CAHIER DE RECHERCHE

DECEMBRE 91 ■ N° 23

LA CODIFICATION DES OBJETS COMPLEXES :

Réflexions théoriques et application à un corpus de 8 000 produits alimentaires

Saadi Lahlou
Joelle Maffre
Valérie Beaudouin

CREDOC

Centre de Recherche pour l'Étude et l'Observation des Conditions de Vie

CRÉDOC

CAHIER DE RECHERCHE

La codification des objets complexes : réflexions théoriques et application à un corpus de 8 000 produits alimentaires

Saadi LAHLOU, Joëlle MAFFRE, Valérie BEAUDOUIN

Ces travaux ont bénéficié d'une subvention de recherches
du Commissariat Général du Plan,
dans le cadre du programme général de recherches du Crédoc

Ils ont été effectués dans
le Département Prospective de la consommation
sous la direction scientifique de Saadi LAHLOU

DECEMBRE 1991

142, rue du Chevaleret
75013 - PARIS

SOMMAIRE

Résumé :

La codification d'objets complexes est un problème récurrent dans le travail statistique. Nous avons mis au point et testé une méthode de codification basée sur une description des objets à l'aide de facettes. La codification utilise pour attribuer des descripteurs aux objets une méthode mixte, combinant des règles d'inférence (codage direct) et la comparaison à des prototypes (codage indirect). Un logiciel (CITOCA) a été mis au point, qui permis de coder rapidement plusieurs milliers de produits alimentaires, de façon compatible avec le système LANGUAL de description des aliments.

Summary :

Codification of complex objects is a permanent issue in statistical work. We have developped and tested a codification method based on a facet description of objects. Descriptors are affected to objects with a mixed method : direct coding with inference rules, and indirect coding by comparison with prototypes. A software (CITOCA) has been developped that allowed coding in a short time several thousands of food products in a way that is compatible with the LANGUAL food description system.

1. L'Observatoire des consommations alimentaires

"Comment pourrait-on classer les verbes qui suivent : cataloguer, classer, classifier, découper, énumérer, grouper, hiérarchiser, lister, numéroter, ordonnancer, ordonner, ranger, regrouper, répartir ?
Ils sont ici rangés dans l'ordre alphabétique.
Ces verbes ne peuvent pas tous être synonymes ; pourquoi aurait-on besoin de quatorze mots pour décrire une même action ? Donc ils sont différents. Mais comment les différencier tous ? Certains s'opposent d'eux-mêmes, tout en faisant référence à une préoccupation identique, par exemple, découper, qui évoque l'idée d'un ensemble à répartir en éléments distincts, et regrouper, qui évoque l'idée d'éléments distincts à rassembler dans un ensemble.
D'autres en suggèrent de nouveaux (par exemple : subdiviser, distribuer, discriminer, caractériser, marquer, définir, distinguer, opposer, etc.), nous renvoyant à ce balbutiement initial où s'énonce péniblement ce que nous pouvons nommer le lisible (ce que notre activité mentale peut lire, appréhender, comprendre). "

Georges Perec, Penser/classer.
pp. 154-155.

La question des nomenclatures, qui est celle de la description structurée des objets, est ancienne. Elle a été principalement débattue sous ses deux aspects philosophiques : ontologique (qu'est-ce que l'objet que l'on décrit) et épistémologique (qu'est-ce que le processus de description du phénomène par l'observateur). Il s'agit de discussions difficiles, souvent érudites, et qui touchent de tellement près aux postulats de notre système de description du monde que les positions prises par les différents auteurs sont, finalement, axiomatiques. Nous ne chercherons pas ici à apporter une contribution philosophique à ce problème éternel. Notre approche, plus restreinte et plus pragmatique, concerne la résolution technique du problème, pour le statisticien.

Le projet sur lequel nous allons confronter nos recherches théoriques à des données empiriques est celui de la base de données de l'Observatoire des consommations alimentaires (OCA). Le département Prospective de la consommation a été chargé par les pouvoirs publics de mettre sur pied cet outil, en collaboration avec le laboratoire de recherche sur la consommation de l'INRA.

"L'Observatoire des Consommations Alimentaires (OCA) a pour vocation de fournir des données quantitatives sur les consommations de groupes d'aliments, de composants

alimentaires, de nutriments, d'additifs et de contaminants. Son rôle devrait être de contribuer à une évolution durable et cumulative des moyens de recueil de l'information sur l'alimentation en France.
Afin de déterminer des seuils, ainsi que les caractéristiques de consommation de certains groupes de consommateurs, on devra estimer la composition d'un régime alimentaire moyen, mais aussi et surtout, évaluer la dispersion des niveaux de consommation au sein de la population.

Pour obtenir ces résultats, l'Observatoire utilisera deux types d'éléments fondamentaux :
- un ensemble de données statistiques sur la consommation des produits alimentaires,
- un ensemble de données analytiques sur la composition des produits (contenu des aliments en nutriments, additifs, contaminants...).

Les consommations des différents groupes d'aliments ou de leurs composants sont obtenues en combinant ces deux ensembles de données.
(...)
Exemple : Le directoire de l'Observatoire désire connaître la consommation d'un composant alimentaire particulier X dans la population française, et les groupes qui en sont sur ou sous-consommateurs.

Confrontée à cette question, l'équipe chargée d'exploiter la base de l'OCA la décompose en 4 interrogations distinctes, qui permettent de répondre progressivement à la question : Qui sont les consommateurs de X ? :
1 - Quels sont les aliments-sources de X ?
2 - Qui consomme ces aliments-source et en quelle quantité ?
3 - Quels sont les dosages de ces aliments-sources en X ?
4 - Combien cette consommation d'aliments-source représente-t-elle d'apport en X ?

A la première et à la troisième question, la base statistique ne peut pas répondre à elle seule, même si elle peut apporter une aide à la recherche d'informations.
La première réponse de la base de l'OCA à ses utilisateurs est donc une question :
- Dans quels aliments-source risque-t-on de trouver le composant X ? (dans quelles grandes catégories d'aliments le trouve-t-on, à quoi sert-il, est-il lié à une phase particulière du processus de production, dans quoi est-il autorisé, bref, quelles sont les caractéristiques qui permettraient de faire un premier repérage de ce qui est susceptible de contenir du X ?).
Munie de ces informations, la base livre une liste des références, repérées et suivies par ses sources, qui sont susceptibles de contenir X, et une liste de produits, sorte d'extrait de nomenclature pertinente pour le sujet qui nous intéresse. Les utilisateurs doivent alors, au mieux de leurs possibilités, fournir les teneurs en X des produits concernés, ou, si possible, des références concernées. Il incombe à l'équipe chargée de la base de s'arranger pour utiliser au mieux les dosages fournis. Naturellement, les résultats obtenus seront plus fins si l'on peut disposer d'un dosage spécifique pour chaque référence (par exemple à partir des fiches de recette) que s'il existe une estimation unique que l'on appliquera à tous les produits[1]."

---

[1] Lahlou, S., Beaudouin, V., Calamassi-Tran, G., Evans, C., Gillet, C., Lion, S., Maffre, J., Verheyden, G., Rapport pour l'Observatoire des Consommations Alimentaires. Etat d'avancement des travaux de la base de données mise en place par le Crédoc à la fin de la deuxième phase : décembre 1991. Crédoc : Département Prospective de la consommation.

La base de l'OCA est constituée à partir des données de consommation tirées de l'enquête sur la consommation alimentaire de l'Institut National de la Statistique et des Etudes Economiques (INSEE), et celles du panel SECODIP (Société d'Etudes de la Consommation, Distribution, Publicité). Pour lire, harmoniser, et exploiter ces données, il est nécessaire de pouvoir les décrire de la manière la plus flexible possible. Plutôt que de se limiter à une simple harmonisation de nomenclatures, qui aboutirait à un nivellement par le bas, nous avons cherché à trouver une solution qui respecte les différents "points de vue" des sources, tout en assurant la possibilité de leur exploitation dans une autre optique.

Nous avons, pour l'Observatoire des Consommations Alimentaires, un certain contexte d'utilisation, délimité par les intérêts des utilisateurs, et nous savons que ces intérêts concernent la consommation, la santé publique, la réglementation, la nutrition, l'orientation de la production... Il nous faut donc disposer d'une *description* des produits alimentaires suivant des dimensions qui soient compatibles avec ces champs d'intérêts.

Or, les descriptions de produits dont nous disposons à l'origine, les "références SECODIP" et les "catégories INSEE", portent sur des dimensions pertinentes pour les utilisateurs des systèmes qui ont recueilli ces informations pour l'utiliser à leur compte : SECODIP et l'INSEE. Le premier opérateur a une vision mercatique, puisque les données servent à déterminer les parts de marché de ses clients industriels et distributeurs, et les profils des acheteurs des produits ; le second, comme son nom l'indique, a une vision statisticienne et économiste destinée d'abord à la connaissance générale et à la comptabilité nationale.

Il nous faut donc mettre au point un système de description des objets (produits, ménages, achats...) fournis par ces sources, qui soit compatible avec les objectifs de l'OCA.

Le cas particulier à propos duquel nous examinerons ces questions est donc celui de la consommation par les ménages des produits alimentaires disponibles sur le marché français ; mais nous espérons que les résultats de nos recherches ont une portée plus générale, et seront utiles à tous ceux qui souhaitent faire des recherches statistiques sur des populations dont la description fait appel à un grand nombre de catégories d'objets complexes.

2. La question des nomenclatures

2.1. Statistiques et nomenclatures

La statistique c'est l'art d'étudier les objets nombreux[2]. Cette notion "d'objets nombreux" mérite qu'on s'y arrête. Pour pouvoir qualifier des objets de "nombreux", il faut qu'on puisse les compter comme une population. Là intervient la première opération de classement, qui considère tous les individus statistiques comme identiques en ce qu'ils sont dans une même classe, la population.

Mais ensuite, le travail du statisticien consiste à décrire, à partir des mêmes individus, différents objets dans lesquels ces individus de base seront considérés soit comme identiques, soit comme différents. Il y a dans la description même d'une population une contradiction. A la fois on considère cette population comme un ensemble d'individus de même nature, et comme une collection d'objets différents.

Tantôt, quand on décrit en tant que tel un objet constitué comme un agrégat d'individus, l'individu observé, indistinct, se fond au profit d'un individu modèle (moyen, médian, typique etc.) qui représente alors l'agrégat. Tantôt, au contraire, pour construire un agrégat et le décrire par rapport à un autre, on utilise les différences des individus qui les constituent.

Voyons par exemple, la description que fait Linné, lors de son voyage en Laponie en 1732, des vêtements des Finnois et des Lapons[3].

> "Les Finnois en Österbotten sont habillés presque comme des Lapons et donc *conveniunt cum iis** .
> *Conveniunt* : 1. bonnet, 2. veste gris clair, 3. pantalons descendant dans les chaussures, 4. bottes, 5. *cingulo cum cultro affixo*** 6. pas de boutons mais des agrafes.
> *Differunt* : 1. *quod careant* *** un col haut, 2. *quod habeant* **** un foulard, 3. le paletot ouvert devant, 4. *cingulo* avec pas plus d'un couteau qui pend, quelques clefs sur le côté derrière, 6. une bande autour du genou.

---

[2] Saadi Lahlou, 1989.
[3] Carl Von Linné. *Voyage en Laponie*,. Editions de la différence. Paris, 1983, pp 128-129.
* ils leur ressemblent
** ceinture où est fixé un couteau
*** en n'ayant pas
**** en ayant

A l'église je les vis sanglés avec une ceinture en drap noir qu'ils enroulaient 2 à 8 fois autour de la taille, ce qui tranchait noir sur gris.
Les femmes sont habillées avec des choses achetées et complètement *diversae* des femmes lapones."

Les premières caractéristiques permettent de décrire "Finnois-et-Lapons" comme une population unique. Dans cette population, le second groupe de descripteurs distingue deux sous-populations. Mais on peut se demander dans quelle mesure les traits sont descriptifs, et dans laquelle ils deviennent définitoires (par exemple pour l'observateur oculaire).

L'utilisation de traits manifestes comme signes d'appartenance à une catégorie et, réciproquement, l'utilisation d'une appartenance comme trait définitoire, traduisent le principe implicite de la classification par similarité des objets (on met dans une même classe des objets analogues). Cette méthode rejoint la façon naïve d'aborder les taxinomies.

De même que l'on rassemble les objets par analogie, on les distingue par leurs différences. Chaque objet est alors ambivalent : en tant que partie d'un agrégat, il perd son individualité, et n'est plus qu'une unité de quantité utilisée pour décrire l'objet qui le contient. Inversement, c'est à partir de ses qualités que l'on décrit les spécificités internes des agrégats.

Le point de vue de l'observateur définit la manière dont est perçu l'objet :
- un point de vue externe à l'agrégat fait que l'objet perd son individualité au bénéfice de l'agrégat ;
- un point de vue interne conduit à l'examen des traits distinctifs, pertinents, propres aux "membres" de l'agrégat.

Par exemple, on parlera "des Britanniques", sans spécifier les particularités de chacun des éléments de cet ensemble. A l'inverse, un Français A décrira un Français B à un autre Français C en énumérant certaines caractéristiques personnelles, et le plus souvent sans mentionner la nationalité. Dans certaines civilisations, il existe des systèmes de description véritablement taxinomiques, comme par exemple la *nisba*, dans le Moyen-Atlas, où les individus sont décrits par degrés successifs à partir de leurs communautés d'appartenance (Geertz, 1983 ; Lahlou, 1990).

Les nomenclatures sont l'outil scientifique qui sert à constituer et à décrire les agrégats. Pour le statisticien, la question des nomenclatures est donc à la base de toute analyse, car elle définit l'espace dans lequel il va pouvoir étudier le réel. Car le statisticien ne manipule que des comptages. Ces comptages sont nécessairement des comptages d'occurrences de types, des types que sont les postes de la nomenclature de description utilisée pour décrire les objets

que l'on compte. Sous la main du statisticien, le réel perd toute son épaisseur, et se réduit aux dimensions définies par les nomenclatures de description.

Cette opération n'est pas neutre : elle projette le phénomène dans un espace de description particulier, structuré par une certaine théorie. C'est d'ailleurs vrai pour tout taxinomiste, qu'il soit statisticien, linguiste, naturaliste, ou, d'une façon générale, chercheur.

On a déjà parlé de Linné. Son utilisation du latin scientifique lui permet de marquer la différence entre la perception naïve et la description classificatoire, et d'introduire ainsi une distance heuristique à l'objet qui est indispensable au chercheur. Cette distance est indispensable car la classification véhicule implicitement d'autres informations qu'une simple description phénoménale de traits. Elle constitue un choix d'analyse des phénomènes, par exemple dans la hiérarchie qu'elle introduit, et reflète alors bien plus qu'un simple choix technique[4]. Il convient alors de se rappeler sans cesse que l'on manipule des types arbitraires et non pas les objets eux-mêmes. La catégorie n'est pas la chose, le nom n'est pas l'objet nommé, et, pour reprendre la célèbre formule de Korzybski, la carte n'est pas le territoire. Tous les procédés, *typographiques* au sens large du terme, qui permettent de rendre visibles ces guillemets épistémologiques autour des objets d'étude, sont les bienvenus, qu'ils utilisent le jargon, le latin, ... ou les italiques.

On pourra comparer avec profit les différences entre les systèmes de description statistiques de la position sociale dans les différents pays. Ainsi, le système des PCS (profession catégorie sociale), organisé selon la double dimension économique et culturelle (ou sociale), est exclusivement français. Les anglo-saxons utilisent une échelle unidimensionnelle en niveaux sociaux. La notion de "cadre" est française et récente. Boltanski (*Les cadres,* Editions de Minuit) en analyse la naissance dans l'entre-deux-guerres.

