SYNAPSE :
un serveur de nomenclatures tous usages

Emile Bruneau

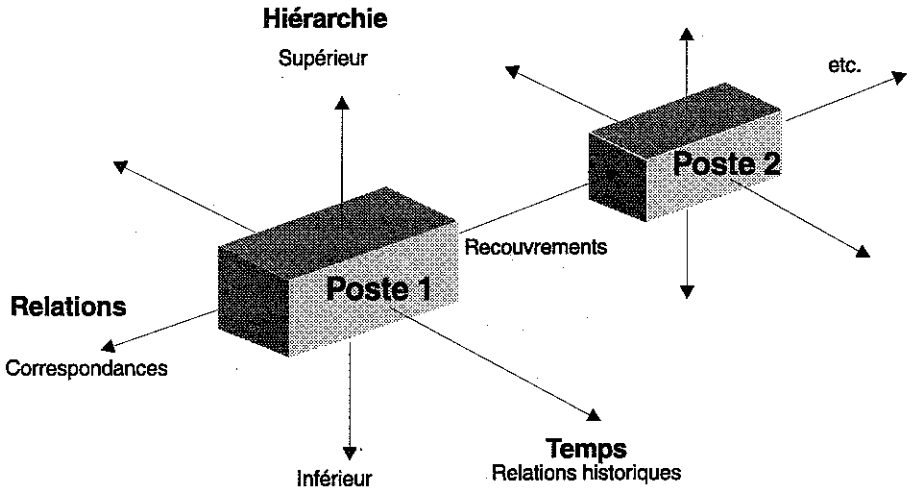
Présentation générale

Notre monde de plus en plus informatisé et "normatif", a besoin de classer et repérer des objets en tous genres, engendrant une multiplication de nomenclatures, codes et listes dans tous les domaines. Ceux-ci répondent à des préoccupations variées (statistiques, administratives, réglementaires ou autres). Bien souvent, une nomenclature (ou une liste ou un code) est définie par rapport à une autre nomenclature ou doit être reliée à une autre pour des besoins de recodification, de comparaison, etc. Enfin, les nomenclatures "vivent", évoluent ; il faut les gérer et les diffuser de la façon la plus efficace possible.

Organisme producteur, gestionnaire et diffuseur de nomenclatures, l'Insee a développé un outil modulaire pour en faciliter la **gestion,** la **coordination** et la **diffusion,** qui innove sur quatre plans :

- 1. Dans la base de données, les nomenclatures sont organisées en réseau. Ceci permet, à partir d'un poste donné, d'accéder à tous les postes qui lui sont liés :

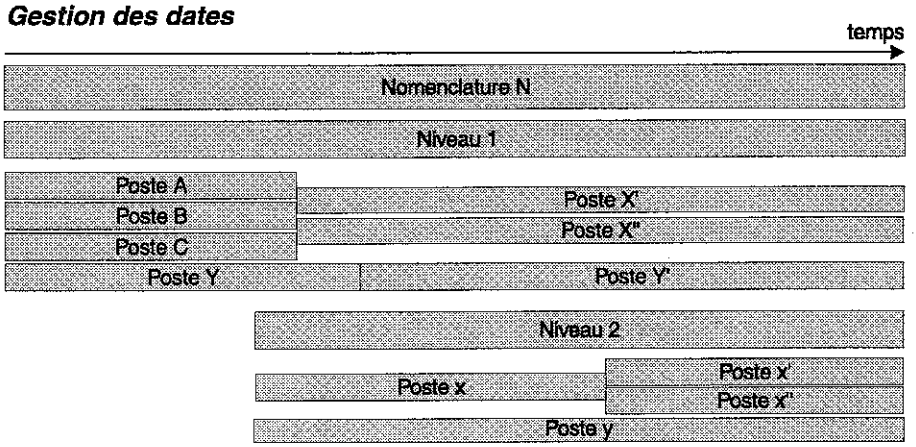
Des tables calculées automatiquement



hiérarchiquement, par correspondance ou recouvrement, ou encore, historiquement (les postes prédécesseurs ou successeurs). Puis, par transitivité, de balayer la base à partir d'un seul point d'entrée, si besoin est ;

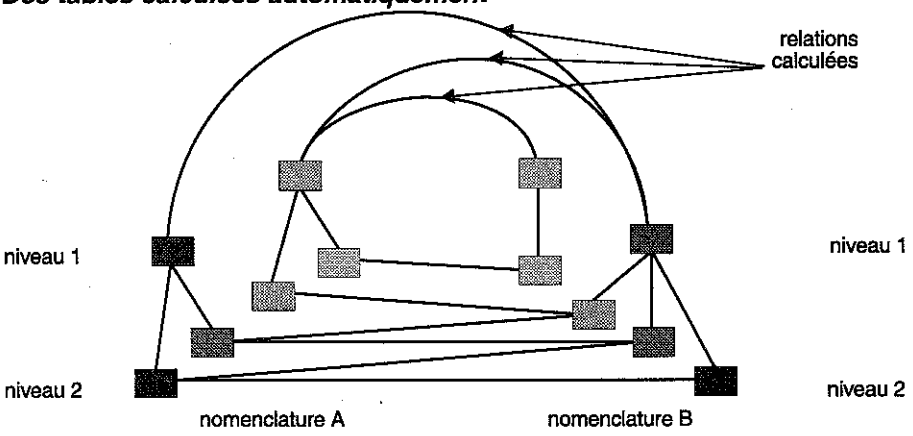
- 2. Les nomenclatures évoluant de façon partielle et sans fréquence fixe, toute entité de la base est datée et chaque élément (nomenclature, niveau, poste, table, relation, intitulé) a sa propre période de validité. Ainsi, lors d'une consultation, d'une édition ou d'une production de fichiers, le choix d'une date de référence quelconque ne sélectionne que les données valides à cette date. Une telle disposition autorise les consultations historiques et limite les dimensions de la base de données, améliorant ainsi les performances du système ;

Gestion des dates



- 3. Les relations entre postes de nomenclatures peuvent être pondérées (à la source et à la cible). Pour les tables calculées par le système (combinaison de plusieurs autres tables), les pondérations sont recalculées en tenant compte des pondérations des tables élémentaires.

