III TERMINOLOGIE ET CONCEPTS MIS EN ŒUVRE

Notre objectif principal est la mise au point de systèmes de détermination (classification et identification) prenant en compte de façon naturelle la diversité, l'interdépendance et la variabilité des caractères observés, et s'accommodant autant que possible des données manquantes si fréquentes dans les domaines biologiques. De cet objectif découle la révision apportée des concepts fondamentaux intervenant dans la classification des êtres vivants (au sens large).

Quiconque s'est intéressé aux productions de la nature, dont les êtres vivants sont les représentants les plus évidents, a perçu que, sous une apparence de diversité et de complexité extrême, se cachait en fait une sorte de plan d'ensemble, une régularité, une logique, un déterminisme, etc.. Les naturalistes sont arrivés à la notion de “système de la nature”, d'un ordre global dans lequel les différents individus se trouvent virtuellement regroupés en “classes”, et ceci à différents niveaux ou “catégories” (Espèces, Genres, Familles, etc.).

Dans ce chapitre, nous exposons notre point de vue sur les concepts utilisés en biologie par rapport à ceux utilisés chez une grande majorité de mathématiciens et philosophes afin de permettre une meilleure compréhension du domaine biologique qui nous intéresse ici.

3.1 Extension et compréhension

3.1.1 L’extension

Deux points de vue de l'**extension** sont possibles selon le sujet d'étude et l'observateur :

3.1.1.1 Point de vue du philosophe et du mathématicien

Ces personnes s'intéressent aux produits de l'activité humaine, c’est pourquoi l'extension est une notion dépendant de la compréhension : on parle d'extension d'un concept par rapport à sa compréhension. Le sujet d'étude est la compréhension (ou intension) à partir de laquelle on cherche une extension.

Pour ces observateurs, l'extension est la sphère plus ou moins grande des êtres ou des espèces auxquels s'applique une condition exprimée par un ou plusieurs attributs. La pensée organise spontanément les choses en classes (ou concepts), d'après leurs caractères communs, et forme les classes les plus étendues en éliminant de plus en plus de caractères. Aussi dit-on que plus l'extension croît, plus la compréhension se restreint.

Par exemple, tant que l'on ne connaissait pas de cygnes noirs, le concept cygne comportait dans sa compréhension l'attribut nécessaire blanc. Son extension comportait tous les cygnes connus (qui étaient tous blancs). Après la découverte de cygnes noirs, le concept cygne a perdu en compréhension l'attribut blanc (qui n'était plus nécessaire désormais) et a gagné en extension les nouveaux cygnes découverts.

L'extension peut être qualifiée de psychique ou abstraite car elle dépend d'une **définition** préalable des classes dans un univers de description donné (PClass et PConcepts [Sutcliffe, 1993]). Dans ce contexte, il peut arriver que l'extension d'un concept soit vide : le concept de licorne par exemple [Sowa, 1984].

En résumé, la classe traduit l'extension d'un concept, elle n'existe que lorsqu'elle a été explicitée : elle constitue l'ensemble des individus qui satisfont à la condition exprimée par son concept dans un univers de description donné.

3.1.1.2 Point du vue des biologistes et des naturalistes

Ces personnes s'intéressent plus aux produits de la nature, la compréhension n'a d'intérêt que si elle traduit une extension concrète. Ainsi l'extension peut être une notion indépendante de la compréhension. Le sujet d'étude est l'extension et l'on considère que les classes préexistent avant même de recevoir une définition. Par exemple, un chien qui passe dans la rue existe indépendamment de sa définition. Chaque classe correspond à une certaine extension (ou couverture) concrète et naturelle dont on veut tirer un enseignement (une compréhension des classes naturelles).

Dans un premier temps, on se contente donc de décrire l'extension ou le contenu (les individus) de la classe sous forme de **descriptions**. Le fait de décrire est déjà en lui-même un enseignement pour le descripteur (celui qui décrit). Il est amené à interpréter des observations multiples et hétérogènes afin de produire des généralisations “de bas niveau” (en ne mesurant que certaines propriétés et en en ignorant d'autres) supposées exactes et dignes de confiance.

Les descriptions doivent tenir compte de la diversité biologique exprimée par la couverture de la classe[1].

[1] Chaque objet de l'extension possède un statut avec différentes modalités que le descripteur peut être amené à envisager : ces informations sont...

On s'efforcera donc de traduire cette diversité dans les descriptions afin de recueillir toute la richesse et la diversité des individus du domaine biologique bien délimité. En effet, chaque individu décrit est un élément représentatif de la classe et a pour extension lui-même : notre approche privilégie ainsi la multiplication des descriptions d'individus appartenant à une même classe (avec des valeurs comprises dans un intervalle de doute ou d'imprécision) plutôt qu'une seule description de “concept” dont l'extension est l'ensemble des individus qui vérifient l'intervalle de variation des valeurs de la description. Cette deuxième approche est celle adoptée par [Vignes, 1991]. La formalisation sous forme d'objets symboliques [Diday, 1987] présentée au chapitre 5 met aussi en lumière cette nuance.

Ainsi comprises, les descriptions forment une base de travail exhaustive pour le traitement et constituent déjà un résultat important pour la transmission du savoir humain.

Dans un deuxième temps, le descripteur cherche à mieux comprendre ses descriptions individuelles.

3.1.2 La compréhension

La compréhension ou **l’intension** est l'ensemble des caractères ou propriétés contenus dans un concept et qui permettent de le définir [Arnauld et Nicole, 1662].

Ainsi vertébré a comme compréhension : animal qui a des vertèbres et comme extension Mammifères, Oiseaux, Batraciens, Reptiles, Poissons. On remarque

---

évolutives : elles changent au fur et à mesure que l'univers dans lequel elles sont utilisées se modifie, ce qui entraine des problèmes de cohérence.

certaines ou incertaines : il peut résider ou non un doute quant à la vérité des informations. Ce doute peut être dû à un manque de confiance dans la source de l'information ou au fait que celle-ci est difficilement accessible à la vérification.

valides ou périmées : elles n'ont pas toujours de valeurs universelles et peuvent être remises en question dans l'avenir.

typiques ou exceptionnelles : chaque objet, qu'il soit considéré comme central ou marginal, porte sa propre originalité et fait ainsi partie intégrante de la couverture de la classe. A ce titre, les cas exceptionnels ont autant d'importance que les cas typiques en biologie, c'est pourquoi les biologistes affirment que dans la nature, l'exception est la règle. Il ne s'agit donc pas de les supprimer !

complètes ou incomplètes : la connaissance disponible sur un objet est généralement incomplète parce qu'elle est implicite et donc généralement oubliée dans la représentation, ou encore parce qu'elle n'est pas encore connue ou qu'elle est difficile à transmettre.

significatives ou fictives : des informations ont un sens pour expliciter des règles de connaissances alors que d'autres ne sont utilisées que pour structurer le domaine de description.

qu'un concept s'étend à d'autant plus d'êtres qu'il réunit moins de caractères comme le montre la figure 3.1 :

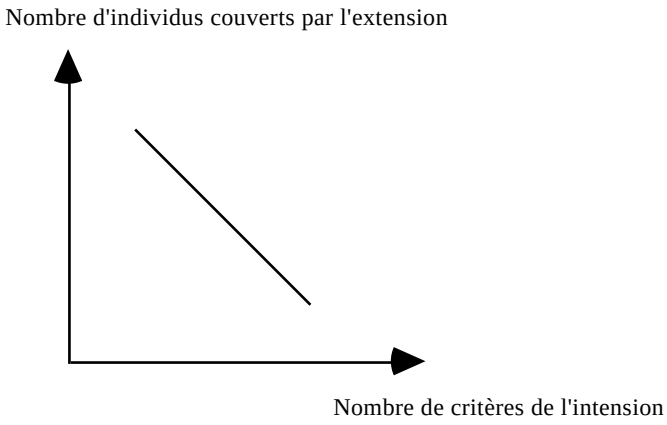

Fig. 3.1 : Rapport entre l'extension et l'intension

Ainsi la compréhension et l'extension sont en raison inverse l'une de l'autre. Animal a une extension plus stricte et une intension plus forte que Vivant, Vertébré a plus de compréhension qu'Animal et Mammifère plus que Vertébré.

D'après cette définition de la compréhension, l'intension est la partie signifiante du concept. Elle énonce certaines propriétés (supposées vraies) permettant de valider des connaissances du domaine. Elle exprime les conditions nécessaires et/ou suffisantes[2] d'appartenance d'un individu au concept.

