dispose d'une série de menus qui le guident dans le

choix et l'exécution de diverses tâches. Les commandes de QMF sont attachées à des clés de fonction, ce qui permet un apprentissage rapide et limite les risques d'erreur.

MVS et VM sont des ensembles de services complémentaires. Il existe divers liens qui permettent de passer facilement d'un système à l'autre. Notamment la transposition d'ÉCORDRE depuis SQL/DS vers le logiciel DB-2 (fonctionnant avec SQL sous TSO) sera prochainement effectuée.

Toutes les fonctions d'ÉCORDRE (création et modification permanente de relations, interrogations et contrôles d'accès) sont activées au moyen d'un langage unique appelé SQL. On accède au système SQL/DS (ou à DB-2) soit directement avec ce langage à partir d'un terminal, soit en intégrant des ordres SQL dans des programmes d'application (par exemple en FORTRAN, PL/1...).

Le langage SQL s'impose comme une norme en matière de bases de données relationnelles. En effet, il est utilisé par plusieurs SGBD relationnels, et est en cours de normalisation par l'Institut américain de normalisation (ANSI). Ce langage est utilisé sur SQL/DS et sur une version plus performante, DB2. (Ne pas confondre Data base 2/DB2 avec DBASE2, logiciel de gestion de fichiers sur micro-ordinateur). « La confusion est d'autant plus risquée que ce dernier se prétend base de données relationnelles alors qu'il n'est qu'un gestionnaire de fichiers (certes très élaboré) et en aucun cas un SGBD » (Jault, 1986).

Le fait de travailler sur gros systèmes offre l'avantage de disposer:
1. de logiciels très performants et très nombreux non seulement pour le fonctionnement d'ÉCORDRE, mais aussi pour l'interprétation par divers moyens statistiques des ensembles de données obtenues sur requête,
2. de moyens d'édition graphique les plus sophistiqués,
3. de moyens télématiques de liaisons avec les réseaux nationaux (tels TRANSPAC), le simple MINITEL ou les liaisons internationales par satellites et câbles (EARN, BITNET, NORTHNET...).

Quel est le prix de ce service ? Ce colloque étant international et la situation à l'intérieur de chaque pays différente, il est difficile de répondre, par exemple, les modes de facturation étant radicalement différents entre les pays de la CEE. Toutefois, il faut indiquer que le chercheur travaillant plein temps avec ce moyen doit disposer d'un ordinateur personnel modeste (14 à 21 000 F ou 2 à 3 000 écus) relié au gros système informatique.

Nous dépensons par an dans nos conditions à peu près le double de cette somme pour disposer de toutes les facilités offertes, sans avoir évidemment à renouveler les grosses machines ni acheter les nouveaux logiciels qui sont implantés au CNUSC à chaque étape du progrès technique. On voit que ceci correspond à un coût modeste... et que l'écologie et ses corollaires appliqués, environique et agronomie, peuvent devenir transdisciplinaires à ce niveau pour un coût réduit, et ceci sans tenir compte du fait que l'on peut travailler en frais partagés.

Développements et rationalité

Les bases de données relationnelles ouvrent de nouvelles voies à l'écologie : le traitement et l'accès aux données peuvent maintenant être maîtrisés par l'écologue. L'utilisateur non-informaticien peut interroger toutes les données antérieurement acquises. De plus, les données étant disponibles, il peut définir leur présentation, les organiser pour des applications avec des logiciels statistiques, mathématiques, graphiques, etc. Enfin, et c'est essentiel, la mise en commun de la partie objective, les DICs, est effectivement possible et aisée.

Dans cette optique, le rôle des informaticiens serait d'assurer la cohérence globale du système, de traiter les problèmes complexes en développant des applications et d'assurer la formation et le suivi des utilisateurs employant ces nouveaux outils.

Au niveau conceptuel, ÉCORDRE peut aborder tout le champ écologique. On a testé avec succès les limites de son utilisation et de son application à partir de nos propres objectifs d'étude mais aussi avec des données issues d'approches qui nous étaient initialement étrangères.

L'évolution d'ÉCORDRE doit se faire à la fois vers une simplification permettant une bonne vulgarisation scientifique et vers une sophistication accrue du système répondant aux désirs des spécialisations pour lesquelles elle doit être conçue. L'avenir d'ÉCORDRE passe donc en premier lieu par la structuration et la formalisation (dynamique) de nos connaissances en travail d'équipe pour parvenir à un système apte à se développer au service de tous. Au stade actuel ce sont les apports sec-

toriels de chacun qui sont pris en charge, « traduits » dans le schéma conceptuel unique, puis chaque équipe retrouve son autonomie de chargement et d'accès pour le domaine qu'elle maîtrise et met en commun dans ÉCORDRE.