Des tables calculées automatiquement



À défaut de pondérations explicites dans les tables élémentaires, le système calcule des pseudo-pondérations en fonction du nombre de relations entretenues par chaque poste (la pondération est égale à 1/n si le nombre de relations est n) ;

- 4. Pour la famille des nomenclatures d'activités et de produits, un module d'analyse linguistique permet d'accéder aux informations par des requêtes en langage naturel. Outre qu'il facilite la consultation (par un public averti ou non), cet outil permet d'effectuer des codifications automatiques de masse (recensements, enquêtes) sans normalisation préalable du langage. Il peut encore aider à la construction de tables de correspondances entre nomenclatures (ou listes ou codes).

Une architecture modulaire

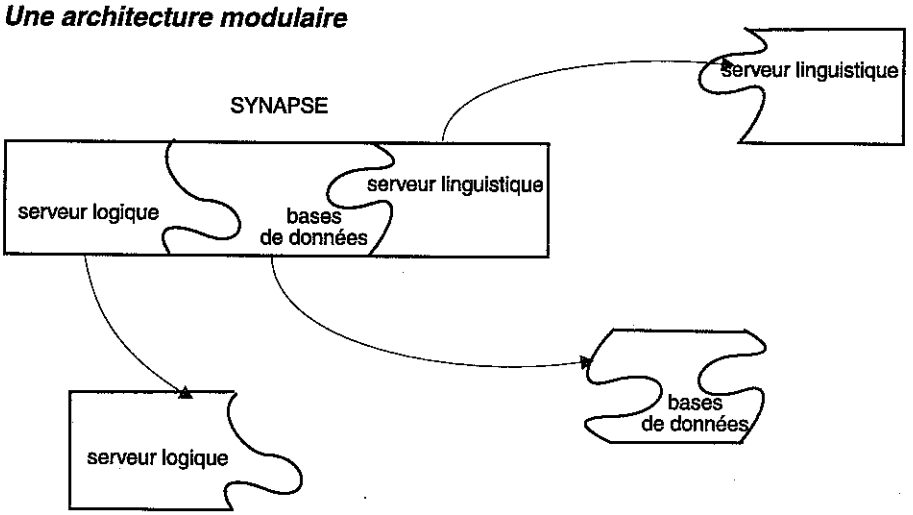
Synapse a été conçu pour répondre à des besoins très divers : gestion et consultation, extraction et édition, recherche conviviale et capacité à prendre en compte l'évolution du langage.

Deux grandes parties structurent le serveur de nomenclatures :

- un serveur logique qui concerne la structure formelle des nomenclatures et les tables de relations, incluant les fonctions d'extraction et de gestion de ces entités ;

- un serveur linguistique qui concerne la structure sémantique des nomenclatures d'activités et de produits, incluant les fonctions de recherche, de codification et de gestion des données linguistiques.

Une architecture modulaire

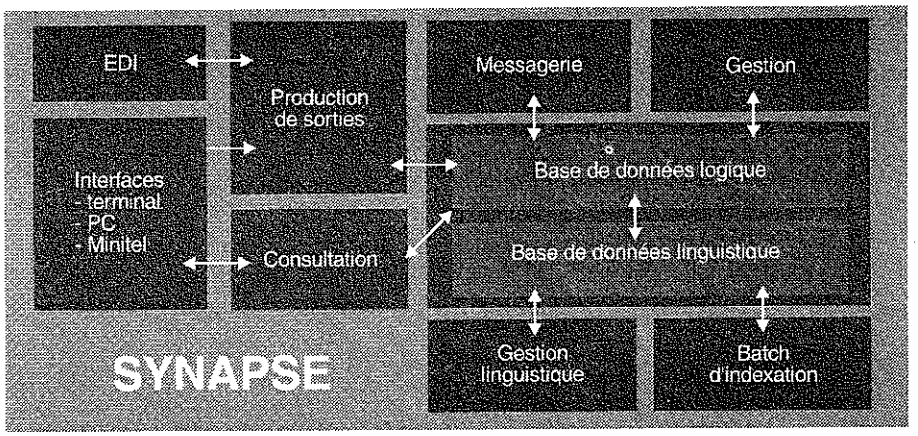


Synapse forme un tout mais chacune des parties (serveur logique et serveur linguistique) est une application en elle-même qui peut fonctionner seule et répondre à des besoins spécifiques. Les bases de données sont aussi des produits intégrables dans d'autres applications (*pour plus de détails, voir le chapitre "Produits dérivés de Synapse"*).

Des fonctionnalités complètes

L'objet même du développement du serveur de nomenclatures était de construire un outil de gestion [1], de coordination et de diffusion de n'importe quel ensemble de nomenclatures et de tables de relations associées. Il fallait donc répondre à tous types de besoins quel que soit l'opérateur. Les fonctionnalités proposées sont donc adaptées aux différents publics potentiellement concernés (nomenclateurs, statisticiens ou simples utilisateurs) connaissant bien, partiellement, peu, voire pas du tout, les nomenclatures.

L'organisation générale de Synapse



Documentation en ligne

Tous les écrans sont documentés par :

- un rappel de l'environnement déjà sélectionné et de la date de référence,
- une aide en ligne (à quoi sert cet écran et comment l'utiliser, définitions, concepts et précisions),

(1) Remarque importante : "gestion" ne signifie pas "construction". La construction de nomenclatures et de tables fait l'objet d'un projet indépendant mais complémentaire du serveur : MIN (Manipulation Informatique des Nomenclatures) sera la boite à outils du nomenclateur qui l'aidera à construire nomenclatures et tables de relations, de façon cohérente et contrôlée, sous contrainte ou non.

