Académie de Montpellier
Université de Montpellier II
Sciences et Techniques du Languedoc

Diplôme d'Etudes Approfondies

"INFORMATIQUE"

DES OUTILS D 'AIDE A L' EXPRESSION ET AU PARTAGE DE LA CONNAISSANCE DANS LE DOMAINE DES SCIENCES DE LA NATURE

Yannis TARER

Date de soutenance : le 2 octobre 1996

Tuteur(s) de stage : Marcel Bouche, Danielle Hérin, Thérèse Libourel

Remerciements

Enfin, je peux m'exprimer en toute liberté et montrer ma reconnaissance de caractères.

Je tiens tout d'abord à remercier Monsieur Marcel Bouche et Madame Thérèse Libourel, tuteurs de mon stage, de m'avoir permis de travailler sous leur direction.

Leurs conseils et les heures passées à travailler de concert ont marqué un point décisif dans ma façon de travailler et d'organiser mon travail.

Je tiens également à remercier Madame Danielle Hérin, tutrice de mon stage, de m'avoir montré le chemin de l'«acquisition de connaissance».

Mes remerciements vont aussi à ma soeur Vanessa et à mon amie Lydie, elles ont su m'apporter le soutien dont j'avais besoin et je les en remercie.

Enfin, tous les étudiants de DEA et thésards du LIRMM ainsi que tous les membres du laboratoire de Zooécologie du CNRS pour leur bonne humeur et leur tolérance.

Table des matières

Introduction **5**

I Contexte de ce mémoire **8**

1 Aspects conceptuels **9**
1.1 Des données et des connaissances 9
1.2 Gestion des données et des connaissances 9
1.2.1 Gestion des données 9
1.2.2 Gestion des connaissances 10

2 Acquis **11**
2.1 ECORDRE 11
2.2 DICs, prélevats, la connaissance sous la forme de faits 12
2.3 La connaissance sous forme d'énonçats [Bou90] 13

3 Objectifs de ce mémoire **15**
3.1 Trouver un formalisme pour représenter la connaissance . . . 15
3.2 Présenter des outils pour exprimer et accéder à la connaissance 16

II Etat de l'art **17**

1 Les principales approches relatives à l'acquisition de connaissances **18**
1.1 Les réseaux sémantiques 18
1.2 Les langages à base de frames 19
1.3 Les logiques terminologiques 20
1.4 Etudes de quelques SRC 21
1.4.1 Les graphes conceptuels 21
1.4.2 CLASSIC 23

2 Les principales approches relatives au partage de la connaissance, à son accès, à sa présentation 25
2.1 L'accès 26
2.1.1 Les moyens d'accès traditionnels 26
2.1.2 Les mécanismes de bases de données 27
2.1.3 Les hypermédias 27
2.2 Internet 28

III Notre approche 29

1 Des modèles pour formaliser et exprimer la connaissance 31
1.1 Les énonçats, une base de connaissance 31
1.1.1 Analyse et exploitation 31
1.1.2 Les normalisations applicables aux énonçats 32
1.1.3 Un formalisme objet pour modéliser concepts et relations 33
1.2 OMT, une méthode de conception de système d'information . 34
1.2.1 Présentation 34
1.2.2 Le modèle objet d'O.M.T. utilisé comme modèle de représentation de la connaissance 35
1.2.3 Convergence entre nos objectifs et le modèle objet d'O.M.T. 36
1.3 Convergence entre PRACs et le modèle objet d'O.M.T. . . . 37
1.3.1 PRACs et DICs 37
1.3.2 Interprétation et PRAC 37
1.3.3 Le PRAC de comparaison typologique et l'héritage . . 37
1.3.4 Le PRAC eusystémique 38
1.3.5 O.M.T., énonçats et DICs 40
1.4 Les organisations, la connaissance sous forme de graphes . . . 40
1.4.1 Description 40
1.4.2 Une représentation graphique puissante 41

2 Java et Internet : alliance pour l'accès à toutes les formes de la connaissance 44

3 Les outils proposés 46
3.1 L'organiseur 47
3.1.1 Le gestionnaire de connaissances 47
3.1.2 L'interface 47
3.2 Le distributeur 48
3.3 Copies d'écrans 49

IV Perspectives 52

1 L'émergence d'ontologies 54
1.1 Définition d'une ontologie 54
1.2 Les travaux réalisés 55
1.2.1 Les ontologies selon une approche constructive 55
1.2.2 Le niveau ontologique, une approche philosophique et linguistique. 56
1.3 Les possibilités offertes par le système 57

2 Un module de traitement automatique de la langue naturelle 58

3 Vers une plus grande interaction des SI et des SRC 59

Conclusion 61

Introduction

Les sciences de la nature sont des sciences dites empirico-formelles. Ceci est dû au fait qu'elles sont construites sur des objets «*extérieurs*». Elles ont donc des **caractéristiques empiriques**, de par leur nature observationnelle et expérimentale, et des **caractéristiques formelles** car elles s'appuient sur, étayent, affirment, ou infirment des théories. Ceci est vérifié par cette remarque de l'Encyclopediae Universalis: «*Or un résultat d'expérience ne consiste nullement en* «**données**» **observationnelles pures**, *il est toujours une* **interprétation** *du* «**donné**».». Ainsi les sciences de la nature constituent un savoir et pas simplement un savoir-faire. Ce savoir afin d'être transmis doit être géré.

L'informatique propose principalement deux types d'outils pour gérer l'information: les **systèmes d'information (SI)**, et les **systèmes de représentation de connaissance (SRC)**. Un SI gère l'information nécessaire à l'organisation, un SRC gère l'organisation (en amont du SI) de la connaissance et sa révision.

Ces deux approches sont complémentaires dans le contexte des sciences de la nature. Dans ce cas, un SI a pour objet la composante empirique (les données), et un SRC la composante théorique c'est à dire un ensemble de termes et de méthodes logiques, mathématiques et descriptifs (soient ceux correspondant au contenu de la ou des théorie). Les besoins actuels de la transmission du savoir imposent la migration de ces deux démarches différentes de la gestion de l'information.

L'évolution des systèmes d'informations vers les systèmes de représentation de connaissance a pour but de permettre une meilleure utilisation de l'information. Elle est décomposée, analysée, puis recomposée sous une forme plus simple et mieux compréhensible par un plus grand nombre d'utilisateurs. Cette gestion différente (par rapport aux SI traditionnels) de l'information montre l'importance, aux yeux des utilisateurs, de disposer d'une bonne perception du domaine concerné par le SI.

Plusieurs utilisateurs interviennent dans les SI: les utilisateurs de base,

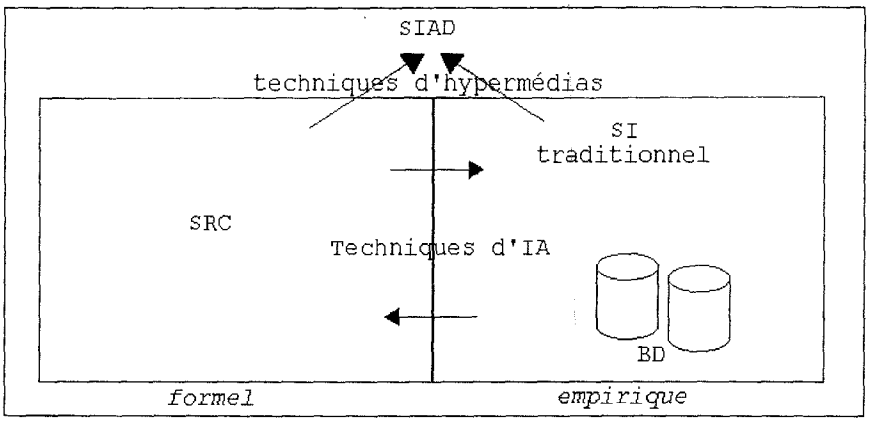


FIG. 0.1 – *Schéma d'un système intelligent d'aide à la décision dans le cadre des sciences empirico-formelles.*

les experts (voire les concepteurs créateurs) du système.

Afin de pallier au manque d'objectivité et aux limites de chacun [1], il faut adjoindre au SI diverses «méthodes» pour permettre à ces divers utilisateurs soit de confronter leur connaissance avec celle représentée au sein du système et qu'il permet de visualiser, soit d'entrer leur connaissance dans le système.

Un SRC doit fournir des services permettant d'**accéder** et de **gérer** la connaissance représentée [Neb90]. L'évolution d'un SI vers des fonctionnalités d'un SRC implique qu'il hérite de ces contraintes.

Ces nouvelles dimensions des SI (accès, acquisition et confrontation de connaissances) soulèvent un certain nombre de problèmes.

Pouvoir entrer de la connaissance, nécessite principalement de pouvoir l'**exprimer** au moyen d'un **formalisme efficace**. Autoriser sa **mise à jour**, voire sa **révision**, permet de la rendre «**vivante**».

Mettre la connaissance à la disposition de tous implique son **accessibilité** sous une forme facilement compréhensible.

Le but du mémoire est de proposer un **formalisme** et des **outils** permettant d'**exprimer la connaissance** dans n'importe quel domaine et de la rendre **accessible** à tous. Ces outils placés au dessus du SI nous permettrons alors d'obtenir un **système d'information et d'aide à la décision (SIAD)** [2]. Nous avons basé notre approche sur les avancées des **méthodes de représentation par objet** et sur les progrès réalisés en matière d'**interface homme-machine** et de **communication à l'échelle mondiale**. Les outils n'ont pu être réalisé que grâce à la volonté internatio-

1. dans le cadre d'une prise de décision, par exemple.

2. La motivation ayant généré ces travaux est liée à la volonté de donner aux décideurs ainsi qu'au reste du public les moyens de raisonner sur des éléments de connaissance grâce au SIAD.

nale d'améliorer les possibilités offertes par les réseaux à vocation mondiale. Ils viennent s'ajouter aux fonctionnalités d'un SI pour lui faire bénéficier des propriétés des SRC. Dans la première partie, nous aborderons certains aspects conceptuels et techniques qui ont aidé à élaborer le sujet, les motivations et les acquis ayant permis notre travail.