Néanmoins, la question de savoir si l'intension prime sur l'extension est un problème philosophique qui dépend de l'observateur et de ses préoccupations. En effet, l'intension précède-t-elle l'extension dans la vision que possède l'utilisateur du domaine étudié ? Il semble naturel que la réponse soit oui pour un psychologue et un mathématicien : la définition ne peut être faite que par l'homme ! Ce à quoi le naturaliste objectera en posant la question suivante : est-ce que les animaux qui existaient au secondaire et que l'on a appelés dinosaures par la suite faisaient partie d'une classe ? Est-ce qu'ils existaient avant que le concept n'apparaisse ? Il semble aussi que oui ! Nous sommes ainsi en présence d'une dualité de point de vue résumée par la figure 3.2 :

[2] Une définition n'est pas forcément nécessaire et suffisante du fait que des personnes différentes ont rarement la même compréhension d'un même phénomène naturel : voir plus loin les définitions des intensions minimales, strictes et généralisées des concepts au § 3.2.2.

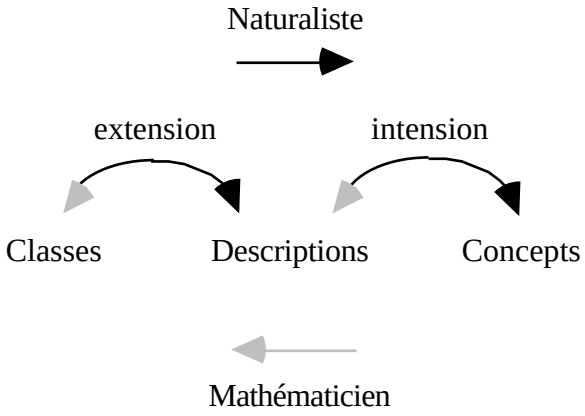

Fig. 3.2 : Mathématiciens et Naturalistes, deux points de vue différents des concepts

Ce schéma un peu caricatural demande un approfondissement dans l'étude des raisonnements différents qu'employent ces deux catégories de personnes :

Le mathématicien a l'habitude d'utiliser un *raisonnement démonstratif* basé sur une valeur de vérité d'une propriété, exprimé par des règles rigoureuses et clarifiées par la logique formelle. Ce type de raisonnement est sûr, à l'abri des controverses et définitif. Inversement, le naturaliste émet des hypothèses qu'il justifie par un *raisonnement plausible*. Ce dernier est hasardeux, il peut être controversé et il est provisoire [Pólya, 1958]. Néanmoins, il est capable de conduire à des connaissances essentiellement nouvelles sur le monde qui nous entoure. C'est pourquoi ces deux types de raisonnement ne sont pas contradictoires comme pourrait le laisser penser le schéma ci-dessus : ils se complètent.

Dans le raisonnement rigoureux, l'essentiel est de distinguer une preuve d'une présomption, une démonstration valable d'une tentative qui a échoué : c'est le savoir démontrer du mathématicien qui prouve la validité de ses concepts. Dans le raisonnement plausible, l'essentiel est de distinguer une présomption d'une autre, l'une plus raisonnable que l'autre : c'est le savoir pressentir du naturaliste qui suggère des classes fiables. Le mathématicien doit donc être capable de deviner une règle ou un théorème mathématique avant de le démontrer, de même que le naturaliste devrait être capable de prouver le bien fondé de ses règles de classification. Il est donc faux d'opposer la démarche d'un naturaliste à celle d'un mathématicien comme voudrait le laisser paraître notre monde contemporain assoiffé de démonstrations et de certitudes.

Dans cette thèse, nous nous plaçons d'abord du point de vue du naturaliste qui considère l'extension comme son sujet d'étude.

Le premier principe de la robustesse est effectivement de bien comprendre le domaine étudié, c'est-à-dire ici d'adopter la terminologie des systématiciens.

Partant d'une classe (ensemble d'individus) dont le contenu (l'extension) est sa couverture, le naturaliste observe et crée le nom de cette classe puis... la définit (en intension) afin de créer le concept associé. Cette démarche constate d'abord la classe avant de procéder à une conceptualisation de ses individus.

Certains mathématiciens comme Euler (1707-1783) ou Laplace (1749-1827) prônent ce point de vue basé sur l'observation. Néanmoins, contrairement au naturaliste pour qui l'observation est le critère le plus élevé (la vérification effectuée dans de nombreux cas bien choisis est la seule méthode de confirmation d'une loi hypothétique dans les sciences naturelles), le mathématicien va plus loin dans son domaine en affirmant que si nombreuses que puissent être des vérifications expérimentales, elles ne suffisent pas à démontrer que la loi supposée est vraie. Cette bifurcation de point de vue tient donc à la nature du domaine étudié (réfutabilité des hypothèses) : la récurrence et la périodicité ne se rencontrent pas dans la nature !

Ensuite, nous nous plaçons du point de vue de l'informaticien dont l'approche est située entre la démonstration et l'observation.

L'informaticien agit au niveau des descriptions : il donne la possibilité avec les outils qu'il développe de normaliser les observations des naturalistes, élevées au rang de descriptions comparables entre elles car utilisant le même schéma de représentation. C'est à partir de ces descriptions que nous allons bâtir des hypothèses "plausibles" par induction et que nous allons les vérifier grâce à l'identification de nouvelles observations. Par les outils que l'informaticien fournit, nous serons capables d'appliquer la méthode hypothético-déductive chère à Popper [Popper, 1973], [Popper, 1978] :

"La méthode de la Science est une méthode de conjectures audacieuses et de tentatives ingénieuses et sévères pour réfuter celles-ci".

Ces descriptions sont un premier niveau d'abstraction : elles constituent le terme commun des deux approches et c'est pourquoi nous les traitons à part au chapitre 4 de cette thèse.

3.2 Classe et concepts

3.2.1 La classe

C'est l'ensemble ou groupe d'individus … (stop, c'est le point de vue du naturaliste) ... possédant tous un ou plusieurs caractères communs et étant les seuls dans ce cas (c'est le point de vue du philosophe ou du mathématicien).

3.2.1.1 Point de vue des mathématiciens

Pour eux comme pour certains philosophes (suivant en cela la tradition d'Aristote), la classe dérive du concept : il s'agit d'un ensemble d'objets qui satisfont une condition prédéfinie nécessaire et suffisante (dans un univers de discours donné) et qui forme ainsi l'extension d'un concept [Sutcliffe, 1993]. Cette sorte de classe peut être nommée *classe conceptuelle* [Niquil, 1993].

Il existe toutefois une partie des mathématiques qui considère les objets sous leur aspect extensif et que l'on peut qualifier d'expérimentale car basée sur l'induction [Euler, 1747]. Néanmoins, la partie la plus importante des mathématiques "modernes" (la théorie des ensembles, la logique formelle, les prédicats) s'intéresse plutôt à leur aspect compréhensif [Frege, 1893] et à la déduction. Le but de cette dernière approche est de calculer l'extension du concept C en définissant une application a_C de l'ensemble des individus observés Ω --> [vrai,faux] qui à chaque individu w de Ω fait correspondre son appartenance au concept C ou non.

$$a_C : \Omega \rightarrow [0,1]$$
$$w \mapsto 1 \text{ si } w \quad C, \ 0 \text{ sinon}$$

Les individus sont ainsi baptisés **instances** du concept s'ils appartiennent au concept[3]. Comme nous l'avons déjà expliqué plus haut (§ 3.1.1.1), l'extension dépend de la compréhension pour certains alors qu'elle est le point de départ pour découvrir une intension pour d'autres.

Donc, pour le mathématicien, la classe n'existe que si elle est *explicitée en intension* (dans le monde des idées) selon un certain point de vue et correspond à un *concept*. Elle peut être qualifiée d'**abstraite**. Prenons garde néanmoins au terme d'existence : une définition n'entraine pas l'existence de la chose définie, les objets mathématiques étant donnés au départ par postulat (les fonctions, les nombres, le cercle, etc.) [Bourbaki, 1974].

3.2.1.2 Point de vue des systématiciens

On trouve la définition suivante de la classe [Larousse] :

(Histoire naturelle) : "Bien que, comme tous les groupes plus vastes que l'espèce, la classe soit un concept en partie abstrait (un niveau taxonomique), on donne à de nombreuses classes une définition tout à fait précise, correspondant au fait que les êtres de cette classe possèdent tous un

[3] Une partie plus récente des mathématiques s'intéresse au degré d'appartenance "flou" des individus à des concepts [Zadeh, 1965] : un élément appartient plus ou moins à un ensemble. En ce qui concerne les spécimens, le naturaliste n'est pas habitué à jongler avec l'incertitude et l'imprécision pour attribuer un individu à un concept, il finit par trancher. Cette caractéristique étant naturelle en biologie, nous n'avons pas étudié plus avant la théorie des possibilités [Dubois & Prade, 1987] pour l'appliquer dans la représentation des connaissances du domaine.

certain caractère et sont seuls à le posséder. Les Insectes ont tous un thorax formé de trois anneaux et portant trois paires de pattes articulées ; les Oiseaux ont tous des plumes ; les Monocotylédones ont toutes un embryon à un seul cotylédon ; les Céphalopodes ont tous des tentacules, etc.."