- une information sur la dernière entité sélectionnée (une brève description, des statistiques et/ou des règles la concernant).

Date de référence

Le choix d'une date de référence est l'élément clé d'une consultation ou d'une production de sortie : toute entité de la base a sa propre période de validité. Ne sont sélectionnées que les entités valides à la date de référence (par défaut, la date du jour de la consultation).

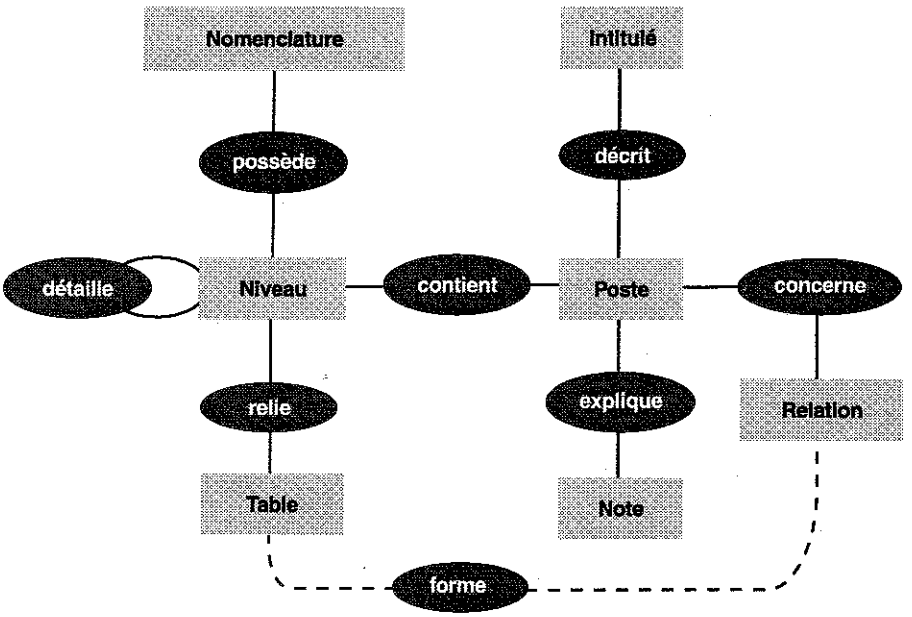
Si par erreur, omission ou méconnaissance, la date choisie est extérieure à la période de validité de la nomenclature ou de la table sélectionnée, le système ramène automatiquement la date de référence à la plus proche date possible (début ou fin de validité).

Consultation

Le but d'une consultation est de pouvoir accéder rapidement et exactement aux informations désirées et à leur environnement.

L'opérateur peut choisir la nomenclature et le niveau à consulter parmi toutes les nomenclatures (et leurs niveaux) présentes dans la base.

Le modèle de données simplifiées



Pour accéder aux postes, trois solutions lui sont offertes :

Il connaît bien la nomenclature : il peut accéder aux postes par leur code ;

- il la connaît moins bien ou cherche un ensemble de postes : il peut définir une liste de postes (par des codets exacts ou en utilisant une formulation abrégée à l'aide d'un joker) ;

Il ne la connaît pas ou veut vérifier un classement : il saisit un descriptif en langage naturel (pour plus de détails à ce sujet, voir la fiche "Serveur linguistique").

Il peut accéder à toutes les informations disponibles sur un poste de nomenclature donné : histoire du poste, notes explicatives, différents intitulés, relations historiques, autres relations.

Si l'opérateur s'intéresse aux tables de relations concernant le niveau choisi, il peut : accéder à toutes les tables existantes dans le système ;

- sélectionner un sous-ensemble de relations en définissant un champ dans le niveau-source et/ou dans le niveau-cible ;

- définir une nouvelle table qui sera construite par le système ;

À partir d'une table, l'opérateur peut accéder aux relations élémentaires qui composent la table, puis aux postes-sources (dans le niveau de départ) ou aux postes-cibles (dans le niveau d'arrivée).

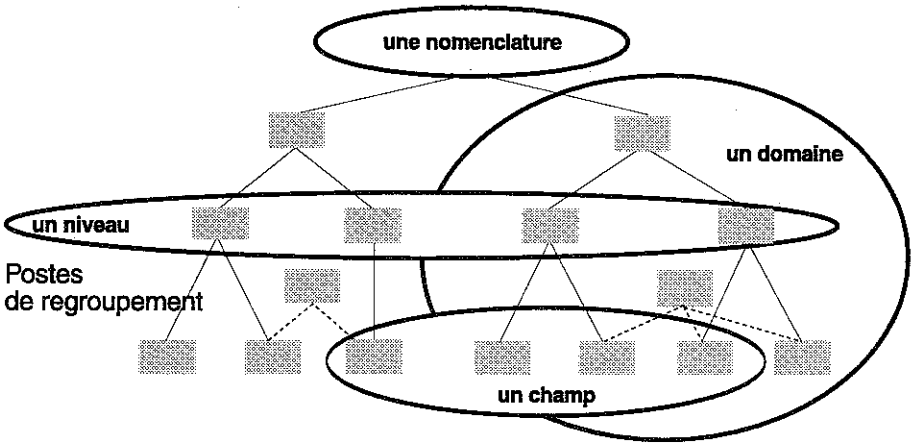
Ainsi, de tables en relations, de relations en postes, de postes en notes explicatives, etc., toute entité de la base liée de près ou de loin aux pôles d'intérêt d'un opérateur peut être consultée lors d'une même session.

Production de sorties

Un opérateur peut vouloir en outre récupérer (sur papiers ou sous forme de fichiers) un ensemble d'informations (visualisées ou non lors d'une consultation).

Il peut donc sélectionner exactement les informations qu'il désire imprimer ou stocker sur fichier. Ceci évite les éditions trop importantes (beaucoup de papier pour peu d'information utile) ou de fastidieux fichiers qu'il faut ensuite retravailler. Ceci permet aussi de répondre immédiatement, sur un support adapté, à toute demande d'informations partielles sur une nomenclature ou une table sans devoir éditer, donner ou vendre l'intégralité d'un document.