La deuxième partie rend compte de l'état des connaissances en matière d'acquisition de connaissances, un processus inhérent à tout système relevant de la représentation de connaissances, et en matière d'accès à la connaissance, un thème de recherche qui concerne surtout les interfaces homme-machine personnelles[3]. La troisième partie développe notre approche des objectifs du mémoire et la façon dont nous les avons atteints.

Enfin, la quatrième partie fait le point sur les résultats obtenus et sur les possibilités d'extension en largeur (développement et amélioration de l'existant) et en longueur (autres thèmes de recherche abordables à partir de notre travail et du sujet).

3. L'adjectif «personnel» implique que ces interfaces sont pour la plupart des applications non accessibles par de grands réseaux de communication

Première partie

Contexte de ce mémoire

Chapitre 1

Aspects conceptuels

1.1 Des données et des connaissances

L'introduction décrit les sciences de la nature comme manipulant les différents types d'objets de connaissance : les objets de la connaissance empiriques (données, faits, ...) et les objets de la connaissance formelle (théories, interprétations, ...). Les chercheurs doivent donc, pour exprimer leur connaissance, gérer un ensemble de données et un ensemble d'interprétations. Le concept de «donnée» étant pris au sens de «données observationnelles pures», elles sont donc assimilables à des données brutes sans hypothèse préalable. Le concept de **DIC**, donnée initiale contrôlée, que nous définissons un peu plus loin, recouvre ce type de données.

Mais l'on ne peut s'arrêter aux données car il faut les rendre compréhensibles, les traduire, leur donner un sens. Dans le discours rationnel, la science est définie comme un «*mode de connaissance critique*» car elle «*élabore des méthodes qui lui permettent d'étendre de façon systématique le champ de son savoir*». [Bou90] et [Bou96] décrivent parmi ces méthodes, les **processus répétitifs d'accroissement des connaissances** (**PRAC**). Ceux-ci sont utilisés dans l'**interprétation** des données (les DICs) qui permet aux chercheurs d'étayer, de réviser ou d'infirmer des théories et des modèles en d'autres termes d'**organiser la connaissance**.

1.2 Gestion des données et des connaissances

1.2.1 Gestion des données

Depuis le début des années soixante-dix, les avancées en matière de gestion de données sont très importantes. Le concept de base de données relationnelle ayant largement franchi les portes des laboratoires de recherche a permis de créer des bases contenant des milliards de données de toute sorte dans le monde entier.

Ces systèmes de gestion de base de données constituent encore un pôle de recherches très actif. Avec l'arrivée de bases de données objet (O2, ObjectStore, Orion, Ontos, Gemstone), actives ou déductives (DATALOG), de nouvelles barrières sont abattues. La gestion des données est de plus en plus satisfaisante et accessible à tous.

1.2.2 Gestion des connaissances

En ce qui concerne la connaissance formelle, elle ne peut être gérée dans des bases de données classiques dont l'efficacité porte sur le stockage de grandes quantités de données. En revanche, le concept de **système de représentation de connaissances (SRC)** né avec les progrès en intelligence artificielle du début des années quatre-vingt est approprié. La différence entre les base de données et les SRC est soulignée dans [BM86] relatée dans [Neb90] : «*Any knowledge representation language must be provided with a (rich)* **semantic theory** *for relating an information base to its subject matter, while a data model requires an (effective)* **computational theory** *for realizing information bases on physical machines*». Ceci est repris en écho dans [Bre95] dans une définition des SRC. Ainsi, ils ont pour but de donner une représentation structurée de l' «*univers du discours*» dans lequel les données (d'une base de données) sont placées, le mot *discours* étant pris au sens d'un «*ensemble d'énoncés se référant à un même objet, à une époque donnée.*». Bien que de nombreuses études, [BM86], aient été menées pour mélanger les deux approches, il n'y a pour l'instant aucun système alliant les performances nécessaires à une exploitation correcte.

L'évolution des SRC a engendré différents systèmes de représentation de connaissances (**réseaux sémantiques, langages de frame et logiques terminologiques**) ayant plus ou moins des formalismes similaires basés sur des descriptions des différents concepts de la connaissance à représenter et des relations qui les lient. Tous ces formalismes sont présentés rapidement dans la deuxième partie de cet exposé(Partie 2).

Chapitre 2

Acquis

2.1 ECORDRE

Dans le domaine de l'environnement, un système d'informations nommé ECORDRE (acronyme de Ecologie et Ordre) a été créé. ECORDRE est défini comme un **système d'information transdisciplinaire**. Il peut être vu comme : «*l'émanation d'un projet informatique dont l'orientation d'action est l'intégration transdisciplinaire de la connaissance écologique et environnementale*» ou encore comme une «*conceptualisation d'un domaine pluridisciplinaire permettant d'utiliser les outils informatiques de façon efficace*» [Bou90].

Quoique des concepts utilisés et nécessaires à ECORDRE aient été forgés préalablement, c'est essentiellement à partir de **1985** que ce système s'est constitué par la reconnaissance du concept de **DIC** et la gestion de DICs de toutes origines (toutes disciplines) dans une base de données relationnelle. Cette base, nommée BASECOL permet une gestion rationnelle séparant les «*faits*» représentant les objets concrets, appelés «**prélevats**[1]», décrits par leurs caractéristiques prenant différentes valeurs, les DICs. Ces dernières n'étant pas liées aux interprétations ne dépendent pas d'hypothèses quelconques.

BASECOL contient des milliards de données sur des domaines divers dont l'écologie des vers de terre et la viticulture.

Mais l'interprétation des DICs absente de BASECOL et même d'ECORDRE, pour le moment ,existe et est nécessaire. C'est dans [Bou90] qu'apparaît le terme d'**énonçat** pour décrire l'**interprétation des DICs**. Ces deux éléments de connaissance , DICs et énonçats sont décrits dans le détail dans les deux chapitres suivants.

1. Le prélevat est un objet concret fournissant au moins une DIC, donnée initiale contrôlée d'une caractéristique.

Exemple : c'est une motte de terre ayant fourni la donnée 5,9 pour la caractéristique pH.

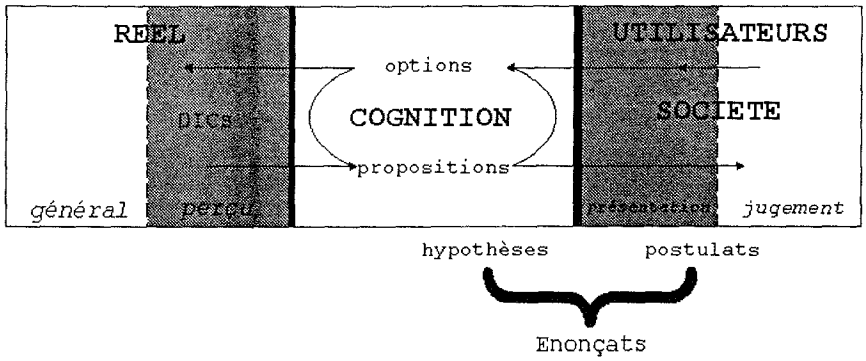


FIG. 2.1 – *Du réel aux utilisateurs*

La figure précédente montre le passage du **réel** dont nous avons une perception réduite (DICs), aux **utilisateurs** par l'utilisation de la cognition et du jugement des utilisateurs et de la **société**. Ces deux dernières démarches permettent la création d'énonçats sous deux formes : les **hypothèses** (options et propositions) et les **postulats** définis au chapitre 1.2.3..

2.2 DICs, prélevats, la connaissance sous la forme de faits

La gestion des DICs, c'est la gestion des données excluant toute interprétation qui est effectuée dans BASECOL d'ECORDRE selon un schéma conceptuel global rattachant ces DICs entre elles mais aussi bien sûr au réel[2] selon **cinq référendaires** :

- le **temps** représentant la durée et l'instant d'obtention de la DIC,
- l'**espace** représentant le milieu d'obtention de la DIC,
- la **composition**,
- le **protocole d'analyse** permettant l'observation quand elle est sophistiquée et,
- l'**obtenteur** soit l'individu appartenant à un organisme obtenant des DICs.

Une DIC n'a donc de sens qu'associée à d'autres DICs selon ces cinq référendaires mais ne résulte pas d'une interprétation. Elle est un élément de connaissance.

2. C'est à dire aux prélevats perçus de ce réel.

Description des graphes conceptuels

Les graphes conceptuels (**GC**) sont «*des graphes munis d'étiquettes ordonnées-qui servent à représenter des connaissances sur lesquels sont définies des opérations de graphes qui servent à raisonner.*». Ainsi définis, les GC permettent de représenter des faits, des buts, des définitions de type, des règles et des contraintes.

Le **support** qui définit la sémantique de ces graphes comprend un ensemble de **types de concepts**, un ensemble de **types de relations**, et un **ensemble de marqueurs** portant sur les concepts, permettant de définir les sommets du graphe. Les arêtes relient toujours un sommet concept à un sommet relation. Une fois créés, les types de relations et de concepts du support sont hiérarchisés respectivement dans des ensembles de types de relations et de types de concepts contenant chacun un plus grand et un plus petit élément.

Afin d'utiliser ces graphes, un certain nombre d'opérations peut être réalisé. L'opération la plus importante est la **projection**. L'une de ses fonction est de permettre de raisonner en utilisant une sémantique logique. Ainsi on peut obtenir la formule de la logique du premier ordre associée à un graphe. L'opération de projection est alors assimilée à l'opération logique appelée **implication**.

Utilisation des GC

Afin de parler de l'utilisation des GC, nous discuterons de l'approche de Rose DIENG et de l'équipe du projet **ACACIA** [Die95].

Notre intérêt se porte surtout sur l'utilisation de «graphes de connaissance». En effet, afin de résoudre les conflits intervenant au cours de l'acquisition entre plusieurs experts, l'auteur compare des graphes modélisant la connaissance et le comportement des experts en question[6].

Ces graphes de connaissance sont des triplets dont la sémantique est la suivante :

- **KG = (C, R, A)** où **C** est un ensemble de concepts, **R** est un ensemble de relations et le 3-uplet **(Cin, rel, Cfin)** appartient à **A** si et seulement si il y a un arc entre **Cin** et **Cfin** dont le label est **rel**. L'on peut après à sa guise définir des types de relations. Dans le cas d'une application à la construction, on aurait les types de relation *lien-électrique, lien-mécanique*, etc.