Remarque : Les systématiciens employent le mot Classe (ou *Classis*) avec une majuscule pour désigner l'une des catégories de la systématique comprise entre les Ordres (*Ordo*) et les Embranchements (*Phylum*). Quoi que traitant ici de systématique, nous n'emploierons jamais le mot classe dans ce sens strict.

Nous l'employerons plutôt comme synonyme de groupe (ou taxon) à un certain niveau hiérarchique [Larousse] :

(Histoire naturelle) : "Subdivision usitée en classification zoologique ou botanique et dont on ne peut pas ou on ne veut pas préciser la valeur hiérarchique: Classe, Ordre, Genre, Embranchement, etc.."

La première définition précédente de la classe, si on l'étend aux différents taxons de la classification linnéenne, donne comme exemples des caractères propres ce qui a été appelé "caractères dominateurs", entièrement caractéristiques d'une classe. Dans les faits, il est rare qu'une classe puisse être ainsi caractérisée par un caractère unique. La diversité biologique que l'on constate dans la nature fait que la définition d'une classe regroupe généralement la conjonction de plusieurs caractères. La définition semble de plus considérer le terme de concept comme synonyme de classe, ce qui ne correspond pas à notre analyse (voir plus bas).

Pour ces raisons, nous considérons que cette définition ne correspond pas toujours à la réalité des choses. Par le terme équivalent de concept, elles apparaissent comme des intensions, des preuves sûres, démontrables au sens mathématique et pas du tout comme des hypothèses plausibles, vraisemblables et à vérifier par de nouveaux faits.

Ces définitions ne sont pas sans rappeler le grand débat sur "l'espèce" [Cuénot, 1936] entre *fixistes* tels G. Cuvier qui croient à la permanence des espèces qui ont été créées séparément et ne passent pas de l'une à l'autre, et *transformistes* tels C. Darwin qui ne sont pas surpris par la variabilité de l'espèce, les variants étant des espèces naissantes sous l'effet de cause extérieures qu'ils subissent. Il est alors impossible de définir les espèces dans cet état d'équilibre momentané. Le point de vue pratique exige néanmoins l'établissement d'une hiérarchie utilisable, ce que permet la systématique moderne avec un matériel écologique et géographique beaucoup plus abondant et des outils d'expérimentation plus performants (microscopes, ordinateurs, etc.).

Une révision de la notion de classe en systématique s'avère donc nécessaire de manière à ce que nous distinguions bien la différence conceptuelle que l'on veut

apporter à la classe par rapport aux concepts : pour le naturaliste, la classe *existe en elle-même* indépendamment de l'homme qui la décrit, elle est *explicitée par son extension*, elle est donc **concrète**, naturelle et unique. Intuitivement, on conçoit bien [Matile *et al.*, 1987] que si l'espèce humaine (la classe des hommes) disparaissait, les autres espèces continueraient à exister dans leur intégrité, indépendamment de leurs observateurs, tout comme certaines d'entre elles ont existé avant l'apparition de l'homme.

Chaque classe naturelle peut être :

1) nommée,
2) définie par son contenu,
3) caractérisée par des traits propres,
4) typifiée,
5) et enfin conceptualisée.

1) On peut s'y référer sans ambiguïté par son **nom** ; c'est un principe, magistralement arrêté par Linné (1735), que la découverte de toute nouvelle classe doit être accompagnée par son auteur de la fixation d'un nom ; cette dénomination doit respecter des règles de nomenclature bien définies (binôme spécifique, loi de priorité, etc.), en particulier pour s'assurer de son unicité.

2) La classe peut être définie concrètement par son **contenu** (sa population), représentée par exemple par l'énumération des individus connus qui composent son effectif. De façon plus pragmatique, on se contente d'un échantillon "représentatif" de la population, qui doit illustrer au mieux la variabilité naturellement présente. Il se peut qu'il y ait parmi les **descriptions d'individus**, à la fois des descriptions d'un seul spécimen (un individu réel) et des descriptions synthétiques de plusieurs spécimens (individu virtuel). Il est clair que l'on ne maîtrise pas toujours dans les descriptions livresques anciennes la nature des individus décrits (réels ou virtuels). Nous affirmons par contre que l'on devrait s'employer à utiliser le mot description uniquement pour décrire des spécimens et non pas décrire une population de spécimens. Par abus de langage, on appelera ces dernières des "descriptions" synthétiques (ou virtuelles) alors qu'elles ont déjà un certain niveau d'abstraction correspondant à des définitions. La distinction entre description et définition permet de montrer la différence entre l'*imprécision* attachée aux valeurs descriptives d'un spécimen (une disjonction de valeurs pour un seul état possible) et la *variation* associée aux valeurs d'un ensemble de spécimens (une conjonction de valeurs décrivant plusieurs états). Dans la pratique de la systématique au MNHN, il sera préférable dans l'avenir de stocker des descriptions correspondant à des spécimens de manière à perdre le moins d'information possible sur les espèces ou autres classes produites. Idéalement, les descriptions devront être complètes et exhaustives !

3) La classe peut être caractérisée, de façon aussi discriminante que possible, par un ensemble de caractères propres à la distinguer des autres classes, dont l'énoncé constitue sa **diagnose**. A côté de la diagnose, volontairement limitée à un minimum de caractères distinctifs, on fait aussi souvent figurer une **définition**, formée par la synthèse des descriptions des individus qui la composent ; cette synthèse, aussi appelée **intension** de la classe, comporte un certain degré de généralisation[4]. Cette généralisation permet de ne pas exclure d'emblée de nouveaux individus qui ne sont pas exactement semblables à ceux déjà admis, mais néanmoins conformes à la diagnose. L'extension originale de la classe est potentiellement élargie à de nouveaux individus de la classe.

L'extension de la classe (sa couverture du point de vue du naturaliste) généralise sa population à tous les individus qui sont ou pourront être reconnus comme lui appartenant. L'extension de la classe, ainsi comprise, est élargie à l'extension du concept : tout individu appartenant à la classe est un représentant du concept (une instance). De la sorte, on tend à rendre équivalentes les définitions en intension et en extension, comme cela semble souhaitable.

4) La classe possède un **type**, que son auteur a choisi pour la représenter de façon unique et définitive. Le type est, à la limite, le seul individu dont l'appartenance à la classe soit certaine. Il faut remarquer que, paradoxalement, il n'est pas attendu que le type soit particulièrement représentatif ; Il est même fréquent qu'il apparaisse par la suite comme extrême par rapport à la gamme de variabilité intra-classe. Il ne faut donc pas confondre le type, purement arbitraire, avec un quelconque prototype ou individu "moyen". Une classe en tant que concept biologique n'existe que si un type lui a été associé. L'*Homo sapiens* est la seule espèce qui ne possède pas de type.

5) Enfin la classe peut être envisagée comme un **concept** une fois qu'elle a été définie, chacun de ses individus apparaissant à la fois comme un représentant du concept et comme un élément objectif (faisant partie de la couverture) ou subjectif (conforme à la définition) de la classe.

Il est aussi important de prendre en compte le fait que les classes sont organisées selon une hiérarchie à multiples niveaux, à laquelle on peut appliquer le nom de "système" (au sens de la systématique, non de la systémique). Chaque niveau peut avoir une signification biologique, mais celle-ci n'est clairement établie que dans le cas du niveau "espèce", pour lequel on peut se référer à un critère biologique (l'interfécondité). Aucun individu ne peut appartenir à plus d'une classe d'un niveau donné (mais l'exception est tout à fait admise en cas de doute sur l'appartenance à l'une ou l'autre de classes voisines). Et tous les individus appartenant à une classe sont des représentants équivalents de son concept. De la

[4] La généralisation s'effectue lorsqu'il s'agit de passer d'une disjonction de descriptions imprécises sur des spécimens d'une classe à une définition réelle qui est la conjonction d'attributs exprimant la variation au sein d'un concept.

sorte, on ne peut parler de “degré d'appartenance flou”, sauf à traduire par là un état incomplet des connaissances et non pas une ambiguïté de fait.

3.2.2 Les concepts

Les concepts sont considérés du point de vue de la compréhension qui désigne l'ensemble des caractères exprimés par le mot, et du point de vue de l'extension, qui désigne l'ensemble des individus auxquels le mot s'applique.