Le classement consiste à assigner à un objet une position dans le système de description du monde de l'observateur. L'observateur qui classifie cherche à transformer des descriptions des objets telles qu'elles lui sont initialement fournies (par ses sens, par des descriptions indirectes, etc...) en coordonnées dans son propre système classificatoire.

Une classification est un système de partitions. Si ce système recouvre le domaine empirique d'où sont extraits les observables, tout objet, par la nature même d'une partition, sera classé dans au moins une sous-partie du domaine. L'attribution d'une classe d'appartenance à un objet observé se fait à partir de certains traits possédés par l'objet.

---

[4] Pour s'en convaincre, on se reportera aux âpres discussions autour de la taxinomie des espèces animales, qui renvoient en fait à des divergences sur les théories de l'évolution des espèces.

La puissance d'une classification réside dans la capacité à pouvoir marquer et distinguer facilement un grand nombre d'objets différents avec un petit nombre de traits et un petit nombre de règles simples.

Nous sommes en face d'un cas particulier du problème plus général de la traduction, qui est épistémologiquement impossible, bien que toujours pratiquée (Mounin, 1963). En fait, ce problème peut se formaliser de la façon suivante :

On cherche, à partir de la description d'un objet A dans une vision du monde Uo, à obtenir une description de ce même objet dans le langage U1.
Or clairement, par construction, il ne peut s'agir de "ce même objet", puisque, justement, il est traduit, et ne contient plus les mêmes traits constitutifs. Les traits utilisés sont différents, par exemple :

- "les animaux courent" (français contemporain)
- "τα ζωα τρεχει" (grec ancien)

ou encore :

- 0010130100000001 BIERE BLONDE FRANCAISE AUTRES MARQUES NON IDENTI LITRE MENAGE CONSIGNE (SECODIP)

-A0195-bois. fermte <= malt / B1217-eau / C0005-pas une partie de plte u d'anim / E0123-lq peu-visq - partic visible / F0001-niveau inconnu de transfo thermique / G0003-pas de méth de cuisson applicable / H0232-frmtation alcoolique / H0323-orge aj / H0319-blé aj / H0246-gazéifié par frmtation / H0229-arô épice ou herbe aromiq naturel aj / J0104-csrvé par fermentation / K0003-sans milieu de cdtionnmt / M0203-bout ou bocal / N0040-verre / P0024-produit de consommation courante / FR-FRANCE / 0 / 1-Recette (LANGUAL/OCA)

On cherche à passer de l'un à l'autre. Le problème de la classification ne peut être résolu sans faire un certain nombre de choix ontologiques. Cette question sera détaillée dans la seconde section de ce papier.

2.2. Nomenclatures et hiérarchie

Les opérations statistiques consistent à rassembler et à comparer les objets. Ces opérations sont antinomiques. On ne s'étonnera donc pas que les méthodes qui favorisent l'une compliquent souvent l'autre.

La description à l'aide de variables

Lorsque l'on utilise la statistique pour chercher des réponses nouvelles, et non pas seulement pour mesurer des objets à travers un cadre déjà connu, il importe au plus haut point de disposer d'un système de description flexible. Le système adopté couramment est celui de la description des objets dits "par variables".

Une variable est un symbole (X, Y, t, M, etc..) qui peut prendre toutes les valeurs d'un ensemble appelé le domaine de la variable. Si ce domaine est réduit à une valeur, la variable est dite constante. L'idée sous-jacente est de décrire l'objet en extension sous chacune des dimensions qui paraît, a priori, pertinente au statisticien. Ce système a l'avantage de la souplesse. Les variables sont un système efficace pour comparer les objets. Par contre, ce système n'est pas économique en termes de classement.

La description par nomenclature

Pour avoir une définition courte de l'objet, et pour se communiquer des résultats, on utilise souvent le système de la "nomenclature", dans lequel il existe à priori un nombre fini de catégories d'objets, dans lesquelles on peut classer tous les objets à classer, d'une façon univoque. Les nomenclatures hiérarchiques autorisent plusieurs niveaux de classement imbriqués.

Définition du Robert électronique :

> NOMENCLATURE [nómäklatyR] n. f.
> ◊ 1. (Déb. XVIIIe). Ensemble des noms, des termes employés dans une science, une technique, un art..., organisés selon les classes d'objets qu'ils désignent (en extension, alors que les terminologies opèrent en compréhension); méthode de classement de ces termes.
> ◊ 2. (1798). Ensemble* des formes (mots, expressions, morphèmes) répertoriées dans un dictionnaire, un lexique et faisant l'objet d'un article distinct.

A l'inverse du système de description par variables, le système de nomenclature hiérarchique est économique en termes de classement, car il contient implicitement une série de règles. Par exemple, supposons que l'on ait une nomenclature à deux niveaux, le niveau le plus élevé contenant deux modalités A et B, et le niveau inférieur 5 modalités, A1, A2, A3, B1, B2. La hiérarchie contient notamment les règles implicites :

si l'objet X possède la caractéristique A1, alors l'objet X possède la caractéristique A

si l'objet X possède la caractéristique A2, alors l'objet X possède la caractéristique A

si l'objet X possède la caractéristique A3, alors l'objet X possède la caractéristique A

si l'objet X possède la caractéristique B1, alors l'objet X possède la caractéristique B

si l'objet X possède la caractéristique B2, alors l'objet X possède la caractéristique B

Une nomenclature bien construite contiendra également implicitement les règles :

si l'objet X possède la caractéristique A, alors l'objet X possède la caractéristique A1, *ou* la caractéristique A2, *ou* la caractéristique A3

si l'objet X possède la caractéristique B1, alors l'objet X possède la caractéristique B1 *ou* la caractéristique B2.

Poussé à l'extrême, ce système de description cherche à englober tous les objets dans une unique variable hiérarchisée. C'est ce qui s'est passé dans les sciences de la nature, comme en Botanique à partir des travaux de Linné. C'est également ce qui se fait dans la description usuelle des produits alimentaires, notamment dans les enquêtes de l'INSEE.

> "Tellement tentant de vouloir distribuer le monde entier selon un code unique ; une loi universelle régirait l'ensemble des phénomènes : deux hémisphères, cinq continents, masculin et féminin, animal et végétal, singulier pluriel, droite gauche, quatre saisons, cinq sens, six voyelles, sept jours, douze mois, vingt-six lettres.

> Malheureusement ça ne marche pas, ça n'a même jamais commencé à marcher, ça ne marchera jamais. N'empêche que l'on continuera encore longtemps à catégoriser tel ou tel animal selon qu'il a un nombre impair de doigts ou des cornes creuses."[5]

Il est possible de renoncer totalement à un tel système qui, si l'on en croit Perec, "ne marche pas" à un niveau global. Mais comme nous avons cependant besoin de classer pour penser, il nous faut alors construire des classifications ad-hoc pour chaque problème spécifique à partir d'une description simplement en termes de variables non hiérarchiques. C'est tout à fait possible, et cela se fait en utilisant des techniques de classification hiérarchique (ascendante, descendante) ou de segmentation. Mais cela ne facilite pas forcément la communication, puisqu'alors chaque classification locale est idiomatique.

Rien n'est facile : nous semblons, en théorie, condamnés à perdre d'un côté ce que nous gagnons d'un autre. Mais ce constat, pour vrai qu'il soit, n'est pas constructif, et nous avons besoin d'un outil, fut-il imparfait. Alors, dans la pratique, confrontés à un vaste ensemble d'objets, sur lesquels on veut réaliser des travaux statistiques très variés, comment peut-on construire un système de description pratique, puissant, et flexible ? Comment réaliser concrètement ce système, et établir des correspondances avec d'autres systèmes descriptifs ?

3. L'utilisation d'un système de description combinatoire pour la catégorisation des produits alimentaires

> "(...) ce monde fut créé par combinaison de possibles qui, dans l'entendement divin, forment éternellement les notes en ségrégation (*sejuncta*), lettres ou nombres, d'une table qui convient à la nôtre en certains rapports. (...)
> Or, nous trouvons tous ces éléments «pêle-mêle», multitude de «détails» infiniment complexe, qui répond originairement à ce que le phénomène mondial a de «multiplex». Science des complexions élémentaires, la combinatoire est, pour Dieu, science des possibles et organon de la constitution du monde, elle est, pour nous, doctrine de déchiffrement de l'univers (...)".
>
> Michel Serres
> Le système de Leibniz
> et ses modèles mathématiques
> pp. 105-107, passim.

[5] Perec G., *Penser/Classer.*, op. cit.

3.1. La description du produit alimentaire

Nous nous intéressons ici à la consommation des "produits alimentaires". Cette expression ne veut rien dire en soi : il y a mille façons de définir un produit alimentaire, comme il y a mille façons de définir tout objet. Parmi toutes ces façons, certaines sont plus pertinentes dans notre contexte (sur cette notion, voir Sperber & Wilson, 1989). Ces façons pertinentes se caractérisent par un usage immédiat pour l'action qui nous intéresse. Il s'agit donc de façons de voir les objets qui soient adaptées à un certain contexte d'utilisation. Ces visions (les informaticiens parleront de "vues externes") sélectionnent, parmi les aspects possibles de l'objet, un certain nombre de dimensions pertinentes, ou facettes, dans chacune desquelles un objet possède une caractéristique, ou descripteur.

Par exemple, si les facettes sont : la famille, la marque, le poids, on pourra décrire un objet comme :

chocolat-noir-en-tablettes /Lindt/ 100-grammes

chacun des termes séparés par des "/" étant un descripteur correspondant à une facette.

Si les dimensions sont : le goût, la couleur, la consistance, on pourra décrire un objet comme :

chocolaté/brun-foncé/fondant

Si les dimensions sont : ingrédient principal, matériau de contact d'emballage, état physique dans les conditions standard de température et de pression, on peut décrire un objet tel que :

cacao/papier-d'aluminium/solide

Rien, évidemment, ne nous permet d'affirmer a priori que ces trois *descriptions* sont des points de vue différents d'un *même objet*.

Malgré sa trivialité apparente, cette affirmation nécessite dans certains cas une analyse en profondeur, qui soulève des problèmes ontologiques extrêmement délicats[6], beaucoup débattus depuis longtemps, et qui portent en dernier ressort sur la validité du postulat de

---

[6] Par exemple : à supposer qu'il existe un tel objet, serait-ce l'échantillon analysé ou la catégorie ? Qu'est-ce qui nous permet de distinguer le type de l'occurrence de type dans le cadre d'une description empirique où le type est le résultat d'une induction ? (Par exemple : "L'analyse a révélé la présence de Listeria"). Le lecteur intéressé par les questions d'identification des variables qualitatives (accord sur les noms de phénomènes) pourra se reporter par exemple à l'argumentation de W. V. O. Quine (1977) sur "lapin/Gavagaï", à Wittgenstein (1961, 1965). Pour les rapports entre mesure et dimensions sur les variables quantitatives, voir par exemple la première partie de G. Monod-Hertzen (1976) et H. Saget (1981).

réalité ("existe-t-il une nature immanente, dont la forme s'impose à l'entendement humain, ou est-ce l'observateur qui crée le monde à l'image de ses structures cognitives ?").

Une des façons d'ouvrir la porte sur l'abîme ontologique est de chercher à préciser l'origine des données : s'agit-il d'échantillons à doser (objet unique et concret), de postes de nomenclatures dans une enquête consommateurs (catégorie formelle), de déclarations par un ménage (énoncés en langue naturelle)... ? Peut-on manger *le poste* "chocolat en tablettes" de l'enquête INSEE ? Peut-on additionner deux tablettes, ou bien additionne-t-on les quantités mesurées des occurrences d'un même type ?

Ces questions naïves ne sont pas innocentes, car, si chacun sait qu'il est interdit d'ajouter des pois et des carottes, il nous faudra cependant bien additionner des *kilos* de pois et de carottes pour parler de kilos de légumes, surtout quand ils ne nous apparaissent que comme composants indifférenciés de l'objet décrit comme :

jardinière-de-légumes/surgelés/Picard/500g

Et il va falloir que nous sachions de quoi on parle pour fournir des statistiques sur les "légumes".

Car nous cherchons, finalement, à répondre à des questions dans un certain univers (par exemple, celui des nutritionnistes), à partir de données obtenues dans un autre (celui des panélistes). Ce que le nutritionniste voit comme assemblage de nutriments, le consommateur le voit comme "nourriture", le statisticien le voit comme "poste de nomenclature", et le panéliste le recueille comme "Unité de Vente au Consommateur". Aucune catégorie n'est exactement traduisible en une autre, car les représentations et les usages qui les sous-tendent sont de nature différente. Là où le consommateur voit "un steak", le nutritionniste distingue cent qualités de viande différentes. Là où le statisticien enregistre "un kilo de pommes de terre", le consommateur réalise cent recettes. Là où le nutritionniste voit "un camembert à 45% de matière grasse", le panéliste repère cinquante marques différentes, et le consommateur quatre gradations de mûrissement. Là où le législateur voit "une bouteille de vin de Bordeaux AOC", le distributeur voit des milliers de références différentes, le prix pouvant varier de un à dix ou plus selon l'année pour un même cru.

Certaines catégories pertinentes pour un opérateur n'existent même pas pour certains autres. Par exemple, les appellations "produits légers de petit déjeuner" (Lahlou, 1988), "garnitures", "grossissants" (Benguigui, 1973) (catégories consommateur), ou "sucres lents" (concept nutritionniste), "vins doux naturels" ou "ultra-frais" (catégories distributeur), etc n'évoqueront rien à certaines classes d'acteurs. Le consommateur pourra ne pas bien distinguer la limite entre les fromages frais et certains fromages blancs, pour lesquels la

réglementation cependant diffère. L'économiste, l'homme de marketing, et d'une façon générale chaque acteur de la filière perçoit des caractéristiques différentes (Lahlou, 1985).

Il est, par construction, impossible de fournir une solution exacte à une question impliquant un choix de catégorisation (par exemple : "combien un consommateur français consomme-t-il de lait ?"). Toute réponse sera une approximation, parce qu'il faudrait s'entendre sur ce qu'est "lait". Faut-il compter la poudre de lait utilisée en biscuiterie ? Et avec quel type d'équivalence de mesure quand il s'agit de poudre de lait *écrémé* ?

Pour préciser en quoi et combien c'est une approximation, il nous faut impérativement expliciter les opérations de traduction que nous faisons d'un univers dans l'autre.

Nous ne pouvons donc pas nous contenter de la commode position scientifique "classique", qui consiste à négliger la relativité des données. Il faut faire appel à quelques notions d'épistémologie, et il est impossible de faire l'économie d'une formalisation un peu rigoureuse. Ce n'est pas le lieu ici de détailler notre position sur ce problème, trop radicalement idéaliste pour ne pas soulever des discussions qui nous éloigneraient de notre objectif immédiat. Comme il faut, néanmoins, adopter une position comportant un minimum de relativisme, nous proposons de camper sur une position acceptable par tous, qui consiste à poser la relativité de toute mesure, et de toute description des phénomènes, sans chercher à trancher sur la qualité ontologique des phénomènes eux-mêmes.