Ces graphes, bien que ne portant pas le nom de graphes conceptuels sont basés sur les concepts développés dans [MC95].

Bien que la modélisation de la connaissance ne soit pas exclusivement représentée par des GC, leur utilisation est primordiale et montre bien leur

6. Dans ce cas, c'est une application à l'accidentologie de la route.

intérêt. Pourtant un défaut qui semble être inhérent à ce formalisme est le fait que «deux sommets-concepts ne puissent être reliés par deux arcs ayant une sémantique contradictoire». Il **ne supporte pas l'incohérence**.

Un problème plus important se pose: celui de l'**arité des relations**. En effet, le principe de base de la création d'un GC impose que l'on fixe l'arité des relations lors de la création des types de relations. Ceci entraîne une certaine rigidité. Les énonçats sont des phrases de la langue naturelle. Il s'y trouve des locutions verbales dont l'arité change. Afin de les modéliser correctement, il faudrait pouvoir construire des relations d'ariré variable. C'est pour le moment impossible dans les GC.

Un autre point gêne l'utilisation des GC, c'est la volonté de mettre à jour une **sémantique logique**. Ainsi, il est supposé que toute connaissance du système a une vérité unique qui ne peut être infirmée. La nature humaine et l'évolution prise à son sens le plus général montrent bien que ce n'est pas toujours le cas.

En conclusion, les graphes conceptuels constituent un **formalisme adéquat** pour **représenter et raisonner** sur de la connaissance avec l'avantage de la représentation graphique utile dans une interface homme-machine. Le rapprochement avec la logique du premier ordre bien que constituant une ouverture pour l'acquisition de la connaissance, et une utilisation jusqu'alors inexploitée de la théorie des graphes semble en contradiction avec l'ambiguité et l'incertitude existant dans le **savoir**. De plus, la rigidité due à la création de types de relations d'arités fixes empêche la modélisation de certaines phrases de la langue naturelle.

1.4.2 CLASSIC

CLASSIC [BMPS+90] est un système de représentation de connaissances basé sur les logiques terminologiques dont la particularité est de tenter d'être «*simple*» et d'avoir des opérations réalisables en un temps «*raisonnable*».

Ses créateurs, reconnaissant les possibles défauts des logiques terminologiques, expliquent les **critères déterminants** pour la **pertinence** de son utilisation. En fonction des traits caractéristiques de CLASSIC, on peut exprimer son intérêt ou l'inverse pour une application.

Les avantages de CLASSIC

Le point principal de cette discussion est le fait que CLASSIC utilise un **langage à objets**. Tous les individus ont donc un **identifiant unique**. Ceci va permettre de retrouver de l'information facilement en utilisant la propriété de **classification automatique** des individus et concepts. On peut ainsi classifier les descriptions des informations en fonction des concepts qui sont représentés.

Cette hiérarchie mise à jour dynamiquement va donc permettre d'utiliser

CLASSIC dans des applications de style **bases de données** en autorisant des **évolutions de schéma**. Ceci est important, par exemple,quand on veut faire une application concernant la **conception ou à la spécification de modèles**

Les défauts de CLASSIC

En revanche, certains traits caractéristiques de CLASSIC lui interdisent certains types d'applications. L'**identification d'un objet** peut donner des résultats faux si l'objet «*a des notions d'égalité basées sur l'identité de ses composants*». C'est-à-dire que l'identification de cet objet par rapport à un autre objet de même type ne doit pas passer par l'identification d'autres objets[7].

La **hiérarchisation automatique** des concepts et des individus et les **tests de cohérence** ont plusieurs conséquences fâcheuses dont voici celles qui nous intéressent:

- le système supporte mal les **exceptions**,
- dans le cas d'applications dont le but est de retrouver des faits, sans idée d'utiliser des mécanismes d'inférence[8], le système pénaliserait l'utilisateur. Le **traitement des données d'entrée et les mises à jour** seraient **trop longues** pour des opérations aussi simples,
- la récupération de l'information est accélérée par la propriété de **chaînage avant** inhérente à de nombreux **systèmes classificatoires**, mais elle entraîne une lenteur lors de la saisie de nouveaux concepts ou individus.

En conclusion, l'utilisation de CLASSIC ou de son formalisme nécessite des **précautions** dues aux propriétés de ce système. Il n'a en outre qu'une **faibles représentation graphique**.

7. les dates sont un exemple
8. c'est notre cas pour l'instant

Chapitre 2

Les principales approches relatives au partage de la connaissance, à son accès, à sa présentation

Le partage de la connaissance est une démarche très ancienne. Elle repose sur deux thèmes très importants de la recherche : **l'accès à l'information**, partie intégrante de la recherche sur les bases de données et les systèmes d'informations, et la **présentation** de celle-ci, une des problématiques des interfaces homme-machine (hypermédias). Bien que ces sujets semblent peu liés, l'accès n'est pas seulement le fait d'obtenir physiquement un corpus de connaissance. C'est aussi la possibilité d'exploiter ce corpus. C'est à dire de **l'utiliser facilement, efficacement et agréablement**, ce qui relève de la présentation.

Depuis la grande bibliothèque, la tour de babel jusqu'à Internet, les moyens d'accès à la connaissance ont toujours été le résultat de la volonté d'un petit nombre d'offrir à une multitude la possibilité de partager la connaissance. Mais cette dernière a évolué entre ces étapes du progrès. Elle a subi des **mutations morphologiques et fondamentales profondes**. Les avancées réalisées ont entraîné une **croissance exponentielle de la quantité de connaissances**. Le petit nombre, bien qu'infime par rapport à la totalité de la population du globe, est devenu très grand. Les moyens de diffusion ayant énormément évolué, de la tablette de marbre à la télévision et aux grands réseaux, il est plus facile d'y accéder.

Ainsi on peut maintenant trouver la connaissance sur des supports très différents : livres, bandes vidéo, CD-ROM.... Mais certains problèmes persistent. Les domaines étudiés sont maintenant extrêmement variés. C'est pourquoi le désir d'acquérir la connaissance sur un thème passe par une spécialisation de plus en plus poussée des intéressés. La recherche d'infor-

mations s'avère plus ardue. Son accès doit donc être facilité par un **langage de recherche de l'information simple et efficace** en ce qu'il doit permettre d'obtenir **l'information pertinente** pour la prise de décision ou le diagnostic par exemple. De plus la volonté de **partager la connaissance** impose de la mettre sous une forme qui soit **agréable, facilement et efficacement exploitable**. Ceci entraîne la question: **les supports actuels sont-ils suffisants et si oui sont-ils bien utilisés?** Nous allons montrer dans les chapitres suivants, l'adéquation partielle des démarches déja effectuées.

2.1 L'accès

2.1.1 Les moyens d'accès traditionnels

De la tradition orale, au support papier l'évolution des moyens d'accès a **suivi et influencé** l'évolution de la connaissance. Dans son étude anthropologique de la population des chercheurs, Bruno LATOUR [Lat95] affirme que seul quatre pourcents des publications sont consultées. Il y a donc, par rapport au support papier, un problème: il est inhérent non seulement au **nombre très important** des publications mais aussi à la **spécialisation croissante** nécessaire à la compréhension de domaines de plus en plus précis. On peut ajouter à cela les contraintes liées au support papier. Ce sont les **exigences écologiques, matérielles** (on ne peut se déplacer facilement avec cinq tonnes d'articles), **physiques** (fragile), **sociales** (les bibliothèques ferment à des heures précises).

A ces exigences viennent s'ajouter des problèmes liés à **l'obtention de la source de connaissance**. Prenons l'exemple d'une bibilothèque où l'on voudrait consulter un livre. Dans ce cas, l'accès à la connaissance se fait de façon indirecte. C'est à dire que l'on n'accède pas directement au contenu d'un livre mais que l'on a une **référence** qui nous mène à ce contenu. La recherche dans un **thésaurus** qui est le meilleur moyen sur papier, bien que très ancien, d'obtenir des informations sur les livres d'une bibliothèque impose la contrainte de poursuivre la recherche de l'information au dela du SI que constitue ce thésaurus.

Les moyens proposés s'avèrent insuffisants à cause de la **quantité d'information à gérer et des variations rapides** de celle-ci. On peut en déduire que la communication avec des moyens traditionnels est insuffisante dans la mesure où elle impose des contraintes aussi bien sociales que structurelles. De plus, l'intérêt des utilisateurs semble se porter sur des supports plus faciles d'accès et moins contraignants tels ceux proposés par les outils informatiques.

2.1.2 Les mécanismes de bases de données

Les bases de données ont été créées pour gérer de grandes quantités d'information. Elles constituent un bon moyen de stocker l'information et de la retrouver. La possiblité de mettre à jour une base de données permet de résoudre le problème de l'augmentation de la quantité de connaissance. On peut rechercher plus simplement l'information grâce à des **langages de requêtes ou des mécanismes de recherches par mots-clés.**

Ce moyen est le premier réellement efficient proposé par l'informatique pour régler le problème de l'accès à la connaissance. Toutefois, si l'on reprend le problème de la bibliothèque, l'utilisation d'une BD de références bibliographiques (obtention d'un numéro ISBN) ne rend pas l'accès direct.

Il faut donc que l'information aussi soit informatisée. Dans les **bases de données documentaires**, on peut accéder directement à des documents techniques en particulier, et à des textes de tous types.

Mais l'obtention d'un document ciblé sur un thème n'est pas l'accès à la connaissance sur ce thème. En effet, l'auteur de ce texte s'impose en dictateur. Il nous énonce sa connaissance sous une forme statique. Et bien souvent, dans un livre de quatre cent pages, seul un paragraphe nous intéresse. Nous subissons donc la dictature du texte et de son auteur. Ici, le problème n'est pas **l'accès physique** mais **l'accès à une information pertinente.**

En ce qui concerne les contraintes sociales et structurelles dont souffre le support papier, les BDs ont apporté des solutions intéressantes. Grâce aux progrès des **systèmes répartis**, il est possible aux utilisateurs du système d'accéder à l'information à partir de plusieurs postes avec des possibilités d'accès concurrents à partir d'endroits différents (sites dispersés). Mais ce sont les seuls utilisateurs du système qui en auront la possibilité. La connaissance n'est donc pas distribuée à qui en a besoin mais à qui en a le «droit».