Coordonner les efforts, accumuler des données non redondantes, faire circuler l'information écologique : voilà l'un des buts principaux de la constitution de la base de données relationnelle « ÉCORDRE ».

L'approche relationnelle apporte en écologie de très nombreux avantages aux usagers d'un système de base de données notamment par :
a) la description simplifiée des données. Elle est plus proche du raisonnement humain, que celui hiérarchique ou en réseau ;
b) l'indépendance entre les données et les programmes de recherche. Toute adjonction ou suppression de données se fera par modification des relations sans la nécessité de toucher les programmes d'applications ;
c) le système de gestion de bases de données relationnel agit « par déduction » à partir de l'analyse des requêtes. Il permet de répondre à toute question formulée correctement faisant appel à une logique déductive ; le contenu des questions détermine quels sont les chemins d'accès automatiques pour extraire de la base les données souhaitées ;
d) l'indépendance des informations par rapport aux structures physiques de stockage. Les liens ou pointeurs ne sont plus fixés une fois pour toutes dans le SGBD mais ils sont fabriqués dynamiquement selon les besoins, donc pas de « navigation » dans le labyrinthe des pointeurs, ni d'obligation de connaître *a priori* la structure détaillée de la base ;
e) des langages de haut niveau pour décrire et manipuler les données. Cela permet l'utilisation des opérateurs logiques de la théorie des ensembles dans le langage d'interrogation pour que le système soit capable de comprendre correctement une question ;
f) des logiciels peuvent fonctionner dans des environnements différents. Cela signifie que, sur les très gros ordinateurs, on trouvera toutes les fonctions décrites antérieurement, alors que, sur des micro-ordinateurs, on trouvera des versions simplifiées ;
g) la standardisation du langage conceptuel, ce qui ne peut qu'avoir une « rétroaction positive » sur la formalisation même des objectifs de recherche.

La circulation de l'information et les systèmes

Au terme de cet article introductif il est nécessaire d'éclairer, même schématiquement l'intérêt du travail auquel oblige une base de données relationnelles : elle exige une formalisation complète et une description exhaustive des données. En contrepartie, elle permet une circulation des informations et une mise en commun « pluridisciplinaire ».

Évidemment l'organisation d'ÉCORDRE se fait avec les logiciels et matériels mis en œuvre. Ces moyens en matériels et logiciels n'ont pas été développés pour l'écologie et il n'est pas usuellement du ressort de l'écologue d'influencer ceux-ci. Par contre, il est du ressort de l'écologue de développer sa discipline par ces moyens, enfin disponibles et adéquats à l'objet de la science qu'il pratique. La recherche écologique et informatique présentée ici s'inscrit dans un contexte d'évolution rapide et la base ÉCORDRE ne saurait être appréciée à partir de points de vue sectoriels ou statiques.

Sectoriellement, dossier par dossier, approche particulière de chaque spécialiste prise isolément, l'intérêt de cette base est nul. Elle ne prend toute sa valeur que par l'échange trans-sectoriel, international si nécessaire, de toutes sortes de données écologiques objectives.

On peut constater que pour l'instant seuls de « gros » logiciels implantés sur des « grosses machines » présentent toutes les propriétés souhaitables et nécessaires aux bases de données pour leur plein usage : contrôle d'intégrité, confidentialité, interconnexion télématique à des réseaux (tel présentement EARN, raccordé à BITNET) mais aussi à des micro-ordinateurs où une part importante du travail peut être faite. C'est même sur de tels appareils (Apple IIe et IBM-XT/AT) que la saisie des données, leurs premiers contrôles sont assurés et transmis après connexion (Soto, 1988). On peut également s'aider de gestionnaires de fichiers élaborés tel DBASE3+. Par interfaçage avec une base de faits, incluse elle-même dans une base de connaissances, ces données peuvent enfin contribuer à un dialogue formalisé en systèmes-experts. Les SGBDR jouent alors un rôle stratégique entre la gestion des données et la gestion des connaissances.

Tant les logiciels que les matériels sont dans ce domaine en évolution rapide et l'intérêt de telles bases ne peut pas être jugé à partir des contraintes actuelles en matériels, connexions, logiciels... ces éléments techniques, rapidement obsolètes, se remplaçant par des matériels et logiciels moins exigeants. Il serait notamment à ce niveau incompréhensible d'opposer « grosse informatique » à « petite », le « gros » actuel étant destiné à devenir petit et de toute façon l'un étant complémentaire de l'autre.