Poirier (1981) donne un aperçu très clair de la manière dont on peut garder une position opérationnelle en faisant à la position idéaliste les concessions indispensables. Par exemple :

> "(...) presque toutes les mesures sont indirectes et se font dans le cadre et par l'intermédiaire de toute une théorie, ce qui ne veut aucunement dire qu'elles soient conventionnelles mais que les réponses se font dans le langage conceptuel des questions, et que le système des questions et des réponses s'adapte solidairement à l'expérience.
> (...) à l'inverse du monde des objets sensibles, qui a une sorte d'unicité absolue, le monde des objets théoriques a une pluralité d'aspects, suivant les systèmes conceptuels à l'aide desquels nous le décrivons ou le construisons. C'est ainsi que nous n'avons qu'une manière de voir un objet, mais diverses manières de le peindre ou de le dessiner, suivant les procédés, les techniques, les styles que nous mettons en oeuvre.
> (...) la réalité objective, c'est ce qu'il est nécessaire ou simplement plausible d'admettre pour sauver les phénomènes, dans un langage cohérent.
> Ce qu'il faut admettre, c'est que plusieurs langages sont possibles pour décrire un objet en lui-même mystérieux, sans qu'il soit possible de dire quel est le vrai, puisqu'aucune expérience ne peut les départager, et que par ailleurs ils correspondent en général à des orientations plutôt qu'à des contenus définis de la pensée."

3.2. Le problème de la transcodification

Nous avons jusqu'ici parlé de la question de la description des phénomènes, des objets naturels. Le problème qui nous occupe ici est un cas un peu particulier de ce cas général, dans la mesure où nous allons nous intéresser à des objets qui sont eux-mêmes des descriptions d'objets. C'est un problème de traduction. Notre matériau de base est en effet constitué de données recueillies par d'autres : nous n'avons qu'une connaissance de deuxième main des phénomènes.

Le problème posé à l'observatoire des consommations alimentaires peut être présenté de la façon suivante. Nous disposons d'une liste d'achats alimentaires, qui sont des dépenses effectuées :

- par un ménage donné
- en un lieu donné
- à une date donnée
- pour acheter un objet qui est lui-même un couple :
  - une certaine quantité donnée
  - d'une "référence SECODIP"

  ou d'un "produit INSEE" donné.

Par exemple, une référence SECODIP sera :

20401102030500/findus/filets poisson/colin lieu noir/de 300 à 499 g

et une référence INSEE :

6121 poissons surgelés

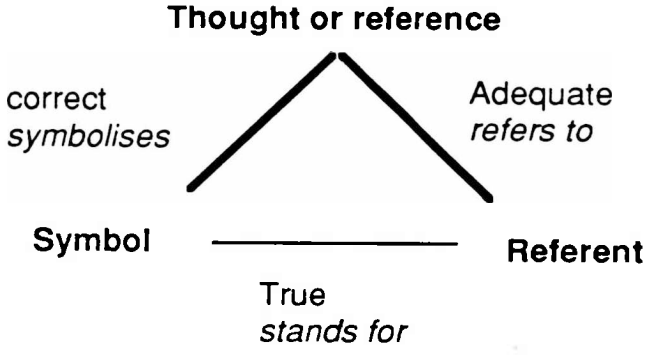
Ces listes ont été élaborées par d'autres observateurs, eux-mêmes en contact direct avec le phénomène indigène. Nous travaillons en chambre, et le rôle du démiurge combinateur de Leibniz est joué pour nous par deux organismes statistiques. Contrairement à ce qui se produit pour le monde naturel dont la structure est occulte (faute de connaître les desseins et le langage du grand Combinateur, s'il existe), le monde de description qui nous sert de matériau d'analyse possède assurément, lui, une logique, et une syntaxe, qui nous sont fournies par les organismes collecteurs.

3.3. Les descriptions de l'aliment

Il va donc nous falloir traduire les descriptions des objets fournies par ces organismes dans un langage pertinent pour la question qui nous occupe. C'est là qu'intervient la problématique épistémologique soulevée plus haut : car nous traduirons des *descriptions*, et non pas des objets. Mais, comme il n'existe pas a priori de correspondance univoque entre les descriptions (SECODIP et INSEE) dont nous disposons, et celles que nous cherchons à mettre en place (les descriptions OCA), il nous faut faire la traduction en nous référant à un objet extérieur, dont chacune de ces descriptions serait la traduction dans les systèmes de recueil qui nous servent de source. C'est à partir de la connaissance de cet objet que nous ferons la traduction dans notre propre univers de description "OCA".

Si nous examinons le triangle sémiotique qui est à la base de la réflexion cognitive contemporaine, et qui a été systématisé en 1923 par Ogden et Richard[7] :

Le triangle sémiotique



cela revient à considérer qu'il existe un *référé* unique ("referent") pour les produits alimentaires que nous étudions. La solution la plus simple consisterait alors à regarder, objet par objet, comment décrire dans un nouveau langage les objets qui nous sont fournis, et faire en quelque sorte de la traduction "mot à mot". Il suffirait alors d'aller dans les linéaires des magasins d'alimentation observer une des occurrences de la "référence" , par exemple :

un paquet de Corn Flakes Kellogs'

pour la coder.

---

[7] Le schéma est tiré de Rastier (1990), cité par Dubois, D. (1991).

Outre le fait que la SECODIP nous décrit plusieurs dizaines de milliers d'objets, ce qui rendrait la tâche assez longue, ce type de résolution de problème ne correspond pas à notre style : il n'est ni théoriquement satisfaisant, ni techniquement économique. Après plusieurs essais infructueux (conception d'un "générateur de nomenclatures", fabrication d'une table "produits Crédoc" centrale dans la base de données, essai de traduction systématique médiatisée par des langages structurés -GENSEC, nomenclature IFLS...-), nous avons fini par mettre au point une méthode qui s'appuie sur une résolution générale du problème de la recodification, puis de son application à notre cas.

C'est celle qui est ici présentée. On commencera par une première partie théorique, donnant les grands principes de la procédure, puis on examinera l'application au cas de la base OCA.

3.3.1. Les mondes de description

Note : La méthode s'appuie sur l'utilisation de la FRC (Formalisation en Relativité Complète), mise au point dans notre département pour décrire les phénomènes sans faire appel au postulat de réalité. Elle permet d'analyser les sociétés concrètes comme systèmes de signes, et ainsi de prendre en compte l'influence de la représentation des acteurs dans la dynamique de ces systèmes. Comme il n'est pas nécessaire d'utiliser tout le formalisme ici, mais seulement son aspect descriptif, nous avons édulcoré les termes techniques qui pourraient désorienter le lecteur. Le chercheur intéressé par un approfondissement pourra se reporter au papier d'origine (Lahlou, 1990) en notant que l'adaptation du formalisme à notre cas particulier a été fait en remplaçant "signe" par "trait", "syplexe" par "combinaison", "U-langage" par "vision du monde".

Dans une certaine vision du monde, les objets sont perçus comme ayant un certain nombre de traits, qui font partie du système de description propre à cette vision du monde. En langue naturelle, ces traits sont signifiés par des mots (noms-de-traits), que l'on appelle encore, dans certains jargons, "libellés" ou "codes", et que nous désignerons du nom général de DESCRIPTEURS.

Par exemple, dans une vision du monde mercatique, le trait dont une traduction possible serait

"a été fabriqué par la société SOPAD-Nestlé"

sera signifié par le libellé

"fabricant : Nestlé"

et, dans un contexte non ambigu, abrégé en

"Nestlé"

"Nestlé" est alors le libellé signifiant le trait. Si, dans un répertoire, les fabricants ont été repérés, non par des caractères alphabétiques, mais par des nombres, et que le code correspondant au trait

"a été fabriqué par la société SOPAD-Nestlé"

est

05489

alors, "05489" sera une autre façon de symboliser le trait en question et sera donc une traduction du libellé "Nestlé" dans un autre langage. On peut encore dire qu'il lui correspondra.

Les objets sont définis par des combinaisons de traits. La description de l'objet en tant que tel se fait par les combinaisons des descripteurs correspondantes. La combinatoire des traits définit une classe d'objets observables, obtenus par la combinaison des traits élémentaires, que l'on appellera univers de cet observateur.

Dans cet univers, seules certaines combinaisons existent en fait, c'est ce que nous appellerons le *monde* de cet observateur. Par exemple, *l'univers* de SECODIP contient des objets tels que :

saumon-fumé/morceaux-N°3/pack-de-12/Kellogs'/allégé/vanille.

qui est bien une combinaison de traits SECODIP. Mais un tel objet ne fait pas partie du *monde* des observables de SECODIP. C'est une description possible, mais qui ne "correspond à rien", de même que

une lampe à refroidir en odeur de licorne

est une description en bon français, mais qui ne correspond pas à un observable.

Un objet comme :

A0152/B1276/C0229/E0121/F0014/G0003/H0151/J0123/K0003/M0001/N0001/P0024

autrement dit

légume-ou-dérivé/tomate-(Lycopersicum-esculentum)/fruit-pelé-sans-trognon-ni-noyau-ni-graine/liquide-très-visqueux-avec-de-petites-particules/transformation-thermique-complète/pas-de-méthode-de-cuisson-applicable/épicé/stérilisé-à-chaud-ou-mis-en-conserve/sans-milieu-de-conditionnement/récipient-ou-emballage-non-spécifié/surface-en-contact-avec-l'aliment-inconnue/produit-de-consommation-courante

existe dans le monde de REGAL (où il porte le n°11549) et de LANGUAL, mais pas dans l'univers de SECODIP. Dans le monde courant, une définition plus compréhensible nous est donnée par son libellé dans REGAL : "sauce tomate en conserve".

Nous cherchons à trouver des règles de traduction, qui nous permettront, si nous trouvons dans SECODIP un objet comme :

> 001000011010200 / THE CEYLAN CARTON AVEC SACHETS AUTRES MARQUES NON IDENTI

de lui attribuer un certain nombre de caractéristiques pour nous intéressantes dans les autres univers, comme par exemple de savoir que ce produit est obtenu à partir du Thea sinensis, et ainsi le transformer en un objet comme :

> A0268 - infusion / B1623 - thé (Camellia ou Thea sinensis) / C0200 - feuille / E0150 - entier de forme naturelle / F0003 - sans transf thermique / G0003 - pas de méth de cuisson applicable / H0138 - déshydraté / J0116 - déshydraté ou séché / K0003 - sans milieu de cdtionnmt / M0197 - sac ou sachet papier / N0039 - carton ou papier / P0024 - produit de consommation courante / LK - SRI LANKA / 0 / 0

Il va pour cela falloir utiliser des informations sur les propriétés des objets, sous la forme de règles.

3.3.2. Nos règles d'inférence

Il nous faut maintenant faire une brève mise au point terminologique sur les règles d'inférence, telles que nous les définissons dans le cadre du formalisme appliqué ici.

Inférence : On appellera inférence un procédé qui, à partir de certaines informations (prémisses), mène à d'autres (conclusions). Une règle d'inférence est une inférence particulière, qui associe à une combinaison donnée de prémisses une certaine combinaison de conclusions.

En général, on utilise ce procédé pour accroître la quantité d'informations dont on dispose sur un objet particulier : les prémisses sont alors les traits accessibles (par observation ou par définition) sur un objet.

Règle d'inférence : Soit un objet R. Soient Rp et Rc deux parties de R. Soit I(Rp, Rc) l'inférence qui associe Rc à Rp. On appellera règle d'inférence sur le domaine de validité D la conjecture selon laquelle :

si
S est un objet inclus dans D
et S a Rp pour partie
alors
S a Rc pour partie

On peut utiliser les règles d'inférence pour vérifier la conformité de certains objets et donc leur appartenance à un domaine particulier D. On peut également les utiliser comme règles de construction, de la manière suivante : on combine Rc à tous les objets ayant pour partie Rp.

En d'autres termes, on applique à l'objet des règles d'inférence qui ont comme prémisses des traits observés sur l'objet, et les conclusions sont alors des traits non évidents que l'objet "doit" posséder également, d'après la règle d'inférence.

Par exemple : soit la règle simple (une seule prémisse, une seule conclusion) :

si
S est inclus dans l'univers de SECODIP
et S possède le trait "morceaux-n°4"
alors
S possède le trait "Sucre"

Les règles d'inférence que nous appliquons ne sont pas, en fait, des règles de déduction, des règles de transformation logiques, qui porteraient sur des situations inférentielles, abstraction faite de la signification particulière des énoncés qu'elle contiennent. En ce sens, nos règles d'inférences sont des outils assez naïfs, simples, relativement proches de la notion d'entraînement.[8]

Notre définition ne cherche pas à être orthodoxe, mais simplement adaptée à notre besoin. Nous avons préféré un concept plus souple que celui de règle de déduction.

Il faut distinguer, pour la question qui nous occupe, différents types de règles d'inférence.

---

[8] Pour une discussion de ces questions, et des différences entre logique de la déduction naturelle et théorie de l'entraînement, voir Ladrière, 1970.

Les règles empiriques

Ce sont les règles induites à partir des régularités empiriques *constatées* par l'observateur dans les combinaisons de traits.

Par exemple,

tout homme est mortel

L'observateur peut les énoncer, mais pas les créer en tant que règles efficaces. Leur validité dépend en effet du bon vouloir de la nature. Rien ne garantit qu'une régularité observée, qui n'est finalement que statistique, se maintienne indéfiniment dans les phénomènes. Le contre-exemple classique est celui de la règle élaborée par un Anglais débarquant à Calais, et qui, ayant rencontré en débarquant deux Françaises rousses, et ces deux là seulement, énonce :

toutes les Françaises sont rousses

La plupart des règles ont ainsi des domaines de validité limités, et les phénomènes pour lesquels la règle est falsifiée (vérité des prémisses mais pas des conclusions) constituent la classe des exceptions. Notons que le fait que l'on utilise des règles bien qu'elles aient des exceptions montre qu'elles gardent une utilité opératoire.

Ainsi, la règle

toute boisson est liquide

est fausse pour le produit "Tang" de General Foods (boisson en poudre). De même ce n'est pas parce qu'un produit est conditionné en bouteille qu'il est forcément liquide (il peut s'agir de gaz, de poudre, d'émulsion, de pâte, ou même de solides (glaces au champagne).

Toute règle empirique est susceptible d'être remise en cause par l'apparition d'un nouvel objet "anormal" dans le domaine. En ce qui concerne l'alimentation, cela veut dire que tous les produits nouveaux menacent potentiellement la validité des règles d'inférence empiriques. On peut considérer, grossièrement, que nous voulons être dans un cadre dont la souplesse excède la logique standard, où l'on pourrait exprimer facilement que : tout ce qui s'intitule "boisson" est liquide, sauf dans certains cas particuliers rares et définis qui seraient traités dans des extensions particulières. C'est là un terrain intéressant, mais difficile. Nous pouvons utiliser des solutions moins élégantes mais plus simples.

Ce qu'il faut retenir, c'est que nos règles empiriques ne peuvent prétendre à une portée générale, et sont véritablement créées par nous à partir de l'observation du matériau.

Les règles syntaxiques

Les règles syntaxiques sont des règles déontiques qui sanctionnent le caractère "bien formé" des combinaisons de descripteurs.

Par exemple, dans le langage de SECODIP, tout objet est décrit par une combinaison de six traits appelés E1, E2, E3, E4, E5, E6.

Référence d'un produit :

Les libellés associés aux éléments E1,...,E6 d'une référence se trouvent dans un fichier dictionnaire[9].

Ces règles sont des règles explicites qui portent non pas sur les propriétés des objets décrits, mais sur le langage utilisé pour les décrire, en l'occurrence les logiques nomenclaturales de SECODIP et de l'INSEE. Dans la mesure où les langages de description bornent nos univers empiriques, nous pouvons accorder à ces règles un statut différent, et être certain qu'elles ne souffriront pas d'exception : tout objet qui ne les respecte pas est une *erreur*, et non pas une exception[10].