Un autre type de problèmes intervient celui de la **compréhension**. On ne peut prétendre connaître quelque chose si on ne l'a pas compris. L'intérêt de cette remarque tient au fait que **l'information pertinente seule n'est pas suffisante** parfois. Elle nécessite souvent l'apport d'informations complémentaires ne serait-ce qu'au niveau de la sémantique.

2.1.3 Les hypermédias

Dans un hypermédia, les informations forment un **réseau** dans lequel on peut naviguer. La lecture de l'information se fait par le **parcours des noeuds** de ce réseau. L'hypermédia a aussi pour vocation de **présenter l'information de façon agréable et explicite.**

L'idée a été émise dans [Bus45], mais l'évolution de ce concept date réellement des années quatre-vingts. Ces mécanismes offrent aux utilisateurs la possibilité de **parcourir de l'information** contenue dans des noeuds

grâce à des liens entre ces noeuds. L'accès est réalisé à travers une **démarche itérative**. Celle-ci correspond mieux au **modèle mental** de l'utilisateur qui recherche une information. On peut en effet agir par **association d'idées** en cliquant sur un mot du texte, une partie de l'image, etc... mis en avant par une sémantique graphique particulière symbolisant le fait qu'il lui correspond une information. Le résultat de cette action est l'obtention d'un document traitant le corpus d'information rattaché à ce mot, ou à cette image, etc....

Cependant l'hypermédia ne possède pas les avantages offerts par une base de données (gestion de l'information en grandes quantités). Il est donc **limité dans son évolution, il est statique**. Il est donc nécessaire d'avoir des mécanismes permettant d'ajouter des informations, de les stocker et de les retrouver parmi une grande quantité.

2.2 Internet

Internet est un réseau développé dans les laboratoires de recherche du département américain de la défense (DARPA). D'abord dédié aux communications entre les différents laboratoires de recherche de la défense, il fut de plus en plus utilisé par les autres laboratoires pour échanger des informations sur leurs travaux.

Internet est aujourd'hui devenu le **grand réseau de communication mondial**. Une information mise sur Internet peut être **accédée de n'importe où à n'importe quelle heure et pratiquement par n'importe qui**[1]. C'est bien le meilleur moyen de diffusion de la connaissance actuel.

Bien que très critiqué en raison de sa **désorganisation**, et de **la lenteur de ses services** à cause de ces mauvaises spécifications de départ, il est utilisé par des millions de personnes dans le monde. La recherche de l'information est réalisée au travers de **moteurs de recherche**. Ceux-ci sont basés sur des mécanismes de bases de données documentaires. Ils ont pour la plupart une représentation textuelle des sites qu'ils permettent d'atteindre. Celle-ci est constituée soit de **mots-clés** entrés par la personne qui s'occupe du site soit de mots-clés déduits d'une analyse automatique du texte du site référencé. L'accés physique est donc simplifié.

En ce qui concerne la **pertinence de l'information**, les sites sont constitués de **pages hypermédias**. Elles permettent donc le parcours de la connaissance par association *simple* d'idées. Ici le terme *simple* est utilisé pour refléter le caractère simpliste des liens entre les noeuds des pages web[2]. La communauté des chercheurs s'intéressant aux hypertextes a en effet depuis longtemps dépassé cette vision des liens hypermédias.

1. ce qui pose quelques problèmes d'éthique
2. ce sont les pages visibles sur les sites INTERNET

Troisième partie

Notre approche

Le chapitre précédent ciblé sur l'acquisition et le partage de la connaissance nous a permis de déterminer l'orientation de notre travail. Le monde de la recherche et plus particulièrement le monde scientifique constitue un «microcosme autarcique». La **transmission du savoir *techno-scientifique*** a lieu par l'intermédiaire de **communications, d'articles et de rapports sur papier**. Le nombre de chercheurs croissant, il est de plus en plus **difficile d'accéder à leur connaissance**. Afin de pallier ce manque nous avons voulu présenter la connaissance sous la forme de graphes, les **organisations** représentant un domaine précis. Ceux-ci, pour une exploitation universelle, doivent être accessible de tous les points du globe et à n'importe quel moment. Le média Iternet en constante évolution semble correspondre au meilleur moyen actuel de diffuser une information. Mais avant d'en arriver là, les objectifs, précisés dans la première partie de ce mémoire, doivent être remplis.

Nous avons d'abord répondu à la nécessité de formaliser la connaissance en nous inspirant de modèles existants et en y ajoutant les spécificités qui nous sont utiles. Afin de valider ces modèles augmentés, nous avons d'une part démontré leur lien avec nos objectifs et avec les propositions énoncées dans [Bou96] et [Bou90] sur les processus répétitifs d'acroissement des connaissances (PRAC), d'autre part concrétisé nos idées en concevant un outil, **EcoInfoSys** annexe à un système d'information capable de permettre à un chercheur d'organiser sa connaissance selon des critères visuels répondant à ces modèles mentaux, et de la distribuer sous la même forme en usant d'un langage de programmation adapté au média utilisé.

Chapitre 1

Des modèles pour formaliser et exprimer la connaissance

1.1 Les énonçats, une base de connaissance

1.1.1 Analyse et exploitation

Les énonçats dont la définition figure au chapitre 3 de la première partie constituent un **ensemble de phrases.**
Exemple : La systémique fait l'étude des systèmes.
Afin d'affiner notre étude de ceux-ci, nous les avons analysés sémantiquement et syntaxiquement de façon à comprendre leur mécanisme de formation et de trouver le meilleur moyen de les représenter.

Leur syntaxe indique qu'ils constitués d'un **ensemble de concepts** organisés autour d'une ou de plusieurs **locutions verbales**. Leur sens est l'établissement des **relations entre les concepts** du domaine représenté et des **définitions** de ceux-ci. Ceci implique une **construction hiérachisée et en réseau.**

C'est à dire que l'énonçat porte sur un concept particulier. Il le définit, l'explique en se servant d'autres concepts. On peut ainsi définir deux catégories de concepts identifiables aux catégories de concepts des représentations de connaissance à objets, les concepts dits **primitifs** non définis et les concepts **définis**. Si l'on considère que les concepts primitifs forment les fondations d'un édifice, les concepts définis sont le rez-de-chaussée et les étages suivants d'ou l'apparence d'une **hiérarchie cumulative**, c'est à dire que le niveau *n* utilise toutes les propriétés du niveau *n-1*. On se trouve alors au cours de la recherche d'une information sur un concept à consulter les énonçats des niveaux inférieurs. Mais l'image s'arrête là. Car certains concepts sont définis à partir de concepts soit n'appartenant pas au niveau directement précédent soit appartenant aux niveaux suivants. Il arrive ainsi que l'on boucle. Si l'on part d'un énonçat contenant la définition d'un

concept, on peut aboutir à la définition d'un concept déja utilisé au cours du parcours. On a donc l'impression de parcourir un réseau de connaissances. Toutes ces remarques nous ont conduit à la **notion de graphe**. Ces objets mathématiques très bien adaptés à l'informatique permettent de décrire des **systèmes dynamiques** dont les étapes assimilées aux sommets peuvent être parcourues **sans ordre défini**. Le concept de graphe et la théorie qui s'y rattache sont donc bien adaptés à la modélisation des énonçats.

Cependant, un problème vient se poser. Si l'on doit définir **une méthode de conception des énonçats**, il faut pouvoir expliquer non seulement leurs utilité, but, sens, mais aussi comment les réaliser de façon à ce qu'ils soient exploitables. En effet, les énonçats fournis au départ montrent bien des difficulté à être représentés graphiquement. En fait, le **langage naturel** utilisé lors de leur conception les rend très différents les uns par rapport aux autres.

Il se trouve que la diversité des mots pour exprimer le même concept entraîne des problèmes. La **synonymie** semble être un des problèmes majeurs. Mais il s'en trouve d'autres liés en particulier au fait que l'auteur d'un énonçat ne le conçoit pas avec la **pensée de modélisation informatique**. Ainsi, bien que définis comme des phrases simples, certains contiennent **l'explication** ou la définition nécessaire accompagnée d'une **explication complémentaire ou de morphologie différente**. Cette dernière, bien qu'intéressante sur le plan de la compréhension, vient compliquer le travail d'une procédure automatique. Nous nous limiterons donc à une procédure manuelle réalisée par le chercheur " organisant sa connaissance ". Il s'agira pour lui de ne représenter de l'énonçat :

- que ce qui est **utile**. C'est-à-dire, ce qui suffit à la **définition** et à l'**explication** du concept de la façon la plus large possible,

- que ce qui est **simple**, c'est à dire que tout énonçat «subdivisible» en énonçats doit être «subdivisé», seuls les **énonçats «indivisibles»**, c'est-à-dire perdant tout sens dans cette «division», doivent subsister.

Ceci dit, la recherche d'une méthode de conception nous a amené à **proposer** un certain nombre de solutions applicables sur des énonçats déja réalisés en langue naturelle ou à la création de ces énonçats que nous appelons **normalisations**.

1.1.2 Les normalisations applicables aux énonçats

Il s'agit là non pas d'énumérer les différentes normalisations mais d'en comprendre l'utilité.

La normalisation sémantique

La première qui est une **normalisation sémantique** est la résolution de la synonymie. Il s'agit de permettre la **minimalisation de la terminologie** utilisée. Elle est bénéfique dans la mesure où elle simplifie les traitements et dans la mesure où l'on ne pourra rencontrer des synonymes de termes déja définis et qui nous soient inconnus.