Définir des sous-ensembles utiles



Les variables qui peuvent être sélectionnées sont les suivantes :

- pour une nomenclature : les niveaux (un ou plusieurs), le champ (la zone de codes, éventuellement disjointe), les variables (intitulés divers, notes, périodes de validité, etc) ;

- pour une table : les intitulés des postes en relation, les pondérations (à la source et/ou la cible), la période de validité des relations, les commentaires éventuels.

Les fichiers produits sont nommés automatiquement et archivés dans une base de travail.

Les éditions, structurées par une couche SGML, sont aujourd'hui adaptées à l'architecture IBM et nécessitent une imprimante Postscript. Pour les autres architectures, le formatage d'édition doit être revu et adapté à l'environnement matériel disponible.

Transferts et échanges

Produire des fichiers informatiques peut encore s'avérer insuffisant. Il est nécessaire de donner la possibilité à des demandeurs extérieurs d'accéder à la base, d'extraire des données et de les transférer automatiquement vers un autre ordinateur dans un format prédéfini.

Dans le sens inverse, il est plus aisé de recevoir automatiquement de telles données (modifications, mises à jour, créations) selon des formats bien définis que d'attendre les informations sur disquette, bande ou papier dans des formats divers voire inattendus.

Pour répondre à ce type de besoins, le serveur de nomenclatures comprendra une fonction EDI (Echanges de Données Informatisés) qui extrait les données demandées, préformate les données reçues, envoie ou reçoit des messages normalisés. Cette fonction ne pourra toutefois être activée que lorsqu'un message international normalisé adapté aux nomenclatures et aux tables aura été dessiné et accepté au plan international. Ce travail se déroule actuellement dans le cadre de l'Edifact-Board et devrait déboucher sur des tests durant l'année 1996.

Gestion

Les nomenclatures évoluent (nouveaux niveaux, changements de postes, d'intitulés, de notes explicatives, introduction de jurisprudence). Il faut pouvoir apporter toutes ces modifications en temps réel, voire par avance. Les fonctions de gestion permettent de mettre à jour toutes les entités par création, modification, cessation ou suppression.

Le système exécute des contrôles de cohérence lors de toute intervention, alerte l'utilisateur sur les conséquences de cette intervention et en exécute lui-même certaines lorsque la logique d'ensemble le permet.

Malgré les contrôles préprogrammés et quelques enchaînements opératoires prédéfinis, cette première version des fonctions de gestion reste encore très "atomisée". Elle sera améliorée par de meilleurs enchaînements et de plus nombreux contrôles, grâce à l'expérience des premières campagnes complètes de mises à jour qu'elle aura permis de réaliser.

Mise en forme des données et chargement

L'exigence de qualité, tant en ce qui concerne la présentation des données que leur exactitude, a amené à mettre au point un ensemble d'utilitaires sur micro-ordinateur pour aider les gestionnaires à formater les données et éviter les rejets du système engendrés par les erreurs.

Les programmes de chargement vérifient la cohérence des données (soit de façon interne soit avec leur environnement) avant d'effectuer la mise en place des données dans la base. Ils préviennent le gestionnaire des écarts, anomalies, omissions et autres incohérences décelées. C'est à l'opérateur qu'il revient de corriger ou de confirmer les données qui provoquent ces messages.

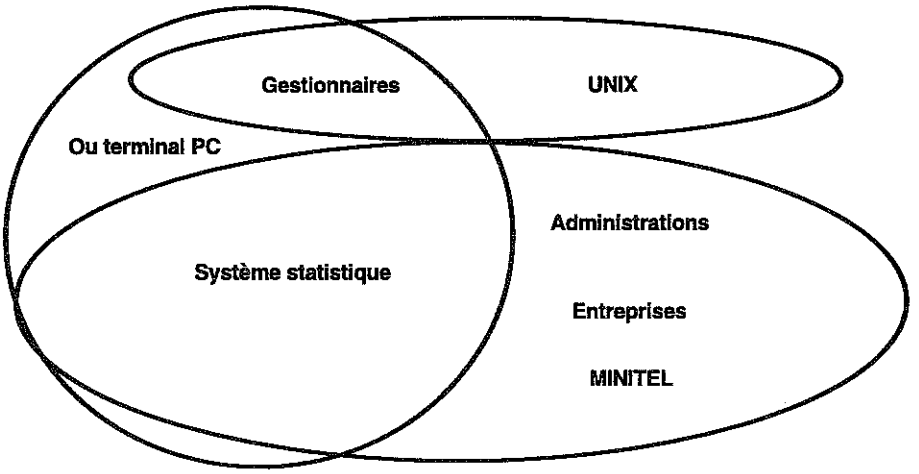
L'intérêt d'un logiciel unique pour gérer et diffuser différentes familles de métadonnées est évident : une seule maintenance, une seule méthode de travail, un système de communication et de consultation homogène pour des opérateurs très divers et, à terme, une interconnexion avec des outils comparables implantés dans des lieux variés.

Interfaces

Les modes d'accès à tout logiciel dépendent des environnements informatiques. Dans le cas du serveur de nomenclatures, ils dépendent des choix d'architecture (IBM et Unix).

Pour ses besoins propres, l'Insee a développé trois interfaces utilisateur différentes (terminal, sous Windows, Videotex) et prévu la possibilité d'un environnement tout Unix, sous X-Motif. Cette diversité permet ainsi des accès via quatre types de matériel : écrans 3270 (dans l'architecture IBM), PC, stations de travail (Unix), Minitel. L'ergonomie est commune aux écrans 3270 et Videotex d'une part, aux PC et stations de travail, d'autre part.

Un exemple de répartition des interfaces



Les nomenclatures, outils complexes au service quotidien des statisticiens, mais aussi d'autres opérateurs (administrations, entreprises, scientifiques) méritaient un développement particulier adapté à leurs spécificités.