Un concept est une abstraction intellectuelle de parties du monde. C'est une idée *abstraite* (obtenue en se bornant à considérer certains caractères des objets, à l'exclusion d'autres caractères pourtant perceptibles) et *générale* (étendant les caractères ainsi considérés à tous les objets qui les possèdent). Tout concept se caractérise par sa *compréhension* (ensemble des caractères considérés dans les objets) et par son *extension* (ensemble des objets auxquels il peut s'appliquer). Compréhension et extension forment donc l'aspect logique du concept une fois élaboré (LConcept). Abstraction et généralisation sont les deux opérations psychologiques par lesquelles il s'élabore (PConcept) [Sutcliffe, 1993].

Chez Aristote, on trouve la notion de *logoi* pour le concept avec deux points de vues : l'un considère les sujets que regroupe la classe correspondante au concept et l'autre est le prédicat qui est la condition d'appartenance d'un sujet à la classe du concept. Il y a trois façons (logoi) de se référer à un concept :

1) par son contenu (l'être),
2) par sa définition (l'essence),
3) par son nom (terme univoque qui abrège la définition).

1) **L'être** est le référent ou l'extension du concept. C'est l'ensemble des instances du concept (les choses existantes auxquelles le concept s'applique).

2) **L'essence** est la condition d'appartenance à la classe. On donne un prédicat ou définition (une condition) ce qui crée le concept en intension (le nom n'est pas forcément présent).

3) **Le nom** du concept est un abrégé ultime de la définition. Il peut faire intervenir la propriété la plus caractéristique pour le résumer (par exemple, la rouille du blé[5], un réfrigérateur, etc.). Néanmoins, le nom est avant tout une commodité, un code de reconnaissance, qui est difficilement utilisable si l'on fait abstraction de sa définition complète (ambiguité). En sciences naturelles, le nom est donné en latin pour lui conférer un caractère universel.

---

[5] Maladie fongique caractérisée par des taches de couleur rouille.

3.2.2.1 Du point de vue naturaliste

Dans notre approche des concepts, nous affirmons leur existence dès lors que nous fixons :

1) une classe,
2) une définition associée à la classe,
3) un univers de discours (un contexte),
4) une capacité d'abstraction intellectuelle plus ou moins élaborée.

1) Pour les biologistes, la classe est une vérité ; elle a une existence naturelle avant même d'être définie en tant que concept.

2) Pour le concept, ce n'est pas le nom qui est important mais bien l'intension qui lui est attribuée (sa définition). Un concept est délimité par la définition de la classe correspondante.

3) La définition de la classe dépend du contexte : il peut exister en effet différents concepts associés à une même classe. Par exemple, le concept de "grand homme" dépend de l'univers de discours pour sa définition. S'agit-il du sens donné à la taille d'un individu ou bien celui de sa réputation ? Napoléon et le Charles de Gaulle ne seraient pas classifiés de la même manière selon le contexte choisi !

Autre exemple : la classe des tomates ne correspond pas à la même définition chez un botaniste et chez un cuisinier : c'est un fruit pour le premier et un légume pour le second.

4) La définition de la classe dépend du niveau de perception. Par exemple, le concept de dinosaures pour un paléontologiste correspond à un stade d'évolution dans la lignée des reptiles alors que le concept de dinosaures pour un enfant peut correspondre à celui d'un monstre sympathique, personnage de dessin animé.

Pour un univers de discours donné et un certain niveau de perception, un concept associé à la classe peut être déterminé. Un concept est déterminé lorsque l'on explicite les caractères compréhensifs du concept [Petit-Robert, 1994].

A chaque concept, on peut associer plusieurs niveaux de définitions de la classe considérée :

◊ Le premier correspond à une **intension généralisée** qui donne des conditions *nécessaires* d'appartenance à la classe. Ces conditions forment une généralisation[6] de la classe et la définition obtenue ne se trouve donc

[6] La généralisation peut être définie comme un ajout d'observable à de l'observé. En effet, le résultat de la généralisation englobe des situations intermédiaires observables, non effectivement observées.

que **partiellement observée**. Tous les individus qui y appartiennent satisfont à cette définition. Néanmoins, il peut y avoir des individus qui n'appartiennent pas à la classe mais qui sont conformes à la définition. Il est nécessaire toutefois d'y attacher un critère de sélectivité (par exemple : couvrir le moins possible de contre-exemples) pour ne pas produire de définition triviale si peu discriminante qu'elle couvre trop de contre-exemples. Ainsi, dans le domaine des Hyalonema, si nous retenions uniquement la couleur blanchâtre du corps de l'éponge pour caractériser une classe, nous serions sûrs de couvrir aussi toutes les autres classes. Autrement dit, il est recherché une intension généralisée aussi discriminante que possible.

◊ Le second niveau est une **intension stricte** exprimant des conditions *nécessaires et suffisantes* d'appartenance à la classe : tout individu qui satisfait à l'intension stricte de la classe en fait partie. Inversement, tout individu qui appartient à la classe satisfait à son intension stricte. Chacune des conditions exprime une régularité intra-classe. L'intension stricte est une intension **observée**, elle est issue d'une simple reformulation[7] de la disjonction des descriptions réelles de la classe (par factorisation, par la prise en compte de connaissances de fond, etc.). Elle est *absolue* car elle ne fait pas intervenir les définitions des autres classes.

Remarque : cette intension stricte est surtout valable pour des objets manufacturés qui sont des productions humaines et dont la reproductibilité des descriptions est assurée : ce sont des clones. Par exemple, une nouvelle pièce de 1F à identifier est conforme à l'intension stricte d'une pièce de 1F. En ce qui concerne les objets biologiques que nous avons à traiter, l'intension stricte n'est pas intéressante car son extension se limite aux individus qui ont servi à la définir, ou à leurs clones ; or les individus naturels diffèrent toujours les uns des autres par quelque caractère objectif (polymorphie).

Autre remarque : l'intension stricte peut être généralisante si les exemples sont imprécisément décrits. En donnant la valeur "argentée ou dorée" à la couleur d'une pièce de 10 F, la disjonction d'imprécision peut être interprêtée comme une conjonction de variation au moment de l'identification d'une autre pièce, ce qui ne permettrait pas de toujours déterminer une pièce de 20 cts. Il y a là un problème crucial rencontré lors de l'interprétation des descriptions au moment de l'apprentissage, ce qui

---

[7] Une reformulation est une formule comprimée de l'intension par réécriture, elle est plus dense, mais elle contient la même information (iso-intension) et le même contenu au niveau de l'extension (iso-extension). Un exemple de reformulation est le suivant :

si (b d) C
si (c d) C } si [d (b c)] C } si (d a) C
a = (b c) (connaissance de fond)

peut conduire à une intension stricte faussement généralisée. L'intension stricte s'applique donc à des descriptions subjectives dont on ne mesure pas toujours l'origine (imprécision ou variation) !

◊ A partir de l'intension stricte, nous pouvons dériver une **intension réduite** ou **diagnose stricte** qui donne le plus petit jeu de conditions nécessaires et suffisantes d'appartenance à la classe. Chacune de ces conditions correspond à une différence inter-classe. Il faut remarquer que cette caractérisation succinte est *relative* aux autres définitions de classes que l'on veut comparer pour être en mesure d'évaluer leurs différences : elle n'est pas absolue du fait qu'elle doit être modifiée à chaque fois qu'une nouvelle classe est prise en considération. Il s'agit en effet d'une "connaissance croisée" (différentielle) dont on a retiré tout ce qui est commun avec les autres définitions de classe. La diagnose, issue d'une intension stricte, est une diagnose **observée**.

◊ le troisième niveau est une **intension modale** ou **typique** donnant des conditions *suffisantes* d'appartenance à la classe. Tout individu (typique) répondant à cette définition "caractéristique" de la classe en fait partie (= modèle de classe). Il peut y avoir néanmoins dans la classe des individus atypiques s'écartant de la définition de cette classe. Pour dériver une intension modale de la classe, on procède de la manière suivante :

On commence par ôter les exceptions de la classe (par exemple enlever les autruches de la classe des Oiseaux parce qu'elles ne volent pas). On forme ainsi une sous-classe épurée ne possédant que des individus typiques de la classe. On construit alors une intension stricte de la sous-classe typique, ce qui produit une intension typique de la classe.

◊ Par réduction de l'intension typique par rapport aux autres classes, on obtient une **diagnose modale ou typique** (on supprime tous les éléments de l'intension modale de la classe qui ne caractérisent pas les autres classes : le résultat est par exemple : les Oiseaux volent). La plupart des "diagnoses" utilisées par les biologistes (surtout les botanistes) sont modales (elles évacuent les exceptions pour gagner en signification) ; elles comportent souvent une part plus ou moins importante de généralisation pour en faciliter la compréhension par le profane.