La normalisation syntaxique

Une autre normalisation possible est la **normalisation syntaxique**. Elle a pour cadre la **pensée de modélisation informatique** qui ne peut s'exprimer pleinement que dans des structures formelles précises. La normalisation syntaxique tend donc à faire d'un énonçat une structure simple et standardisée comme le montre la figure suivante :

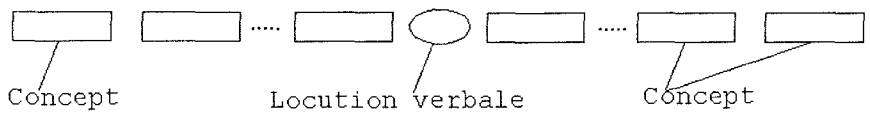


FIG. 1.1 – *Schéma standard d'un énonçat.*

Cette forme facilite l'utilisation d'un formalisme graphique et de type objet pour représenter l'énonçat.

1.1.3 Un formalisme objet pour modéliser concepts et relations

Ce formalisme objet doit être apte à modéliser les énonçats sous leur forme standardisée. Si la représentation des concepts, souvent abordée dans les SRC, est simple à mettre en oeuvre, la locution verbale inhérente à l'énonçat doit être analysée pour avoir une idée précise de son traitement.

Nous avons pu classer les locutions verbales en trois groupes :

- l'**association** : c'est la relation qui lie deux concepts. Il existe des types d'association particuliers : l'**héritage** et la **connexion** explicités dans les paragraphes suivants. Nous avons souhaité garder le terme *association* pour décrire les relations simples entre concepts. C'est à dire les liens simples tels celui d'un terme à sa définition, d'un objet à sa figure, d'une DIC à son attribut,

- l'**héritage** : c'est l'association représentée dans les **hiérarchies de type objet**. Elle implique qu'une classe de niveau n-1 hérite des propriétés des classes de niveau n, ces propriétés sont partagées par toutes

les classes de niveau n-1. Elles se distinguent entre elles par des propriétés *particulières* ou *secondaires* qui leur sont propre. Cette hiérarchie est dite cumulative, chaque niveau ascendant étant défini par les propriétés communes aux classes sous-jacentes,

- la **connexion** : c'est l'association qui lie un «**tout**» à ses **parties** ou un **ensemble** à ses **sous-ensembles**. Elle implique aussi l'**échange d'information** (énergie, matière, bit, ...) existant entre ces différents composants. Elle correspond à la notion d'**agrégation** ou de **composition** utilisé dans les formalismes objet classiques à la différence non négligeable qu'elle déclare **modéliser les relations entre les agrégats ou composants** ce qui est absent des formalismes cités précédemment. Un exemple expliquant ce qu'est la connexion est celui du **robot**. Un robot, nous le choisissons rudimentaire dans un souci de clarté, est composé de boulons, vis, et engrenages interagissant dans le but de le faire fonctionner. Autrement dit, si l'on a un ensemble de boulons, vis et engrenages séparés les uns des autres et sans interaction, on ne peut avoir un robot. Ainsi la notion d'agrégation ou de composition est insuffisante telle qu'elle est définie généralement, car elle ne semble pas tenir compte de la relation entre les agrégats ou composants.

A propos de la connexion, il faut souligner que le terme a déja été utilisé pour décrire une relation " tout-partie " spatio-temporelle [Gua95]. Les chercheurs s'intéressant aux **objets complexes et composites** ont réalisé une classification des différentes relations de composition dans laquelle on retrouve cette notion. Nous affirmons que l'utilisation de ce terme dans notre propos ne correspond pas à celle citée dans ces articles en ce que nous ne nous limitons pas au contexte spatio-temporel mais à tous les contextes possibles.

Au moyen de cette **classification tripartite** des relations entre concepts, nous pouvons choisir un formalisme objet pour représenter les énonçats. C'est le **modèle objet d'O.M.T.**.

1.2 OMT, une méthode de conception de système d'information

1.2.1 Présentation

O.M.T. (Object Modelling Technique) est une méthode de modélisation objet. Issue des progrès de la recherche en matière de conception de programmes et des avancées des langages à objet, elle a été développpée par **J. RUMBAUGH** [RBP+91] en 1991 dans les laboratoires de la General Electric aux U.S.A..

Les méthodes d'analyse et de conception classiques sont basées sur une vision tridimensionnelle de la modélisation objet (données, architecture, comportement) autour d'un noyau qui est le système lui-même. O.M.T. ne déroge pas à la règle en fonctionnant selon trois points de vue (modèles objet, dynamique et fonctionnel) qui chacun relèvent d'un aspect important du système. Le **modèle objet** contient un ensemble de représentations graphiques et de méthodes pour décrire la structure statique des objets et leurs relations. Le **modèle dynamique** décrit les interactions entre objets et les changements au cours du temps. Le **modèle fonctionnel** donne un comportement général au système, c'est à dire qu'il décrit ce qui arrive quand le système fonctionne. Ces trois modèles sont construits au cours de la phase d'analyse. La phase de conception permet la description et la construction de l'architecture générale du système. Les trois modèles sont ensuite affinés les uns par rapport aux autres.

O.M.T., bien que récent, est déja très utilisé. Ses processus itératifs et ses formalismes graphiques puissants facilitent son utilisation et son évolution.

1.2.2 Le modèle objet d'O.M.T. utilisé comme modèle de représentation de la connaissance

Pour comprendre le choix de ce modèle, nous le détaillons afin de bien prouver son adéquation à nos spécifications. Comme précisé plus haut, il décrit la structure statique des objets et de leurs relations. C'est-à-dire qu'il est supposé qu'un système (informatique ou autres) est consitué d'objets de types différents ayant des relations les uns par rapport aux autres.

Le modèle propose trois types de relations, l'**association**, l'**agrégation** qui est un type d'association particulier modélisant la relation entre un «tout» et ses parties, et la relation d'**héritage**. Ces trois types sont familiers aux utilisateurs de langages à objet et compréhensibles pour des opérateurs de bases de données. Ils sont directement issus du **modèle entité-association**. A l'aide de ces trois groupes de relations et des objets, on peut construire un **diagramme d'objets** permettant de **visualiser** la structure du système grâce à un **formalisme graphique puissant**.

Chaque classe d'objet est représentée dans un rectangle. Les relations entre deux classes ont elles aussi des représentations graphiques propres. Ainsi la relation de spécialisation/généralisation (héritage) est décrite par un arc coupé en son milieu par un triangle. L'association est un arc étiqueté. Enfin, l'agrégation se constitue d'un losange portant sur l'agrégat et d'un arc allant vers les parties. La figure suivante montre un exemple de ses trois relations:

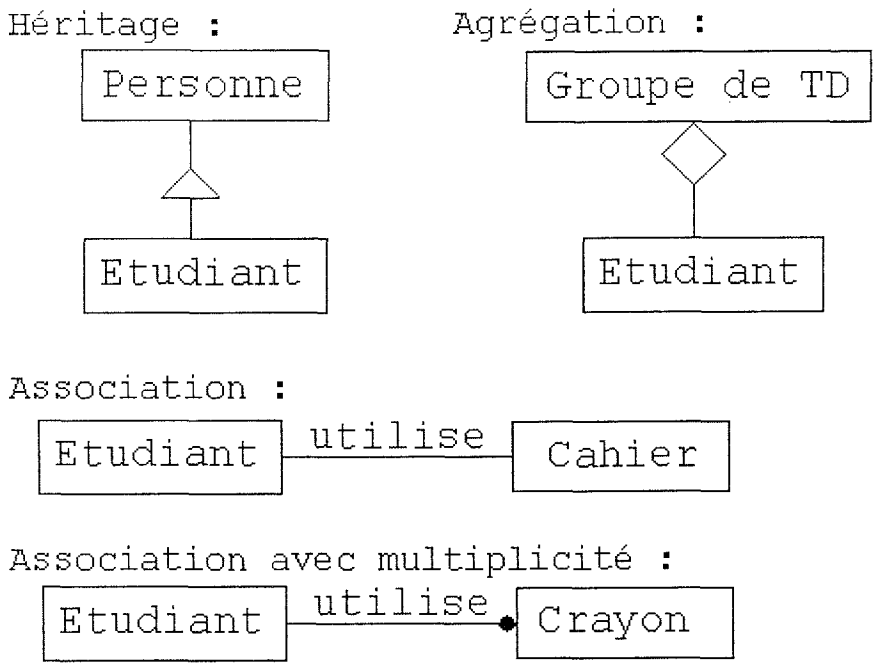


FIG. 1.2 – *Formalisme graphique du modèle objet d'O.M.T..*

1.2.3 Convergence entre nos objectifs et le modèle objet d'O.M.T.

Ce formalisme nous convient à double titre dans la mesure où il respecte des conventions graphiques et où il nous permet de représenter les concepts développés dans les chapitres précedents, c'est-à-dire les concepts du domaine décrit par les énonçats et leurs relations. En effet, l'énonçat «un étudiant utilise un cahier et des crayons[1]» peut être représenté graphiquement en utilisant les conventions vues à la figure précédente.

La sémantique utilisée par les concepteurs d'O.M.T. différant de la nôtre, il faut en montrer les correspondances. Les relations d'héritage et d'association sont les mêmes. En revanche, la relation d'agrégation décrite dans la méthode de conception de RUMBAUGH correspond à la connexion. Elle souffre d'ailleurs dans O.M.T. de la même insuffisance que dans les autres modèles objet existant. Mais le graphisme utilisé n'ayant pas de liaison particulière avec la sémantique de la relation, c'est-à-dire que l'on aurait pu choisir un cercle ou un pentagone que cela serait revenu au même, on le considère valable.

Un lien supplémentaire est nécessaire à l'adoption définitive du modèle

1. Le pluriel du concept «crayon» est obtenu au moyen de la **multiplicité**. Celle-ci permet de définir une quantité ou une relation de valuation sur le concept.

objet d'O.M.T.. C'est celui entre les PRACs et le modèle.

1.3 Convergence entre PRACs et le modèle objet d'O.M.T.

1.3.1 PRACs et DICs

Dans [Bou90], l'auteur établit la nécessité de distinguer **deux types d'éléments de connaissance** : les **DICs** observées sur les objets d'étude des sciences exactes, et leur **interprétation** formalisée en **énonçats** [Bou96]. Mais ces derniers ne s'organisent que selon **deux et seulement deux** PRACs : le PRAC de **comparaison typologique** et le PRAC **eusystémique**. Il faut ici bien entendre cette distinction. Les DICs sont des faits observés **sans aucune interprétation**, mais reliés au réel par des relations de type **association** au sens défini au chapitre 1 section 1.1.3 de cette partie. Les relations représentées au sein des énonçats sont d'un ordre différent.