Les codes et les listes, objets voisins, sont aussi concernés par cette boîte à outils de gestion et de diffusion. Il s'avère que d'autres objets de structure comparable, peuvent aussi être pris en compte :

- les organigrammes (structures, objets, responsables et fonctions) ;
- les "dictionnaires" de métadonnées (tels le dictionnaire Edifact, les normes ISO, etc.).

Le produit "serveur de nomenclatures" a donc une vocation plus large que celle décelée par l'analyse initiale des besoins associés à la gestion des nomenclatures d'activités et de produits dans un environnement international mouvant.

Le serveur linguistique

La partie linguistique de Synapse a été construite dans trois buts :

- faciliter l'accès aux nomenclatures d'activités et de produits à des opérateurs non avertis ;
- apporter une assistance à la codification pour les autres opérateurs ;
- autoriser des codifications automatiques de masse.

Un quatrième intérêt a été confirmé par les essais :

- aider à la construction de tables de relations en utilisant le langage et pas seulement des choix logiques exogènes.

Qu'est ce que le serveur linguistique ?

C'est une application qui analyse des descriptifs écrits en langage courant, les transforme en formules et compare ces formules aux index des descriptifs définissant les postes des nomenclatures (intitulés et notes explicatives).

Deux différences fondamentales avec les systèmes à mots-clés et la conséquence importante qui en découle sont à souligner :

- le serveur linguistique ne recherche pas, pour les comparer, des chaînes de caractères mais des concepts (un mot pouvant couvrir plusieurs concepts) ;
- les descriptifs peuvent inclure (et il est souhaitable qu'il en soit ainsi) des mots de liaison, des prépositions et de la ponctuation. Ce sont ces mots qui, vides de sens pour les systèmes à mots-clés, donnent tout leur sens aux descriptifs ;
- conséquence : les réponses aux requêtes ne sont pas les postes dont les intitulés comportent un ensemble (ou un sous-ensemble) de mots-clés communs avec ceux des requêtes, mais ceux dont le contenu sémantique défini par les intitulés, les notes explicatives, voire la jurisprudence, est le plus proche.

Comment fonctionne le serveur linguistique ?

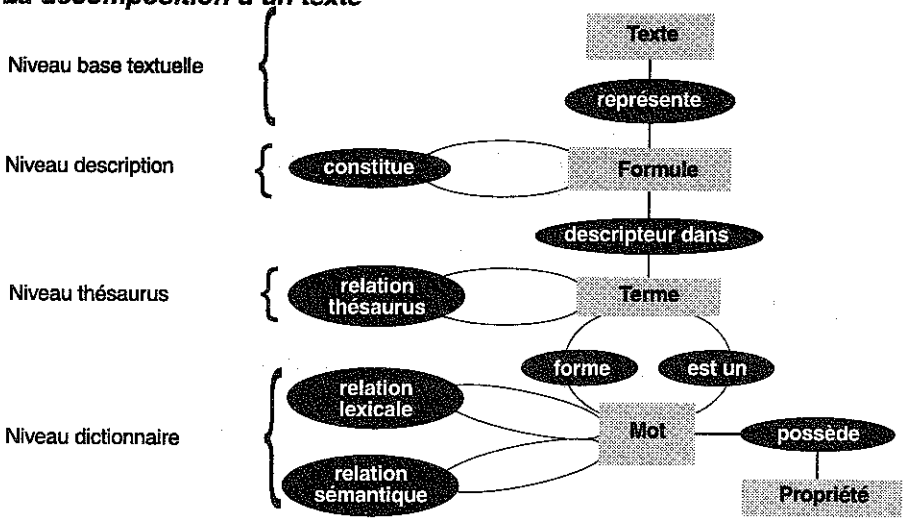
Les composantes de bases sont :

- des référentiels (dictionnaire, thésaurus, base d'index) ;
- des algorithmes (grammaires, cheminements dans le réseau des nomenclatures).

Tout descriptif est analysé de la façon suivante :

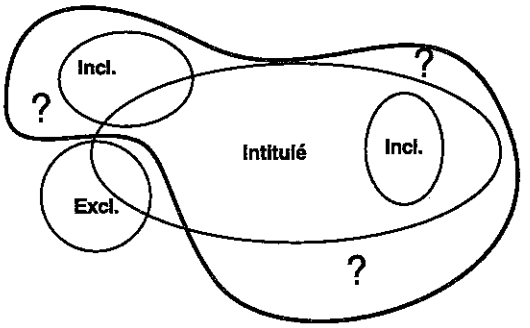
- un texte est transformé en une formule comprenant une ou plusieurs interprétations ;
- une formule est décrite par des termes (ou concepts) qui forment le réseau sémantique du thésaurus ; des attributs précisent la vocation des termes dans les formules (et donc dans les descriptifs) ;
- les termes de thésaurus sont associés entre eux par des liens de domaine, de proximité et de généricité-spécificité ;
- les termes de thésaurus sont associés aux mots du dictionnaire qui possèdent des propriétés (utilisations, domaines, liens de synonymie).

La décomposition d'un texte



Pour définir le champ sémantique des postes, tous les descriptifs connus (intitulés, notes explicatives, jurisprudence) sont analysés et indexés sur les postes correspondants. Les exclusions renvoient bien vers les bons postes même si ceux-ci ne comportent pas les descriptifs correspondants en inclusion. Cet ensemble forme la base d'index.