Par exemple, prenons les Orchidées qui est une des Familles la plus importante du règne végétal : de manière générale, elle est caractérisée par l'absence d'albumen dans les graines, la mycotrophie (vie en symbiose avec le mycelium des champignons) et des fleurs entomophiles (attirant les insectes) très zygomorphes (avec un plan de symétrie) [Guignard, 1989].

3.2.2.2 Du point de vue mathématique

Pour formaliser ce que l'on vient de dire, donnons les définitions suivantes :

Soient $\Omega = \{w_1,...,w_n\}$, l'ensemble des spécimens ou individus observés,
$\Pi \quad \Omega$,l'ensemble de tous les individus observables,
$P(\Omega)$, l'ensemble des parties de la population observée Ω.

Soit F, une fonction de représentation de Π --> O, O désignant l'espace d'observation, qui à chaque individu observable ω de Π fait correspondre sa description potentielle $y(\omega) = \delta \quad O$:

$$F : \Pi \rightarrow O$$
$$\omega \mapsto F(\omega)$$

Soit y, une fonction de représentation de Ω --> Δ, Δ désignant l'espace de description des individus observables ($\Delta = F(\Pi) \quad O$), qui à chaque individu observé w de Ω fait correspondre sa description $d = y(w) \quad \Delta$:

$$y : \Omega \rightarrow \Delta$$
$$w \mapsto y(w)$$

Soit une classe observée $C \quad P(\Omega)$. Pour chacune, on peut associer une définition $D = y(C)$, $D \quad P(\Delta)$. En notant b_D la fonction d'appartenance à la classe D :

$$b_D : \Delta \rightarrow [0,1]$$
$$d \mapsto 1 \text{ si } d \quad D, \ 0 \text{ sinon}$$

D représente la somme (ou disjonction) des descriptions observées de chaque individu de la classe : $D = \sum d.\ d \quad D$ est aussi appelé un exemple de la classe D, un contre-exemple est donc un élément de $\Delta \backslash D$.

On obtient ainsi le schéma de la figure 3.3 présenté dans [Diday, 1993] :

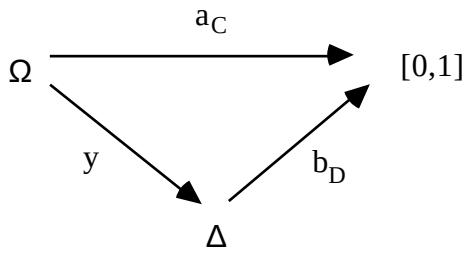

Fig. 3.3 : Le triangle des fonctions entre individus et leurs descriptions

avec la propriété : $w \quad \Omega, a_C(w) = b_D(y(w)) = b_D \circ y(w)$

Pour résumer le formalisme, on peut présenter le schéma de la figure 3.4 :

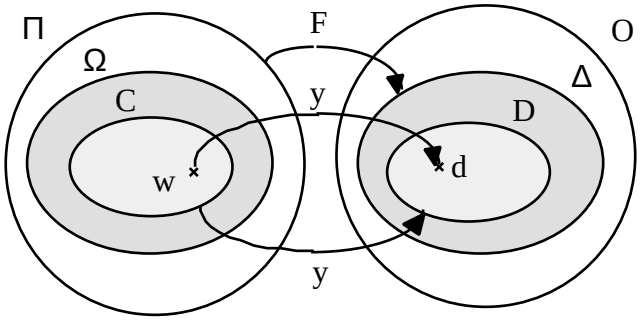

Fig. 3.4 : Schéma du formalisme de modélisation des données

Par exemple : si w_i = “o” (cf. symbole de la figure 3.5), alors $y(w_i)$ D_1,
si w_i = “x” alors $y(w_i)$ D_2.

Le schéma de la figure 3.5 est une illustration des trois niveaux de définition précédents :

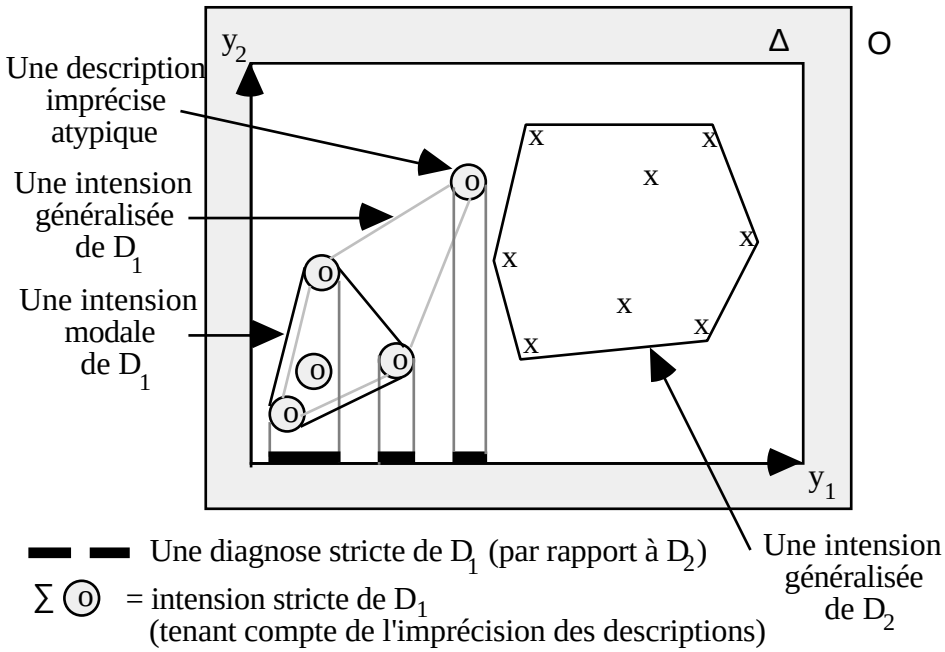

Fig. 3.5 : Les trois intensions de la classe

3.3 Classement et classification

3.3.1 Classer et le classement

Classer consiste dans un premier sens à regrouper des individus ou des objets afin de former des classes. Chacune d'elles se voit attribuer un nom (une étiquette). Classer est une action en deux étapes : à partir d'un tas d'individus, on effectue un tri en répartissant les objets selon leurs ressemblances et différences (on établit une partition des objets), puis on étiquette chaque groupe ainsi formé par un nom de code. Il existe un second sens au verbe classer qui est celui de déterminer : assigner la classe à laquelle appartient une chose, un individu. Nous préférerons employer le terme déterminer pour la seconde acception.

Le classement, selon les deux sens attribués au verbe classer, permet dans un sens de constituer des regroupements nommés d'objets *a priori* afin de former des classes concrètes (définies en extension par les objets qu'elles possèdent) et, dans l'autre, à retrouver le nom d'un nouvel individu *a posteriori* par rapport aux classes déjà formées. Le classement *a priori* est une démarche exploratoire sur un ensemble d'objets dont on ne perçoit aucune définition en l'état (ou dont la définition n'a pas d'intérêt immédiat). Une personne naïve dans un domaine est capable d'effectuer ce classement.

Le classement *a posteriori* permet l'identification des objets entre eux de manière globale en partant de la classe. Il s'agit d'un processus de comparaison directe des objets entre eux qui ne nécessite pas forcément l'usage de descriptions de ces objets, et moins encore d'une quelconque définition de ces individus.

3.3.2 Classifier et la classification

Classifier, c'est conceptualiser des classes, c'est-à-dire les créer par classement, puis les définir, et les nommer éventuellement. Classifier est une des fonctions essentielles de l'intelligence humaine : elle repose sur un plus grand niveau d'expertise que le classement. Cette notion est souvent confondue avec déterminer ou identifier en intelligence artificielle où l'on parle de classifier des observations lorsqu'il s'agit de trouver le nom de la classe auxquels elles se rapportent. En effet, pour certains statisticiens et mathématiciens, la classification veut dire la même chose que le classement *a posteriori*.

La classification, prise dans le sens des systématiciens ("classification des êtres vivants") est la faculté de former un classement (en partitionnant), puis

pour un regroupement donné d'individus, de formuler une définition de ce groupe. Le résultat s'appelle une classification. Il s'agit de représenter les caractéristiques de chaque classe : on établit ainsi des classes abstraites définies en intension par des concepts (et non plus par des objets). De plus, la classification cherche à hiérarchiser les classes selon leur degré de généralité afin de former différents niveaux taxonomiques.

Comme on l'a déjà vu au chapitre 1, la classification en analyse des données n'est pas nécessairement conceptuelle : aucune définition des classes n'est extraite à partir des données.