L'intérêt des règles syntaxiques est qu'elles se rapprochent des règles de déduction en ce qu'elles sont formelles, et portent sur les objets quel que soit leur contenu sémantique. En d'autres termes, elles ne dépendent pas du contenu particulier des objets. Elles permettent alors de faire des transformations systématiques d'objets.

---

[9] Pour une description détaillée, voir les travaux de C. Gillet dans Lahlou et coll. 1991, op. cit. annexe 7 : "décodage de la référence SECODIP".

[10] Naturellement, il s'agissait ici d'un point de vue théorique. Dans la pratique, malheureusement, il s'est avéré que les règles syntaxiques de SECODIP avaient souffert des exceptions, dues à une saturation des potentialisés de la combinatoire de description E1...E6 par la prolifération du nombre de références. Nous n'insisterons pas ici sur cet aspect des choses, mais il est bon de le mentionner, tant il est vrai que, dans tout système de codage qui s'est longtemps confronté à la réalité, il existe en général des transgressions adaptatives des règles originelles. Ces transgressions ne sont pas toujours connues de ceux qui ont conçu le système, ou qui l'utilisent rarement de manière concrète. Ainsi, nous conseillons vivement à ceux qui s'attaquent à des problèmes similaires au nôtre de vérifier si, au long de la chaîne de transformation des données, certains opérateurs (codeurs, informaticiens...) n'ont pas rajouté des règles, ou modifié les règles existantes, pour réussir à rentrer au moins en apparence dans le moule de la syntaxe officielle la description de certains objets atypiques.

3.3.3. Classement indirect et typicalité

Dans un monde donné, pour opérer une classification, on peut procéder par application de règles de classement qui ont pour prémisses des traits observés. C'est ce que nous appellerons le classement direct.

Mais on peut également faire un classement "indirect", qui utilise non plus les propriétés de l'objet lui-même, mais ses relations avec d'autres objets déjà classés. On pourra alors classer B "de la même façon que l'on a classé A", sous certaines conditions reliant B et A. Il suffit alors de disposer de relations entre les observables, par exemple une distance, une mesure de similarité.

Cette méthode, quoique peu immédiate sur le plan théorique, est très couramment appliquée dans la pratique. Son succès tient, pensons-nous, à ce qu'il est relativement plus facile de fabriquer des "comparateurs", fonctions ou processus dont les arguments portent sur une même classe d'objets, que des "identificateurs", qui sont capables de décrire des traits sur un objet. A tel point que les "identificateurs" rencontrés empiriquement en biologie sont en général des "comparateurs" à un modèle donné : l'identification est alors (comme on pouvait étymologiquement le prévoir) une re-connaissance. Il est peut-être même possible que l'identification par comparaison soit, après tout, la méthode canonique, si, comme le prétend Bateson l'information est l'expression d'une différence. Et, de fait, l'examen détaillé des procédures concrètes de ce que nous appelons "classement direct" risque de découvrir un enchaînement de classements indirects.

Quoi qu'il en soit, ce mécanisme efficace doit être envisagé dans un système de classification qui se veut opératoire. Nous pouvons alors fonder sur ce mécanisme ce que nous appellerons la codification typicaliste[11], qui consiste à nous référer à un prototype pour coder un objet donné. Cette approche a par ailleurs l'avantage de correspondre à une stratégie cognitive naturelle (voir par exemple les travaux d'Eleanor Rosch).

---

[11] On pourrait également adopter des méthodes de classification comparatives, qui consisteraient à se baser sur l'objet le plus proche pour classer. Ce type de méthode peut se passer d'un jeu de prototypes de référence, il est plus élégant, et permet de déterminer d'un même mouvement les classes et les classements (comme en analyse relationnelle, en classification par "clustering", etc.). Cependant, la nécessité d'une validation de nos codages par nos collègues d'autres organismes, et le désir de disposer dans un premier temps d'une sorte d'armature fixe pour structurer notre champ d'objets, nous ont amenés à préférer une approche typicaliste, dans laquelle les prototypes de référence ne peuvent appartenir qu'à un ensemble prédéfini de modèles. C'est cette distinction qui nous amène à utiliser le terme un peu pédant d'approche "typicaliste" plutôt que celui, plus familier mais sujet à confusion, de "comparative".

Ce prototype de référence n'existe que dans notre vision du monde, c'est en fait un modèle. Mais il peut éventuellement être considéré comme parfaitement réalisé en une occurrence particulière. Ainsi, on pourra considérer que

"sucre blanc en morceaux n°3"

est parfaitement réalisé dans

"Beghin-Say sucre blanc en morceaux n°3 boîte carton 1 kg".

Si nous disposons d'une description de ce prototype dans le langage qui nous intéresse, nous pouvons alors partir de celle-ci pour décrire d'autres objets qui sont proches de ce prototype, en modifiant à la marge les descripteurs qui ne conviennent pas, un peu à l'image des criminologues qui cherchent à reconstituer des visages à partir de bandes photos des yeux, nez, menton, etc. Le fait de savoir "en gros" à quoi ressemble l'objet qu'on cherche à décrire permet de gagner un temps considérable.

4. L'application à l'OCA

4.1. Les outils

A l'issue de cette exploration théorique, nous disposons de plusieurs outils.

Le premier est un formalisme : la description de type combinatoire des aliments, qui paraît plus flexible et évolutive qu'une approche en termes de nomenclature. Cette dernière ne permet la codification que lorsque l'objet a été complètement identifié comme forme globale, et n'autorise l'agrégation que sous une unique hiérarchie préétablie, tandis que la description à facettes autorise autant de regroupements qu'il existe de facettes, et permet la codification partielle, applicable à des objets dont certains aspects seulement sont connus.

Ce premier outil consiste en un choix de syntaxe. Notons que la syntaxe combinatoire est la moins contraignante et donc la plus générale : elle englobe toutes les autres. Cet avantage se paye par une générativité également exceptionnelle.

Le second outil est celui des règles d'inférences, qui nous permettent de traduire sur une grande échelle des descriptions d'objets d'un langage dans un autre.

Le troisième est celui de l'approche typicaliste, qui permet de traduire les descriptions de produits en s'appuyant sur des modèles de produits particuliers.

4.2. L'univers de description : LANGUAL étendu

Après divers essais de nomenclatures, nous avons conclu que seule une approche de codification combinatoire pouvait fournir la flexibilité et la capacité à évoluer.

Une codification combinatoire consiste à définir un certain nombre de descripteurs élémentaires, sorte de thesaurus de traits qui servent à décrire les objets. Pour faciliter la codification, l'archivage et la recherche, ces descripteurs sont eux-mêmes classés en grandes catégories, dimensions ou "facettes", qui correspondent à de grandes classes de propriétés. Par exemple, des *descripteurs* comme "solide", "liquide", "semi-liquide", "émulsion", "gaz" seront rassemblés dans une *facette* appelée "ETAT PHYSIQUE". Comme on l'a déjà souligné plus haut, un système de codification à facettes est plus flexible et plus évolutif qu'une nomenclature, et en particulier il n'est pas hiérarchique a priori. On peut l'utiliser comme générateur de nomenclatures, en croisant les facettes, et en attribuant à chacune un rang de priorité. On choisit ainsi la hiérarchie adaptée au problème particulier qui nous occupe alors.

Une fois décidé ce choix d'un système de description à facettes, nous avons eu à choisir entre créer notre propre système OCA de codification multifactorielle, ou utiliser le système de codification LANGUAL, mis au point par le Center for Food Safety and Applied Nutrition de la FDA. On trouvera dans la bibliographie des références d'articles sur ce système.

Dans un premier temps, nous avons cherché, sans a priori, à construire notre propre système de codification. Il s'est avéré qu'un grand nombre des facettes résultantes étaient finalement assez proches des facettes LANGUAL, par un effet de convergence fonctionnelle que l'on observe assez souvent dans la recherche scientifique, et même en biologie (c'est ainsi que l'oeil du poulpe est très semblable à celui des mammifères, malgré une phylogénèse extrêmement différente). Devant cette convergence, il aurait été stupide de ne pas s'aligner sur LANGUAL, déjà adopté en France par le CIQUAL (Centre Informatique sur la Qualité des Aliments), et, à l'étranger, aux USA, au Danemark, en Hongrie. Cette standardisation permet de bénéficier des dizaines d'années-hommes consacrées au développement de ce système descriptif, d'enrichir nos bases de données avec les codifications réalisées par nos

collègues étrangers (plusieurs milliers), et surtout de pouvoir échanger et comparer des données de sources différentes.

Nous avons cependant été amenés à créer quelques facettes nouvelles, ce qui est d'ailleurs conforme à l'esprit de LANGUAL, qui préconise une adaptation fine par enrichissement de la description, pourvu que les descripteurs de base soient remplis. Par ailleurs, nous avons suggéré une modification du système de codification des *recettes* des aliments, dans un sens qui nous paraissait techniquement meilleur et théoriquement plus puissant pour des raisons qui seraient trop longues à expliciter ici. Il se trouve que cette proposition converge avec les améliorations envisagées par les promoteurs et les utilisateurs actuels de LANGUAL.

En résumé, disons que, non seulement la base de l'OCA est LANGUAL-compatible, mais que le CREDOC s'est engagé dans une coopération active avec le CIQUAL et le CFSAN pour l'entretien et l'enrichissement du système.

4.2.1. L'univers de LANGUAL

Le système LANGUAL est un ensemble de descripteurs composés d'un code et d'un libellé organisés selon seize catégories ou facettes :

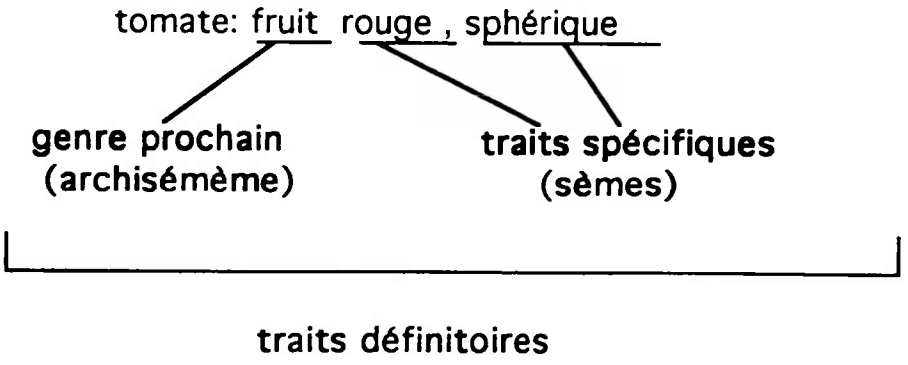
Type de produit	Facette A
Origine de l'ingrédient principal	Facette B
Partie utilisée d'une plante ou d'un animal	Facette C
Etat physique ou forme	Facette E
Degré de transformation thermique	Facette F
Méthode de cuisson	Facette G
Modification technologique	Facette H
Méthode de conservation	Facette J
Milieu de conditionnement	Facette K
Récipient ou emballage	Facette M
Surface en contact avec l'aliment	Facette N
Utilisateur	Facette P
Durée et conditions de stockage	Facette S
Période ou année de production	Facette T
Origine géographique	Facette R
Caractéristiques complémentaires	Facette Z

Chaque facette comprend un nombre variable de codes qui correspondent à toutes les "valeurs" que peut prendre la facette. A titre d'exemple voici un extrait du contenu de la facette E (Etat physique ou forme) :

E0130 - liquide
  E0102 - liquide très-visqueux
    E0139 - liquide très-visqueux sans particule visible
    E0121 - liquide très-visqueux avec de petites particules
    E0138 - liquide très-visqueux avec des morceaux solides
  E0109 - liquide peu-visqueux
    E0123 - liquide peu-visqueux sans particule visible
    E0114 - liquide peu-visqueux avec de petites particules
    E0149 - liquide peu-visqueux avec des morceaux solides
E0001 - état physique ou forme inconnu
E0108 - état physique ou forme multiple
E0103 - semi-liquide
  E0110 - semi-liquide avec des morceaux solides
  E0135 - semi-liquide à consistance lisse
E0144 - semi-solide
  E0134 - semi-solide avec des morceaux solides
  E0119 - semi-solide à consistance lisse

L'information est résumée par un code composé d'un identifiant de la facette et d'une suite alphanumérique. C'est un principe partiellement hiérarchique qui organise chaque facette. La structure ouverte de ce système permet l'ajout ou la modification de codes sans que la structure globale en soit affectée.

Les facettes correspondent à des classes de traits définitoires. La codification des produits s'apparente en effet à la définition lexicographique en vue de la constitution de dictionnaires. Prenons une définition du petit Robert :



La définition en LANGUAL de la tomate se présente comme suit:

11529	**TOMATE FRAICHE**
A0152	légume
B1276	tomate (Lycopersicum esculentum)
C0139	fruit non pelé avec trognon ou noyau ou graine
E0150	entier de forme naturelle
F0003	sans transformation thermique
G0003	pas de méthode de cuisson applicable
H0003	aucun traitement appliqué
J0001	traitement de conservation inconnu
K0003	sans milieu de conditionnement
M0001	récipient ou emballage non spécifié
N0001	surface en contact avec l'aliment inconnue
P0024	produit de consommation courante
R0	
S0	
T0	
Z0	

Dans un cas comme dans l'autre, le nom est décrit par une combinaison de traits. Mais une définition en langage naturel répond à la seule règle de l'économie (seuls les traits pertinents apparaissent), tandis que la description en LANGUAL comporte toujours seize rubriques - certaines pouvant éventuellement ne pas être remplies. La structure de la description est orientée vers des usages spécifiques propres à certains univers (nutritionniste, médical, technologique...).

La souplesse du système nous permet d'apporter une information plus ou moins exhaustive selon les cas et d'avoir la possibilité de remplir progressivement les facettes.

Chaque facette contient un nombre variable de descripteurs, qui peuvent éventuellement être hiérarchisés (jusqu'à 10 dans la facette B).

Nombre de descripteurs par facettes dans LANGUAL, fin 1991

Facettes	Nombre de postes
A Type de produit	196
B Origine de l'ingrédient principal	1435
C Partie utilisée d'une plante ou d'un animal	163
E Etat physique ou forme	46
F Degré de tansformation thermique	7
G Méthode de cuisson	33
H Modification technologique	264
J Méthode de conservation	49
K Milieu de conditionnement	39
M Récipient ou emballage	112
N Surface en contact avec l'aliment	47
P Utilisateur	33
R Origine géographique	371
S Durée et conditions de stockage	3
T Période ou année de production	2
Z Caractéristiques complémentaires	136
Total	2936

Nous avons réalisé un programme pour faire de la codification assistée par ordinateur. Ce programme s'intitule CITOCA (pour Codification par Inférence et Typicalité de l'OCA) ; il s'inspire dans ses fonctionnalités du logiciel "Autofactor" réalisé par le CFSAN. Ce programme, développé sur le logiciel 4D, est portable sur Macintosh. Il nécessite un moniteur de 19 pouces ou plus et une capacité de stockage de 20 mégaoctets pour pouvoir tourner ; il est recommandé de disposer d'un processeur rapide (68030).

4.3. La méthode de codification : un mélange d'inférence et de typicalité

4.3.1. Processus de codification

Chacune des deux approches (inférentielle et typicaliste) pourrait, en principe, suffire à résoudre notre problème pourvu qu'il soit formalisé en termes combinatoires. Cependant, le général trouve sa limitation dans le particulier, et les produits alimentaires, comme tout ensemble, sont une collection de cas particuliers. L'approche inférentielle cesse d'être

économique lorsque les règles s'appliquent seulement à un faible nombre d'objets, qu'on aura alors plus vite fait de coder un par un. Inversement, l'approche typicaliste nécessite, pour chaque objet, de déterminer s'il peut être rapporté à un prototype existant, ou s'il est nécessaire de créer un nouveau prototype. Elle est en principe moins économique que l'approche inférentielle qui, elle, est facilement automatisable.