Petit retour sur la définition d'un poste

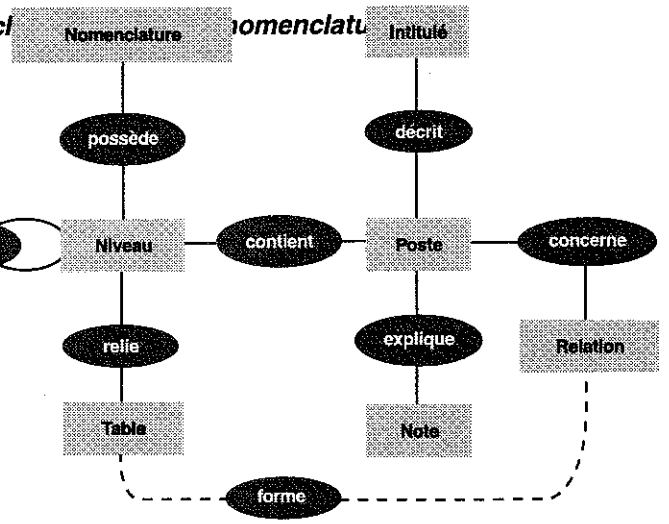


- Un intitulé " balaie" grossièrement le contenu d'un poste et ne cherche qu'à le nommer
- Les notes explicatives précisent le contenu et définissent pour partie ses frontières tant en inclusion qu'en exclusion.
- Des zones d'ombre subsistent, que seule la jurisprudence peut combler au fil du temps

Toute formule représentant le texte d'une requête est comparée à l'ensemble des formules de la base d'index (avec un algorithme d'optimisation). Après comparaison, suivant une échelle de proximité sémantique calculée par un autre algorithme, le serveur renvoie en écho, pour chaque interprétation trouvée, les postes répondant à la requête.

Si des index équivalents à (ou proches de) la formule de la requête sont trouvés dans d'autres nomenclatures, le système utilise les relations entre postes (réseau logique) pour répondre dans la nomenclature et au niveau choisis.

Exemple : recl

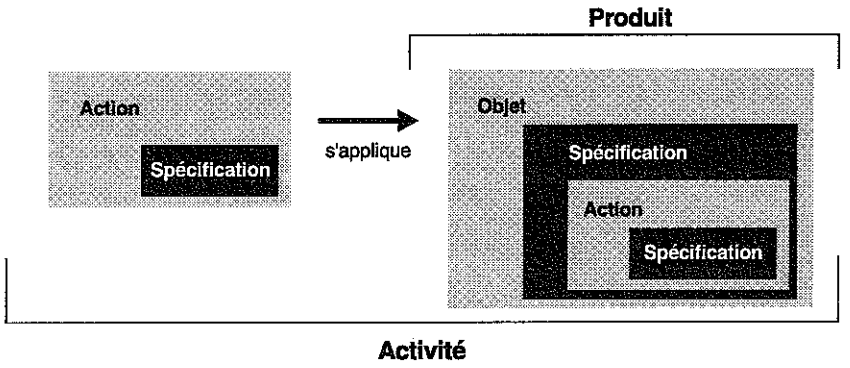


En cas d'interprétations multiples pour une même requête :

- lors d'une consultation, l'opérateur supprime l'ambiguïté en choisissant la bonne interprétation ;

- lors d'une codification automatique, le système retient la plus probable, à savoir la plus proche sémantiquement de la requête.

Descriptifs et multilinguisme



Dans la quasi-totalité des langues européennes (au moins), les descriptifs d'activités et de produits suivent le même schéma :

- une action (éventuellement spécifiée) soit décrit seule une activité, soit la décrit en s'appliquant à un produit ;
  exemples : commerce, décolletage, commerce de gros, montage pour tiers de X.
- un produit est décrit par un objet éventuellement spécifié. Cette spécification peut elle-même être une action (spécifiée ou non) ;
  exemples : meubles, machines-outils à bois, matériel agricole, équipements pour automobile, pièces détachées de Y.

La conjugaison des deux parties (action + objet) entraîne la formulation de descriptifs parfois complexes (sans compter les synonymes, les périphrases ou les inversions);
exemple : conception, fabrication et réparation de machines automatiques destinées à la conservation des aliments.

Cette formulation générique des descriptifs d'activités et de produits conduit à une économie d'échelle importante en cas d'utilisation d'une autre langue. En effet, la structuration des concepts du thésaurus comme les formules représentant les descriptifs des nomenclatures sont indépendantes de la langue. En revanche, le vocabulaire (dictionnaire) et ses propriétés ainsi que les grammaires dépendent de la langue.

Le serveur linguistique inclut un thésaurus de type multilingue qui évite de repartir à zéro si l'on décide de prendre en compte une nouvelle langue et permet de limiter les travaux à ceux spécifiques à cette nouvelle langue :

- développement des grammaires ;
- constitution d'un dictionnaire adapté ;
- traduction du thésaurus ;
- indexation de nouvelles nomenclatures
  (hors les nomenclatures internationales et européennes).

L'ensemble des fonctionnalités et des outils linguistiques étant en place, l'économie globale lors du traitement d'une nouvelle langue se situe entre 50 et 70% suivant les langues.

Comment évolue le serveur linguistique ?

Une boîte à outils de gestion linguistique et d'indexation est intégrée au serveur linguistique.

Elle permet de

- compléter le dictionnaire ;
- d'enrichir le thésaurus ;
- d'indexer de nouvelles nomenclatures ;
- de surindexer, modifier ou corriger les index des nomenclatures déjà chargées ;
- d'analyser la construction des formules en vue de corriger les index, les grammaires, voire les descriptifs...

Une retombée du serveur linguistique est de permettre aux nomenclateurs de corriger des intitulés ou des notes explicatives qu'ils ont eux-mêmes rédigés : ce que le serveur ne comprend pas peut bien sûr être lié à une mauvaise indexation ou à l'imprécision des grammaires mais est aussi très souvent dû à une mauvaise rédaction des descriptifs.

On trouvera en fin de fiche "quelques réflexions, questions et conseils" concernant les serveurs linguistiques.

Quels résultats faut-il attendre ?

Fût-il un système-expert, le serveur linguistique ne saurait remplacer l'analyse d'un humain surtout si celui-ci est expert en nomenclatures. Par ailleurs, le système travaillant à partir d'une base de connaissances, la qualité de celle-ci dépend de son enrichissement au fil du temps et pas seulement de son état initial. Ceci dit, ses performances n'en sont pas pour autant médiocres.