Dans toute science, il est nécessaire de classifier les phénomènes et les objets que l'on veut étudier et ceci est particulièrement vrai dans les sciences qui étudient les êtres vivants. Une classification vraiment scientifique des végétaux et des animaux doit être naturelle et non artificielle, c'est-à-dire fondée non sur des caractères arbitrairement choisis pour une raison de commodité ou d'utilité quelconque, mais sur les caractères les plus importants du point de vue de la structure anatomique des êtres et de leurs grandes fonctions physiologiques. Les classifications de l'histoire naturelle se proposent d'indiquer le degré de ressemblance et de différence réelle, et non pas apparente et superficielle, de chaque être avec tous les autres. Certains auteurs affirment (d'autres nient) que ces ressemblances sont l'expression d'une *parenté* généalogique entre les espèces et qu'une bonne classification doit tendre à mettre en évidence la *phylogénie* des groupes, c'est-à-dire la suite des formes que l'évolution leur a fait parcourir.

La classification est la partie noble du classement. Elle consiste à ranger dans un même groupe (une classe au sens du biologiste) et à désigner du même nom des faits, des objets ou des êtres qui possèdent en commun certains caractères. Elle suppose l'analyse, la comparaison, mais plus encore la faculté de faire abstraction des différences individuelles. La formation d'une idée générale est un acte de classification. Cette formation s'appuie sur la capacité à décrire les individus, de les classer et de les nommer avec une étiquette, puis de les définir par une intension : cette capacité est le propre de l'expert du domaine. La figure 3.6 synthétise ce que l'on vient de dire :

	Acteur	**Action**	**Moyen**	**Résultat**	**niveau d'expertise**
extension	enfant	répartir	tri	partition	- -
↓	naïf	classer	étiquette	classement (classes)	- +
intension	expert	classifier	critères	classification (concepts)	++

Fig. 3.6 : Schéma de comparaison des termes employés en systématique

La classification s'accompagne de la **caractérisation** des classes (obtenues de manière expérimentale ou artificielle) : elle recherche les critères représentatifs (ou caractéristiques) de la classe (par confirmation des ressemblances intra-classe) et les critères de différenciation (ou de discrimination) des classes (par élimination des différences inter-classe). Elle permet d'*expliciter les classes* à partir des descriptions d'individus (explicitant elles-mêmes les individus des classes). La classification procède par **généralisation inductive** des descriptions, elle est une démarche **synthétique**. Cette synthèse permet de créer des connaissances nouvelles que l'opérateur espère meilleures pour comprendre son domaine.

Deux sortes de classification "artificielle" sont évoquées parmi les méthodes d'apprentissage des descriptions qui nous intéressent :

1) La première sorte procède à partir de descriptions d'un échantillon du domaine étudié sans connaissance préalable du nom associé à chacune d'elles. Ces descriptions sont appelées **observations** en apprentissage automatique car elles ne possèdent pas d'identification associée (on parle aussi d'apprentissage sans professeur). Le but consiste ici à découvrir les classes et/ou les concepts cachés dans les observations.

Ce type de démarche classificatoire, classique en analyse des données (méthodes factorielles [Benzecri, 1973], nuées dynamiques [Diday, 1971]), et en taxonomie numérique [Sneath & Sokal, 1973], est aussi appelé **catégorisation** [Napoli, 1992] ou classification conceptuelle [Fisher, 1985]. Il procède par **agrégation** des observations selon leurs ressemblances avec certaines mesures de similarité puis **caractérisation** en interprétant les classes obtenues par un ensemble de caractères propres permettant de définir les concepts associés.

Le regroupement conceptuel est le même type de classification dans le secteur de l'intelligence artificielle et qui tient compte en plus de connaissances sur le domaine [Stepp & Michalski, 1986].

2) La seconde sorte de classification opère à partir d'**exemples** ou de **cas** qui sont des descriptions d'individus observés auxquelles l'expert a attribué un nom (une étiquette ou bien encore une identification associée après classement) : là, on connaît le concept à apprendre (la maladie, l'espèce, etc.). Ce type de classification avec professeur (ou supervisé) est encore divisé en deux sortes :

Le premier, qualifié de "descendant", est appelé **discrimination** à partir d'exemples et procède par **segmentation** des cas selon leurs différences en fonction de certains critères: fonction coût [Hunt, 1966], gain d'information [Quinlan, 1979], réduction d'impureté [Breiman *et al.*, 1984], etc..

Le second utilise une stratégie ascendante guidée par les données dont l'algorithme de l'étoile avec les systèmes AQ [Michalski, 1983] est le représentant le plus typique.

Quels que soient les modes de classification, elles ont pour point commun de partir de descriptions d'échantillons (pré-classés ou non) pour représenter les concepts à apprendre. Le schéma de la figure 3.7 synthétise les différentes interprétations des termes que nous adoptons dans cette thèse. En outre, nous affirmons que les descriptions sont issues d'observations concrètes et que par conséquence, nous ne parlerons pas de "descriptions" de concepts au sens de [Lebbe, 1991] et [Vignes, 1991] : nous parlerons plutôt de définitions de (associées à des) concepts.

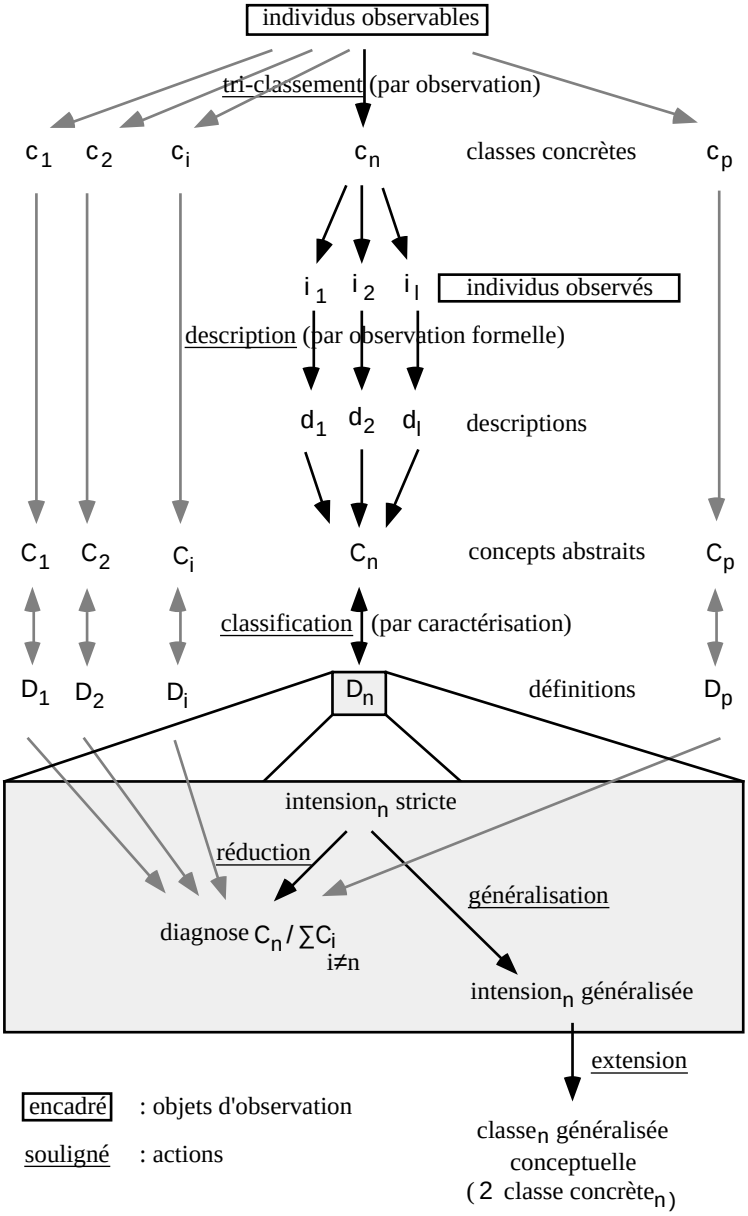

Fig. 3.7 : Notre conception des différents termes employés dans cette thèse

3.4 Détermination et identification

Comme pour la classification, **la détermination** peut avoir une double signification opposée : d'une part, on parle de la détermination d'un concept lorsqu'il s'agit de le définir ou de le caractériser ("déterminer un concept" est alors équivalent à "classifier"). D'autre part et de façon plus courante, le mot est employé pour désigner l'action inverse de la classification : c'est une démarche qui permet de déduire l'appartenance d'un individu à une classe en utilisant sa définition en intension : cette démarche est **analytique**. Dans ce sens, il n'y a pas de détermination possible sans classification préalable. Nous souhaitons bien distinguer les deux aspects inductif et déductif de la démarche scientifique dans cette thèse. C'est pourquoi nous emploierons la détermination dans le sens déductif opposé à la classification inductive.