1.3.2 Interprétation et PRAC

Un PRAC est un processus dynamique d'interprétation où l'apport de nouvelles DICs permet de reprendre l'interprétation en accroissant la connaissance par cet apport, mais aussi par une représentation meilleure, plus complète. La connaissance interprétée aboutit toujours à des résultats hypothétiques. Ils sont en effet fondés sur l'application consécutive de deux familles d'hypothèses :

- les **options**, c'est à dire le processus ayant conduit à retenir une problématique de recherche, puis dans celle-ci un objet d'étude, et enfin dans cet objet d'étude un ensemble de données objectives (les DICs) supposées adéquates. Ensuite cet ensemble de données sera l'objet des propositions,

- les **propositions**, organisant ces ensembles en matrices, admettant que les fonctions mathématiques choisies sont pertinentes, que celles-ci combinées à nouveau aboutissent finalement à des **estimations** permettant d'obtenir les résultats de l'interprétation.

1.3.3 Le PRAC de comparaison typologique et l'héritage

Le PRAC de **comparaison typologique** [Bou90] porte sur des objets, les prélevats, **identifiables** grâce aux valeurs, les DICs, des attributs caractérisant ces objets. Ces comparaisons de prélevats visent, à dégager des classes, et souvent un prélevat «**typique**». L'exigence opérationnelle est que ces prélevats soient **identifiables**, c'est-à-dire ne soient pas des entités vagues, imaginées et non décrites par des DICs.

Ici les prélevats sont **identifiés**, comme précédemment, mais **connectés** : ce sont des composants eusystémiques formellement décrits en deux catégories : les **compartiments** et les **cinèses**. Les compartiments correspondent à des portions spatio-temporelles quelconques **identifiées**, c'est à dire aux composants de l'eusystème au sens classique du terme. Les cinèses sont les **échanges d'information** (énergie, matière, bit, ...) entre ces compartiments.

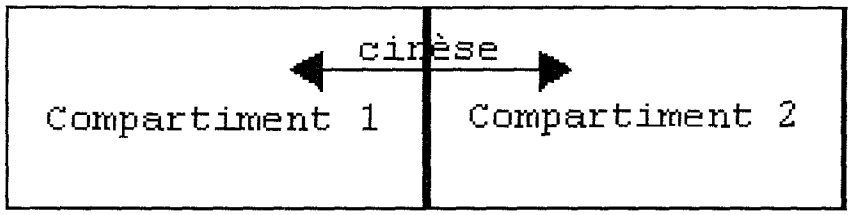


FIG. 1.3 – *Schéma d'un eusystème minimal.*

Par exemple, deux composants : une ampoule électrique et une douille d'ampoule connectées par une cinèse : le courant alternatif. Les objets ne sont pas seulement composés mais ils sont aussi agencés en fonction de leur connexion.

Le formalisme O.M.T. s'applique bien à ce PRAC si l'on pose toutefois l'exigence de connexion commme étant la caractéristique reliant les éléments des agrégats ainsi assemblés.

Remarquons qu'une **hiérarchie eusystémique**[2] peut également être dégagée de l'étude et de la représentation d'eusystèmes, tels des écosystèmes ou des robots, mais l'abord de celle-ci ne se fait pas commme précédemment par le *bas* d'une hiérarchie ascendante. Précisons ici que la construction d'une hiérarchie eusystémique n'est guidée que par l'ordre chronologique de la perception du système. Ainsi si l'on a les compartiments et les cinèses, on les associe pour former le système (similaire à la composition). Si l'on a une connaissance entière du système, on peut le décomposer (similaire à la décomposition) en ces compartiments et ces cinèses. Le choix est donc indifférent de la nature ou de la représentation préalable du système. Le choix est **pragmatique**.

On peut commencer en considérant les composants (gros systèmes, micro-ordinateurs, réseaux) d'un *eusystème informatique matériel.* Une première possibilité est de le décomposer analytiquement en ses composants. Mais il existe un deuxième abord qui serait de partir d'un transistor d'une puce que l'on relierait peu à peu par connexion aux (eu)systèmes où cette puce joue un rôle.

2. hiérarchie de composition dans les modélisations objet

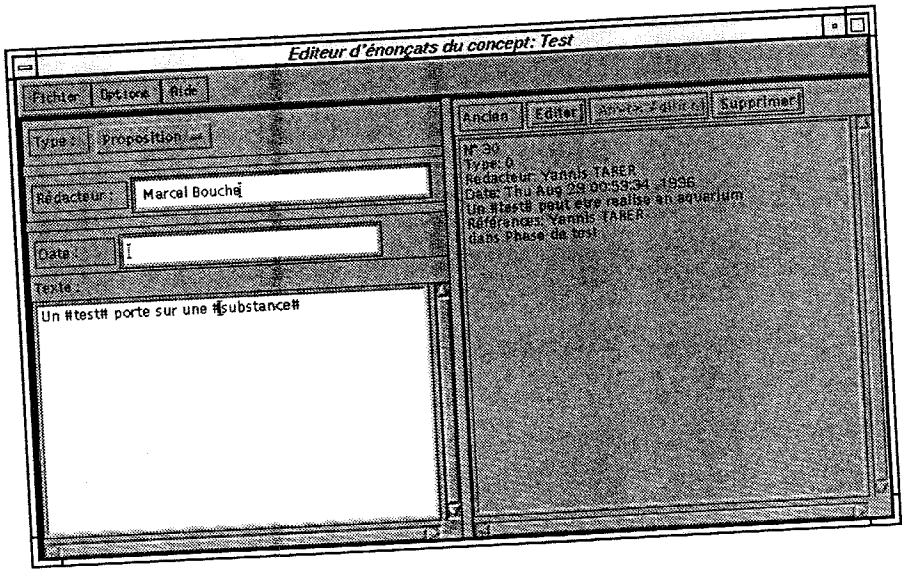


FIG. 3.4 – *Editeur d'énonçats du concept Test.*

Quatrième partie

Perspectives

Bien que le sujet de ce mémoire soit l'**expression et le partage de la connaissance**, la proposition de départ comportait bien d'autres thèmes de l'intelligence artificielle que les limites temporelle et physique ne nous ont pas permis d'aborder.

De nombreuses perspectives peuvent être envisgées à partir du système existant :

- l'**émergence d'ontologies**,
- le **traitement de la langue naturelle**
- l'**interaction entre SI et SRC**,

De plus, les outils proposés devant fonctionner **réellement**[4], il semble obligatoire d'en faciliter l'usage (**manuel online**[5]) et d'en étendre les fonctionnalités.

4. Ce ne sont pas des prototypes, ils doivent être opérationnels à la mi-novembre.
5. Ce manuel a été réalisé sous forme d'organisations

Cette approche comme d'autres utilise la logique pour obtenir l'assurance de la cohérence du domaine. Comme nous l'avons dit précédemment, il semble dangereux pour l'ouverture du système de contraindre la connaissance représentée en lui affectant une valeur de vérité. Toutefois, il faut souligner et c'est le plus important que notre but n'est pas de réutiliser la connaissance, il est de la **présenter, de la critiquer ou d'en permettre la critique**. Ainsi la construction d'une ontologie consiste pour nous, en partie, à déterminer un ensemble de concepts d'un domaine et de relations entre ces concepts menant à une **vision globale**, une **perception globale** de ce domaine.

1.2.2 Le niveau ontologique, une approche philosophique et linguistique.

Dans son article, Nicola GUARINO [Gua95] reprend le thème des ontologies. Il revient en fait à une démarche plus proche de la **philosophie** qui définit une **ontologie au sens platonicien du terme** comme «*la science de l'être en tant qu'être*». Il aborde donc l'ontologie d'un point de vue très différent de celui de GRUBER.

Ceci s'explique non seulement par la différence des définitions mais aussi par la différence des objectifs. En effet, l'auteur se propose d'utiliser l'ontologie dans le **processus de modélisation de la connaissance**. Le but de l'article lui-même est d'établir les relations entre l'ontologie et la représentation de connaissances. Pour cela, il insère une distinction dans les modèles existants se basant sur une structure en quatre niveaux, *logique, épistémologique, conceptuel et linguistique*, établie par [Bra79]. Cette distinction est dénommée le **niveau ontologique**. Il s'agit d'un niveau se situant entre le niveau *conceptuel* dont les entités primitives (relations conceptuelles) sont définies de la façon dont elles sont perçues dans le réel, et le niveau *logique* dont les entités primitives (notions de prédicats, de fonctions) sont définies de façon générale. L'introduction du **niveau *épistémologique*** [Bra79] avait pour but de casser le passage entre réalité et abstraction avec un langage formel de représentation de connaissances. Mais cet effort est considéré par GUARINO comme insuffisant dans la mesure où il ne permet pas de dire quand on doit **déclarer** un objet comme un **concept**, un **rôle** ou un **attribut**.

Le niveau ontologique correspond à ce manque. Il permet d'établir correctement la **sémantique du formalisme de représentation**. C'est-à-dire qu'il permet de classer les concepts et les relations en différents groupes permettant d'aider l'utilisateur dans ses choix de conceptualisation.

GUARINO ne s'attache donc pas à la façon de raisonner sur la connaissance entrée dans un système à base de connaissances. Mais il voit ce système comme imparfait tant qu'il ne possède pas une sémantique claire et bien établie de la modélisation conceptuelle: «*In practice, formal ontology*

can be intended as the theory of a priori distinctions: among entities of the world...; among meta-levels categories used to model the world (concepts, properties, qualities, states, roles, parts...)»

1.3 Les possibilités offertes par le système

«Most current candidate ontologies are hidden, since they are part of the infrastructure for some effort in communicating among people, or beetween people and systems they share.»

WIEDERHOLD [Wie94] explique que l'ontologie provient souvent **naturellement** d'un **effort de communication**. Dans **EcoInfoSys**, l'effort réalisé a pour but de donner une perception globale d'un domaine. Elle doit être suffisante pour pouvoir raisonner à partir des éléments les plus fins de la connaissance actuelle de ce domaine.