De fait, une première approche strictement inférentielle nous a rapidement conduit à constater des conflits de règles d'inférence. Une partie de ces conflits a pu être résolue en créant des méta-règles qui donnaient la priorité à certaines règles sur d'autres, puis des règles de priorité conditionnelle. Il est apparu que ce type d'approche butait rapidement sur le nombre d'exceptions. Une solution eut été d'abandonner la logique standard pour utiliser la logique des défauts, qui a été développée précisément pour résoudre de telles difficultés[12]. Compte tenu de l'état de l'art dans ce domaine et du temps imparti, cette option nous a semblé trop hasardeuse. Nous avons donc opté pour une approche mixte, qui combine inférence et typicalité, et ce en trois temps. Ce mélange méthodologique est formellement moins élégant, mais plus efficace, et finalement, justifiable sur le plan épistémologique, puisque :

> "Lorsque les conduites de recherche se heurtent à des obstacles, à des complications, à des difficultés imprévus et que plusieurs solutions se présentent pour les surmonter, compte tenu des exigences de la logique (cohérence de la théorie et des modèles) et des contraintes expérimentales, il faut *choisir* : et que choisit-on ? La solution la plus *idoine*. En excluant l'idée de choix, on réduit la méthode à un algorithme dont la mise en oeuvre sans lutte, sans histoire, sans imagination, sans risque ni fantaisie, ne saurait produire des découvertes que par accident. La méthode débute avec l'engagement du chercheur qui prend le risque de ses choix. Il appartient au discours de la méthode de mentionner le rôle du *sujet* en recherche : en définitive, c'est le *sujet* qui se prononce sur la meilleure convenance et qui assume la responsabilité de ses choix. La poursuite de la recherche validera ou réfutera *l'idonéité* de la stratégie qu'il a décidée et engagée." (Morel, 1981)

C'est donc dans un souci d'efficacité que seront combinées les deux types d'approches.

4.3.2. Approche inférentielle

En deux mots, il s'agit d'attribuer aux objets de façon automatisée, tous les descripteurs qui peuvent l'être.

Il nous faut consigner certaines informations qui n'entrent pas dans le cadre prévu par le système de description Langual. Il en est ainsi pour la marque du produit alimentaire, la

---

[12]Dans la logique du raisonnement par défaut de Reiter, on noterait pour tenir compte du "défaut" Tang : "P est une boisson : P n'est pas en poudre/P est liquide".

quantité ("quantité unitaire") et le nombre d'unités ("par combien") dont se compose un produit. Aucune facette Langual n'a été prévue pour enregistrer ce type d'informations. Nous avons créé pour ces indicateurs des facettes spécifiques (facette OCA) que nous avons complétées à l'aide de règles d'inférence élémentaires.
Sur un marché donné, on repère la partie du libellé qui a trait à la marque, à la quantité et au "par combien". Pour la marque, on attribue les codes de la SECODIP. Pour les deux autres facettes, on uniformise préalablement les valeurs des facettes en les exprimant dans les unités internationales de mesure. Ensuite, sur un marché donné, à chaque expression de la mesure, on attribue un code donné. Ainsi, pour chaque référence alimentaire, les codes seront inférés à partir du libellé SECODIP.

Des règles d'inférence seront également utilisées pour compléter ou corriger les codages obtenus par l'approche typicaliste décrite dans le paragraphe 4.3.3.. Par exemple :

Si une référence SECODIP appartient au marché des "glaces à emporter"
et que le libellé qui lui est associé contient l'expression "bac plastique",
alors on attribuera à la facette M(Récipient ou emballage) la valeur M0184 ;

ou bien :

Si une référence SECODIP appartient au marché des "jus de fruit, sodas, boissons aux fruits"
et que le libellé qui lui est associé contient les expressions "allégé" ou "light",
alors on attribuera à la facette P (utilisateur) la valeur P0045 ("aliment de régime à teneur en calorie contrôlée").

Les règles d'inférence sont donc utilisées en amont de la codification typicaliste : pour coder des facettes non prévues par Langual et en aval pour contrôler la cohérence des codages, pour faire en sorte que des libellés synonymes soient codés de façon identique.

4.3.3. Approche typicaliste

Elle consiste à coder les objets en s'inspirant d'un objet proche, prototypique, dont le codage est déjà fait.

Pour les facettes purement Langual, le principe de base est de reprendre des codages effectués par d'autres organismes et de les attribuer aux références SECODIP, en les modifiant à la

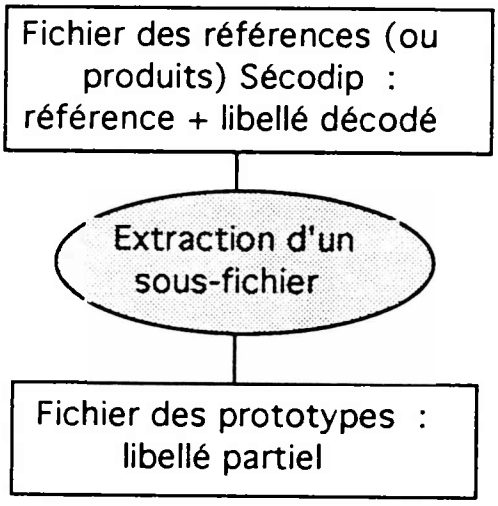
marge si nécessaire. Nous bénéficions ainsi du savoir-faire d'équipes qui travaillent dans ce domaine depuis une dizaine d'année.

La codification se fait marché par marché.

4.3.3.1. Première étape : constitution d'un fichier de prototypes

A partir du fichier SECODIP des références alimentaires d'un marché donné, on crée un sous-fichier qui ne tient compte ni de la marque, ni de la quantité (quantité unitaire, par combien). Ces derniers ont été préalablement codés à l'aide de règles inférentielles. On ne retient qu'une sous-partie des traits (E1, E2,... E6) qui constituent la référence complète et on conserve les libellés correspondants. On obtient donc un fichier de prototypes de produits dont la taille peut être jusqu'à dix fois plus faible que celle du fichier initial.

Constitution du fichier des prototypes



Voici un extrait du fichier des prototypes du marché des yaourts et desserts :

FLAN NAPPE FLAN NAPPE CARAMEL
FLAN NAPPE CREME CARAMEL
FLAN NAPPE BIM BOUM
FLAN NAPPE FEE FLAN
FLAN NAPPE CREME RENVERSEE
FLAN TRADITIONNEL
YAOURTS LAIT ENTIER NATURE
YAOURTS LAIT ENTIER NATURE POT VERRE / NON SUCRE
YAOURTS LAIT ENTIER AVEC FRUIT
YAOURTS LAIT ENTIER AROMATISE

YAOURTS NORMAUX NATURE
YAOURTS NORMAUX NATURE SUCRE
YAOURTS NORMAUX NATURE POT VERRE / NON SUCRE
YAOURTS MAIGRES NATURE

4.3.3.2. Deuxième étape : codification des prototypes

Pour coder ces produits-types, nous disposons à ce jour de deux précieuses aides à la codification :
- le fichier REGAL (Répertoire Général des Aliments) composé de 1217 produits libellés en français codés en LANGUAL par le CIQUAL (Centre Informatique sur la Qualité des Aliments);
- le fichier du CFSAN qui comprend 3 585 produits (principalement rencontrés sur le marché américain).

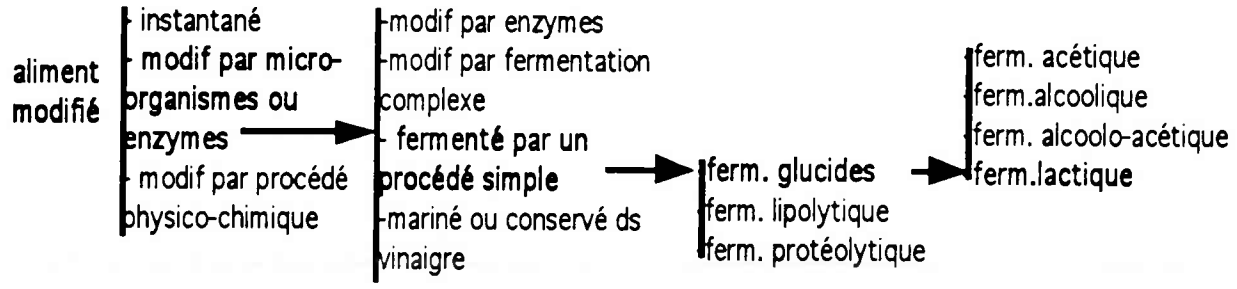
Pour un prototype donné, nous recherchons à l'aide de son libellé ou d'une fraction de celui-ci s'il existe un libellé similaire dans le fichier CIQUAL. Dans l'affirmative, on attribue au prototype les codes du produit CIQUAL, puis on corrige à la marge les facettes mal adaptées aux spécificités de notre produit et l'on passe au produit suivant.
Mais si le produit n'apparaît pas dans le CIQUAL, on effectue la même recherche dans le fichier CFSAN, ce qui implique la traduction préalable du libellé. Nous ne nous étendrons pas sur les difficultés nées de l'imparfaite coïncidence entre les référents liés à des appellations françaises et américaines apparemment voisines. S'il existe un produit apparenté, comme précédemment, on attribue les codes et on les modifie si nécessaire.
Enfin, si les deux fichiers n'ont su répondre à notre attente nous attribuons nous-mêmes des codes. Cela nécessite des recherches bibliographiques et sur le terrain (lecture des étiquettes de composition des produits sur les linéaires de supermarchés...).

Pour la modification à la marge des facettes ou pour l'attribution manuelle des codes, nous avons intégré l'ensemble des descripteurs dans le système en conservant les différents niveaux de la hiérarchie. Ceci nous permet de circuler aisément à l'intérieur des facettes.

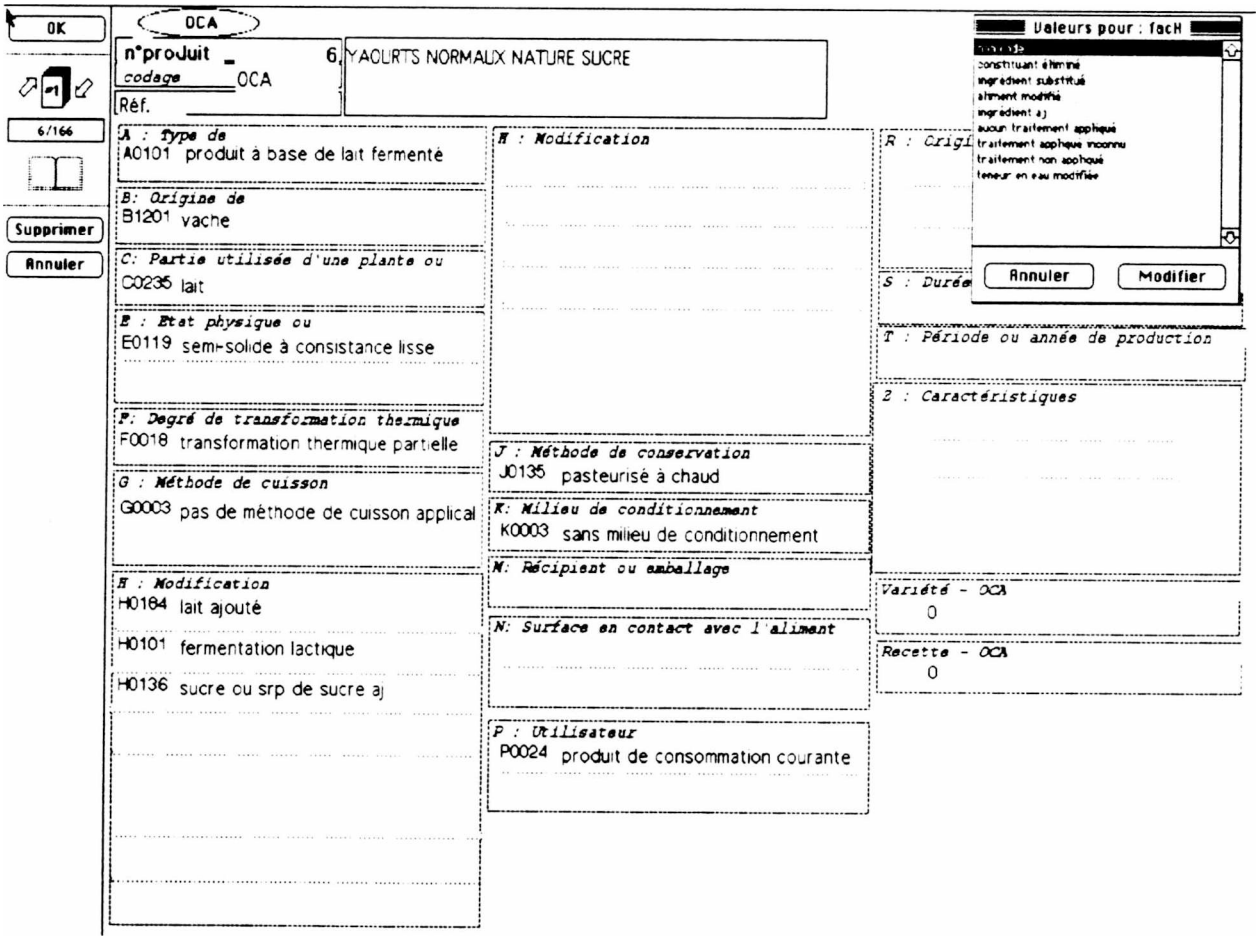
Par exemple pour un yaourt, il faut indiquer qu'il y a eu fermentation lactique. Pour cela on chemine dans la facette H (modification technologique). En sélectionnant un descripteur du niveau n, on obtient toutes les valeurs du niveau n+1 qui sont classées sous ce descripteur. On peut ainsi choisir le niveau de la hiérarchie en fonction de nos connaissances sur le produit.

Ainsi, si la nature de la fermentation est inconnue, nous choisirons comme descripteur "modification par micro-organismes ou enzymes".