De plus, la détermination ne doit pas être confondue avec l'identification : déterminer permet de trouver le nom de la classe ou le concept associé à la nouvelle observation. Le procédé permettant de passer d'un indéterminé (individu ou spécimen que l'on peut observer et/ou décrire) à un déterminé (indéterminé affecté à une classe d'identification) est nommé détermination.

L'identification s'applique plus au domaine de l'extension contrairement à la détermination qui concerne le domaine de l'intension : dans le langage courant, identifier est employé plus souvent pour trouver le nom d'un individu (la plupart du temps un humain), ou un code qui permet de se référer à l'identité de quelque chose. On dit plutôt "identifier un individu" pour dire que l'on a trouvé son identité, plutôt que "déterminer un individu". Inversement, on parlera de "déterminer la classe d'un individu" lorsque l'on utilisera une définition de son concept. Pour résumer :

identification	=>	nom d'un individu (ex : Lee Oswald)
détermination	=>	classe d'un individu (ex : Homo sapiens)

d'où : détermination d'un individu = identification de sa classe.

Alors que la classification est affaire de spécialistes, il est fréquent que la détermination soit conduite par un "béotien" en la matière, comme ce douanier qui doit déterminer s'il a devant lui un animal protégé ou non par la convention de Washington, ou lors d'un recensement écologique où il est nécessaire de distinguer (et de désigner) les différentes espèces en présence.

Toute détermination se fait par référence à un corpus de connaissances préexistant, qu'il soit organisé (clef de détermination, système expert, etc.) ou non (livres, connaissance résultant d'un apprentissage plus ou moins empirique).

Il faut aussi remarquer qu'une détermination ne conduit pas toujours à un résultat certain, du fait d'inexactitudes ou d'imprécisions soit dans les connaissances de référence soit dans la possibilité ou la capacité d'observer correctement l'individu à déterminer. De plus, la précision attendue pour une détermination doit être adaptée à l'utilisation prévue du résultat ; les applications dans le domaine scientifique sont bien sûr les plus exigeantes.

Selon les cas, plusieurs situations de détermination peuvent se rencontrer, isolément ou en concours.

3.4.1 Détermination par comparaison directe

Ce premier mode de détermination exige la disposition d'une collection de référence (herbier, jardin botanique par exemple) ou d'un substitut (flore où les différentes espèces sont figurées). Il suffit (non sans mal néanmoins !) de comparer visuellement l'indéterminé avec chacun des référents disponibles, afin de sélectionner celui qui correspond le mieux ; le nom de ce référent est alors adopté comme l'identification recherchée.

Du fait que cette méthode n'astreint pas à décrire, la qualité du résultat est étroitement dépendante des dons d'observation du déterminateur. Tout tient en effet en sa capacité de juger de “l'identité” entre deux individus, qui ne sont pourtant jamais semblables s'agissant de créatures de la nature. Comme aucun contrôle n'est possible, puisqu'aucune connaissance n'est *a priori* pré-requise, elle peut conduire à des erreurs quand l'œil n'est pas suffisamment exercé.

Elle constitue par contre l'ultime confirmation pour le spécialiste, pour lequel la comparaison visuelle directe avec le type demeure l'épreuve de vérité irremplaçable. Le type est l'unique spécimen désigné comme le référent absolu de chaque classe lors de la création de celle-ci ; il n'existe pas de classe dépourvue de type, sauf celle de l'Espèce *Homo sapiens* peut-être pour des raisons éthiques.

3.4.2 Détermination par comparaison avec des descriptions

Ce deuxième mode nécessite d'abstraire le spécimen indéterminé, en en faisant la description plus ou moins complète. La seule observation n'est plus suffisante. En effet, la comparaison va se faire non plus avec des référents concrets, mais avec des descriptions jouant le rôle de référents abstraits. Chaque classe naturelle est pourvue, outre son type, d'une description ou d'une diagnose (description différentielle) ; chaque flore ou chaque faune constitue ainsi un recueil de descriptions, équivalent en quelque sorte de la collection de référence utilisée pour la comparaison concrète.

On procède par élimination progressive. Pour chaque caractère examiné, on met de côté tous les référents incompatibles. Quand tous les caractères ont ainsi été explorés, soit les référents restant en lice appartiennent à la même classe, et celle-ci devient la classe de détermination, soit ils se répartissent dans plusieurs classes et la détermination est incomplète. S'il ne reste aucun référent, il y a une erreur quelque part, soit dans la description de l'indéterminé, soit dans celle des référents, soit dans l'affectation des référents aux différentes classes ; à moins qu'il ne s'agisse de quelque chose de nouveau, ne se rapportant à rien de connu.

3.5 Apprentissage et raisonnement

L'apprentissage est en lui-même une activité intelligente de l'être humain. Le but de l'apprentissage automatique effectué par une machine est de simuler l'apprentissage humain à l'aide de différents mécanismes de raisonnement.

Le raisonnement agit sur des connaissances dont on constate plusieurs niveaux de généralité : faits particuliers, définitions de concepts (règles), méthodes de résolution d'un problème, méta-connaissances, etc.. De plus, ces connaissances sont structurées dans notre cerveau selon un modèle. Pour être capable de simuler le raisonnement, il faut être en mesure de représenter ces différentes sortes de connaissances. On constate de même que ces connaissances évoluent avec le temps, dans le sens d'un enrichissement (espéré). Pour Michalski (1986), l'apprentissage est “lié à la construction ou modification des représentations de ce que l'on expérimente”.

Si l'on veut doter les machines de capacités d'apprentissage, il faut absolument prendre en compte la définition d'une structure pour représenter l'espace des connaissances, ainsi que des moyens d'y accéder pour les modifier ou pour en générer de nouvelles.

Classiquement, les systèmes experts ont utilisé le formalisme des règles de production pour modéliser les connaissances d'un expert. L'acquisition des connaissances s'effectue par l'intermédiaire d'un cogniticien qui aide l'expert à expliciter ses règles de décision. Ensuite, l'apprentissage met en place un mode de raisonnement par **déduction** à partir de ces règles explicites et de faits nouveaux qui leur sont présentés. Le système expert infère des conclusions dont les résultats valides seront ajoutés dans la base de connaissances.

Nous considérons l'apprentissage comme le processus de classification (discrimination) qui permet de généraliser des cas spécifiques pour construire une définition abstraite (des règles de décision) en fonction d'un “bon” critère de classification. Il s'agit d'apprentissage où le raisonnement se fait d'abord par **induction**. Ensuite, comme pour les systèmes experts classiques, on déduit à

partir de ces nouvelles connaissances qu'un nouveau cas est couvert par cette définition abstraite.

Les généralisations “de haut niveau” extraites à partir des cas sont utiles pour comparer des concepts différents, les valider les uns par rapport aux autres (notamment par rapport à ceux élaborés de manière classique), mais aussi pour identifier rapidement une nouvelle observation. Ce raisonnement nécessite donc une classification préalable.

Une autre forme de raisonnement logique, introduite par Peirce (1965), est l'**abduction**. Elle est l'opération qui consiste à choisir une hypothèse explicative obtenue en faisant la trace arrière des règles du domaine, compte tenu des conclusions supposées vraies. Par exemple, soit la règle suivante (*modus ponens*) qui permet de déduire que si l'on observe du feu, alors on a de la fumée :

R : x {lieux}, feu(x) => fumée(x)

Dire qu'il n'y a pas de fumée sans feu, c'est faire de l'abduction : on fait l'hypothèse qu'il y a un feu du fait que l'on observe de la fumée et que l'on connait R. La déduction est le raisonnement inverse exprimé par la règle R. Pour l'induction, on doit observer qu'à chaque fois qu'il y a un feu quelque part, on observe aussi de la fumée à ces endroits, et on construit donc la règle générale R.

Une autre forme de raisonnement fait aujourd'hui l'objet de recherches actives : elle repose sur les exemples eux-mêmes sans chercher à les généraliser. L'idée consiste à interpréter une nouvelle observation à l'aide d'un cas similaire extrait du système et choisi comme guide [Bareiss, 1990]. C'est le principe du **raisonnement par cas**.