Pour mesurer les possibilités du serveur, l'Insee a retenu une batterie de 7 tests (8 pour les versions ultérieures) regroupant 8300 descriptifs de complexité variable, codés dans trois nomenclatures (NAF : Nomenclature d'Activités Française, SH : Système Harmonisé, CPA : Classification des Produits associée aux Activités). Ces descriptifs sont des activités bien formulées ou mal rédigées, des produits et des raisons sociales incluant une indication d'activité.

Les tests ont été ventilés en requêtes "complexes", "simples" et "hétérogènes". Deux environnements ont été pris en compte : la consultation (un seul écran de treize réponses possibles est consulté ; la bonne réponse s'y trouve) et la codification automatique (seule la première réponse rendue est intéressante). Quatre indicateurs ont été retenus :

$$\text{efficacité} = \frac{\text{nombre de descriptifs codés}}{\text{nombre total de descriptifs}}$$

$$\text{qualité} = \frac{\text{nombre de descriptifs bien codés}^{*}}{\text{nombre de descriptifscodés}}$$

performances:

- temps moyen par réponse ;
- nombre moyen de réponses.

Les résultats sont les suivants : (avec la version 2.0 livrée en juillet 1994)

Corpus en %	Efficacité en %	Qualité en %	
		Consultation	Codification
Complexe	82,2 à 85,7	63,1 à 70,0	40,3 à 45,5
Hétérogène	87,8	73,3	55,6
Simple	97,6 à 97,8	87,0 à 95,3	65,7 à 73,9

* parmi 13 en consultation ou en premier en codification

Corpus	Temps moyen en sec.	Nombre moyen de réponses
Complexe	à 1s	3,4 à 5,1
Hétérogène	de 1/10	4,6
Simple	à 13 s	3,9 à 5,2

À souligner que des tests portant sur deux corpus complexes et réalisés par des "spécialistes" (non experts) de la NAF et un système de mots-clés employé dans le Répertoire des entreprises donnent les résultats suivants :

Corpus	Efficacité en %	Qualité en %	
		Consultation	Codification
Spécialistes	93,4 et 100	///	58 et 67
Mots-clés	62,0 et 91,4	56 et 62	25 et 29

S'il ne faut donner aucune valeur statistique à une telle comparaison, elle n'en est pas moins un "indicateur des performances" du serveur linguistique.

Les améliorations prévues (aménagement des grammaires, des algorithmes et des index, indexation de nouvelles nomenclatures, introduction de jurisprudence) n'auront plus qu'un effet mineur sur l'efficacité (taux de réponse). En revanche, la qualité tant en codification qu'en consultation doit encore gagner 10 à 15 points grâce à ces améliorations pour aboutir aux résultats attendus suivants :

Corpus	Efficacité en %	Qualité en %	
		Consultation	Codification
Complexe	85 à 90	75 à 80	50 à 60
Hétérogène	90 à 95	80 à 85	60 à 70
Simple	98 à 99	95 à 100	75 à 85

Il ne faudra pas espérer mieux, dans un premier temps au moins.

Quelques réflexions, questions et conseils

Pour avoir investi dans un système linguistique, l'Insee n'est pas devenu pour cela expert du domaine. L'expérience du développement du serveur linguistique a toutefois permis d'accumuler des informations qui peuvent servir à d'autres :

- les problèmes linguistiques concernant les descriptifs d'activités et de produits ont été clairement identifiés ;

- les questions à poser concernant les méthodes, les algorithmes et les performances des systèmes linguistiques sont bien répertoriées ;

- les possibilités et les limites de tels systèmes ont été analysées de façon réaliste ;

- la batterie de tests pour juger des résultats (y compris de serveurs analogues) est variée sinon complète ;

- les problèmes de coûts de développement comme de maintenance sont connus.

Toutes ces informations peuvent être communiquées sur demande par l'Insee.

Produits dérivés de Synapse

L'architecture modulaire du serveur de nomenclatures développé par l'Insee permet de constituer quatre produits adaptés aux besoins d'utilisateurs ayant des préoccupations différentes dans le domaine des nomenclatures :

- une application de gestion et de diffusion de n'importe quel ensemble de nomenclatures ;

- une application de gestion et de diffusion adaptée aux nomenclatures d'activités et de produits ;

- une application de consultation en langage naturel et de codification automatique concernant les nomenclatures d'activités et de produits ;

- une application intégrant les deux précédentes (Synapse).

Chacun de ces produits fait l'objet d'une fiche décrivant le produit et les services associés.

Les trois produits applicatifs peuvent être installés sur les architectures suivantes :

- pour Synapse et le serveur logique : sur IBM 3090 et/ou serveur Unix de type RS 6000, chargé (si besoin) avec les données sur les nomenclatures d'activités et de produits ;

- pour le serveur linguistique : sur serveur Unix de type RS 6000 uniquement.

D'autres configurations matérielles sont possibles. Dans ce cas, le matériel, suffisamment dimensionné pour accueillir la ou les applications choisies, doit être fourni. Un coût de portage est à prévoir.

Les interfaces-utilisateur peuvent être de type Windows ou de type terminal (écran 3270) pour l'ensemble des fonctionnalités (consultation, production de sorties, gestion).

Produit 1 : Serveur logique (application de gestion et de diffusion d'un ensemble quelconque de nomenclatures et de tables de relations)

Cette application permet de gérer et consulter n'importe quel ensemble de nomenclatures, codes ou listes ainsi que les tables de relations qui leur sont associées.

Consultation

L'accès aux postes peut être réalisé : soit à partir de listes de codets, soit par les codets, soit à partir des tables de relations (hiérarchiques ou de correspondance).

Les postes sont décrits par un ou plusieurs intitulés (officiels, normalisés, en anglais, etc.) et par des notes explicatives structurées, officielles ou non (jurisprudence).

Toutes les entités (nomenclatures, niveaux, postes, intitulés, tables, relations) sont datées avec une période de validité propre. Le choix d'une date de référence permet de sélectionner les informations valides à cette date.