De la même façon, la base de données, tout comme le dénombrement d'oiseaux ou la pesée d'une récolte, n'est pas une fin en soi, mais un moyen d'élaboration dans un contexte intégré en rapide évolution, notamment en écologie par le développement de l'aide à la modélisation actuellement en cours, dont on peut voir une prémisse *in* Gouze et Vignard (1984) et Pave (1987).

Il ne s'agit pas au niveau de la base de données d'un problème de traitement biométrique de données, qui dépend des objectifs spécifiques d'interprétation, mais de conceptualisation de la logique d'acquisition des données par l'écologue et de mise en forme des ensembles de données et de formalisation des relations par l'informaticien.

Gérant des données, la base de données ÉCORDRE n'a aucun rôle vis-à-vis de la qualité intrinsèque de chaque donnée, ni des interprétations que l'on fera sur ces données. On peut y gérer de mauvaises mesures, ou également après avoir sélectionné des séries de données correctes, tirer des conclusions erronées sur ces ensembles. Ceci n'est pas du ressort d'une SGBD, mais bien de l'intelligence naturelle.

Par contre, l'utilisation d'ÉCORDRE rend possible le partage des connaissances entre spécialistes en langage naturel, ce qui pose aux écologues le problèmes de partager les mêmes concepts avec la même sémantique. En rendant possible une écologie effectivement et opérationnellement globale ÉCORDRE renvoie les écologues à leurs responsabilités.

Si le choix actuel des logiciels et matériels peut effectivement limiter ou faciliter le travail de l'écologue, dès à présent il est possible par celui-ci de « dé-sectorialiser » ses pratiques, et, tout en restant un bon spécialiste d'un domaine restreint appliqué ou fondamental, de contribuer à la croissance de nos connaissances objectives sur l'écosphère. Ainsi les problèmes généraux, tels ceux soulevés par le Programme international géosphère-biosphère (PIGB), seront-ils alimentés en données à toute échelle d'espace-temps et de composition analytique.

Comme le dit joliment Claude Jault (1986), les SGBDR et les bases qu'elles permettent donnent le « libre accès à l'information » ; si les écologues savent prendre cette mesure, ils pourront valoriser une masse considérable de données de façon infiniment plus souple et efficace que présentement.

En tout cas, la stratégie générale de recherche s'en trouve profondément modifiée (figure 4).

Devant une nouvelle problématique environnementale ou agronomique, les objectifs de recherche débuteront non plus comme au stade I, primitif, par l'acquisition de données mais par l'exploitation des données préexistantes. Ce stade II, de gestion

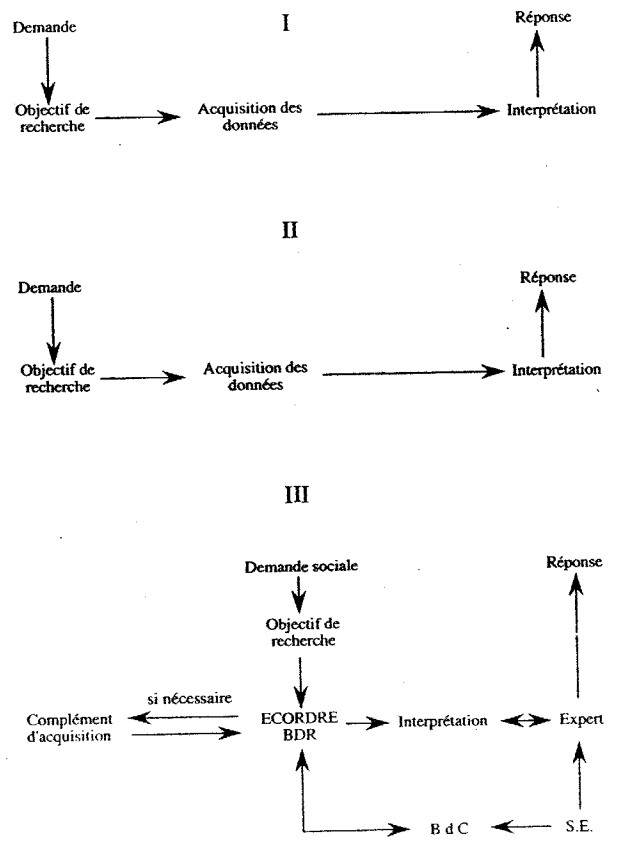

Fig. 4. Modification de la stratégie générale de recherche induite par les SGDBR : exemple d'évolution de mise en œuvre des connaissances en fonction des moyens de l'intelligence artificielle. I : stade primitif sans I.A., II : stade avec gestion par BDR, III : stade avec système-expert interfacé avec une BDR.