Raisonner consiste à comparer la proximité des cas avec la nouvelle observation par une mesure de distance. Il ne nécessite donc qu'un classement des individus au préalable (individus pré-classés par un nom de classe). Pour résumer, nous donnons la figure 3.8 suivante :

raisonnement	entrée	sortie
déduction	prémisses + règles	concepts
induction	prémisses + classes	règles + concepts
abduction	règles + concepts	prémisses
“par cas”	prémisses + classes	classes

Fig. 3.8 : Les modes principaux de raisonnement en apprentissage automatique

En définitive, l'aspect très important du raisonnement en apprentissage automatique doit être la mise en œuvre concertée dans les algorithmes, de mécanismes symboliques logiques issus des recherches en intelligence artificielle (représentation des connaissances, règles de généralisation, stratégies de contrôle, etc.) et de méthodes numériques performantes (distances, mesures de proximité, entropie, etc.) propres à l'analyse des données et aux statistiques. Cette nécessité est à l'origine du développement des recherches sur le traitement des connaissances “symboliques - numériques” en apprentissage [Kodratoff, Diday, 1991].

3.6 Individus, instances et objets

L'individu est considéré de manière extensive, synonyme d'un élément d'un groupe ou d'une classe. Dans l'idéal, un individu est un être réel, une entité tangible et distincte. Il s'agit d'un sujet unitaire correspondant à un spécimen en biologie. Seul un individu peut être décrit, et ce n'est que dans un sens généralisé que l'on peut parler de “description de classe”. Dans ce contexte, l'individu est synthétique et correspond à un ensemble d'éléments distincts comme par exemple l'Espèce avec ses différents spécimens.

L'instance est l'individu passé, présent et à venir qui appartient à un concept (le petit chien à naître fait partie du concept de chien) alors que l'individu existe indépendamment de celui-ci. Pour résumer :

l'individu appartient à la classe l'instance appartient au concept

Du point de vue mathématique, l'individu fait partie d'une **population observable** notée Π que l'observateur cherche à décrire. Une fois observé, l'individu devient objet d'observation noté **w**. Une fois décrit, l'objet a une description notée **d(w)**. L'observateur ou le descripteur (celui qui décrit) s'est approprié l'individu (le sujet) qui est devenu un objet de description (observé ou décrit). La **population observée** est notée Ω.

L'objet prend différentes significations selon le point de vue et l'échelle d'observation auxquels l'observateur se place : du point de vue d'une “description de classe”, l'objet est pris comme un élément de cette classe, c'est-à-dire un **individu**. Par contre, si l'on se place à l'échelle d'une description individuelle, l'objet correspond à un **composant** de l'individu (ou partie “individualisable”). Tout dépend donc du point de vue ! Pour résumer :

Un objet	=	une entité descriptive d'un individu
Un individu	=	une entité descriptive d'une classe

Pour illustrer cette distinction, considérons l'ensemble (taxon) des Mammifères : en se plaçant du point de vue de la “description de cette classe”, l'objet sera par exemple une baleine ou un éléphant particulier. Par contre, en considérant la description d'un individu de la classe des Mammifères, l'objet sera l'une des entités descriptives de cet individu, à savoir sa tête, son tronc, ses jambes, etc..

Dans cette thèse, nous nous plaçons dans le second cas de figure : nous souhaitons acquérir des descriptions d'individus dont les objets sont les différents composants de ces individus à analyser.

Entre individu et classe, la relation qui lie ces deux notions est celle d'appartenance de l'individu à la classe : l'individu w est un élément de l'ensemble C. Par opposition, deux classes emboitées sont liées par la relation d'inclusion ensembliste.

3.7 Synthèse des concepts utilisés dans cette thèse

Dans ce chapitre, nous avons indiqué les différents points de vue des utilisateurs systématiciens et mathématiciens sur les mots clé tels que le classement, la classification, la classe, le concept, etc.. Dans la figure 3.9, nous regroupons les différents termes employés et nous les organisons de manière à faire ressortir les relations qu'ils entretiennent :

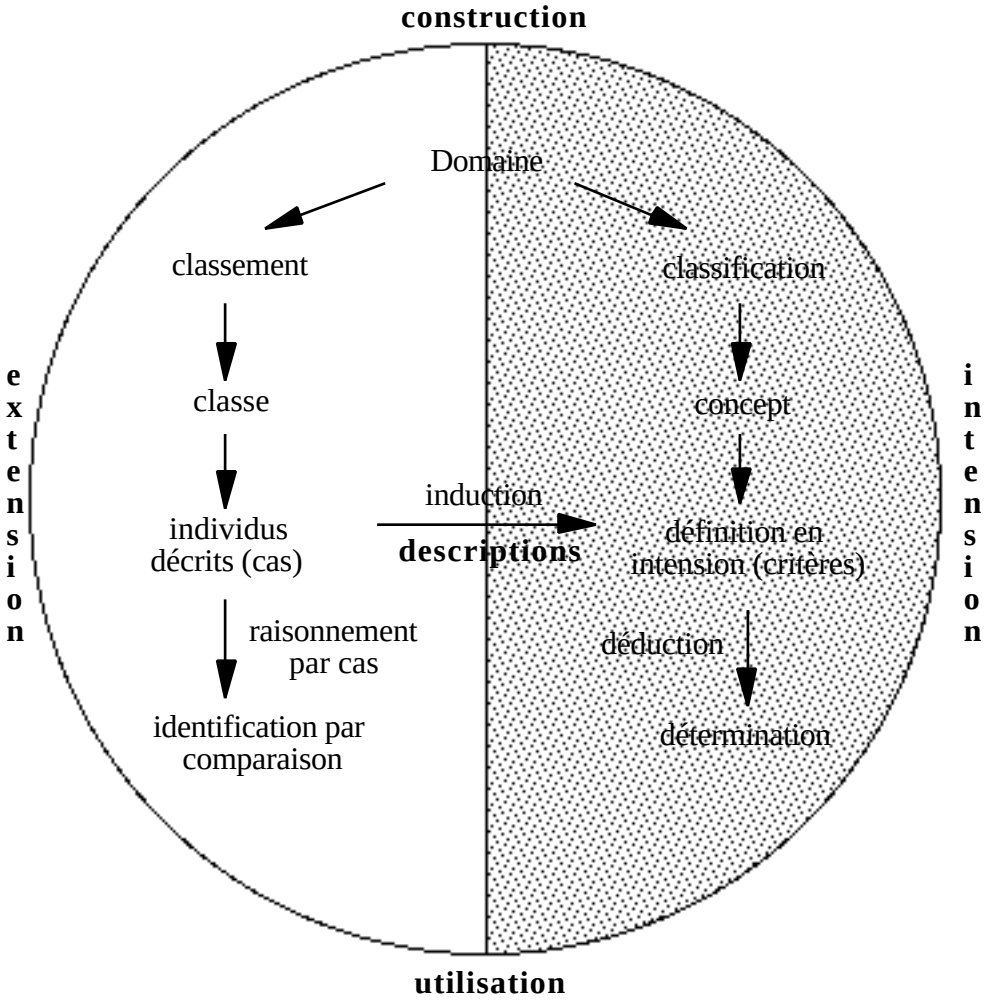

Fig. 3.9 : Relations entre les concepts utilisés

Nous pouvons analyser un domaine naturel sous deux angles différents. La partie grisée correspond plus à la vision du mathématicien contemporain. Il raisonne dans le monde des idées, c'est pourquoi les notions abstraites de classification et de concept lui sont plus familières. Il utilise plus naturellement la déduction pour résoudre un problème de détermination. Contrairement à lui, le naturaliste raisonne au niveau du monde réel (partie non grisée). Partant d'une

classe (ensemble d'individus) dont le contenu (l'extension) est sa couverture, il observe et crée le nom de cette classe puis la définit (en intension) afin de créer le concept associé. Mais avant de généraliser, le systématicien aura au préalable décrit beaucoup d'échantillons pour se familiariser avec son domaine. La construction d'hypothèses par induction n'est néanmoins pas seulement la démarche des sciences expérimentales, ainsi que nous l'affirment Euler et Pólya au niveau des mathématiques : la découverte de règles résulte d'un aller et retour permanent entre des observations et des hypothèses sur ces observations. Nous affirmons que l'informaticien peut contribuer de manière originale à l'amélioration des règles apprises en se positionnant au niveau des descriptions entre les observations et les règles. Par exemple, il peut les rendre comparables entre elles du fait qu'elles utilisent le même schéma de représentation, celui du modèle descriptif.

Les descriptions sont au centre des préoccupations des différents opérateurs (mathématiciens, psychologues, biologistes, etc.) souhaitant faire de la classification et de la détermination d'objets. Elles permettent d'expliciter un individu, c'est-à-dire que le fait de connaître la description d'un individu rend celui-ci explicite. Si elles ne sont pas forcément nécessaires pour faire du classement ou comparer les objets entre eux, elles sont néanmoins le support de la transmission du savoir car elles expriment la richesse et la diversité des observations du monde réel. A ce titre, les descriptions jouent un rôle central en sciences naturelles comme nous le verrons au chapitre 4. Il sera donc très important pour l'informaticien de voir comment les rendre le plus robuste possible.