Cheminement hiérarchisé dans la facette H
(modifications technologiques)



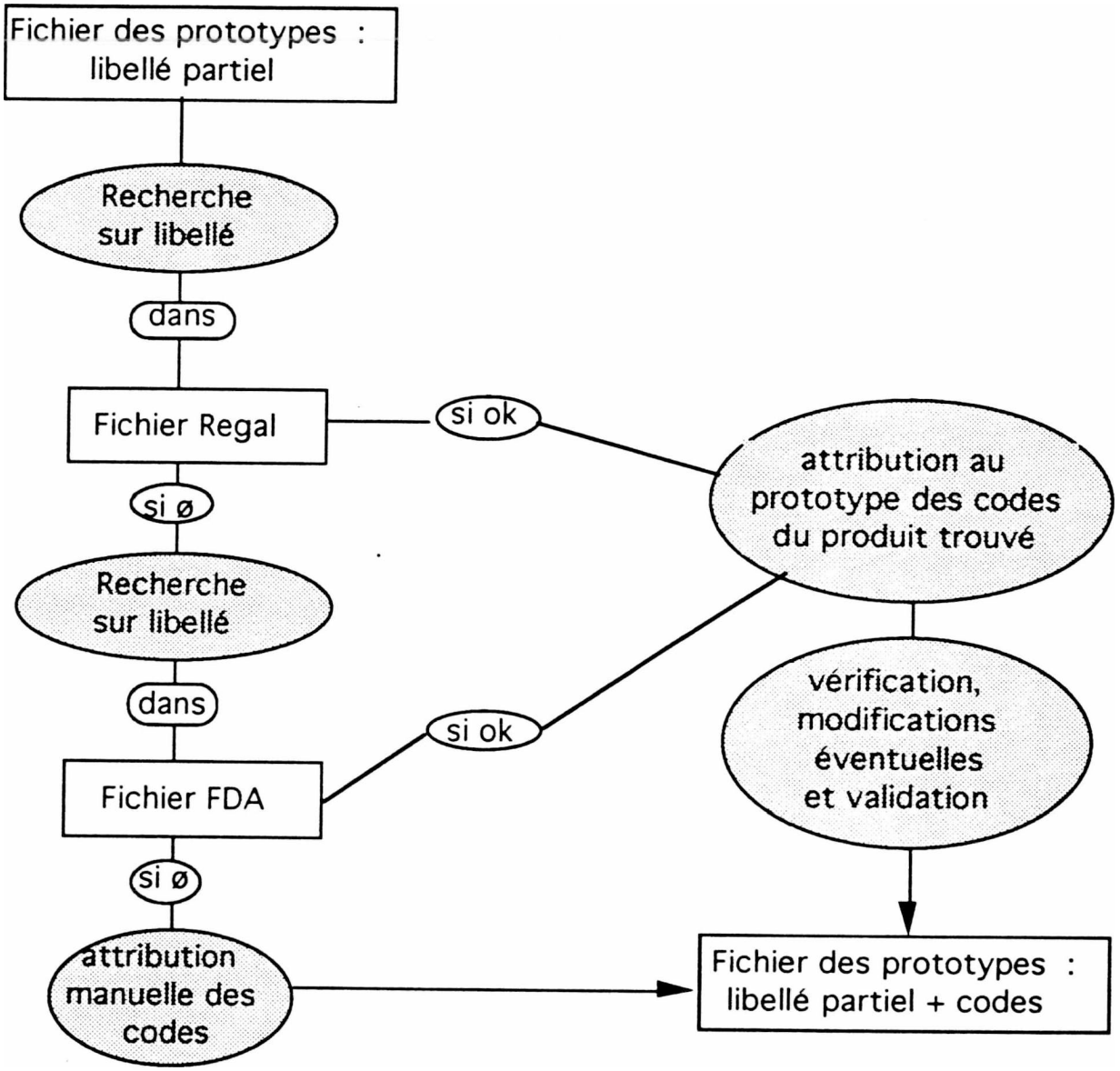
Pour faciliter le contrôle et éventuellement la modification des codes, la fiche du produit nous est présentée comme suit (ci-dessous, une copie d'écran, réduite 5 fois).



Nous avons alors accès, grâce à la fenêtre située en haut à droite, à la liste hiérarchisée des descripteurs d'une facette. A titre d'exemple, si l'on sélectionne le descripteur "aliment

modifié", apparaîtront les sous-descripteurs de ce libellé. En "cliquant" sur un des items proposés, on obtient une nouvelle liste, contenant les sous-rubriques de cet item, et ainsi de suite jusqu'à ce qu'on atteigne le plus bas niveau hiérarchique ou que l'on choisisse un item. On pourra reproduire le cheminement présenté dans le tableau précédent. Une fois trouvé le descripteur adéquat, il s'inscrit dans la case correspondante.

Schéma de la codification des prototypes



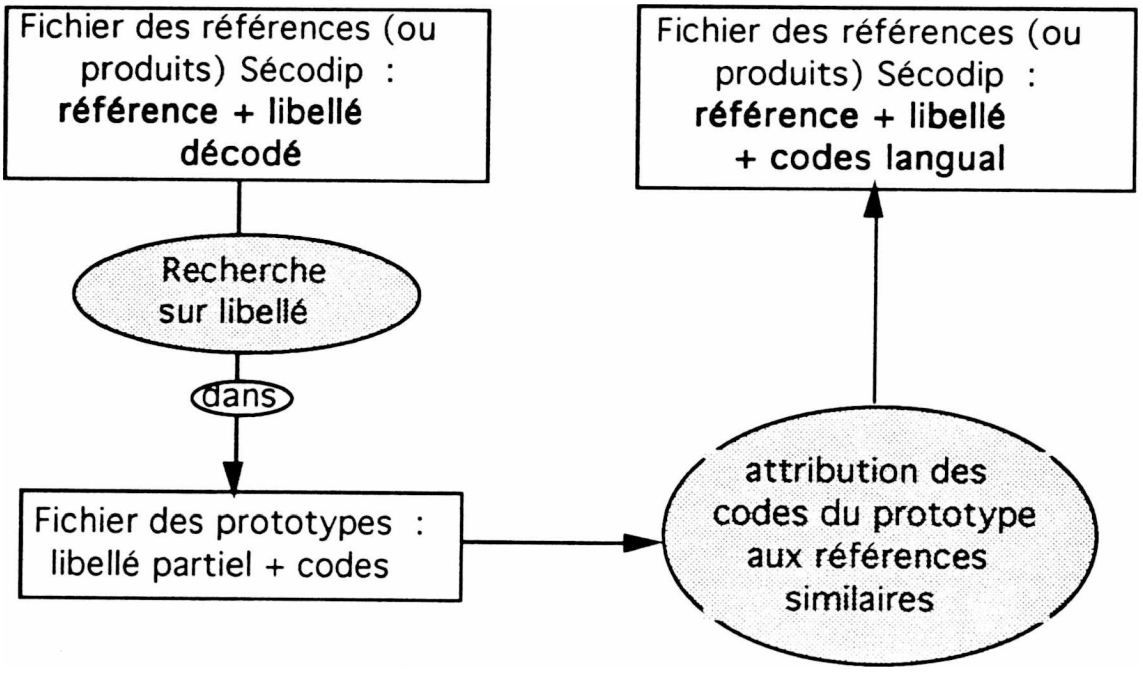
4.3.3.3 Troisième étape : codification des références

Une fois le marché des prototypes préparé, on reprend le fichier initial pour coder toutes les références.

Les facettes justiciables d'un traitement inférentiel (MARQUE, PARCOMBIEN, QUNIT...) sont traitées à part avec un système de règles de transcodification.

Ensuite, une procédure de codification par groupe permet de remplir de façon identique toutes les facettes des produits qui ne diffèrent que par la marque ou par d'autres indicateurs ignorés par LANGUAL. A un groupe de références donné, on attribue les codes du prototype correspondant. Eventuellement, certains descripteurs sont modifiés à la marge à partir de la description du prototype pour décrire chaque référence de façon ad-hoc.

Codification des "références" SECODIP



Enfin, un fichier composé de la référence et de l'ensemble des codes est transféré vers la base de données OCA.
Les références codifiées seront alors envoyées pour validation au CIQUAL.

5. Exemples de codification

On présente ici deux exemples réels de codification, l'un portant sur le marché des pains préemballés en tranches, l'autre sur celui des glaces à emporter.

5.1. Les pains préemballés

Le marché des "pains préemballés en tranches" comporte 175 références différentes, et combine pour les décrire les éléments E1 à E6 habituellement utilisés par SECODIP pour établir une référence unique. E1 porte sur les marques. Les postes E2 à E6 ont des significations variables selon les marchés et même des significations variables à l'intérieur d'un même marché.

Ces éléments prennent pour le marché des "pains préemballés en tranches" les significations suivantes :

E1 = marques
E2 = noms commerciaux (King corn, Raisinella...)
E3 = frais/longue durée
E4 = quantité unitaire
E5 = variété
E6 = nombre d'unités dans l'emballage

5.1.1. Règles d'inférence

Nous avons dans un premier temps établi les règles d'inférence pour coder les quantités unitaires et le nombre d'unités dans l'emballage.
Chacune des références est alors affectée, le cas échéant, d'un code correspondant à la facette quantité unitaire, "QUNIT", suivant les règles suivantes (si code = X, alors libellé = Y) :

Code	*Libellé*
63	Moins de 250 g
65	250 - 320 g
73	321 - 449 g
81	450 - 500 g
83	Plus de 500 g

et à la facette "PAR COMBIEN" :

Code	*Libellé*
4	X 3
5	X 4
6	X 5
9	X 6

Il s'agit, dans ce cas, des règles les plus simples possibles.

5.1.2. Approche typicaliste

5.1.2.1. Constitution d'un fichier de prototypes

Nous avons ensuite retenu les éléments E3 et E5 pour constituer les prototypes du marché et obtenu alors la liste suivante de 10 prototypes :

PAIN EN TRANCHES PREEMBALLE PAIN FRAIS MIE
PAIN EN TRANCHES PREEMBALLE PAIN FRAIS BRIOCHE
PAIN EN TRANCHES PREEMBALLE PAIN FRAIS SEIGLE
PAIN EN TRANCHES PREEMBALLE PAIN FRAIS CAMPAGNE- FROMENT
PAIN EN TRANCHES PREEMBALLE PAIN FRAIS COMPLET
PAIN EN TRANCHES PREEMBALLE PAIN FRAIS AU SON
PAIN EN TRANCHES PREEMBALLE PAIN FRAIS AUTRE PAIN PREEMBALLE
PAIN EN TRANCHES PREEMBALLE LONGUE DUREE MIE
PAIN EN TRANCHES PREEMBALLE LONGUE DUREE BRIOCHE
PAIN PREEMBALLE HAMBURGER HOT DOG

Le codage de ces prototypes s'est fait au moyen du logiciel CITOCA.

5.1.2.2. Codification des prototypes

Le traitement du premier de ces produits, ici, le

pain en tranches préemballé pain frais de mie

se fait d'abord en interrogeant un fichier de prototypes pré-codés, le fichier "CIQUAL" pour repérer les produits semblables existants.
La recherche s'effectue sur le libellé, partiel ou complet selon le cas.

Une première recherche dans le fichier CIQUAL sur le mot "pain" permet de voir que le CIQUAL a codé 42 produits qui comportent ce mot,

7001	Pain
7002	Mie de pain
7003	Croûte de pain
7004	Pain grillé
7030	Pain, fabrication domestique (Algérie)
7100	Pain de campagne
7101	Pain brié
7102	Pain de Beausaine
7110	Pain complet
7115	Pain au son
7120	Pain au seigle pur
7125	Pain de seigle et froment
7130	Pain de méteil avec 50% de seigle
7160	Pain sans sel
7200	Pain de mie
7205	Pain de mie artisanal
7210	Pain de mie artisanal au lait
7220	Pain de mie industriel
7225	Pain viennois
7400	Pain grillé industriel
7405	Pain grillé industriel, sans sel
7410	Pain braisé industriel
7415	Pain braisé industriel, sans sel
7710	Pain au lait
7720	Pain aux raisins
7730	Pain au chocolat
7731	Pain au chocolat feuilleté
7732	Pain au chocolat brioché
7733	Pain au chocolat et aux oeufs
7734	Pain au chocolat et au lait
23025	Pain d'amandes
23200	Pain d'épices, sans spécification
23210	Pain d'épices fondant
23220	Pain d'épices fourré
23230	Pain d'épices de Dijon, préparation artisanale
23231	Pain d'épices de Reims, préparation artisanale
23550	Gâteau au pain noir, préparation artisanale
23564	Pain aux dattes et aux noix, préparation artisanale
23565	Petits pains au lait de Clunie Rock, préparation artisanale
23569	Pain d'épices aux raisins secs, préparation artisanale
24651	Pain de Gênes, préparation artisanale
24653	Portugais, petit pain de préparation artisanale

dont 4 sont des "pains de mie"

7200 Pain de mie
7205 Pain de mie artisanal
7210 Pain de mie artisanal au lait
7220 Pain de mie industriel

codés de la façon suivante :

7200 - Pain de mie	**7205 - Pain de mie artisanal**	**7210 - Pain de mie artisanal au lait**	**7220 - Pain de mie industriel**
A0178 - pain	A0178 - pain	A0178 - pain	A0178 - pain
B1421 - blé tendre (Triticum aestivum)	B1421 - blé tendre (Triticum aestivum)	B1421 - blé tendre (Triticum aestivum)	B1421 - blé tendre (Triticum aestivum)
C0208 - grain(e) sans enveloppe et sans germe	C0208 - grain(e) sans enveloppe et sans germe	C0208 - grain(e) sans enveloppe et sans germe	C0208 - grain(e) sans enveloppe et sans germe
E0105 - entier façonné épais de 1,5 à 7 cm	E0105 - entier façonné épais de 1,5 à 7 cm	E0105 - entier façonné épais de 1,5 à 7 cm	E0105 - entier façonné épais de 1,5 à 7 cm
F0014 - transformation thermique complète	F0014 - transformation thermique complète	F0014 - transformation thermique complète	F0014 - transformation thermique complète
G0005 - cuit ou rôti au four	G0005 - cuit ou rôti au four	G0005 - cuit ou rôti au four	G0005 - cuit ou rôti au four
H0136 - sucre ou sirop de sucre ajouté	H0136 - sucre ou sirop de sucre ajouté	H0136 - sucre ou sirop de sucre ajouté	H0136 - sucre ou sirop de sucre ajouté
H0221 - corps gras ajouté	H0221 - corps gras ajouté	H0221 - corps gras ajouté	H0221 - corps gras ajouté
H0256 - fermenté au niveau des glucides	H0256 - fermenté au niveau des glucides	H0256 - fermenté au niveau des glucides	H0256 - fermenté au niveau des glucides
		H0184 - lait ajouté	
J0003 - sans traitement de conservation	J0003 - sans traitement de conservation	J0003 - sans traitement de conservation	J0003 - sans traitement de conservation
K0003 - sans milieu de conditionnement	K0003 - sans milieu de conditionnement	K0003 - sans milieu de conditionnement	K0003 - sans milieu de conditionnement
P0024 - produit de consommation courante	P0024 - produit de consommation courante	P0024 - produit de consommation courante	P0024 - produit de consommation courante

Ces quatre produits sont décrits de façon identique par les différentes facettes LANGUAL.

Nous choisirons donc un de ces quatre produits comme base de codage pour le prototype que nous devons traiter. Et, pour indiquer qu'il s'ait de pain *tranché*, on complètera la facette E (état physique ou forme du produit) par l'affectation du code suivant :

E0137 - tranché

Le codage LANGUAL du "pain frais de mie en tranches préemballé" sera alors le suivant :

PAIN EN TRANCHES PREEMBALLE PAIN FRAIS MIE
A0178 - pain
B1421 - blé tendre (Triticum aestivum)
C0208 - grain(e) sans enveloppe et sans germe
E0105 - entier façonné épais de 1,5 à 7 cm E0137 - tranché
F0014 - transformation thermique complète
G0005 - cuit ou rôti au four
H0136 - sucre ou sirop de sucre ajouté H0221 - corps gras ajouté H0256 - fermenté au niveau des glucides
J0003 - sans traitement de conservation
K0003 - sans milieu de conditionnement
P0024 - produit de consommation courante

Ces prototypes sont chargés dans le fichier des références SECODIP.

5.1.2.3. Codification des références ou produits SECODIP

Une fois l'ensemble des prototypes traité nous passons aux références SECODIP réelles.

Le marché du "pain frais de mie", qui comprend 75 références, est un cas très simple. Par rapport au prototype, les références ne contiennent pas de renseignements supplémentaires susceptibles d'être retenus. Les codes affectés au prototype "pain frais de mie" seront donc attribués tels quels à l'ensemble des références. Les facettes M et N qui décrivent l'emballage du produit et les matières en contact avec l'aliment ne seront pas davantage renseignées dans le cas des références réelles. Les postes "non spécifié" ou "inconnu" seront alors utilisés.