des données, permet ainsi d'optimiser l'éventuel échantillonnage qu'il faudra conduire. La base de données ÉCORDRE, ne marque qu'une étape vers un stade III, où l'intelligence artificielle prendra en compte la connaissance des experts qui deviendra électroniquement accessible devant tout nouvel objectif de recherche ou demande sociale dans une base de connaissances (à système-expert orienté objet).

Références

Bouché M.-B., 1975. « Discussion d'écologie. I) Introduction. II) Obtention de données écologiques et uniformisation spatiale », *Bull. écol.*, 6 (1) : 23-32.

Bouché M.-B., 1980. « Discussion d'écologie. IV. L'écologie », *Doc. labo. zooécol. sol*, 6 : 1-20.

Bouché M.-B., Aupetit P., Philippe M.-A., 1988. « Étude agroenvironnementale intégrée du Marais Poitevin », *Doc. pédozool.*, 1 (3) : 1-126.

Bouché M.-B., sous presse. *Écologie opérationnelle assistée par ordinateur*, 2 tomes, éd. Masson, Paris.

Delobel C., Adiba M., 1982. *Bases de données et systèmes relationnels*, Dunod, Paris, 1-449.

Escoufier Y., 1987. *Méthodes statistiques basées sur un usage intensif de l'ordinateur*, rapp. techn. 8702, unité Biométrie, ENSAM, INRA/USTL, Montpellier, 1-16.

Frontier S., 1982. *Stratégies d'échantillonnage en écologie.* Masson, Paris, 1-494.

Gardarin G., Valduriez P., 1985. *Bases de données relationnelles – analyse et comparaison des systèmes*, éd. Eyrolles, Paris.

Gouze J.-L., Vignar P., 1984. « EDORA : un système intelligent d'aide à la modélisation », *AMSE Review*, 2 (3) : 9-26.

Haeckel H., 1866. *Generelle Morphologie der Organismen*, 2 tomes, G. Reiner (ed.), Berlin, tome 1 : 1-463, tome 2 : 1-574.

Jault C., 1986. *Les bases de données relationnelles – ou le libre accès aux informations*, Les éditions d'organisation, Paris, 1-109.

Miranda S., Busta J.-M., 1987. *Introduction aux bases de données*, éd. Eyrolles, Paris.

Pave A., 1987. « Systèmes experts : application à la modélisation en biologie », *MATAPLI*, 11 : 5-19.

Scherrer B., 1984. *Biostatistique*, Morin (édit.), Montréal, 1-850.

Scherrer B., sous presse. *Application et optimisation de l'échantillonnage probabiliste en écologie.*

Soto P., 1985. « Données écologiques et système relationnel de bases de données – notions élémentaires pour la constitution de la base de données ÉCORDRE », *publication ZES-INRA/CEPE-CNRS*, n° 166.

Soto P., 1988. « Transfert de fichiers IBM-APPLE », *Electron. Applic.*, 58 : 61-64.

Spandau L., 1989. *Angewandte Oekosystemforschung im Nationalpark Berchtesgaden*, Ed. Nationalpark Berchtergaden, Forschungsbericht 16., 2e édition, ISBN 3-922325-15-17, 1-88.

Trabalka J.R., Reichle R.E., 1983. *The changing carbon cycle. A global analyzis*, Springer Verlag NY, 1-592.

Remerciements

Les auteurs tiennent à remercier R. Gimilio (ministère de l'Environnement), P. Aupetit, M.-A. Philippe, R. Aliaga, J.-Cl. Lefeuvre qui ont permis, par la réunion d'une masse importante de données écologiques, de toute nature (pédologiques, botaniques, entomologiques, agrotechniques, hydrauliques, pédoclimatiques, géohydrologiques...) de tester la faisabilité d'une base écologique exhaustive. Évidemment, ces remerciements s'adressent également à tous ceux qui, trop nombreux pour être nommés, ont participé aux collectes et analyses de ces données.

Ils adressent également un remerciement particulier à G. Long pour sa lecture critique et constructive du manuscrit.

Cette recherche a bénéficié des aides du Marais Poitevin et du ministère de l'Environnement (contrat n° 86-336).