Nous observons donc l'apparition naturelle d'une ontologie lors de la conception d'une organisation. Cette constatation est liée au fait que **nos exigences de départ sont ontologiques**. En effet, nous pouvons nous situer partiellement par rapport aux critères de GRUBER et à des critères propres à nos objectifs :

- la **clarté**, elle est le résultat de notre volonté de permettre l'accès à tous à la connaissance. Nous avons donc été contraints d'utiliser des définitions aussi larges que possibles et une représentation indépendante du niveau symbolique (selon NEWELL [Neb90]),
- l'**extensibilité**, elle est inhérente à tout bon SRC,
- la **charge ontologique minimale**, elle découle de la granularité minimale imposée aux éléments de connaissance, et des normalisations imposées aux énonçats (minimalisation de la terminologie (Cf. partie 3 chapitre1)).

D'un autre point de vue, si l'on reprend la définition de GUARINO, l'existence dans notre modélisation d'un niveau ontologique peut être dévoilée. Elle fait aussi partie des spécifications de départ. Mais son utilisation n'est pas voulue. L'obtention d'un formalisme et la preuve de l'adéquation de celui-ci avec des phénomènes " naturels " d'organisation de la connaissance impliquent l'**explicitation précise des relations existants entre les concepts**. Ainsi **on sait** quand on a une association, un héritage ou une connexion.

Dans l'avenir, nous espérons exhiber ces différents types d'ontologie, les ontologies construites de GRUBER et la science décrite par GUARINO, de façon tangible.

Chapitre 2

Un module de traitement automatique de la langue naturelle

Il est clair qu'en facilitant l'expression de la connaissance par la création d'un module de **traitement de la langue naturelle** et en automatisant la création des concepts et des relations à partir des énonçats, nous parviendrions à de meilleurs résultats.

Dans notre approche, nous avons abordés plusieurs **problèmes liés à l'utilisation de la langue naturelle**. Malgré tout, nous ne cherchons pas à **proscrire son emploi**, c'est pourquoi nous nous proposons de réaliser un module de traitement de la langue naturelle dont l'objectif serait l'analyse des énonçats afin :

- d'en **extraire automatiquement les concepts «importants»** que nous pourrions ainsi éviter de rentrer manuellement. Ce processus serait tout de même réalisé sous la surveillance de l'utilisateur organisant sa connaissance.

- d'en **extraire automatiquement les différentes locutions verbales** dans le but de constituer le **classement tripartite** nécessaire à la **décision du type d'association entre deux concepts**.

Chapitre 3

Vers une plus grande interaction des SI et des SRC

La volonté de faire évoluer les SI et les SRC nécessite d'appréhender les applications possibles de ce processus. Nous pouvons les citer rapidement, nous ne retenons que celles qui sont intéressantes dans notre cadre :

- l'**intégration bases de données (BD) bases de connaissance (BC)** qui a pour but de permettre l'utilisation conjointe des mécanismes de ces deux types de bases. Les chercheurs travaillant sur ce sujet tentent en particulier de **mettre à jour des méthodes de génération de requêtes** dont les résultats (dans l'un des deux formalismes ou dans un formalisme fédérateur) contiennent des informations collectées dans les BD et dans les BC,

- l'**acquisition de connaissances** qui nécessite dans le cadre des sciences empirico-formelles de **rechercher l'information** permettant de **générer de nouvelles connaissances** dans de grandes bases de données. Cette recherche peut elle-même être facilitée par l'amélioration de l'approche suivante appelée «**data mining**»,

- les SRC peuvent aider à l'**optimisation des requêtes** permettant de retrouver l'information dans un SI afin de **réduire le temps de recherche**.

Ce besoin d'intégration illustré par l'activité de la recherche dans ce domaine est lié à la prise de conscience des **limites des deux paradigmes de gestion des objets de la connaissance**. Comme cité plus haut, un SI et un SRC ont des objectifs et des qualités différents. Mais certains de leurs défauts peuvent être effacés grâce à la volonté de les faire évoluer l'un vers l'autre.

Dans notre cas, nous avons **deux priorités**. La première exprime la **nécessité d'interfacer** la base de donnée BASECOL. En effet, certains

énonçats peuvent être traduits en **formules mathématiques ou statistiques**. Celles-ci utilisent des variables prenant leurs valeurs dans l'ensemble des DICs. Il est donc important de pouvoir faire des **requêtes** sur la BD permettant d'obtenir les résultats de ces formules (**intégration BD et BC**). La deuxième découle de la première. Les résultats des requêtes après **interprétation** pourraient donner naissance à de nouveaux énonçats (**acquisition de connaissances**) au moyen des mécanismes d'inférences adéquats.

Mais il ne s'agit pas que de la génération de nouveaux énonçats. Nous pensons que pour **gérer l'ensemble des concepts, relations et énonçats** des mécanismes de base de données s'imposent. Ainsi nous projetons d'utiliser les nouvelles possiblités de Java (**Java DataBase Connectivity**) concernant les BD pour réaliser des mécanismes de requête sur les attributs des concepts (s'il y a un énonçat ou pas, ou s'il hérite d'un autre concept), des relations et des énonçats.

Chapitre

Conclusion

Comme nous l'avons précisé en introduction, notre intention était de faire évoluer un SI (ECORDRE) vers un SRC. Notre travail a plus précisément consisté à enrichir le système initial d'un module de représentation de connaissances (acquisition, représentation, accès).

Les approches de la représentation de connaissances vues dans la deuxième partie (réseaux sémantiques, langages à base de frames et logiques terminologiques) témoignent de la volonté de représenter la connaissance à partir de son état «naturel», telle qu'elle est présentée par les experts.

Dans ce cas, c'est l'expert lui–même, dans un but d'efficacité sur son propre travail, qui a été amené à s'interroger sur la nature des éléments de connaissance qu'il manipulait. Il s'agit ici de reconnaître les concepts réellement «premiers» de la connaissance à représenter.

Ceci a conduit à couper clairement l'univers de la connaissance en trois secteurs distincts (le réel, la cognition et la société dont font partie les utilisateurs) décrits outre les relations par deux objets : les DICs et les énonçats.

Les énonçats ont servi de base à notre travail. Leur analyse sémantique et syntaxique nous a permis de concevoir une approche *originale* et *réalisable*.

Le résultat est EcoInfoSys qui se décompose en : l'organiseur et le distributeur. Le premier s'appuie sur une modélisation et une représentation de la connaissance dans le but de faciliter son expression et son partage. Cette modélisation est réalisée à partir du formalisme du modèle objet d'O.M.T.. Nous avons établi l'adéquation de ce formalisme et des processus " naturels " d'organisation de la connaissance (les PRACs). Au travers du concept d'organisation, nous avons traduit nos exigences de visualisation et de compréhension. Le deuxième est basé sur des techniques d'hypermédias. Elles nous ont fourni des méthodes pour présenter et expliciter la connaissance contenue au sein du système et pour permettre à tous les types d'utilisateurs de raisonner sur des éléments de connaissance afin d'avoir une compréhension globale d'un domaine.

Au cours de la conception de ces prototypes nous avons pu constater que

leur champ d'application pouvait être élargi à la conception d'ontologies ou pouvait s'élargir grâce à la description précise d'un niveau ontologique. L'apport de mécanismes d'inférences permettant de déduire de la connaissance contenue dans le système, de nouvelles connaissances est un facteur de l'intérêt que nous portons à l'avenir de ce système.

La poursuite de notre travail se base d'une part sur une modélisation poussée des relations entre les aspects empiriques, les aspects formels et les possibilités d'aide à la décision à partir de ces différents aspects. D'autre part, il est important de continuer à comprendre les processus permettant de déterminer les plus petits éléments de la connaissance formelle dans un domaine pour mieux la transmettre à tous et entraîner ainsi des prises de conscience, base nécessaire du progrès.

Bibliographie

[Ber93] M. Bernstein. Enactment in information farming. In *Hypertext 93 proceedings*, 1993.

[BM86] M.L. Brodie and J. Mylopoulos. Knowledge bases versus databases. In *On knowledge base management systems*, pages 83–86, 1986.

[BMPS+90] R.J. Brachman, D.L. McGUINESS, P.F. Patel-Schneider, L. Alperin Resnick, and A. Borgida. Living with classic : When and how to use a kl-one-like language. In *Principles of Semantic Networks*. Morgan KAUFMAN, Juin 1990.

[Bou90] M. Bouche. *Ecologie opérationnel assitée par ordinateur*. 1990.

[Bou96] M. Bouche. Integrologie, ... 1996.

[Bra79] R.J. Brachman. On the epistemological status of semantic networks. In *Associative Networks : Representation and Use of Fnowledge by Computers*, pages 3–50, New-York, 1979. Academic Press.

[Bre95] P. Bresciani. The challenge of integrating knowlege representation and databases. 1995.

[BS85] R.J. Brachman and J.G. Schmolze. An overview of the kl-one knowledge representation system. In *Cognitive Science*, volume 9, pages 171–216, Avril 1985.

[Bus45] V. Bush. As we may think. *Atlantic Monthly*, pages 101–108, 1945. in Un système hypertexte pour l'acquisition de connaissances de S. Fraisse.

[Die95] R. Dieng. Conflict management in knowledge acquisition. *Artificial Intelligence for Engeneering Design, Analysis and Manufacturing*, pages 337–351, 1995.

[FFPR95] A. Farquhar, R. Fikes, W. Pratt, and J. Rice. Collaborative ontology construction for information integration. 1995.

[Gra90] A. Le Grand. *Conception d'un Poste de Travail d'Aide à l'Acquisition de Connaissances.* PhD thesis, LIRMM, 1990.

[Gre95] F.D. Greco. Introduction to java: a lok at what java promises... and delivers. Web, 1995.

[Gru93] T.R. Gruber. Towards principles for the design of ontologies used for knowledge sharing. In *International Workshop on Formal Ontology*, Mars 1993.

[Gua95] N. Guarino. Formal ontology, conceptual analysis and knowledge representation. *Journal of Human and Computer Studies*, 1995.

[HS95] A. Van Hoff and S. Shaio. Hooked on java, 1995.

[Lat95] B. Latour. *Le métier de chercheur, regard d'un anthropologue.* éd. INRA, 1995.