100001010100 PAIN EN TRANCHES PRE. PAIN FRAIS MIE MOINS DE 250 G AUTRES MARQ. NON IDENTI
100001020100 PAIN EN TRANCHES PRE. PAIN FRAIS MIE 250 A 320 G AUTRES MARQUES NON IDENTI
100001030100 PAIN EN TRANCHES PRE. PAIN FRAIS MIE 321 A 449 G AUTRES MARQUES NON IDENTI
100001040100 PAIN EN TRANCHES PRE. PAIN FRAIS MIE 450 A 500 G AUTRES MARQUES NON IDENTI
100001050100 PAIN EN TRANCHES PRE. PAIN FRAIS MIE + DE 500 G AUTRES MARQUES NON IDENTI
200001010100 PAIN EN TRANCHES PRE. PAIN FRAIS MIE MOINS DE 250 G SANS MARQUE
200001020100 PAIN EN TRANCHES PRE. PAIN FRAIS MIE 250 A 320 G SANS MARQUE
200001030100 PAIN EN TRANCHES PRE. PAIN FRAIS MIE 321 A 449 G SANS MARQUE
200001040100 PAIN EN TRANCHES PRE. PAIN FRAIS MIE 450 A 500 G SANS MARQUE
200001050100 PAIN EN TRANCHES PRE. PAIN FRAIS MIE + DE 500 G SANS MARQUE
10100001020100 PAIN EN TRANCHES PRE. PAIN FRAIS MIE 250 A 320 G CARREFOUR
10100001040100 PAIN EN TRANCHES PRE. PAIN FRAIS MIE 450 A 500 G CARREFOUR
10200001040100 PAIN EN TRANCHES PRE. PAIN FRAIS MIE 450 A 500 G CONTINENT
10400001040100 PAIN EN TRANCHES PRE. PAIN FRAIS MIE 450 A 500 G EUROMARCHE
10500001040100 PAIN EN TRANCHES PRE. PAIN FRAIS MIE 450 A 500 G PRODUIT SINCERES
10600001040100 PAIN EN TRANCHES PRE. PAIN FRAIS MIE 450 A 500 G SUPER M

10700001040100 PAIN EN TRANCHES PRE. PAIN FRAIS MIE 450 A 500 G CORA
10900001040100 PAIN EN TRANCHES PRE. PAIN FRAIS MIE 450 A 500 G RECORD
10900001050100 PAIN EN TRANCHES PRE. PAIN FRAIS MIE + DE 500 G RECORD
11000001040100 PAIN EN TRANCHES PRE. PAIN FRAIS MIE 450 A 500 G RADAR
11100001020100 PAIN EN TRANCHES PRE. PAIN FRAIS MIE 250 A 320 G CASINO
11100001040100 PAIN EN TRANCHES PRE. PAIN FRAIS MIE 450 A 500 G CASINO
11400001020100 PAIN EN TRANCHES PRE. PAIN FRAIS MIE 250 A 320 G AUCHAN
11400001030100 PAIN EN TRANCHES PRE. PAIN FRAIS MIE 321 A 449 G AUCHAN
11400001040100 PAIN EN TRANCHES PRE. PAIN FRAIS MIE 450 A 500 G AUCHAN
11500001040100 PAIN EN TRANCHES PRE. PAIN FRAIS MIE 450 A 500 G MONOPRIX/LA FORME
11600001020100 PAIN EN TRANCHES PRE. PAIN FRAIS MIE 250 A 320 G FORZA
11600001040100 PAIN EN TRANCHES PRE. PAIN FRAIS MIE 450 A 500 G FORZA
12000001040100 PAIN EN TRANCHES PRE. PAIN FRAIS MIE 450 A 500 G CORSO
13600001040100 PAIN EN TRANCHES PRE. PAIN FRAIS MIE 450 A 500 G MAMMOUTH
14000001040100 PAIN EN TRANCHES PRE. PAIN FRAIS MIE 450 A 500 G ED
14100001040100 PAIN EN TRANCHES PRE. PAIN FRAIS MIE 450 A 500 G INTERMARCHE
14400001020100 PAIN EN TRANCHES PRE. PAIN FRAIS MIE 250 A 320 G RALLYE
14400001030100 PAIN EN TRANCHES PRE. PAIN FRAIS MIE 321 A 449 G RALLYE
14400001040100 PAIN EN TRANCHES PRE. PAIN FRAIS MIE 450 A 500 G RALLYE
14500001040100 PAIN EN TRANCHES PRE. PAIN FRAIS MIE 450 A 500 G GRANDE CONFIANCE
19600001020100 PAIN EN TRANCHES PRE. PAIN FRAIS MIE 250 A 320 G AUTRES DISTRIB. NON
19600001030100 PAIN EN TRANCHES PRE. PAIN FRAIS MIE 321 A 449 G AUTRES DISTRIB. NON
19600001040100 PAIN EN TRANCHES PRE. PAIN FRAIS MIE 450 A 500 G AUTRES DISTRIB. NON
19600001050100 PAIN EN TRANCHES PRE. PAIN FRAIS MIE + DE 500 G AUTRES DISTRIB. NON
20100001010100 PAIN EN TRANCHES PRE. PAIN FRAIS MIE MOINS DE 250 G JACQUET
20100001020100 PAIN EN TRANCHES PRE. PAIN FRAIS MIE 250 A 320 G JACQUET
20100001030100 PAIN EN TRANCHES PRE. PAIN FRAIS MIE 321 A 449 G JACQUET
20100001040100 PAIN EN TRANCHES PRE. PAIN FRAIS MIE 450 A 500 G JACQUET
20100001050100 PAIN EN TRANCHES PRE. PAIN FRAIS MIE + DE 500 G JACQUET
20200001020100 PAIN EN TRANCHES PRE. PAIN FRAIS MIE 250 A 320 G DUROI
20200001040100 PAIN EN TRANCHES PRE. PAIN FRAIS MIE 450 A 500 G DUROI
20200001050100 PAIN EN TRANCHES PRE. PAIN FRAIS MIE + DE 500 G DUROI
20200101040100 PAIN EN TRANCHES PRE. PAIN FRAIS MIE GRAND MOELLEUX 450 A 500 G DUROI
20200101050100 PAIN EN TRANCHES PRE. PAIN FRAIS MIE GRAND MOELLEUX + DE 500 G DUROI
20200201050100 PAIN EN TRANCHES PRE. PAIN FRAIS MIE LUXBREAD + DE 500 G DUROI
20400101010100 PAIN EN TRANCHES PRE. PAIN FRAIS MIE BRIOCHISSIMO MOINS DE 250 G TURNER
20400101020100 PAIN EN TRANCHES PRE. PAIN FRAIS MIE BRIOCHISSIMO 250 A 320 G TURNER
20400101040100 PAIN EN TRANCHES PRE. PAIN FRAIS MIE BRIOCHISSIMO 450 A 500 G TURNER
20400201010100 PAIN EN TRANCHES PRE. PAIN FRAIS MIE RAISINELLA MOINS DE 250 G TURNER
20400201020100 PAIN EN TRANCHES PRE. PAIN FRAIS MIE RAISINELLA 250 A 320 G TURNER
20400201040100 PAIN EN TRANCHES PRE. PAIN FRAIS MIE RAISINELLA 450 A 500 G TURNER
20400201050100 PAIN EN TRANCHES PRE. PAIN FRAIS MIE RAISINELLA + DE 500 G TURNER
20400401010100 PAIN EN TRANCHES PRE. PAIN FRAIS MIE KING CORN MOINS DE 250 G TURNER
20400401020100 PAIN EN TRANCHES PRE. PAIN FRAIS MIE KING CORN 250 A 320 G TURNER
20400401040100 PAIN EN TRANCHES PRE. PAIN FRAIS MIE KING CORN 450 A 500 G TURNER
20400401050100 PAIN EN TRANCHES PRE. PAIN FRAIS MIE KING CORN + DE 500 G TURNER
20500001010100 PAIN EN TRANCHES PRE. PAIN FRAIS MIE MOINS DE 250 G HARRY'S
20500001020100 PAIN EN TRANCHES PRE. PAIN FRAIS MIE 250 A 320 G HARRY'S
20500001040100 PAIN EN TRANCHES PRE. PAIN FRAIS MIE 450 A 500 G HARRY'S
20500001050100 PAIN EN TRANCHES PRE. PAIN FRAIS MIE + DE 500 G HARRY'S
20500101040100 PAIN EN TRANCHES PRE. PAIN FRAIS MIE AMERICAN BREAD 450 A 500 G HARRY'S
20500101050100 PAIN EN TRANCHES PRE. PAIN FRAIS MIE AMERICAN BREAD + DE 500 G HARRY'S
20500301040100 PAIN EN TRANCHES PRE. PAIN FRAIS MIE EXTRA TENDRE 450 A 500 G HARRY'S
20500301050100 PAIN EN TRANCHES PRE. PAIN FRAIS MIE EXTRA TENDRE + DE 500 G HARRY'S
20500401050100 PAIN EN TRANCHES PRE. PAIN FRAIS MIE KING CORN + DE 500 G HARRY'S
20600001040100 PAIN EN TRANCHES PRE. PAIN FRAIS MIE 450 A 500 G A.V.F.
20700001040100 PAIN EN TRANCHES PRE. PAIN FRAIS MIE 450 A 500 G BLEOR

Le traitement de ces 75 références prend alors moins de deux minutes.

5.2. Les glaces à emporter

Ce marché contient 1229 références.

5.2.1. Règles d'inférence

Les attributions pour les facettes "QUNIT" et "PAR COMBIEN" ont été faites dans un premier temps. L'utilisation de règles d'inférences automatisées ne pose pas ici de difficulté. Les libellés correspondant à la "quantité unitaire" ou au "par combien"ont préalablement été uniformisés. Ensuite à chacun des types de libellé est associé un code unique.

Un extrait du fichier de la détermination des régles d'inférence

245: GLACES A EMPORTER						OCA 8 - PAR COMBIEN	OCA 9 - QUNIT
Référence SECODIP	Eléments E2 à E6						
	LIB2 LIB3	LIB4	LIB5	LIB6			
5030000	. SPECIALITES PORTIONS	BOITE DE 3 à -				3 X 3 et moins	
8033000	. BATON	BOITE DE 3 à -	70 ML ET -			3 X 3 et moins	4000 7 cl et moins
5040000	. SPECIALITES PORTIONS	BOITE DE 4				5 X 4	
6040000	. POTS - 100 CC	BOITE DE 4				5 X 4	4002 10 cl
8043000	. BATON	BOITE DE 4	70 ML ET -			5 X 4	4000 7 cl et moins
8043100	. BATON	BOITE DE 4	+ DE 70 ML			5 X 4	4001 Plus de 7 cl
9040000	. CORNET	BOITE DE 4				5 X 4	
11040000	. BUCHETTES	BOITE DE 4				5 X 4	
15040000	. BARRE	BOITE DE 4				5 X 4	
25040000	. MYSTERE	BOITE DE 4				5 X 4	
27040000	. LIEGEOIS	BOITE DE 4				5 X 4	
54040000	. OMELETTE NORVEGIENNE	BOITE DE 4				5 X 4	
5080000	. SPECIALITES PORTIONS	BOITE DE 5 à 6				7 X 5 - 6	
8083000	. BATON	BOITE DE 5 à 6	70 ML ET -			7 X 5 - 6	4000 7 cl et moins
8083105	. BATON	BOITE DE 5 à 6	+ DE 70 ML	BATONNETS ADULTES		7 X 5 - 6	4001 Plus de 7 cl
9080000	. CORNET	BOITE DE 5 à 6				7 X 5 - 6	
1150102	. CREME GLACEE FAMIL.	2 LITRES à +	SANS NOTION PARFUM	AUTRES EMBALLAGES			4032 200 cl et plus

5.2.2. Approche typicaliste

5.1.2.1. Constitution d'un fichier de prototypes

Sur ce marché les éléments E1 à E6 de la référence SECODIP prennent les significations suivantes :

E1 = marques

E2 = glaces / sorbet

E3 = variété

E4 = conditionnement

E5 = parfum

E6 = emballage

Nous avons retenu comme significatifs, hors marques et conditionnements, les éléments E3 et E5 pour définir les prototypes de ce marché. Dans un premier temps, 41 produits différents devront être codés et serviront ensuite de base au traitement de l'ensemble des 1229 références. La diminution de la taille du fichier à traiter est ici particulièrement importante.

Fichier des prototypes du marché des "glaces à emporter"

Glaces	CREME GLACEE FAMILIALE	SANS NOTION DE PARFUM
Glaces	CREME GLACEE FAMILIALE	PLUSIEURS PARFUMS
Glaces	CREME GLACEE FAMILIALE	VANILLE
Glaces	CREME GLACEE FAMILIALE	CHOCOLAT
Glaces	CREME GLACEE FAMILIALE	PRALINE
Glaces	CREME GLACEE FAMILIALE	PISTACHE
Glaces	CREME GLACEE FAMILIALE	CAFE
Glaces	CREME GLACEE FAMILIALE	CITRON
Glaces	CREME GLACEE FAMILIALE	FRAISE
Glaces	CREME GLACEE FAMILIALE	FRUIT ROUGE
Glaces	CREME GLACEE FAMILIALE	FRUIT EXOTIQUE
Glaces	CREME GLACEE FAMILIALE	FRUIT TEMPERE
Glaces	CREME GLACEE FAMILIALE	AUTRE PARFUM
Glaces	SORBET FAMILIAL	SANS NOTION DE PARFUM
Glaces	SORBET FAMILIAL	PLUSIEURS PARFUMS
Glaces	SORBET FAMILIAL	PRALINE
Glaces	SORBET FAMILIAL	PISTACHE
Glaces	SORBET FAMILIAL	CAFE
Glaces	SORBET FAMILIAL	CITRON
Glaces	SORBET FAMILIAL	FRAISE
Glaces	SORBET FAMILIAL	FRUIT ROUGE
Glaces	SORBET FAMILIAL	FRUIT EXOTIQUE
Glaces	SORBET FAMILIAL	FRUIT TEMPERE
Glaces	SORBET FAMILIAL	AUTRE PARFUM
Glaces	BUCHE GLACEE	
Glaces	AUTRES SPECIALITES A PART	
Glaces	SPECIALITES EN PORTIONS	
Glaces	POTS - 100 CC	
Glaces	COUPE + 100 CC	
Glaces	BATON	70 ML ET -
Glaces	BATON	+ DE 70 ML
Glaces	CORNET	
Glaces	BOUCHEES GLACEES	
Glaces	BUCHETTES	
Glaces	BOULES	
Glaces	KOUKOULINA INDIVIDUEL	
Glaces	BARRE	
Glaces	MYSTERE	
Glaces	LIEGEOIS	
Glaces	VACHERIN	
Glaces	OMELETTE NORVEGIENNE	

5.1.2.2. Codification des prototypes

Dans le fichier du CIQUAL n'apparaît aucun produit du type "crème glacée". Nous nous reportons donc au fichier américain de la FDA. La recherche s'effectue sur le libellé "ICE". Cinq produits sont sélectionnés. Trois sont effectivement des crèmes glacées.