La structuration en réseau permet de circuler dans la base de données en suivant les relations (hiérarchiques, de correspondance, de recouvrement ou historiques) entre les postes.

(suite produit 1)

Production de fichiers

L'application permet d'extraire et de mettre en forme des ensembles d'informations sur les nomenclatures ou les tables correspondant exactement aux besoins du demandeur (définition de champs, sélection de niveaux, choix de variables). Une interface d'édition et une imprimante connectée permettent d'imprimer ces informations. (1)

Un module d'Échanges de Données Informatisés (EDI : transferts automatiques d'informations entre ordinateurs) sera développé dès que des messages Edifact normalisés pour les nomenclatures existeront.

Gestion

Toutes les entités peuvent être créées, supprimées, modifiées.

Un ensemble de fonctions permet de mettre en forme les données en vue de leur chargement.

Des contrôles permettent de vérifier la cohérence des données chargées et modifiées.

Un module de calcul de tables permet de construire des tables de relations, synthèses de plusieurs autres tables.

Bases de données

Trois bases ORACLE sont préformatées pour accueillir les données sur les nomenclatures :

- une base de gestion destinée aux opérateurs chargés d'introduire et gérer les nomenclatures et les tables,

- une base de diffusion, copie stabilisée de la précédente, destinée aux consultations et demandes de fichiers,

- une base de travail destinée à l'application elle-même pour sauvegarder des fichiers de façon temporaire.

(1) Un module d'édition sur imprimante "Postscript" comprenant une couche SGML fonctionne sur IBM3090 dans l'architecture retenue par l'Insee. Ce module doit être modifié si l'environnement d'édition est différent.

Produit 1bis : Serveur logique adapté aux nomenclatures d'activités et de produits

Cette application est le serveur logique (voir Produit 1) chargé d'un ensemble de nomenclatures d'activités et de produits qui comportera à terme (fin 1995) : 28 nomenclatures (dont 11 internationales ou européennes), 70 niveaux, 140 tables, 100 000 postes, 175 000 liens (en 300 fichiers environ).

Toutes les entités sont décrites. Pour chaque poste, la base contient les informations suivantes (quand elles existent) : intitulés (officiels et normalisés en français, officiels en anglais) et notes explicatives (structurées, officielles ou non).

Dans les tables, des "pondérations" mesurent les parts de recouvrement ou de correspondance des postes en relation partielle.

Produit 2 : **Serveur linguistique** (application de consultation en langage naturel et de codification automatique des nomenclatures d'activités et de produits).

Cette application permet d'accéder à des nomenclatures d'activités et de produits à partir de descriptifs écrits en langage naturel (sans vocabulaire spécialement normalisé) mais de façon plus précise qu'avec les systèmes à mots-clés.

Consultation

Le système permet d'accéder aux postes de toute nomenclature (décrite par des intitulés et des notes explicatives, et indexée).

Les requêtes sont analysées par une grammaire capable de détecter les ambiguïtés (à lever par l'utilisateur). Les réponses sont classées par ordre de vraisemblance.

Pour améliorer la précision des réponses, le système utilise les liens sémantiques avec le vocabulaire utilisé ainsi que le réseau logique des nomenclatures. (1) (2)

On peut, en posant des questions portant sur des produits, obtenir en réponse les différentes activités concernant ces produits.

(1) Le serveur linguistique est actuellement disponible uniquement en français mais peut être enrichi par d'autres langues, le thésaurus étant de type multilingue.
L'addition d'une langue implique la constitution d'un dictionnaire, la formulation d'une grammaire, la traduction du thésaurus et l'indexation des nomenclatures supplémentaires chargées.
(2) Un analyseur morphologique corrige les fautes de frappe ou d'orthographe. Les phénomènes de synonymie et et de généricité-spécificité sont pris en compte.

Codification automatique

Dans le cadre de travaux de grande ampleur (fichiers, enquêtes, recensements), le serveur linguistique peut codifier automatiquement des descriptifs d'activités ou de produits ou être un outil d'aide à la codification (présélection de postes puis validation manuelle).

Il est aussi un outil d'aide à la construction de tables de relations (analyse de descriptifs, codification puis validation manuelle).

Gestion linguistique

Le serveur linguistique inclut un module de gestion des données linguistiques (dictionnaire et thésaurus) et un module d'indexation des nomenclatures (intitulés, notes explicatives, jurisprudence). Ainsi peut-on enrichir le langage accepté et améliorer la définition des postes. (3)

Bases de données

Le serveur linguistique comporte deux bases de données : une base d'index (les liens entre descriptifs indexés et les postes des nomenclatures) et une base de connaissances.

Ce dernier ensemble d'informations constitue le réseau sémantique et lexical du domaine concerné.

Le thésaurus est structuré par environ 12 000 concepts décrivant les activités et les produits ; le dictionnaire renfermera à terme environ 20 000 mots couvrant le domaine et rattachés au thésaurus.

Deux restrictions limitent (volontairement) la richesse du vocabulaire pris en compte : les jargons professionnels ne sont pas intégrés (sauf cas particuliers) ; le domaine de la chimie est couvert par des mots que l'on peut définir mais s'arrête en deçà des formules.

Dictionnaire et thésaurus seront enrichis des mots ou des concepts nouveaux que l'évolution de l'activité économique et du langage introduira.

(3) Les grammaires de requête et d'indexation ainsi que les algorithmes de recherche peuvent aussi être gérés, mais restent de la compétence des linguistes.

Produit 3 : Synapse (application complète concernant les nomenclatures d'activités et de produits)

Cette application intègre les applications logique et linguistique (pour plus de détail, voir "fiches-produits"1bis et 2) et répond à l'ensemble des besoins d'utilistateurs concernant :

- la consultation,
- la diffusion,
- L'extraction d'informations,
- La codification auromatique,
- la gestion,

des nomenclatures d'activités et de produits internationales, européennes, nationales ou particulières.