[LBWR94] G.L. Lohse, K. Biolsi, N. Walker, and H.H. Rueler. A classification of visual representations. *Communications d'ACM*, 37(12):36–49, 1994.

[MC95] M. Mugnier and M. Chein. Représenter des connaissances et raisonner avec des graphes. Technical report, LIRMM, 1995.

[MI95] C.C. Marshall and F.M. Shipman III. Spatial hypertext : Designing for change. *Communications d'ACM*, (8):88–97, 1995.

[Min75] A framework for representing knowledge. In *The Psychology of Computer Vision*, pages 211–277. McGraw-Hil, 1975. in P. WINSTON.

[MMS+96] E. Mesrobian, R. Muntz, E. Shek, S. Nittel, M. Larouche, and M. Kriguer. Oasis : An open architecture scientific information system. In *RIDE 96*, 1996.

[Neb90] B. Nebel. *Reasoning and Revision in Hybrid Representation Systems.* Springer–Verlag, 1990.

[Pop34] K.R. Popper. *La logique de la découverte scientifique.* Ed. Payot, Paris, 1934.

[PS89] P.F. Patel-Schneider. A four-valued semantics for terminological logics. *Artificial Intelligence*, pages 319–351, Juin 1989.

[Qui66] M.R. Quillian. PhD thesis, Carnegie Institute of Technology, 1966.

[RBP+91] J. Rumbaugh, M. Blaha, W. Premerlani, F. Eddy, and W. Lorensen. *Object-oriented Modelling and Design.* PRENTICE HALL, 1991.

[Ses] Sesmat. Dialectique... in Dictionnaire de la philosophie ed Larousse.

[TT96] P. Trigano and B. Talon. Aide à l'acquisition de connaissances sémantiques : des cadres sémantiques aux graphes conceptuels. Projet d'article pour la revue RIA, 1996.

[Wie94] G. Wiederhold. Interoperation, mediation, and ontologies. 1994.

[Woo75] W.A. Woods. What's in a link : Foundations for semantic networks. In Academic Press, editor, *Representation and Understanding : Studies in Cognitive Science*, pages 35–82. New York, 1975. in D.G. Bobrow and A.M. Collins.

Plan général

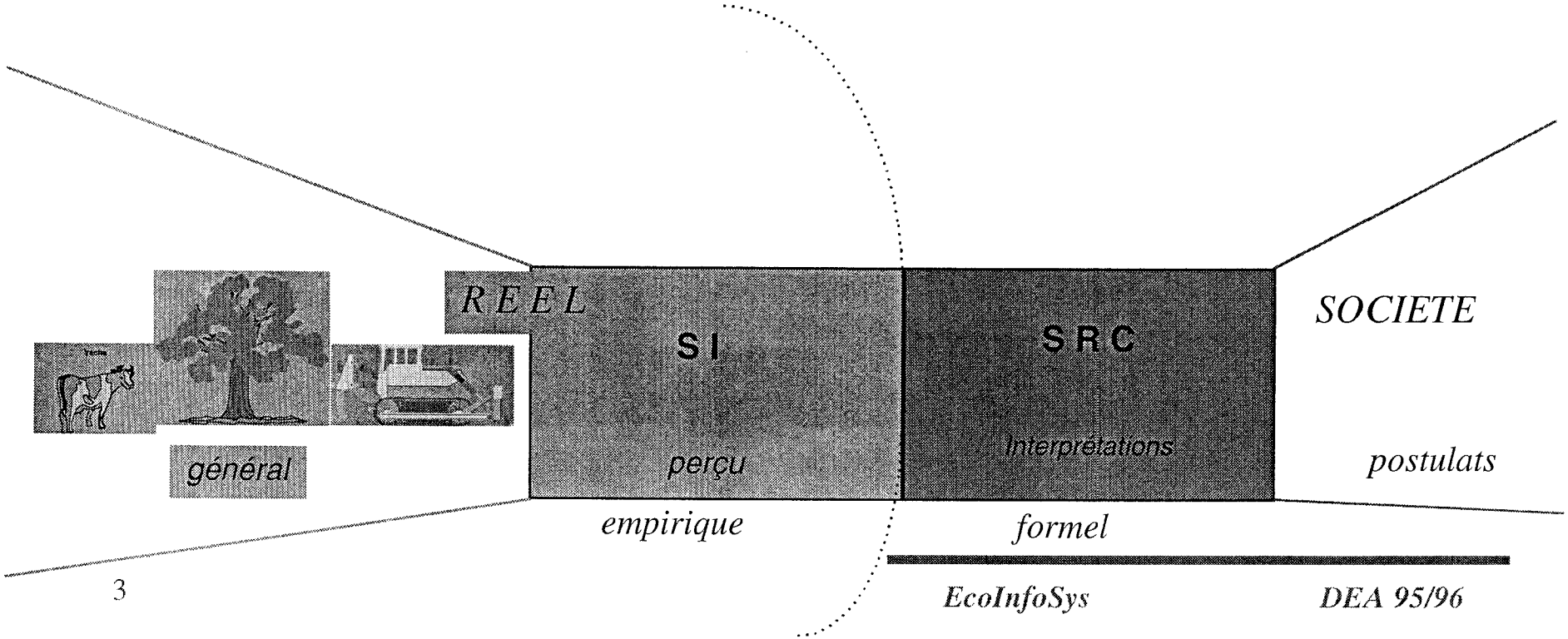
- **Motivations et contexte**
- **Objectifs**
- **Systèmes existants**
- **Solution proposée**
- **Perspectives**
- **Conclusion**

Motivation et contexte

Sciences de la nature : sciences empirico-formelles

➡ **Deux types d'éléments de connaissance**

- **Données observationnelles et expérimentales**
- **Interprétations**

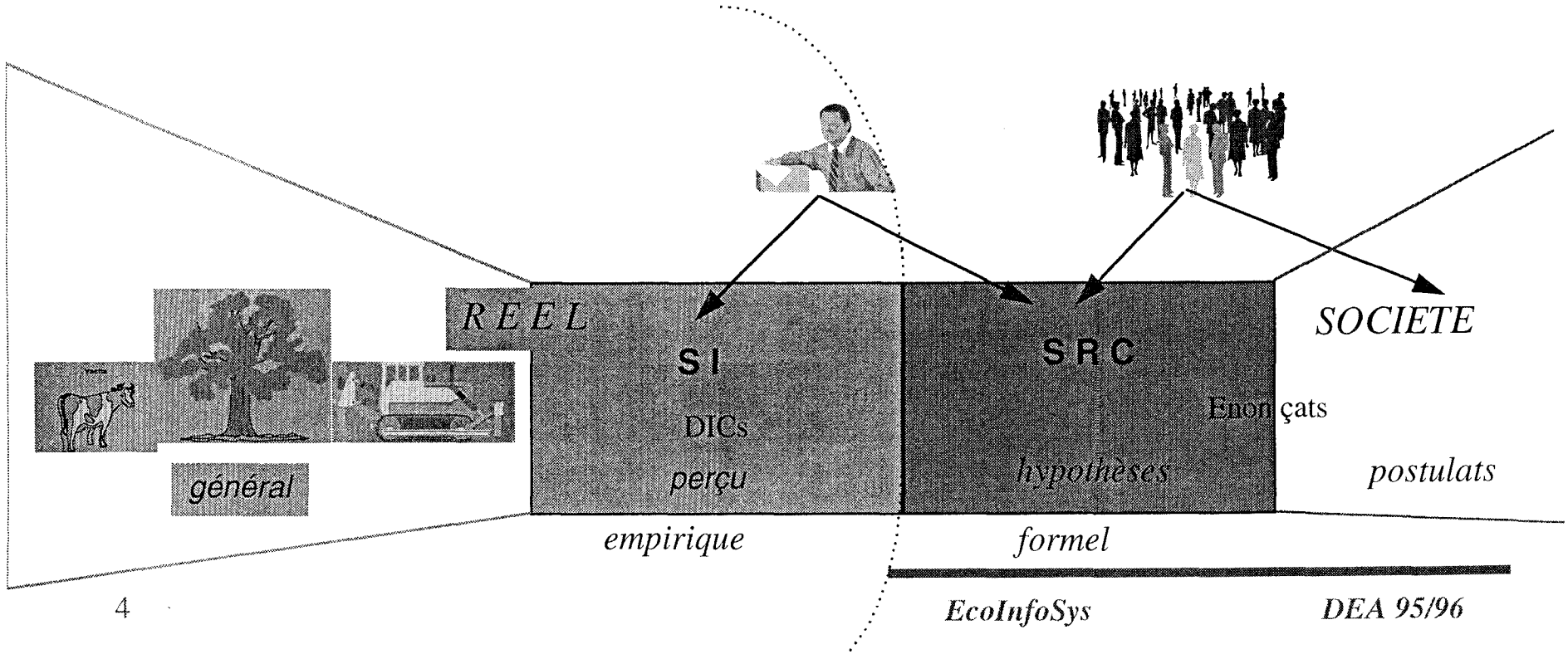


Motivation et contexte

ECORDRE: système d'information transdisciplinaire

➠ **Deux types d'éléments de connaissance**

- **DICs**
- **Enonçats**



Objectifs

Faire évoluer un SI vers un SRC

- **SI gère l'information nécessaire à l'organisation**
  - ➟ permet de gérer l'information en grande quantité
  - ➟ permet de retrouver l'information pertinente

  connaissance empirique

- **SRC gère l'organisation de la connaissance et sa révision**
  - ➟ accès et gestion de la connaissance
  - ➟ mécanismes d'inférence

  connaissance formelle

Objectifs

Un formalisme pour exprimer la connaissance

- **Connaissance sous forme de phrases**
  - ➟ perception globale d'un domaine
  - ➟ recherche de l'information pertinente

Des outils pour : *- acquérir de la connaissance*
- accéder à la connaissance

- **Connaissance technoscientifique**
  - ➟ accès, confrontation de la connaissance
  - ➟ présentation de la connaissance

ECOINFOSYS
Expression et partage de la connaissance
dans le cadre des sciences de la nature

DEA Informatique 1995/96

Yannis TARER

2 Octobre1996

Equipe SIC & BD