1061	VANILLA ICE CREAM 10% FAT
1062	VANILLA ICE CREAM 16% FAT
1063	FRENCH VANILLA ICE CREAM
1064	VANILLA ICE MILK
1065	VANILLA ICE MILK SOFT SERVE

Elles sont codées de la façon suivante :

1061	VANILLA ICE CREAM 10% FAT
A0227	crème glacée
B1201	vache
C0113	lait u sous-prd
E0139	lq très-visq - partic visible
F0018	transf thermique partielle
G0003	pas de méth de cuisson applicable
H0178	aéré ou fouetté
H0100	ext ou conc d'arô ou d'épice aj
H0136	sucre ou srp de sucre aj
J0135	pasteurisé à chaud
K0003	sans milieu de cdtionnmt
M0001	rcp ou embal non spécifié
N0001	surf contact alim inconnue
P0024	produit de consommation courante

1062	VANILLA ICE CREAM 16% FAT
A0227	crème glacée
B1201	vache
C0113	lait u sous-prd
E0139	lq très-visq - partic visible
F0018	transf thermique partielle
G0003	pas de méth de cuisson applicable
H0178	aéré ou fouetté
H0100	ext ou conc d'arô ou d'épice aj
H0136	sucre ou srp de sucre aj
J0135	pasteurisé à chaud
K0003	sans milieu de cdtionnmt
M0001	rcp ou embal non spécifié
N0001	surf contact alim inconnue
P0024	produit de consommation courante

1063	FRENCH VANILLA ICE CREAM
A0227	crème glacée
B1201	vache
C0113	lait u sous-prd
E0139	lq très-visq - partic visible
F0018	transf thermique partielle
G0003	pas de méth de cuisson applicable
H0178	aéré ou fouetté
H0100	ext ou conc d'arô ou d'épice aj
H0185	jaune d'oeuf aj
H0136	sucre ou srp de sucre aj
J0135	pasteurisé à chaud
K0003	sans milieu de cdtionnmt
M0001	rcp ou embal non spécifié
N0001	surf contact alim inconnue
P0024	produit de consommation courante

Les deux premiers produits ont été codés de façon identique, le troisième ne diffère des autres que par l'ajout de jaune d'oeuf. Nous retiendrons les codes du premier produit comme référence et nous les modifions éventuellement pour qu'ils soient parfaitement adaptés à notre produit.

Les attributions effectuées pour le prototype sont donc les suivantes :

GLACES CREME GLACEE FAMILIALE PLUSIEURS PARFUMS"

A0227	crème glacée
B1201	vache
C0113	lait ou sous-produit
E0139	liquide très visqueux sans particule visible
F0018	transformation thermique partielle
G0003	pas de méthode de cuisson applicable
H0178	aéré ou fouetté
H0100	extrait ou concentré d'arôme ou d'épice ajouté
H0136	sucre ou sirop de sucre ajouté
J0136	congelé
K0003	sans milieu de conditionnement
M0001	récipient ou emballage non spécifié
N0001	surface en contact avec l'aliment inconnue
P0024	produit de consommation courante
1	Recette

5.1.2.3. Codification des références ou produits SECODIP

Une fois les prototypes codés, nous reprenons la liste complète des produits SECODIP. Voici un extrait du fichier des références réelles sur le marché des "Glaces à emporter" :

00100001120102	GLACES CREME GLACEE FAMILIALE SANS NOTION DE PARFUM 500 ML ET MOINS AUTRES EMBALLAGES AUTRES MARQUES NON IDENTI
00100001120201	GLACES CREME GLACEE FAMILIALE PLUSIEURS PARFUMS 500 ML ET MOINS BAC PLASTIQUE AUTRES MARQUES NON IDENTI
00100001120301	GLACES CREME GLACEE FAMILIALE VANILLE 500 ML ET MOINS BAC PLASTIQUE AUTRES MARQUES NON IDENTI
00100001120401	GLACES CREME GLACEE FAMILIALE CHOCOLAT 500 ML ET MOINS BAC PLASTIQUE AUTRES MARQUES NON IDENTI
00100001120501	GLACES CREME GLACEE FAMILIALE PRALINE 500 ML ET MOINS BAC PLASTIQUE AUTRES MARQUES NON IDENTI

Une référence (ou produit) SECODIP est donc codée comme suit :

00100001120201	GLACES CREME GLACEE FAMILIALE PLUSIEURS PARFUMS 500 ML ET MOINS BAC PLASTIQUE AUTRES MARQUES NON IDENTIFIEES"

A0227	crème glacée
B1201	vache
C0113	lait ou sous-produit
E0139	liquide très visqueux sans particule visible
F0018	transformation thermique partielle
G0003	pas de méthode de cuisson applicable
H0178	aéré ou fouetté
H0100	extrait ou concentré d'arôme ou d'épice ajouté
H0136	sucre ou sirop de sucre ajouté
J0136	congelé
K0003	sans milieu de conditionnement
M0187	récipient rigide avec couvercle en plastique
N0036	plastique
P0024	produit de consommation courante
1	Recette

Les modifications pour adapter le codage du prototype à la référence réelle concernent l'emballage (facette M) et la surface en contact avec l'aliment (facette N).

5.3. Avancée de la codification au 31/12/91

Une vingtaine de marchés alimentaires (entre autres les marchés susceptibles de contenir des édulcorants) ont été traités. Cela représente près de huit mille produits codés.

N° de marché	Nom du marché	Nbre de références
11	Boissons à base de bière	49
12	Sucre	198
15	Fromages frais	512
16	Yaourts et desserts	1 014
19	Edulcorants	136
21	Barres céréalières	136
22	Bière	470

24	Chocolats en tablettes	387
25	Jus de fruit sodas et boissons aux fruits	1 212
55	Gâteaux de pâtisseries pré-emballés	182
96	Desserts à préparer et tout prêts en conserve	99
212	Pain préemballé	173
215	Fromages frais à tartiner	119
218	Poudres chocolatées	105
223	Pâtes à tartiner chocolatées	27
226	Apéritifs	206
245	Glaces à emporter	1 158
249	Thés et infusions	425
255	Miel	548
269	Lait frais	495
	20 marchés	7 649

La codification continue. Notre objectif est de coder une trentaine de milliers de produits en 1992.

6. Conclusion

La codification d'objets complexes est un problème récurrent dans le travail statistique. Le plus souvent, les difficultés théoriques sous-jacentes peuvent être ignorées sans inconvénient dans la pratique empirique.
Le cas particulier de l'Observatoire des Consommations Alimentaires, en ce qu'il nécessite une codification à la fois très fine et très flexible, ne permettait pas de faire l'économie d'une ontologie relativiste dans les principes de codage, et a été l'occasion d'une réflexion plus approfondie, ainsi que de la mise au point d'une méthode de codification mixte, originale mais transposable dans son principe à d'autres cas. Cette recherche nous a permis d'appliquer à un problème concret certains acquis du formalisme relativiste mis au point dans le département.

Une exploration théorique nous a amenés à reconnaître, dans la façon dont la question de la codification se pose au statisticien, le problème plus général de la traduction, en ce que le

statisticien travaille souvent sur une description du réel (matériau de seconde main) plutôt que sur les phénomènes eux-mêmes. Cette situation est spécifique en ce qu'il existe alors une première série de règles que suit le matériau, à savoir les règles du premier langage de description.
Nous avons également pu déterminer deux grands principes de classification, la classification directe, qui part de la description de l'objet et utilise des règles d'attribution à une classe, et la classification indirecte, qui consiste à classer les objets non pas de manière absolue, mais relative, en les comparant (part exemple à l'aide d'une distance) à d'autres objets.

Nous avons finalement sélectionné trois outils. Le premier est celui de la combinatoire, qui consiste à décrire l'objet par une série prédéfinie de facettes, qui ne peuvent chacune prendre des valeurs que sur un ensemble de descripteurs (fini, mais éventuellement extensible). Les deux autres outils sont ce que nous avons reconnu comme étant les grands principes sous-jacents aux méthodes classificatoires : l'inférence et la typicalité. Chacune a ses avantages et ses inconvénients.

L'inférence est économique car d'application large, systématisable, et peut se mettre en oeuvre à partir de la reconnaissance de seulement une partie de l'objet. Elle permet un codage partiel de nombreux objets à la fois. Par contre, elle est très vulnérable aux "défauts", c'est à dire aux objets atypiques, ainsi qu'aux conflits de règles.
La typicalité est une approche cognitive naturelle, elle permet un codage total ou presque pour des objets pris individuellement. Elle permet de gérer les exceptions sans difficulté, et ne provoque pas de conflits. Par contre elle est peu économique, et difficilement automatisable.

Notre méthode combine ces trois outils. Les objets sont formalisés dans un système combinatoire, à facettes. Les facettes qui peuvent être remplies à l'aide de règles d'inférence le sont, puis on cherche à attribuer, par groupes autant que possibles, les objets à un prototype. On remplit ainsi les facettes qui sont identiques à celles du prototype, puis on attribue "manuellement" les facettes restantes.

La méthode a été concrétisée par un logiciel, CITOCA, qui permet de coder les produits alimentaires à partir de leurs libellés, et d'aboutir à une codification compatible avec le système LANGUAL de description des aliments.

Outre l'efficacité de la méthode et du logiciel pour résoudre notre problème de codification particulier, nous pensons que les outils conceptuels ici mis en évidence et mobilisés peuvent être réemployés avec profit pour la codification d'objets complexes. Le dosage des différentes méthodes devra sans doute être revu en fonction du matériau empirique utilisé.

Il nous semble enfin que le mélange de codification directe et indirecte, et, à un autre niveau, d'inférence et de typicalité, est présent, souvent de manière occulte, dans toute tentative classificatoire. Expliciter ces mécanismes est à la fois une saine précaution épistémologique, et un intéressant levier technique, dans la mesure où l'on peut alors améliorer, de façon raisonnée, les règles de codage.

7. Bibliographie

BENGUIGUI, G., "Les besoins des objets de consommation et les groupes sociaux", *Epistémologie sociologique*, n° 15-16, 1973.

BOLTANSKI, L. *Les cadres*, Paris, Editions de Minuit, 1982.

COMTET, L. *Analyse combinatoire*. Paris, PUF, collection Sup, 1970.

ECO, U. *Lector in fabula : le rôle du lecteur ou la coopération interprétative dans les textes narratifs*, Grasset, 1985.

FEINBERG, M., IRELAND-RIPERT, J., FAVIER, J.-C., "Langual : un langage international pour la description structurée des aliments", *Sciences des Aliments*, 11 (1991) 189-210

GEERTZ, C. *Savoir local, savoir global. Les lieux du savoir*, PUF, Paris, 1986. 1ère éd. am. 1983

LADRIERE, J. L'explication en logique. in : L'explication dans les sciences. Colloque de l'Académie Internationale de Philosophie des Sciences, avec le concours du Centre international d'Epistémologie génétique, Genève 25-29 septembre 1970. Paris, Flammarion, 1973. pp. 19-56.

LAHLOU, S. Le produit nouveau : un concept flou. Consommation N°2, 1985-86.

LAHLOU, S. Les comportements alimentaires des Français. Rapport Crédoc, 1988.

LAHLOU, S. Rappels de statistique : le minimum. Document de travail, Département prospective de la consommation, Crédoc, 1989.

LAHLOU, S. La système-compatibilité. Cahiers de recherche du Crédoc, n° 4, 1990.

LAHLOU, S. BEAUDOUIN, V., CALAMASSI-TRAN, G., EVANS, C., GILLET, C. LION, S., MAFFRE, J., VERHEYDEN, G. "Rapport pour l'Observatoire des Consommations Alimentaires. Etat d'avancement des travaux de la base de données mise en place par le

Crédoc à la fin de la deuxième phase : décembre 1991". Crédoc : Département Prospective de la consommation.

VON LINNE, C. *Voyage en Laponie*, Editions de la différence. Paris, 1983.

MONOD-HERTZEN, G. *L'analyse dimensionnelle et l'épistémologie*, Coll. recherches interdisciplinaires. Editions Maloine-Doine, Paris, 1976.

MOREL, B. "Réflexions philosophiques à partir de la «méthodologie ouverte»". in J. Parain-Vial, ed. Les difficultés de la quantification et de la mesure. Actes du colloque de l'Université de Dijon «Méthodologie comparée des sciences» Maloine, Paris, 1981. pp. 251-263

MOUNIN, G. *Les problèmes théoriques de la traduction*, Gallimard, Paris, 1963.

PEREC, G. *Penser/Classer*. Hachette, Paris, 1985.

POIRIER, R., Nature et réalité de l'objet. in J. Parain-Vial, ed. Les difficultés de la quantification et de la mesure. Actes du colloque de l'Université de Dijon «Méthodologie comparée des sciences» Maloine, Paris, 1981. pp. 275-293.

PORTE, J., "Recherches sur la théorie générale des systèmes formels et sur les systèmes connectifs", Paris, Gauthier-Villars, Louvain E. Nauwelaerts. Collection de logique mathématique, série A, XVIII. 1965.

QUINE, W. V O. *Le mot et la chose*. Flammarion, Paris, 1977.

RASTIER, F. La triade sémiotique, le trivium et la sémantique linguistique, nouveaux actes sémiotiques, 9, PULIM (Limoges) 1990.

ROSCH, E.R. Universals and Specifics in Human Categorization, in R. Brislin, S. Bosner & W. Lonner eds. Cross Cultural Perspectives on Learning, Halsted, New-York, 1975.

ROSCH, E.R., MERVIS, C.B., GRAY, W., JOHNSON, D.N., BOYES-BRAEM, P. Basic Objects in Natural Categories, Cognitive Psychology, 1976, 8, pp. 382-439.

SAGET, H. Nature et limites de la quantification. in J. Parain-Vial, ed. Les difficultés de la quantification et de la mesure. Actes du colloque de l'Université de Dijon «Méthodologie comparée des sciences» Maloine, Paris, 1981. pp. 265-274.

SAUSSURE, F. de. *Cours de linguistique générale*, Payot, Paris, 1985.

SERRES, M. *Le système de Leibniz et ses modèles mathématiques*, PUF, Paris, 1968.

WITTGENSTEIN, L. *Tractatus Logico-philosophicus*, Gallimard, Paris, 1961.

WITTGENSTEIN, L. *Le brun et le cahier bleu*, Trad fr. Gallimard, 1965.

* *

Center for Food Safety and Applied Nutrition, *Langual Vocabulary User's Manual,* Rev. December 1990.

Center for Food Safety and Applied Nutrition, *Langual an automated method for describing, capturing and retrieving data about food*, Codata conference, Columbus, Ohio, USA, 1990.

CAHIER DE RECHERCHE

Récemment parus :

Pratiques exemplaires ou exemples de pratiques : l'évaluation dans le secteur social aux Etats-Unis - Analyse de monographies présentées dans "Evaluation Review" et dans "Evaluation and the health professions", par Patricia Croutte, Michel Legros, N° 17, Juillet 1991.

Etude de l'opinion et enquêtes de référence : Aspects théoriques, méthodologiques et informatiques (Soutenance : Avril 1988), par Anastassios Iliakopoulos, N° 18, Septembre 1991.

Entre école et emploi : les transitions incertaines, par Denise Bauer, Patrick Dubéchot, Michel Legros, N° 19, Septembre 1991.

Price expectations of french households : A test on INSEE panel data (1972 - 1988), par François Gardes, Jean-Loup Madre, N° 20, Octobre 1991.

Chômeurs au fil du temps, par Isa Aldeghi, N° 21, Novembre 1991.

Deux analyses lexicales : Les améliorations à apporter au fonctionnement de la société - L'image du milieu professionnel, Enquête "Conditions de vie et Aspirations des Français", par Laurent Clerc, Ariane Dufour, N° 22, Janvier 1992.

Président : Bernard SCHAEFER Directeur : Robert ROCHEFORT
142, rue du Chevaleret, 75013 PARIS - Tél. : (1) 40.77.85.00

CRÉDOC

Centre de recherche pour l'Étude et l'Observation des Conditions de Vie